DISSERTATION

GEOSTATISTICAL MODELS: MODEL SELECTION AND PARAMETER ESTIMATION UNDER INFILL AND EXPANDING DOMAIN ASYMPTOTICS

Submitted by

Andrew A. Merton

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctorate of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2006

COLORADO STATE UNIVERSITY

October 30, 2006

WE HEREBY RECOMMEND THAT THE DISSERTATION GEOSTATISTICAL MODELS: MODEL SELECTION AND PARAMETER ESTIMATION UNDER INFILL AND EXPANDING DOMAIN ASYMPTOTICS PREPARED UNDER OUR SUPERVISION BY ANDREW A. MERTON BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTORATE OF PHILOSOPHY.

Committee on Graduate Work

N. Scott Urquhart (Committee Member)

Paul F. Doherty (Committee Member)

Jennifer A. Hoeting (Co-Adviser)

Richard A. Davis (Co-Adviser)

F. Jay Breidt (Department Chair)

ABSTRACT OF DISSERTATION

GEOSTATISTICAL MODELS: MODEL SELECTION AND PARAMETER ESTIMATION UNDER INFILL AND EXPANDING DOMAIN ASYMPTOTICS

The research presented in this dissertation was originally motivated by the application of spatial models to the field of ecology. Often ecologists are interested in either (1) identifying significant relationships between the response of interest and candidate explanatory variables or (2) generating maps of the mean response and making predictions at unobserved locations. In the former case the scientist is trying to identify significant explanatory variables and understand the underlying relationship(s) with the response. For the latter case the scientist desires to make inference about the model parameters and/or produce predictions at unobserved locations and quantify the variability of these predictors. General linear models have proven to be very effective at addressing these two problems. However, over the past 10 to 20 years with the advent of global positioning systems (GPS), satellite imagery, etc., the ease of obtaining geo-referenced data has increased several fold. As a consequence scientists are now eager to incorporate spatial dependency into the models. A statistical model for a continuous process with geo-referenced data, henceforth referred to as a geostatistical model, provides a powerful means to investigate both motivating questions by accounting for potential spatial relationships.

Coupled with any modeling exercise is the ability to compare competing models. This is especially important when trying to identify significant explanatory variables. A common anecdotal observation in the ecological sciences is that "neighbors at close proximity tend to be more similar than neighbors separated by large distances." A

strength of the geostatistical model is its ability to prescribe just such a relationship. Although the importance of accounting for spatial correlation has been discussed in other contexts (Cressie, 1993), the effect of spatial correlation on model selection has not been fully explored. We begin in Chapter 2 by developing the AIC statistic for geostatistical models. Roughly speaking, AIC is a measure of the loss of information incurred by fitting an incorrect model to the data. The AIC statistic can be broken down into two components: the first component is a measure of the quality-of-fit and is a function of the likelihood function and the second component is a penalty factor that increases with increased model complexity. Evaluating the likelihood function for geostatistical models is computationally expensive because, in general, there does not exist a closed form for the parameter estimates. Thus optimization requires systematic searching of the parameter space to identify the maximum likelihood estimates. This becomes more and more taxing with increased sample size and/or model complexity.

Traditionally in geostatistical modeling, the AIC statistic is used to identify the best subset of explanatory variables assuming independent residuals. Having selected a subset of the explanatory variables, one proceeds to investigate the nature of the correlation structure of the model residuals. If the independence assumption appears to be met the researcher is done. If there appears to be correlation among the residuals, a suitable family is chosen to model the covariance function, the parameters of the trend surface are updated, followed by an updating of the covariance parameters. This process proceeds iteratively until some convergence criterion is met. We refer to this procedure as independent AIC. A deficiency associated with independent AIC is that the importance of one or more explanatory variables may be masked by the covariance structure. Indeed, the presence of one or more additional explanatory variables may reduce or eliminate the presence of correlation in the residuals. Thus we suggest that (possible) correlation in the error process must

be incorporated into the model selection process. Through a series of simulations we demonstrate that inclusion of spatial dependence during model selection can greatly improve the probability of identifying the correct model. We also demonstrate that sampling pattern and signal-to-noise ratio impact model selection where we define the signal-to-noise ratio as the ratio of the variability of the mean structure (large scale variability) to the variability of the noise process (small scale variability). Performance comparisons are made between independent and spatial AIC as well as a non-information based procedure, minimum description length (MDL). MDL has the distinct advantage that the researcher need not assume that there exists a true model. Instead, MDL attempts to minimize the amount of "storage space" required to adequately describe the data set. Similar to information-based criteria methods it reduces the data into two components: quality-of-fit and a penalty term that increases with model complexity.

We follow with two examples that implement spatial AIC and illustrate the flexibility of the method. The first example attempts to identify the best subset of explanatory variables for species abundance data while assuming the noise process is Matérn. We then construct a simulation study using the selected model to compare the performance of spatial AIC to independent AIC. A second example examines water chemistry response variables collected along a stream network in Maryland. An important complication is that the distance between observation locations can now be defined in one of several ways: Euclidean distance and hydrological distance (restricting movement to "within" the network). Hydrological distance can be further categorized as either symmetric or asymmetric depending on whether or not one accounts for the direction of (water) flow. Ver Hoef et al. (2006) demonstrate that the exponential function, among others, can be used for each of these distance measures, although a carefully constructed weight matrix is required for asymmetric hydrologic distance. Model selection proceeds using spatial AIC and

then selected models (one for each distance measure) are compared by evaluating the mean square prediction error (MSPE) for a randomly selected subset of the data that were withheld from model selection/fitting.

The derivation of the spatial AIC statistic requires standard asymptotic assumptions. For example, we assume that the parameter estimates for the large scale variation parameters and the correlation parameters are asymptotically unbiased and normally distributed with asymptotic covariance equal to the inverse of the Fisher Information. This motivates Chapters 3, 4, and 5 which set out to show, among other things, that the maximum likelihood estimator (MLE) of the spatial parameters are normally distributed. Since the underlying process is assumed to have a continuous domain, collection of additional observations of the response can proceed in one of two ways: collect additional observations within the current domain (infill) or collect new observations outside the current domain (expansion of the domain). Conceivably the former method of increasing the sample is always available to the scientist, although it may not be practical. However, the latter method is often restricted in the sense that for real applications the domain of interest is finite and hence one cannot expand the domain indefinitely. In either case one can approximate the distribution of the finite sample MLEs with random variables generated over infinite domains with the sampling distributions per unit length. Thus we show that the standard asymptotic assumptions required for the derivation of spatial AIC hold with respect to expansion of the domain with and without infill for the exponential correlation function in one-dimension. Simulation results indicate that the same can be said for both the exponential and the Matérn class of correlation functions in two-dimensions.

We begin in Chapter 3 by considering the Ornstein-Uhlenbeck process, the continuous analogue of the discrete first order autoregressive Gaussian process (AR(1)) in one-dimension. This process is characterized by the range parameter θ which

describes the strength of the correlation between two locations as a function of the distance between them. We develop the asymptotic distribution for (1) the MLE of θ when the observations are equally spaced, (2) a weighted least squares estimate for θ when observations are randomly spaced, and (3) the MLE of θ for randomly spaced observations. For each we provide simulation results that corroborate the theoretical results. Chapter 4 proceeds to the two-dimensional case where we explore the asymptotic properties of the MLE of θ for the exponential correlation function through simulation. Chapter 5 investigates the asymptotic behavior of the MLE $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ for the Matérn class of correlation functions in both one- and two-dimensions. Coupled with each of these analyses is the concept of sampling design by which we mean the prescribed manner in which sampling locations are selected. We demonstrate that infill and expansion of the domain both impact the distribution of the MLE and, for the exponential case, there appears to exist an optimal sampling design for a fixed domain and sampling effort (where sampling effort refers to the total number of observations).

The development of spatial AIC along with the (partial) verification of the underlying asymptotic assumptions provides researchers with a powerful tool for conducting model selection and model fitting in the geostatistical framework. Applications are numerous and include diverse fields of study including the ecological, geological, atmospheric, and oceanographic sciences.

Andrew A. Merton
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Fall 2006

ACKNOWLEDGEMENTS

The following people were instrumental toward completion of my thesis. Their patience, guidance, enthusiasm and persistence was greatly needed as I progressed, and on occasion regressed, over the past few years. Each contributed in their own way and I'd like to think that I've been able to absorb the best from each to bring forward into my future studies...

A special thank you to Jennifer Hoeting and Richard Davis, my thesis advisers, who have graciously shared their wisdom with me these past few years. They have given me a wonderful experience and I sincerely hope to carry forward their lessons into my new career.

Thank you to Scott Urquhart who provided brilliant input regarding the presentation of material and wonderful anecdotes about the world of statistics. Also I'd like to thank Paul Doherty whose insight about issues concerning the scientist "in the field" has highlighted the importance of a collaborative approach to solving ecological problems.

The work reported here was developed under STAR Research Assistance Agreements CR-829095 awarded by the U.S. Environmental Protection Agency to the Space Time Aquatic Resource Modeling and Analysis Program (STARMAP) at Colorado State University. This presentation has not been formally reviewed by the EPA. The views expressed here are solely those of the author. The EPA does not endorse any products or commercial services mentioned in this report.

DEDICATION

For Melanie...

I can't thank you enough for all of the support that you have given me these past many years. It was you who encouraged me to return to school to obtain my doctorate and you have always been there to celebrate my successes and brace me through the rough patches. Your patience has been unwavering and I can't imagine completing my degree without you. I will love you always...

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION AND MOTIVATION

Over the past 10 to 20 years the development and application of spatial models has grown at a remarkable rate. Much of this stems from the relative ease of being able to collect geo-referenced data with the introduction of global positioning systems (GPS), remote imaging technology (sattelite imagery), etc. as well as improved computing capabilities for fitting highly complex models. Examples are plentiful and include varied topics such as the presence of pollutants within stream networks and across larger bodies of water, the composition of soil throughout agricultural or forested regions, ozone levels across urban environments, and the migration of a bird species across North America. What all of these examples share in common is that each observation can be referenced to a specific geographic location. Hence scientists can begin to ask questions with respect to location: "Is there a relationship between the response and the geographic location?", "Are neighboring locations at close proximity (significantly) more similar than observations separated by large distances?", "Is there an effective range outside of which observations of the response are approximately independent?", etc. The first question is akin to asking how does the mean response vary over space. The remaining two questions are associated with the anecdotal observation that "observations of the response at close proximity tend to be more similar than observations separated by large distances." This notion appears frequently in ecological studies and suggests that the response is positively correlated at small distances. The geostatistical model provides a framework

in which one can easily incorporate spatial (geographic) dependencies along with potential explanatory variables.

Often the scientist is interested in developing a model to answer one of two motivating questions: "Which explanatory variables (covariates) are not *significant* with respect to the response?" and "Which subset of the explanatory variables provide the most precise predictions for unobserved locations?" The first question involves identifying which covariates are most influential with respect to the response. In this case the scientist may not necessarily be interested in the model form per se, rather she wants to classify which factors are significant. For the second question the scientist is chiefly concerned with developing a model for interpolating, or kriging in the geostatistical lexicon, the data to generate maps of the predicted response and to quantify the corresponding standard error. In either case the geostatistical model is a powerful tool for investigating these questions. By fitting the geostatistical model to different subsets of explanatory variables one can make comparisons and identify the subset that best meets the desired goal.

1.1 The Geostatistical model

We begin by assuming that the data can be modeled by a random field defined over a continous domain $\mathcal{D}$. The linear model for the random field is given by

$$Y(s) = \mathbf{X}'(s)\boldsymbol{\beta} + \delta(s), \tag{1.1}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ is a $p-$vector of the unknown regression coefficients, $\boldsymbol{X}(s) = (1, X_1(s), \ldots, X_{p-1}(s))'$ is $p-$vector of covariates, and $\delta(s)$ is a stationary, isotropic Gaussian random field with covariance function $\mathrm{Cov}(s_i, s_j) = \sigma^2\rho(||s_i - s_j||; \boldsymbol{\theta})$. Here σ^2 is the variance of the process, $\rho_(||\cdot||; \boldsymbol{\theta})$ is a family of autocorrelation functions parameterized by the $k-$vector $\boldsymbol{\theta}$, and $||\cdot||$ denotes the Euclidean distance between two sampling sites. For a fixed sampling effort of size n, we observe the

response and covariates at locations $s_i, i \in \{1, 2, \ldots, n\}$. Therefore the model for the complete partial realization, i.e., the collection of all n sampling locations, is

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{1.2}$$

where $\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2\mathbf{\Gamma}$ is the $n \times n$ Gaussian variance-covariance matrix of the residuals. Here σ^2 is the variance of the process and $\mathbf{\Gamma}$ is the correlation matrix such that the (i, j)th element equals $\rho(||s_i - s - j||; \boldsymbol{\theta})$ where is $\rho(\cdot; \boldsymbol{\theta})$ is a family of autocorrelation functions parameterized by the k-vector $\boldsymbol{\theta}$ and where $||\cdot||$ denotes the Euclidean distance between two locations. Note that if $\mathbf{\Gamma} = I_n$ we recover the general linear model with independent errors.

The model form above, adopted from (Cressie, 1993), can be used to predict the response at unobserved locations through interpolation, known as kriging in the spatial context. This approach requires fitting the complete model; that is to say one must fit the regression coefficients β, the correlation coefficients θ, and the variance component σ^2. Hence for a particular set of explanatory variables we can estimate all of the model parameters and interpolate (krige) the response at any location $s \in \mathcal{D}$. But which, if any, subset of the explanatory variables should be included in the model?

1.1.1 Parameter Estimation

For the majority of this dissertation we estimate the model parameters by optimizing the likelihood function. We adopt this method because of its importance with respect to model selection using information-based selection criterion. Since the random field is assumed to be Gaussian, the likelihood function of (1.2) as a function of the partial realization $\mathbf{Y}$ is

$$\mathcal{L}(\beta, \theta, \sigma^2; \mathbf{Y}) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)'\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\beta)\right\}. \tag{1.3}$$

Defining the log-profile function as $\ell(\cdot;\mathbf{Y}) = \log\left(\mathcal{L}(\cdot;,\mathbf{Y})\right)$ and substituting $\Sigma = \sigma^2\Gamma$ into (1.3), we obtain

$$\ell(\beta,\theta,\sigma^2;\mathbf{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log|\Gamma| - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)'\Gamma^{-1}(\mathbf{Y}-\mathbf{X}\beta). \tag{1.4}$$

Optimization of (1.4) requires searching over a parameter space of $p+k+1$ dimensions which can be computationally expensive and may fail to converge for poorly defined initial quesses of the parameter values. For this particular form of the log-likelihood function, one can concentrate out both the regression parameter vector β and the variance paramter σ^2 by setting the partial derivatives of $\ell(\cdot;\mathbf{Y})$ with respect to $\boldsymbol{\beta}$ and σ^2 to zero and solving. Hence,

$$\frac{\partial\ell}{\partial\beta} = 2\mathbf{X}'\Gamma^{-1}\mathbf{X} - 2\mathbf{X}'\Gamma^{-1}\mathbf{Y} = 0$$

which implies

$$\beta(\theta) = \hat{\beta} = \left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\mathbf{Y}. \tag{1.5}$$

Furthermore, after substituting (1.5) into (1.4), we find

$$\frac{\partial\ell}{\partial\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{Y}-\mathbf{X}\hat{\beta})'\Gamma^{-1}(\mathbf{Y}-\mathbf{X}\hat{\beta}) = 0$$

implying

$$\sigma^2(\theta) = \hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y}-\mathbf{X}\hat{\beta})'\Gamma^{-1}(\mathbf{Y}-\mathbf{X}\hat{\beta}). \tag{1.6}$$

Thus (1.4) can be written in terms of the correlation parameter vector $\boldsymbol{\theta}$ alone, referred to as the *profile* log-likelihood function, such that

$$\ell_{\text{profile}}(\theta;\hat{\beta},\hat{\sigma}^2,\mathbf{Y}) \propto \frac{n}{2}\log(\hat{\sigma}^2) + \frac{1}{2}\log|\Gamma| + \frac{n}{2} \tag{1.7}$$

so that now we have an objective function to maximize of dimension k instead of $p+k+1$.

Optimization of (1.7) can still be computationally expensive due to the presence of a determinant term and matrix inversion. For particular sampling designs,

such as along a fixed lattice in one-dimension, the exact likelihood can be computed very efficiently using the spectral density and fast fourier transforms (FFT) (Brockwell and Davis, 1991). This methodolgy can be extended to two- or more dimensions as well as to sampling designs where the sampling locations no longer fall on a lattice. For the latter case the sampling locations are "snapped" to a fine lattice and the likelihood is approximated using, say, Whittle's approximation (Fuentes, 2005). Large data sets can also be computationally challenging. As the number of sampling locations n increases, the number of operations required to invert the $n \times n$ correlation matrix increases somewhere between $\mathcal{O}(n^2 \log(n))$ and $\mathcal{O}(n^3)$, depending on the computational method. Johns et al. (2003) propose various methods for coping with very large data sets such as avoiding matrix inversion by solving the implied set of linear equations. For many of the correlation functions commonly considered, the correlation between locations s and t goes to zero when they are separated by a large distance. Hence by keeping track of where the "zeros" are located throughout a large correlation matrix one can avoid unnecessary multiplications. Theses approximation methods, among others, are powerful tools for estimating the likelihood but they do not, in general, achieve the exact optimal solution, i.e., the maximum likelihood estimate. Therefore we have have restricted our efforts to methods that solve for the exact MLEs. This is especially important when discussing the asymptotic distribution of the parameter estimates (see Chapter 3).

1.2 Model selection

A general philosophy for choosing a model is that we would like to incorporate information that we believe influences the response variable while acknowledging that we do not know everything associated with the response. These unknowns could be quantities that we did not (or could not) measure, complex variable interactions, heterogeneity, etc. Thus the error process is included in the model to

"account" for these unknowns. The problem becomes increasingly complicated when we consider that there may be competing models each using a different subset of known variables, a different error structure, or some combination of the two. Therefore a measure is required that will allow one to compare models. We consider the Akaike Information Criterion (AIC) (Akaike, 1974) as applied to the geostatistical model. Roughly speaking, AIC is a measure of the loss of information incurred by fitting an incorrect model to the data. The AIC statistic can be broken down into two components: the first component is a measure of the quality-of-fit and is a function of the likelihood function and the second component is a penalty factor for the introduction of additional parameters to the model. Hence, we derive heuristically a modified form of AIC to account for the addition of the correlation coefficient vector θ.

Typically modelers use the AIC statistic to identify the best subset of explanatory variables assuming independent residuals. Having selected a model the modeler then investigates the nature of the correlation structure of the model residuals. If the independence assumption appears to be met, the modeler is done. If there appears to be correlation among the residuals, a suitable family is chosen to model the covariance function, the parameters of the trend surface are updated, followed by an updating of the covariance parameters. This process proceeds iteratively until some convergence criterion is met. We refer to this procedure as independent AIC. A deficiency associated with independent AIC is that the importance of one or more explanatory variables may be masked by the covariance structure. Indeed, the presence of one or more additional explanatory variables may reduce or eliminate the presence of correlation in the residuals. Thus we suggest that (possible) correlation in the error process must be incorporated into the model selection process. Since the class of models being considered here are Gaussian, it is relatively straightforward to derive the likelihood function and optimize it to achieve the maximum likelihood

estimates (MLEs) of the model parameters. Thus, unlike the independent AIC procedure, all of the model parameters are fit simulatancously, i.e., spatial correlation is accounted for during model fitting. We refer to this procedure as spatial AIC.

1.3 Asymptotic Distribution of the Model Parameters

The development of the spatial AIC statistic relies on standard asymptotic assumptions. For example, it is assumed that the relations $\hat{\beta} - \beta$ and $\hat{\theta} - \theta$ converge in distribution to mean zero normal random variables with covariance matrix equal to the inverse of the Fisher information. In the spatial context one can obtain additional observations of the response in two distinct ways. First one can assume that new observations are available outside the current domain $\mathcal{D}$. Hence new observations are obtained by "expanding the domain". However, since the underlying process is assumed to be continuous one can also obtain additional observations within the current domain. This is referred to as "infill". For the former case the domain can be extended indefinitely while for the latter, as the total sampling effort increases, the mean distance between locations must go to zero. Therefore it is necessary to investigate the asymptotic properties of the MLE under both paradigms.

Coupled with the asymptotic distributions of the model parameter is the notion of sampling design (or sampling pattern). For example, one common sampling procedure is to take observations on a fixed lattice in one-, two-, or even three-dimensions. A second method is to "cluster" sampling locations by randomly selecting parent sites and then subsampling within a fixed radius of each parent location. Although not a primary objective of this dissertation, it is important to explore the relationship between sampling design and the distribution of the model parameters. In recent years there has been an ongoing effort to try to develop optimal sampling design by, for example, minimizing the Fisher information (Xia et al., 2005; Zhu and Zhang, 2005; Zhu and Stein, 2005).

1.4 Outline of Dissertation

Chapter 2 begins by developing the AIC statistic for geostatistical models. We demonstrate through simulation the importance of incorporating spatial dependency during model selection by comparing the independent and spatial AIC methods. Furthermore we introduce the non-information based model selection method minimum description length (MDL) and tabulate its performance as well. Different sampling designs and signal-to-noise ratios are used throughout the simulation studies where the signal-to-noise ratio is defined as the variability of the mean surface to the variability of the error term (noise process). Two model selection examples using real-world data are also presented: the first models the presence of the Orange-throated whiptail lizard found in southern California (Ver Hoef et al. (2001)) and the second performs model selection for stream water chemistry data located in Maryland. The former example demonstrates the utility of spatial AIC for a highly clustered sampling design with numerous potential explanatory variables. The latter example illustrates the flexibility of spatial AIC with respect to different distance measures: distance between two locations along a stream network depends on whether or not travel is restricted to "within" the network.

As stated above, the derivation of the spatial AIC statistic requires standard asymptotic assumptions for the model parameter estimates. Chapter 3 proceeds to derive the asymptotic distribution of the range parameter for a mean zero stationary process with exponentially correlated residuals in one-dimension. We begin by presenting Uhlenbeck-Ornstein process, a stochastic differential equation, which is the continuous analogue of the stationary first order autoregressive model. Analytical results with respect to expansion of the domain and infill, supported by simulation studies, are derived for three cases: (1) the maximum likelihood estimator for equispaced sampling locations, (2) the weighted least squares estimate for randomly spaced sampling locations, and (3) the maximum likelihood estimator for

randomly spaced sampling locations. We illustrate the impact of sampling design on the asymptotic distributions and provide evidence that there does exist an optimal sampling design for the exponential case. We extend the empirical results of Chapter 3 to two-dimensions in Chapter 4. We begin by defining sampling designs in two-dimensions that are analagous to the one-dimensional sampling designs. Next we illustrate the distribution of the MLEs with respect to both expanding domain and infill asymptotics.

In Chapter 5 we revisit the Matérn class of correlation functions first presented in Chapter 2. The Matérn class is a large class of correlation functions that includes the exponential correlation function as a special case. We summarize the results from a series of simulation studies for both one- and two-dimensions using various sampling designs and introduce the statistic integrated mean square error (IMSE) as a candidate measure for comparing sampling design strategies. We close the dissertation by citing the major findings and suggesting future work in Chapter 6.

Chapter 2

MODEL SELECTION FOR GEOSTATISTICAL MODELS

2.1 Introduction

Ecologists and scientists in other fields typically consider a number of plausible models in statistical applications. Formal consideration of model selection in ecological applications has dramatically increased in recent years, perhaps in part due to the publication of the book by Burnham and Anderson (1998, 2002). Concurrently, the wide availability of inexpensive global positioning systems and other advances in technology have allowed for the collection of vast quantities of data with geo-referenced sample locations. As a result, models for spatially correlated data are becoming increasingly important. We consider these two problems together, spatial modeling and model selection. The importance of accounting for spatial correlation has been discussed in other contexts (Cressie, 1993), but the effect of spatial correlation on model selection has not been fully explored.

A general philosophy for choosing a model is that we would like to incorporate information that we believe influences the response variable while acknowledging that we do not know everything associated with the response. These unknowns could be quantities that we did not (or could not) measure, complex variable interactions, heterogeneity, etc. Thus an error process is often included in the model that "accounts" for these unknowns. For example, we may suspect that the abundance of a certain species is dependent on the availability of a certain type of vegetation and the predator to prey ratio. But we must acknowledge that other variables,

perhaps unmeasurable, are likely to play an important role such as the availability of fresh water or the prevalence of a certain disease. The model that we construct should account for these unknown influences. This is the main role of an error term in any such modeling exercise. The problem becomes more complicated when we consider that there may be competing models each using a different subset of known variables. For example, perhaps there are two types of vegetation that the species will eat. Is either vegetation species a better predictor of abundance or should some combination of the two be used? In other words, which subset of explanatory variables and error structure together provides the best model? To attempt to answer this question, we adopt a geostatistical model (Cressie, 1993) which can be used to predict a response at unobserved locations. This approach, also referred to as kriging, involves the fitting of an autocorrelation function which describes the relationship between observations based on the distance between the observations. This method allows for any number of the explanatory variables observed at the sample locations to be included in the model to improve the overall predictions.

Typically, spatial correlation is ignored in the selection of explanatory variables which can have a dramatic effect on the choice of model. For example, the importance of particular explanatory variables may not be apparent when spatial correlation is ignored. To address this problem, we consider the Akaike Information Criterion (AIC) as applied to a geostatistical model. We provide simulation results that show that using AIC for a geostatistical model is superior to the standard approach of ignoring spatial correlation in the selection of explanatory variables. This idea is demonstrated via two real-world problems: a model for the abundance of the orange-throated whiptail lizard found in southern California and a model for stream water chemistry in Maryland. The principle of minimum description length (MDL) applied to the variable selection problem is also investigated and simulation results are provided for comparison. The water chemistry example extends the

model selection methodology to non-traditional measures of distance, e.g., in-stream distance.

This chapter proceeds as follows. In Section 2.2 we describe a geostatistical model and methods for parameter estimation. In Section 2.3 we develop the AIC and offer a heuristic derivation of the AIC in context to geostatistical models and discuss fitting spatial models and other model selection issues. Simulation results in Section 2.4 and the examples in Section 2.5 underscore the importance of accounting for spatial correlation in the selection of explanatory variables.

2.2 The Geostatistical Model

Suppose we are interested in the abundance of the orange-throated whiptail lizard in a specific region in southern California. (Analysis results of this data set are given in Section 2.5.) Assume that we have collected information at each of 150 sites spread across the area of interest. Our data set consists of the average number of lizards observed per day, the percent coverage of vegetation, the abundance of ants (a primary food source), and a geo-reference for each site, such as latitude and longitude. It is not feasible to collect data at all possible locations, thus we are assuming that these 150 sites are representative of the entire area of interest. Let $Z(s_i)$ denote the average abundance of lizards at site i where $i = 1, \ldots, 150$. Thus the vector $\boldsymbol{Z} = (Z(s_1), \ldots, Z(s_{150}))'$ is a partial realization of the continuous random field over this finite area, D. In other words, we are assuming that at any given site s within the domain D, the average abundance of the lizards is a function of a specific set of variables that can be observed along with some random noise.

A model for the continuous random field at any location $s \in D$ is given by

$$Z(s) = \mathbf{X}'(s)\boldsymbol{\beta} + \delta(s), \tag{2.1}$$

where $\boldsymbol{X}(s) = (1, X_1(s), \ldots, X_{p-1}(s))'$ is a p vector consisting of the constant 1 and $p-1$ explanatory variables observed at location s, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ is a p vector

of the unknown model coefficients, and $\delta(s)$ is the unobserved "regression" error at location s. For example, $X_1(s)$ and $X_2(s)$ may be the percent coverage of vegetation and the abundance of ants at location s, respectively. For computational ease we will assume that the error process $\delta(s)$ is a stationary, isotropic Gaussian process with mean zero and covariance function $\text{Cov}(\delta(s_i), \delta(s_j)) = \sigma^2\rho(||s_i - s_j||; \boldsymbol{\theta})$. Here σ^2 is the variance of the process, $\rho_{\boldsymbol{\theta}}(|| \cdot ||)$ is a family of autocorrelation functions with a parameter vector $\boldsymbol{\theta}$ of length k, and $|| \cdot ||$ denotes the Euclidean distance between two sites. Thus, we assume that the correlation between any two sites is only a function of the distance between them. In deciding among the covariates, we must also choose an appropriate autocorrelation function. As will be demonstrated below, these two issues are inextricably linked. The isotropic assumption could certainly be omitted if the process is directional by nature, e.g., weather.

The autocorrelation function must satisfy certain mathematical conditions in order to be valid. This restricts our selection to one of a number of standard autocorrelation families. Most readers should be familiar with the independent error process associated with multilinear regression. In this case one is assuming that the errors are identically distributed and independent of one another and independent of location. For geospatial data, it is reasonable to assume that observations that are nearby will have similar response values, so we seek to model this relationship via the autocorrelation function. A rich family of autocorrelation functions is the Matérn family (Handcock and Stein, 1993; Stein, 1999a). The Matérn autocorrelation function has the general form

$$\rho_{\boldsymbol{\theta}}(d) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{2d\sqrt{\theta_2}}{\theta_1}\right)^{\theta_2} \mathcal{K}_{\theta_2}\left(\frac{2d\sqrt{\theta_2}}{\theta_1}\right), \qquad \theta_1 > 0,\ \theta_2 > 0, \tag{2.2}$$

where $\mathcal{K}_{\theta_2}(\cdot)$ is the modified Bessel function of order θ_2 (Abramowitz and Stegun, 1965). The "range" parameter, θ_1, controls the rate of decay of the correlation between observations as distance increases. Large values of θ_1 indicate that sites that

are relatively far from one another are moderately (positively) correlated. The parameter θ_2 can be described as controlling behavior of the autocorrelation function for observations that are separated by small distances. The Matérn class includes the exponential autocorrelation function when $\theta_2 = 0.5$ and the Gaussian autocorrelation function as a limiting case when $\theta_2 \to \infty$. The Matérn class is very flexible, being able to to strike a balance between these two extremes, thus making it well suited for a variety of applications. Figures 2.1 and 2.2 illustrate the flexibility of the Matérn autocorrelation function. Notice that for small distances the correlation between sites is large and decreases as distance increases.

The autocorrelation function given in (2.2) can be further adapted to include the possibility of measurement error, often referred to as nugget in the spatial context. A mixture model that incorporates measurement error in these spatial models is considered in Thompson (2001). To minimize the complexity of the current discussion, we have chosen not to include a nugget effect in our simulations or analysis of the lizard data example. The stream water chemistry analysis includes a nugget parameter. It should be noted that selection of the form of the autocorrelation function can be easily incorporated into the model selection process. For example, one could assume that the autocorrelation function is Matérn but allow the selection process to determine whether or not a nugget should be included.

2.2.1 Estimation

The model in (2.1) is often referred to as a geostatistical model or a universal kriging model. For a particular subset of explanatory variables and structured error process, we are now tasked with estimating the parameters $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\theta}$. Estimation of the parameters of this model can proceed using one of several likelihood based approaches (Cressie, 1993; Haining, 1990; Smith, 2000) or a Bayesian approach (Handcock and Stein, 1993; Thompson, 2001). Here we consider the former. Note

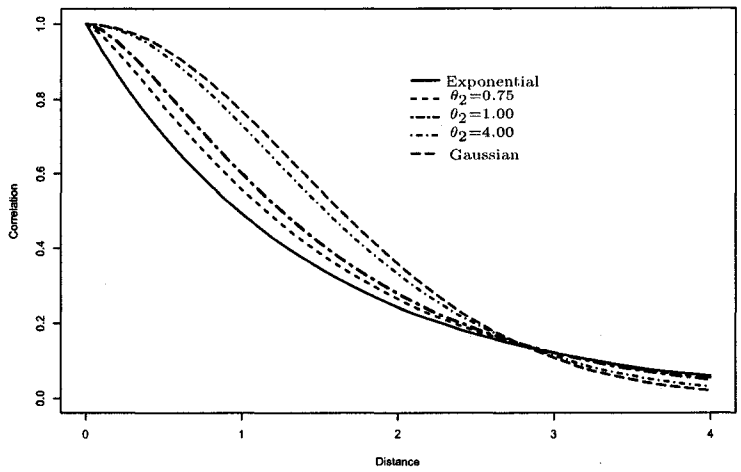

Figure 2.1: Matérn autocorrelation function for several parameter values. The horizontal axis is the distance between points and the vertical axis is the correlation between two points at a given distance. We used a fixed range parameter, $\theta_1 = 2.00$, with various smoothness parameter values, θ_2. Note that the exponential autocorrelation is equivalent to the Matérn autocorrelation function with $\theta_2 = 0.50$ and that the Gaussian autocorrelation function corresponds to the limiting case such that $\theta_2 \to \infty$.

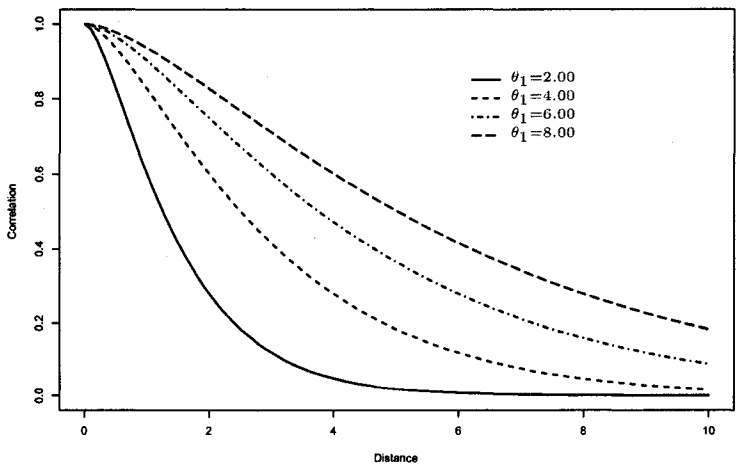

Figure 2.2: Matérn autocorrelation function for smoothness parameter $\theta_2 = 1.00$ and various range parameter values, θ_1.

that both approaches can be computationally challenging to implement for large sample sizes.

Using the assumption that the error process is Gaussian, the log-likelihood of the parameters in equation (2.1), $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2)$, based on the observed data, $\boldsymbol{Z}$, is given by

$$\ell\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2; \boldsymbol{Z}\right) = -\frac{1}{2}\log\left|\sigma^2\boldsymbol{\Gamma}\right| - \frac{1}{2\sigma^2}\left(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta}\right)'\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta}\right), \tag{2.3}$$

where $\boldsymbol{\Gamma} = [\rho_{\boldsymbol{\theta}}(||s_i - s_j||)]$ represents the matrix of correlations between all pairs of observations, $i, j = 1, \ldots, n$. By concentrating out $\boldsymbol{\beta}$ and σ^2, the profile likelihood can be easily computed which can often accelerate optimization of the likelihood. That is, by maximizing the likelihood with respect to $\boldsymbol{\beta}$ and σ^2, we obtain $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\boldsymbol{X}'\boldsymbol{\Gamma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}$ and $\hat{\sigma}^2 = \hat{\sigma}^2(\boldsymbol{\theta}) = \left(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)/n$. The resulting log profile likelihood is

$$\ell_{profile}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{Z}\right) = -\frac{1}{2}\log|\boldsymbol{\Gamma}| - \frac{n}{2}\log\left(\hat{\sigma}^2\right) - \frac{n}{2}. \tag{2.4}$$

Maximizing (2.4) yields the maximum likelihood estimates for the parameters of the spatial autocorrelation function, $\boldsymbol{\theta}$, which can in turn be used to compute the maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2.

An alternative approach for parameter estimation is the restricted maximum likelihood (REML) approach of Patterson and Thompson (1971). Cressie (1993, p. 93) supports the use of REML over maximum likelihood as a method of estimation when the number of explanatory variables is large. For model selection, most procedures involve a component consisting of the maximized likelihood function. Since REML does not maximize the likelihood, we do not consider REML here further. However, once a model has been selected, the researcher is free to re-estimate the model parameters using, for example, REML.

2.3 Model Selection for Geostatistical Models

Model selection is a critical ingredient in nearly any model building exercise. Depending on one's philosophical perspective, which is often driven by the modeling objective, there are a myriad of procedures for selecting an optimal model subject to a particular criterion. The introductions in the books by McQuarrie and Tsai (1998) and Burnham and Anderson (2002) give excellent accounts of the various philosophies underpinning model selection. It is important, however, to adopt a model selection paradigm that reflects the ultimate objective of the modeling process. For example, an explanatory model that establishes useful relationships between explanatory and response variables may not necessarily perform as well as a predictive model and vice versa. Section 2.3.1 develops the Akaike Information Criterion (AIC) for spatial models of the form (2.1) while Section 2.3.2 discusses spatial model fitting. Section 2.3.3 contains a brief discussion of the concept of Minimal Description Length (MDL) and further remarks on model selection issues.

Returning to our working example of the whiptail lizard, the current question at hand is which model should be selected? Should we include both of the potential explanatory variables, just one, or perhaps neither? What is the most appropriate form of the autocorrelation function? What is required is a quantitative measure of how closely each of the candidate models coincides with the true model. We may also wish to penalize less parsimonious models. We suggest that AIC, extended to spatial models, accomplishes these goals.

2.3.1 AIC for Spatial Models

There are often two points of view taken in model selection. The first presumes that there exists a true finite-dimensional model from which the data were generated. For example, one might hypothesize the true model to be linear in which there exists an explicit linear relationship between the explanatory variables and the response.

In this case, the key modeling objective is to identify the correct set of covariates that comprise the model. The second modeling perspective, which seems particularly well suited for ecological data, is that the "truth" and consequently, the underlying true model, is essentially infinite dimensional and we have no hope of identifying all the requisite factors that go into the process under study. In other words, reality cannot be expressed as a simple "true model" because, as Burnham and Anderson (1998) observe, "[Ecological] systems are complex, with many small effects, interactions, individual heterogeneity, and individual and environmental covariates (being mostly unknown to us)." Thus, the goal is to find the best approximating finite dimensional model to this infinite dimensional problem.

Under the first scenario, consistency should be a minimum requirement of a model selection procedure. That is, as more data are acquired, the model selection procedure should ultimately choose the correct model with probability one. In the second situation where the true model is infinite dimensional, a model selection procedure ought to choose a finite dimensional model that is closest to the true model in some sense. The Akaike Information Criterion (Akaike, 1973) is one procedure that is designed to achieve this second goal.

AIC was developed as an estimator of the Kullback-Leibler Information. Roughly speaking AIC is a measure of the loss of information incurred by fitting an incorrect model to the data. To describe the main idea behind AIC, let $\boldsymbol{Z}$ be an n-dimensional random vector with true probability density function f_T and consider a family $\{f(\cdot;\psi), \psi \in \boldsymbol{\Psi}\}$ of candidate probability density functions. The Kullback-Leibler information between $f(\cdot;\psi)$ and f_T is defined as

$$I(\psi) = \int -2\log\left\{\frac{f(\boldsymbol{z};\psi)}{f_T(z)}\right\} f_T(\boldsymbol{z})d\boldsymbol{z}. \tag{2.5}$$

Applying Jensen's inequality, we see that

$$\begin{aligned} I(\psi) &= \int -2\log\left\{\frac{f(\boldsymbol{z};\psi)}{f_T(\boldsymbol{z})}\right\} f_T(\boldsymbol{z})d\boldsymbol{z} \\ &\geq -2\log\left\{\int \frac{f(\boldsymbol{z};\psi)}{f_T(\boldsymbol{z})} f_T(\boldsymbol{z})d\boldsymbol{z}\right\} \\ &= -2\log\left\{\int f(\boldsymbol{z};\psi)d\boldsymbol{z}\right\} \\ &= 0, \end{aligned}$$

with equality holding if and only if $f(\boldsymbol{z};\psi) = f_T(\boldsymbol{z})$ almost everywhere with respect to the true model f_T.

By treating $I(\psi)$ as the information loss associated with $f(\cdot;\psi)$, the idea is to minimize $I(\psi)$ over all candidate models $\psi \in \boldsymbol{\Psi}$. Unfortunately this is not possible without knowing f_T, thus we need to adopt a strategy that is not dependent on the unknown density f_T.

First rewrite the Kullback-Leibler information in the following manner;

$$\begin{aligned} I(\psi) &= \int -2\log\left\{\frac{f(\boldsymbol{z};\psi)}{f_T(\boldsymbol{z})}\right\} f_T(\boldsymbol{z})d\boldsymbol{z} \\ &= \int -2\log\left\{f(\boldsymbol{z};\psi)\right\} f_T(\boldsymbol{z})d\boldsymbol{z} + \int 2\log\left\{f_T(\boldsymbol{z})\right\} f_T(\boldsymbol{z})d\boldsymbol{z} \qquad (2.6) \\ &= \Delta(\psi) + \int 2\log\left\{f_T(\boldsymbol{z})\right\} f_T(\boldsymbol{z})d\boldsymbol{z}. \end{aligned}$$

The first term, defined as the Kullback-Leibler index, can be written as $\Delta(\psi) = E_T\left\{-2\log L_Z(\psi)\right\}$ where the expectation is taken with respect to the true density and $L_Z(\psi)$ is the likelihood based on the candidate model corresponding to ψ using the data $\boldsymbol{Z}$. Note that the second term in (2.6) is a constant and plays no role in the minimization of $I(\psi)$. While it is generally not possible to compute either $\Delta(\psi)$ or $\Delta(\hat{\psi})$, where $\hat{\psi}$ is the maximum likelihood estimate of ψ, we can strive to find a model that minimizes an unbiased estimate of $E_\psi(\Delta(\hat{\psi}))$, where E_ψ represents the expectation operator relative to the candidate density $f(\cdot;\psi)$.

It is shown in Section 2.3.1.1 that the quantity

$$\text{AIC} = -2\log L_Z(\hat{\psi}) + 2n\frac{p+k+1}{n-p-k-2} \tag{2.7}$$

is an approximately unbiased estimate of the expected Kullback-Leibler information evaluated at $\hat{\psi}$, where p is the number of explanatory variables (including an intercept term), k is the number of parameters associated with the autocorrelation function, and n is the number of observed sites. Henceforth we will refer to this version as the spatial AIC which includes a measure of the quality of fit of the model (first term) and a penalty factor for the introduction of additional parameters into the model (second term). The traditional (corrected) AIC statistic for this model is

$$\text{AIC}_\text{c} = -2\log L_Z(\hat{\psi}) + 2(p+k+1). \tag{2.8}$$

Notice that for large n the penalty factors, $2n(p+k+1)/(n-p-k-2)$ and $2(p+k+1)$ are nearly equivalent. The spatial AIC statistic has a more severe penalty for larger order models which helps counterbalance the tendency of AIC to over fit models to data.

The principle of AIC is to select a combination of explanatory variables and a model for the autocorrelation function which minimize either spatial AIC or AIC_c. It is worth remarking that in many classical situations, such as linear regression or time series modeling, spatial AIC and AIC are not consistent order selection procedures. In other words, as the sample size increases there is a positive probability that a model selected by spatial AIC or AIC_c does not correspond to the true model. Nevertheless, these statistics should produce good estimates of the Kullback-Leibler Information for which they were formulated (Hurvich and ling Tsai, 1989).

2.3.1.1 Derivation of spatial AIC

To give a heuristic derivation of the spatial AIC statistic in the spatial model setup of (2.1), we follow the development in Brockwell and Davis (1991, p. 303).

Suppose $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ are two independent realizations from model (2.1) at fixed locations $(s_1, \ldots, s_n)$ with true parameter value $\boldsymbol{\psi}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\theta}_0, \sigma_0^2)'$. Let $f(\cdot; \boldsymbol{\psi})$ be a candidate Gaussian density function corresponding to the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)'$. Then by the independence of $\boldsymbol{Y}$ and $\boldsymbol{Z}$,

$$E_{\boldsymbol{\psi}}\left[\Delta(\hat{\boldsymbol{\psi}})\right] = E_{\boldsymbol{\psi}}\left[E_{\boldsymbol{\psi}}\left\{-2\log L_Y(\hat{\boldsymbol{\psi}})|\boldsymbol{Z}\right\}\right] = E_{\boldsymbol{\psi}}\left[-2\log L_Y(\hat{\boldsymbol{\psi}})\right],$$

where L_Y is the likelihood based on $\boldsymbol{Y}$ and $\hat{\boldsymbol{\psi}}$ is the maximum likelihood estimate of $\boldsymbol{\psi}$ based on $\boldsymbol{Z}$. Using properties of the Gaussian density function and the representation $\hat{\sigma}^2 = (\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\left(\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\theta}})\right)^{-1}(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n$, we have

$$-2\log L_Y(\hat{\boldsymbol{\psi}}) = -2\log L_Z(\hat{\boldsymbol{\psi}}) + \hat{\sigma}^{-2}S_Y(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - n, \tag{2.9}$$

where $S_Y(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\left(\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\theta}})\right)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. The goal is find an unbiased approximation for $E_{\boldsymbol{\psi}}\left[\hat{\sigma}^{-2}S_Y(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})\right]$ of equation (2.9).

Using a second order Taylor series to expand $S_Y(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ in a neighborhood of $(\boldsymbol{\beta}, \boldsymbol{\theta})$, we obtain

$$\begin{aligned} S_Y(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \simeq\ & S_Y(\boldsymbol{\beta}, \boldsymbol{\theta}) + \left((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - (\boldsymbol{\beta}, \boldsymbol{\theta})\right)' \frac{\partial S_Y(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial(\boldsymbol{\beta}, \boldsymbol{\theta})} \\ & + \frac{1}{2}\left((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - (\boldsymbol{\beta}, \boldsymbol{\theta})\right)' \frac{\partial^2 S_Y(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial(\boldsymbol{\beta}, \boldsymbol{\theta})\partial(\boldsymbol{\beta}, \boldsymbol{\theta})'}\left((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - (\boldsymbol{\beta}, \boldsymbol{\theta})\right). \end{aligned} \tag{2.10}$$

To evaluate the expected value of the terms in (2.10), we assume that standard asymptotics hold for the MLE $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2)'$. These are

(i) $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})'$ is approximately normal with mean $(\boldsymbol{\beta}, \boldsymbol{\theta})'$ and asymptotic covariance matrix given by the inverse of the Fisher information, I_n,

(ii) for large n, I_n^{-1} can be approximated by

$$V(\boldsymbol{\beta}, \boldsymbol{\theta}) = \left\{-\frac{1}{2\hat{\sigma}^2}E_{\boldsymbol{\psi}}\left[\frac{\partial^2 S_Y(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial(\boldsymbol{\beta}, \boldsymbol{\theta})\partial(\boldsymbol{\beta}, \boldsymbol{\theta})'}\right]\right\}^{-1},$$

(iii) for large n, $n\hat{\sigma}^2 = S_Z(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ is distributed as $\sigma^2\chi^2(n-p-k)$ and is independent of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})'$, where k is the dimension of the parameter $\boldsymbol{\theta}$ associated with the correlation function for the noise process $\{\delta(s)\}$.

Using the independence of $\boldsymbol{Y}$ and $\boldsymbol{Z}$, we find that

$$\begin{aligned} E_\psi S_Y(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) &\simeq E_\psi S_Y(\boldsymbol{\beta}, \boldsymbol{\theta}) + \Big((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - (\boldsymbol{\beta}, \boldsymbol{\theta})\Big)' \Big\{V(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})\Big\}^{-1} \Big((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - (\boldsymbol{\beta}, \boldsymbol{\theta})\Big) \\ &\simeq \sigma^2 n + \sigma^2(p+k). \end{aligned}$$

Hence, from the last two terms of (2.9), we have

$$\begin{aligned} E_\psi(\hat{\sigma}^{-2} S_Y(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})) - n &= E_\psi(\hat{\sigma}^{-2}) E_\psi(S_Y(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})) - n \\ &\simeq \left(\sigma^2 \frac{n-p-k-2}{n}\right)^{-1} \sigma^2(n+p+k) - n \\ &= 2(p+k+1)\frac{n}{n-p-k-2}. \end{aligned}$$

The quantity

$$AIC = -2\log L_Z(\hat{\psi}) + 2(p+k+1)\frac{n}{n-p-k-2} \tag{2.11}$$

is an approximately unbiased estimate of the expected Kullback-Leibler information evaluated at $\hat{\psi}$.

The argument given above for spatial AIC relied on the validity of standard asymptotic theory for the maximum likelihood estimates of the parameters in the spatial model (2.1). In order for these results to hold, it is likely that an increasing sample size that both fills in and expands the domain under study is required. In the statistics literature, this is often referred to as infill and increasing (expanding) domain asymptotics. Unfortunately asymptotic theory for maximum likelihood estimates for unequally spaced data is not fully developed. Chapter 3 explores whether or not these assumptions hold by first developing complete asymptotic results for a regular lattice in one-dimension for an AR(1) process. The result is then generalized to observations collected at random (and independent) intervals in one-dimension.

2.3.2 Spatial Model Fitting

Traditionally, the fitting of the model (2.1) is accomplished in two steps (see, for example, Venables and Ripley (1999, pp. 439–444)). In the first step, explanatory variables for modeling the large scale variation are chosen via a model selection technique such as Akaike's Information (Corrected) Criterion (AIC_c) in (2.8) (Sugiura, 1978; Hurvich and ling Tsai, 1989). Second, the residuals from the model are examined for spatial correlation and a suitable family of correlation functions is chosen. The estimates of the parameters in the trend surface are updated using generalized least squares followed by maximum likelihood estimation of the parameters of the covariance function using the updated residuals. This two step estimation process is repeated until some suitable convergence criterion is attained. Since a correlation function is not identified in the selection of the explanatory variables in Step 1, AIC_c is implemented under the working assumption of independent residuals (Cressie, 1993; Haining, 1990).

A limitation of the model selection procedure described above is that it ignores potential confounding between explanatory variables and the correlation in the spatial noise process $\{\delta(s)\}$. Although it is convenient to select explanatory variables for the model before fitting a covariance function to the residuals, it is generally not a good idea to separate these two steps. The inclusion of one or more important explanatory variables may remove or reduce the correlation structure of the residuals from the model. For example, Ver Hoef et al. (2001) 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.

Model selection techniques for spatial models need to include the correlation structure to determine the best set of predictors. By computing the spatial AIC statistic described in Section 2.3.1 for all members of the candidate models, one can find a single "best" model or a set of models which fit the data well. This method attempts to strike a balance between the competing forces of large scale variability as modeled via the explanatory variables with small scale variability as modeled through the correlation in the residuals.

During model selection, the researcher should restrict their attention to maximum likelihood estimates of the model parameters. This is because the spatial AIC statistic (see Section 2.3.1) is derived using the principle of maximum likelihood. Alternative parameter estimation techniques, such as REML or the method of moments, do not maximize the likelihood function, precluding them from model selection. In principle, the profile likelihood achieves the maximum and is therefore appropriate for model selection. However, once a model has been selected, the researcher is free to re-estimate the model parameters using, for example, REML to improve the individual parameter estimates.

2.3.3 Other considerations

In Section 2.3.1 the spatial AIC statistic derived for the geostatistical model (2.1) required that the true model was a member of the family of candidate models, all of which were finite dimensional. However, in many applications (McQuarrie and Tsai, 1998; Burnham and Anderson, 2002), the AIC selection procedure enjoys additional optimality properties regarding the choice of a finite-dimensional model when the true model is in fact infinite dimensional. This includes the notion of efficiency for prediction in time series models and optimal signal-to-noise ratios for linear models (McQuarrie and Tsai, 1998).

AIC and other information-based criteria such as BIC and HQ have an objective function consisting of two pieces (Kass and Raftery, 1995; McQuarrie and Tsai,

1998). The first is related to $-2\times$log-likelihood, which is a measure of the quality of fit of a model, and the second is a penalty factor for the introduction of additional parameters into the model. The principle of minimum description length (MDL), an idea developed by Rissanan in the 1980s, e.g., Rissanen (1986), also contains two similar pieces, but is motivated by different ideas. MDL attempts to achieve maximum data compression by the fitted model.

The idea behind MDL is to decompose the code length of the "data" into two pieces (see the survey paper by Lee (2001) for more details). Roughly speaking, the code length of the "data" is the amount of memory required to store the data. Typically the code length of the data can be decomposed into the sum of the code length of the fitted model and the code length of the data given the fitted model, i.e.,

$$L(\text{"}data\text{"}) = L(\text{"}fitted\ model\text{"}) + L(\text{"}data\ given\ fitted\ model\text{"}).$$

Here $L(\text{"}fitted\ model\text{"})$ might be interpreted as the code length of the model parameters and $L(\text{"}data\ given\ fitted\ model\text{"})$ as the code length of the residuals from the fitted model. It follows that a more complex model is chosen provided there has been a compensating decrease in the code length of the residuals. According to the MDL principle, the best model is the one producing the shortest code length for the data. The attraction of this procedure is that the data is being compressed in the most efficient manner possible and the notion of a true model at any level is not required.

The code length of the fitted model based on the MLE, $\hat{\psi}$, can be approximated by $L(\text{"}fitted\ model\text{"}) \simeq \frac{1}{2}(p+k+1)\log_2 n$. The code length of the data given the model based on $\hat{\psi}$ is approximated by $\log_2 L(\hat{\psi})$. Adding these terms together and rescaling, the minimum description length is defined as

$$\text{MDL} = \frac{1}{2}\left(-2\log(L_Z(\hat{\psi})) + \log(n)(p+k+1)\right).$$

The only difference between the value of the spatial AIC statistic and 2·MDL is the magnitude of the penalty term coefficient. For spatial AIC, the leading coefficient is of order 2 compared to $\log(n)$ for 2·MDL. For sample sizes greater than 8, the penalty for 2·MDL is larger. For example, when $n = 100$, $p = 4$, and $k = 2$ the penalty coefficients are 2 and 4.60, respectively. MDL generally selects more parsimonious models, i.e., models with fewer explanatory variables.

Bayesian model averaging is an alternative approach to model selection and prediction (Hoeting et al., 1999). The idea of Bayesian model averaging is to average across several models instead of selecting one model. In computing the average, each model is weighted by its posterior model probability, a measure of the degree of model support in the data. Empirical and theoretical results over a broad range of model classes indicate that Bayesian model averaging can provide improved out-of-sample predictive performance as compared to single models. For the geostatistical model in (2.1), Thompson (2001) showed that Bayesian model averaging can offer improved predictive performance as compared to the single models that are selected when spatial correlation is ignored. However, the gains are modest in the simulations that were explored.

2.4 Simulations

To explore the impact of ignoring spatial correlation on model selection, we carried out two simulations to evaluate the model selection methodology proposed in Section 2.3.2. Of primary importance was comparing which explanatory variables were selected using the traditional independent AIC model selection procedure (which ignores spatial correlation) to those selected using the spatial AIC approach and the MDL approach described in Section 2.3.3. We also conducted simulation studies to characterize the strength of the predictive ability when spatial correlation was included in the selection process of explanatory variables.

2.4.1 Simulation 1

For the first simulation we generated five possible explanatory variables, $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$, and $\boldsymbol{X}_5$. Each explanatory variable was independently generated from a standardized Student's t distribution with 12 degrees of freedom, $\boldsymbol{X}_i \sim \sqrt{\frac{10}{12}} t_{12}$ for $i = 1, \ldots, 5$. The explanatory variables were held fixed for all simulations.

For a fixed set of $n = 100$ sites randomly located over a 10×10 square, data were simulated from the model

$$\boldsymbol{Z} = 2 + 0.75\boldsymbol{X}_1 + 0.50\boldsymbol{X}_2 + 0.25\boldsymbol{X}_3 + \boldsymbol{\delta}, \tag{2.12}$$

where $\boldsymbol{\delta}$ was a Gaussian random field with mean zero, $\sigma^2 = 50$, and autocorrelation Matérn with parameter vector $\boldsymbol{\theta} = (4.0, 1.0)'$. Five hundred replicates were simulated with a new Gaussian random field generated for each. The largest component of the "signal" in (2.12) is associated with X_1 which is three times the "strength" of X_3. Thus one expects that the majority of models selected should at least include X_1.

With five possible explanatory variables, there are $2^5 = 32$ possible combinations of explanatory variables, including the intercept-only model. For each realization, we computed the AIC_c statistic for all 32 possible models. For the traditional method, the AIC statistic was calculated using (2.7) with $k = 0$. We refer to this as the independent AIC approach. The spatial AIC results were calculated using (2.7) with $k = 2$. Further details on the simulation set-up and additional simulation results are given in Thompson (2001).

2.4.1.1 General Results

Table 2.1 compares the models selected by the spatial and independent AIC approaches. When independence is assumed, the AIC statistic selects the true model

Table 2.1: Model Selection Results for the Random Pattern with $\boldsymbol{\theta} = (4.0, 1.0)'$. Independent AIC, Spatial AIC, and MDL report the percentage of simulations that each model was selected. Of the 32 possible models, the results given here include only those with 10% or more support for one of the models.

Variables in Model	Spatial AIC	Independent AIC	MDL
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$	56.0	2.4	40.4
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_5$	14.4	0.2	4.2
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$	10.8	0.2	0.8
$\boldsymbol{X}_1, \boldsymbol{X}_2$	10.2	8.4	46.4
Intercept only	0.0	26.8	0.0
$\boldsymbol{X}_1$	0.4	14.2	1.2
$\boldsymbol{X}_2$	0.0	13.8	0.2

$(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$ only 12 out of 500 simulations (2.4%) while the intercept-only model is selected 134 out of 500 simulations (26.8%). Over all 500 simulations, the independent AIC statistic selected models that included both explanatory variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ only 15.8% of the time. These results provide a vivid example of the drawbacks of the standard model selection approach for spatially correlated data. In total, the first explanatory variable is in 40.2% of the selected models, and the second explanatory variable is included in 35.4% of the models.

The spatial AIC method outperformed the independent AIC method. The true model was selected in 56.0% of the simulations. When the true model was not selected, spatial AIC tended to overestimate the number of parameters in the model, selecting models with one or two extra variables (28.4%). In contrast to indpendent AIC independence, the first explanatory variable is in 100% of the selected models and the second explanatory variable is included in 98.6% of the models.

Figure 2.3 illustrates the necessity of including spatial correlation during model selection. The top panel lists the models from smallest to largest average AIC over all 500 simulations. The horizontal axis list the variables included in the model

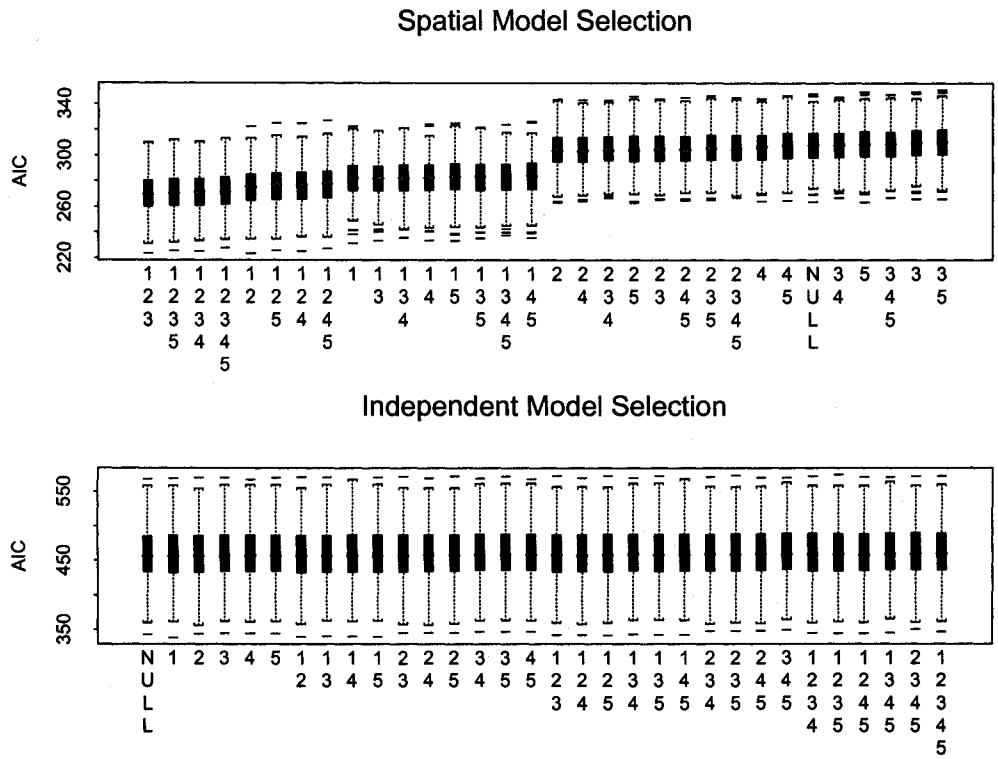

Figure 2.3: AIC values for the spatial and independent AIC selection strategies. Note that the models for the spatial AIC method have been ordered from smallest to largest average AIC over all 500 simulations. The horizontal axis lists the variables included in each model where NULL refers to the intercept-only model.

where NULL refers to the intercept-only model. Note that the model with the smallest average AIC is the true model $(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$. All of the first 16 models listed include X_1, while the first eight models also include X_2. In sharp contrast, the boxplots for the independence assumption during model selection are virtually identical. Although the models are listed from most to least parsimonious, any rearrangement would look nearly identical. The lack of trend in this plot illustrates that ignoring spatial dependence during variable selection may lead to selection of an inappropriate model.

Table 2.1 also illustrates MDL's ability to select the appropriate model when spatial correlation is accounted for during variable selection. Although it only selects the "true" model for 40.4% of the simulations, it selects the model containing only $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ 46.4% of the time. These results are consistent with the idea that MDL more strongly penalizes models with a large number of explanatory variables and thus tends to select more parsimonious models. Also note that MDL selects one of three models for more than 90% of the simulations.

To further evaluate the performance of the spatial AIC strategy, we performed additional simulations using different values of the Matérn correlation function parameters, $\boldsymbol{\theta}$. For the first experiment, we fixed the range parameter, $\theta_1 = 4.0$, and varied the smoothness parameter, $\theta_2 = (0.50, 0.75, 1.00, 4.00)$. For the second experiment, the smoothness parameter was fixed, $\theta_2 = 1.00$, and the range parameter was varied, $\theta_1 = (2.0, 4.0, 6.0, 8.0)$. For each experiment, 100 sets of random observations were generated according to (2.12) at the same sampling locations as in the previous simulation study, and the two model selection techniques were compared.

Table 2.2 illustrates how varying the smoothness parameter, θ_2, influenced model selection. As θ_2 was increased from 0.5, which corresponds to an exponential autocorrelation function, the AIC for the spatial model approach tended to pick the true model more frequently. In addition, the tendency of overfitting was enhanced. Note that this second result is not necessarily undesirable because the inclusion of additional explanatory variables often improves the overall predictive capabilities of the model. In contrast, assuming independence during model selection tended to lead to under fitting, often with one or no explanatory variables. In fact, the traditional approach appeared to be invariant over all values of θ_2. Table 2.3 shows similar results when the range parameter, θ_1, was varied. Inclusion of spatial correlation lead to the correct model or an over fit being selected as θ_1 was increased. Ignoring spatial dependence during model selection tended to under fit the model (often the intercept-only model was selected).

Table 2.2: Model selection results for the Matérn family with various smoothness parameter values. For all simulations the range parameter is fixed at $\theta_1 = 4$. Listed is the percentage of simulations that each model was selected using the spatial AIC and independent AIC methods. Of the 32 possible models, the results given here include only those with 10% or more support for one of the models.

Variables in Model	$\theta_2 = 0.50$		$\theta_2 = 0.75$		$\theta_2 = 1.00$		$\theta_2 = 4.00$	
	spat	ind	spat	ind	spat	ind	spat	ind
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$	15	3	35	1	56	2	62	3
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_5$	1	0	3	0	14	0	14	0
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$	3	0	7	0	11	0	18	1
$\boldsymbol{X}_1, \boldsymbol{X}_2$	22	7	20	4	10	8	0	8
Intercept only	4	30	0	32	0	27	0	22
$\boldsymbol{X}_1$	17	17	12	20	0	14	0	20
$\boldsymbol{X}_2$	4	12	0	11	0	14	0	8

Table 2.3: Model selection results for the Matérn Family with various range parameters values. For all simulations the range parameter is fixed at $\theta_2 = 1.00$. Listed is the percentage of simulations that each model was selected using the spatial AIC and independent AIC methods. Of the 32 possible models, the results given here include only those with 10% or more support for one of the models.

Variables in Model	$\theta_1 = 2$		$\theta_1 = 4$		$\theta_1 = 6$		$\theta_1 = 8$	
	spat	ind	spat	ind	spat	ind	spat	ind
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$	25	2	56	2	66	5	71	4
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_5$	6	0	14	0	11	0	10	0
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$	7	0	11	0	19	2	14	1
$\boldsymbol{X}_1, \boldsymbol{X}_2$	22	8	10	8	0	14	0	21
Intercept only	0	22	0	27	0	25	0	12
$\boldsymbol{X}_1$	14	20	0	14	0	15	0	23
$\boldsymbol{X}_2$	1	8	0	14	0	11	0	9

2.4.1.2 Impact of Sampling Pattern

Thompson (2001) demonstrated that the use of spatial AIC appears to be enhanced when the sampling pattern includes both some closely spaced and more distant pairs of sample locations. Note that all of the simulations in Section 2.4.1.1 used the same 100 sampling locations that were randomly placed over a 10×10 square. To explore the impact of the sampling pattern, the main simulation was repeated using four additional sample patterns: highly clustered, lightly clustered, regular, and grid. Figure 2.4 displays all five patterns. For each sampling pattern (including the random pattern) 100 new realizations of the Gaussian random field were generated using (2.12) with $\boldsymbol{\theta} = (4.00, 1.00)'$ and both the spatial and independent AIC statistics computed for each.

Table 2.4 summarizes the models that were most frequently selected. The highly and lightly clustered patterns selected the true model in over 65% of the simulations. There is evidence that as the sampling pattern provides less information at small distances, the selection of the correct set of explanatory variables becomes more challenging. Indeed, for the grid design the correct model was only selected in 16% of the simulations.

The independent AIC approach gave similar results to those for the random pattern given in Table 2.1. Over all five sampling patterns, the independent AIC approach selected the correct model in less than 1% of the simulations and the model with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ was selected in 5% of the simulations.

2.4.1.3 Mean Square Prediction Error

Another measure of the importance of including spatial correlation during model selection is the concept of mean square prediction error (MSPE). We can evaluate MSPE for the simulated data because we know the true underlying model (2.12). MSPE is the average squared difference between the actual and predicted

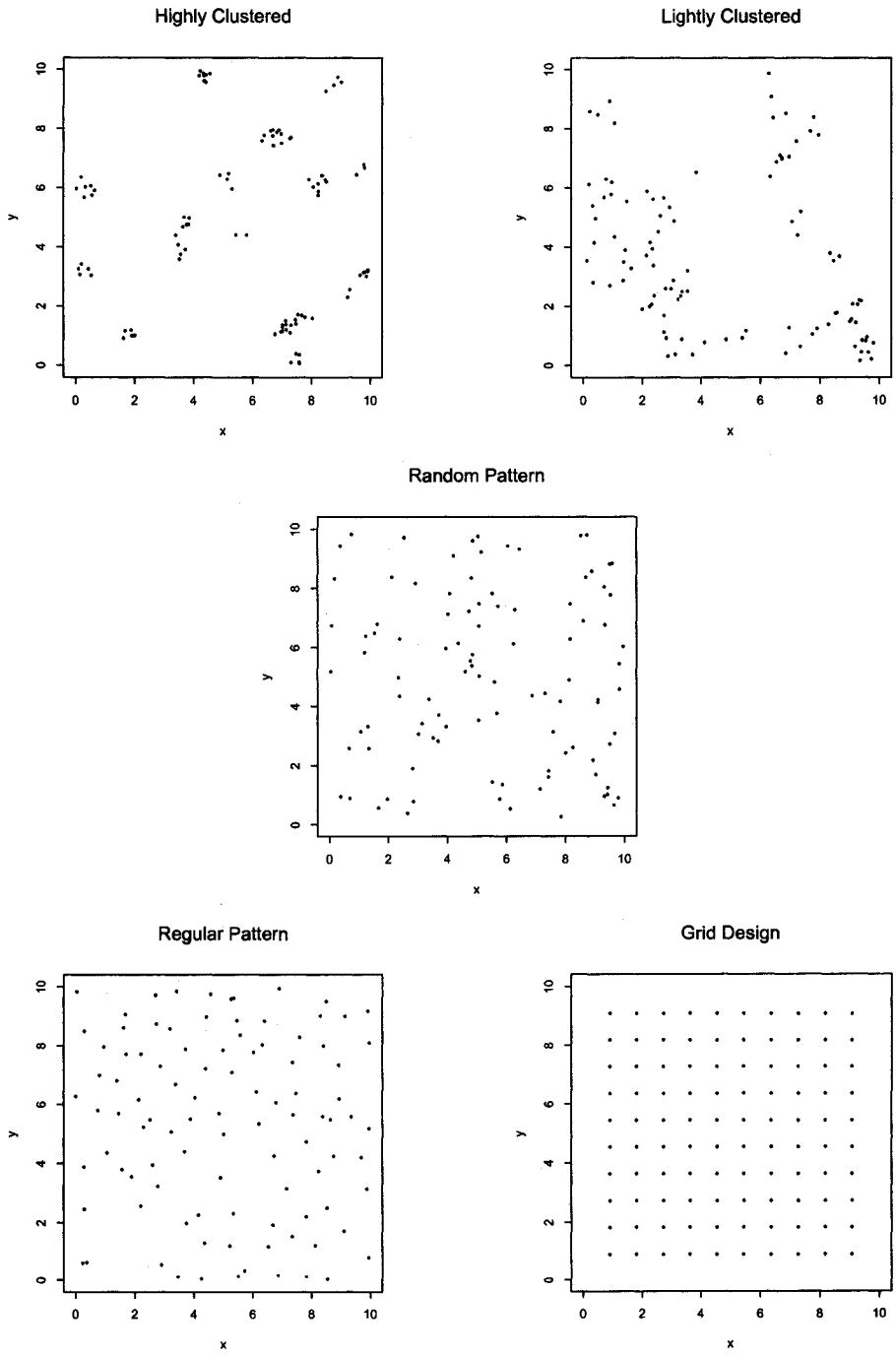

Figure 2.4: Five Sampling Patterns

Table 2.4: Spatial AIC model selection results for five different sampling patterns. Each column reports the percentage of simulations that each model was selected. Of the 32 possible models, the results given here include only those with 10% or more support for at least one of the sampling patterns.

Variables in Model	Highly Clustered	Lightly Clustered	Random	Regular Pattern	Grid Design
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$	73	65	46	43	16
$\boldsymbol{X}_1, \boldsymbol{X}_2$	0	2	18	21	35
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$	12	13	8	8	3
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_5$	10	13	11	7	7

values at a new collection of locations such that

$$\text{MSPE} = \frac{1}{n}\sum_{j=1}^{n}(Z_j - \hat{Z}_j)^2. \tag{2.13}$$

Here $\hat{Z}_j$ is the universal kriging predictor for the j^{th} prediction location using the maximum likelihood estimate of the parameter vector $\boldsymbol{\psi}$ and Z_j is the true value at location j. Small values of MSPE indicate predicted values are close to the true values on average, where an MSPE of exactly zero corresponds to perfect prediction. What we expected to see was an MSPE that was systematically smaller for spatial AIC compared to independent AIC.

Returning to the random sampling pattern, 100 new locations were randomly selected over the 10×10 square. At each of these new locations the response was simulated using (2.12) [conditioned on the original 100 observations]. Next, $\hat{\boldsymbol{Z}}$ was computed using the selected model and the corresponding maximum likelihood estimates for $\boldsymbol{\beta}, \sigma^2$, and $\boldsymbol{\theta}$ for both spatial and independent AIC. Finally, MSPE was calculated for each of the 500 simulations.

Figure 2.5 illustrates the improvement made by incorporating spatial correlation into the model selection process. The mean MSPE for the spatial AIC method was 4.57 compared to 5.50 for the independent AIC selection method (an improvement

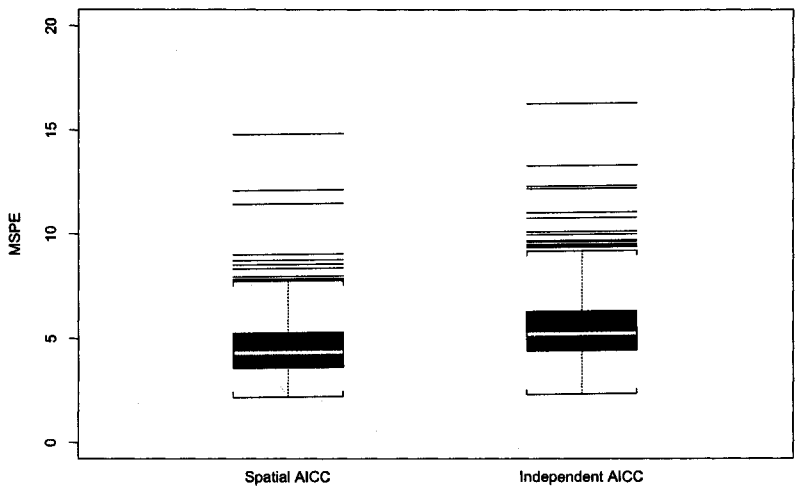

Figure 2.5: Comparison of the Mean Squared Prediction Error (MSPE) for both selection methods based on 500 simulations.

of 16.9%). Over the set of 500 simulations, the two methods selected the same model only 11 times. When these simulations were removed from the data set, the improvement in mean MSPE increases to 17.3%. It should be noted that when spatial correlation was ignored altogether, i.e., independent error structure assumed *after* model selection, the mean MSPE was 39.6.

2.4.2 Simulation 2

The results from the first simulation study clearly demonstrate that (possible) spatial correlation should be accounted for during model selection. But there is a deficiency in the set-up of this simulation; the ratio of the "signal" of the mean trend is small compared to that of the noise process. Note that the explanatory variables, $\boldsymbol{X}_i, i = 1, \ldots, 5$ are realizations of a mean zero white noise process. Figure 2.6 illustrates a single realization of a random field generated by this process. Note that the surface is very "rough" and that locations at close proximity can have

remarkably different response values. Many explanatory variables do not behave this way. For example, the observed canopy density at two locations in a forest separated by 10 meters are unlikely to be very different, i.e., positively correlated. As the distance between locations increases the observed canopy densities will become nearly independent. Furthermore, the mean surface of simulation 1 is itself a white noise process because the sum of two or more iid processes is also iid. Therefore, the only "signal" present in simulation 1 was imposed by the error process, i.e., the correlation among the residuals. Hence it is not surprising that the independent AIC method was unable to select the correct explanatory variables.

To generate a more interesting mean surface it was decided that the explanatory variables should be "smoother", i.e., have a non-zero trend surface. For example, one can define $\boldsymbol{X}$ as a quadratic function, i.e., a two-dimensional "smooth" surface. We chose to let each $\boldsymbol{X}_i$ be an independent realization of the model $bX_i \sim \mathcal{N}(0, \tau^2 \text{Matérn}(4,1))$ where $\tau^2 = 1$. Figure 2.7 is one such realization smoothed over a 10×10 lattice. Note that for locations at close proximity the values of the response surface are similar. Also observe that the correlation structure of each $\boldsymbol{X}_i$ is identical to that of the noise process $\boldsymbol{\delta}$. (This was done to simplify computing S/N. See below.)

2.4.2.1 Signal-to-Noise Ratio

We define the signal-to-noise ratio as the square root of the ratio of the variability of the mean structure (large scale variability) to the variability of the noise (small scale variability). We write this as

$$\begin{aligned} \text{S/N} &= \left(\frac{\text{Var}\left(\sum_{i=1}^{n} (Z_i - \delta_i)\right)}{\text{Var}\left(\sum_{i=1}^{n} \delta_i\right)} \right)^{1/2} \\ &= \left(\frac{\text{Var}\left(\sum_{i=1}^{n} (\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi})\right)}{\text{Var}\left(\sum_{i=1}^{n} \delta_i\right)} \right)^{1/2} \\ &= \left(\frac{\text{Var}\left(\mathbf{1}'\boldsymbol{X}\boldsymbol{\beta}\right)}{\text{Var}\left(\mathbf{1}'\boldsymbol{\delta}\right)} \right)^{1/2}, \end{aligned} \tag{2.14}$$

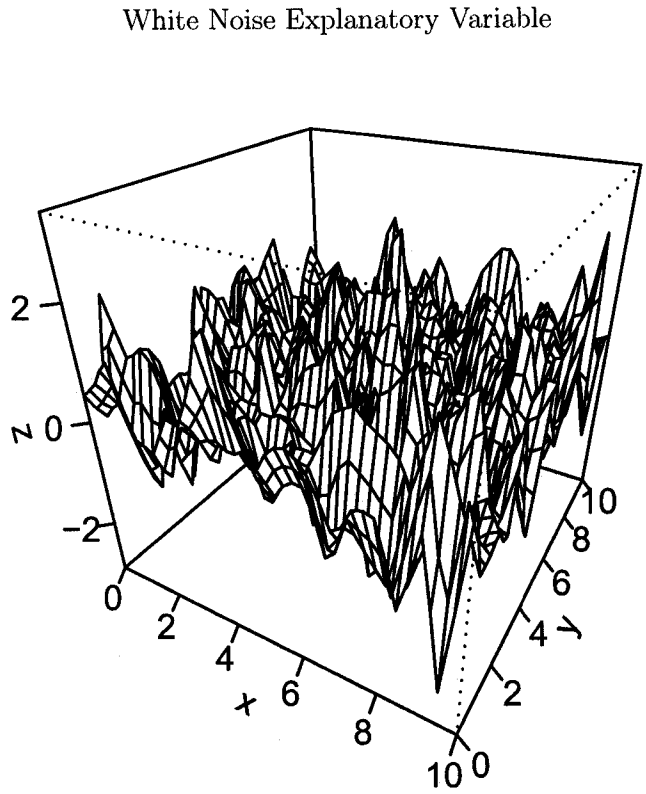

Figure 2.6: A (smoothed) realization of a white noise explanatory variable $\boldsymbol{X}$ such that $\boldsymbol{X} \sim \sqrt{\frac{10}{12}} t_{12}$.

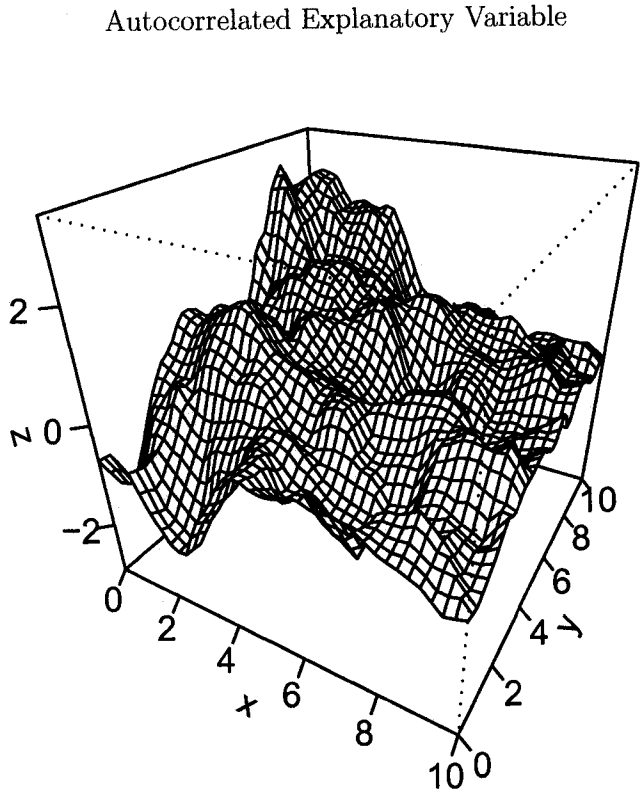

Figure 2.7: A (smoothed) realization of an autocorrelated explanatory variable $\boldsymbol{X}$ such that $\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 \text{Matérn}(4, 1))$, where $\sigma^2 = 1.0$.

where $\mathbf{1}'$ is an n-vector of ones. As an exercise, let's compute the S/N for the first simulation. Recall that the explanatory variables are independent and identically distributed, i.e., $\boldsymbol{X}_i \overset{\text{iid}}{\sim} \sqrt{\frac{10}{12}} t_{12}$. This implies

$$\mathrm{E} X_{ji_1} X_{ki_2} = \begin{cases} 1 & \text{if } i_1 = i_2 \text{ and } j = k, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the numerator (inside the radical) is $\mathrm{Var}\,(\mathbf{1}'\boldsymbol{X}\boldsymbol{\beta}) = n(\beta_1^2 + \beta_2^2 + \beta_3^2)$ and is function of $\boldsymbol{\beta}$. The denominator can be evaluated as follows:

$$\begin{aligned} \mathrm{Var}\,(\mathbf{1}'\boldsymbol{\delta}) &= \mathrm{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\delta_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\,(\delta_i\delta_j) \\ &= \sigma^2\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[\rho(||\boldsymbol{s}_i - \boldsymbol{s}_j||;\boldsymbol{\theta})\right], \end{aligned}$$

where $\mathbf{s}_i$ and $\mathbf{s}_j$ are sampling locations. Therefore the denominator is a function of the spatial parameters, $\boldsymbol{\theta}$. Evaluating the numerator and denominator with the parameter values $\boldsymbol{\beta} = (2.00, 0.75, 0.50, 0.25)', \sigma^2 = 50$, and $\rho(\cdot;\boldsymbol{\theta}) = \text{Matérn}(\cdot; 4.0, 1.0)$, we find that S/N does not exceed 0.50 for any of the sampling patterns. This is quite low and indicates that the contribution by the explanatory variables to the overall process is small. This provides a possible explanation as to why independent AIC did so poorly.

Replacing the original $\boldsymbol{X}_i$'s with the autocorrelated explanatory variables, the numerator of S/N is written as

$$\mathrm{Var}(\mathbf{1}'\boldsymbol{X}\boldsymbol{\beta}) = (\beta_1^2 + \beta_2^2 + \beta_3^2)\sum_{i}^{n}\sum_{j}^{n}\mathrm{E}\left[\rho(||\boldsymbol{s}_i - \boldsymbol{s}_j||;\boldsymbol{\theta})\right].$$

Combining this result with the denominator result simplifies the evaluation of the signal-to-noise ratio to $\text{S/N} = ((\beta_1^2 + \beta_2^2 + \beta_3^2)\sigma^{-2})^{1/2}$. Hence, for fixed $\boldsymbol{\beta}$, the S/N can be set to any value by simply scaling σ^2.

It should be noted that the reported signal-to-noise ratios in Section 2.4.2.2 are approximate. To compute S/N we are assuming that each $\boldsymbol{X}_i$ is random. For this problem, and indeed for real world applications, we observe a single realization of each explanatory variable (held fixed for all subsequent simulations). Hence, to compute the actual S/N one would need to evaluate (2.14) using the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\theta}}$.

2.4.2.2 Model Selection as a Function of S/N

Following the same procedure outlined in Section 2.4.1, a random Gaussian field was generated using (2.12) and the autocorrelated explanatory variables and evaluated at each of the 100 locations (for the random sampling pattern only). We performed model selection using both spatial and independent AIC for 100 separate realizations at each level, S/N = $\{0.25, 0.50, 1.00, 2.00, 4.00\}$.

Table 2.5 clearly illustrates the impact of S/N on model selection. For small S/N, spatial AIC identifies that either (or both) $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the only important significant explanatory variables (under fitting). As the S/N increases, spatial AIC selects the "true" model more often and, consistent with the nature of AIC, favors overfitting the model for large S/N. Contrast this with independent AIC which favors the full model at all levels of S/N. For large S/N, independent AIC consistently selects larger models.

2.5 Examples

In this section we present results using actual data. The first example explores the abundance of the orange-throated whiptail lizard *Cnemidophorus hyperythrus* of southern California. These data were previously analyzed by Hollander et al. (1994) and Ver Hoef et al. (2001). It turns out that the sampling locations for the lizard data are highly clustered. We use our model selection results to set up an additional simulation to further explore the importance of using spatial AIC

Table 2.5: Model selection results for five different levels of S/N using the random sampling pattern. Each column reports the percentage of simulations that each model was selected. Of the 32 possible models, the results given here include only those with 10% or more support for at least one S/N level.

Variables in Model	Signal-to-Noise Ratio, S/N									
	0.25		0.50		1.00		2.00		4.00	
	spat	ind	spat	ind	spat	ind	spat	ind	spat	ind
X_1, X_2, X_3	6	7	23	8	61	13	77	16	77	17
X_1, X_2, X_3, X_4, X_5	0	13	2	17	4	27	5	29	5	33
X_1, X_2, X_3, X_5	1	6	5	9	9	16	10	20	10	21
X_1, X_2, X_3, X_4	1	7	4	9	7	16	8	24	8	24
X_1, X_2	12	1	43	2	16	2	0	1	0	0
X_1	22	3	7	0	0	0	0	0	0	0
X_2	14	1	2	0	0	0	0	0	0	0
Intercept only	13	0	0	0	0	0	0	0	0	0

for model selection. The second example is a model selection exercise using water chemistry data collected throughout the state of Maryland. The general goal of this analysis was to select the best model for each response variable. The key extension in the second example is that the definition of "distance" between sites, and hence the autocorrelation structure, is also in the set of competing models. All of the analyses presented thus far have involved Euclidean distance ("as the crow flies"). Stream networks provide a unique challenge in that movement can be restricted to within the stream network ("as the fish swims" or "as the water flows").

2.5.1 Orange-throated Lizard of Southern California

We applied the model selection strategy to the whiptail lizard data previously analyzed by Hollander et al. (1994) and Ver Hoef et al. (2001). The data set consists of abundance data for the orange-throated whiptail lizard of southern California. A total of 256 locations in 21 regions were used for trapping. Each observation consists of the average number of lizards caught per day at each location. After removing sites where no lizards were caught (to maintain distributional assumptions), a total

of 148 observations remained for the abundance analysis. Figure 2.8 shows that the pattern of the sites where the lizards were observed was highly clustered. A log transformation was applied to the response, average number of lizards caught per day, to allow for the use of a Gaussian random field.

There are a total of 37 explanatory variables available including information on vegetation layers, vegetation types, topographic position, soil types, and abundance of ants. This corresponds to approximately 2^{37} or 1.374×10^{11} possible models [without interactions]. To make the analysis tractable, the number of explanatory variables was reduced to six. See Ver Hoef et al. (2001) for further details about preliminary explanatory variable selection.

The subset of explanatory variables used in the analysis were *Crematogaster* ant abundance (3 categories - low, medium, and high), log percent sandy soil, elevation, a binary indicator variable describing whether or not the rock was bare, percent cover, and log percent chapparal plants. Ant abundance is a categorical variable and has five (5) unique modeling subsets. This leads to a total of $5 \times 2^5 = 160$ competing models.

All 160 unique models were fit to the data using the strategy outlined in Section 2.3.2. We assumed a Matérn autocorrelation structure (without nugget) for each model. For comparison, the independent AIC model selection approach was also applied to the data set. Table 2.6 summarizes the top 3 models selected when employing each strategy. For each model the corresponding rank (by AIC) under the opposing strategy is also listed. The two methods select very different models. When spatial dependence is incorporated into the selection process, very parsimonious models are chosen and are consistent with the results of Ver Hoef et al. (2001). The traditional approach leads to much more complicated models. By initially assuming independent covariates, the selection process is trying to compensate for correlation in the error structure by incorporating too many explanatory variables.

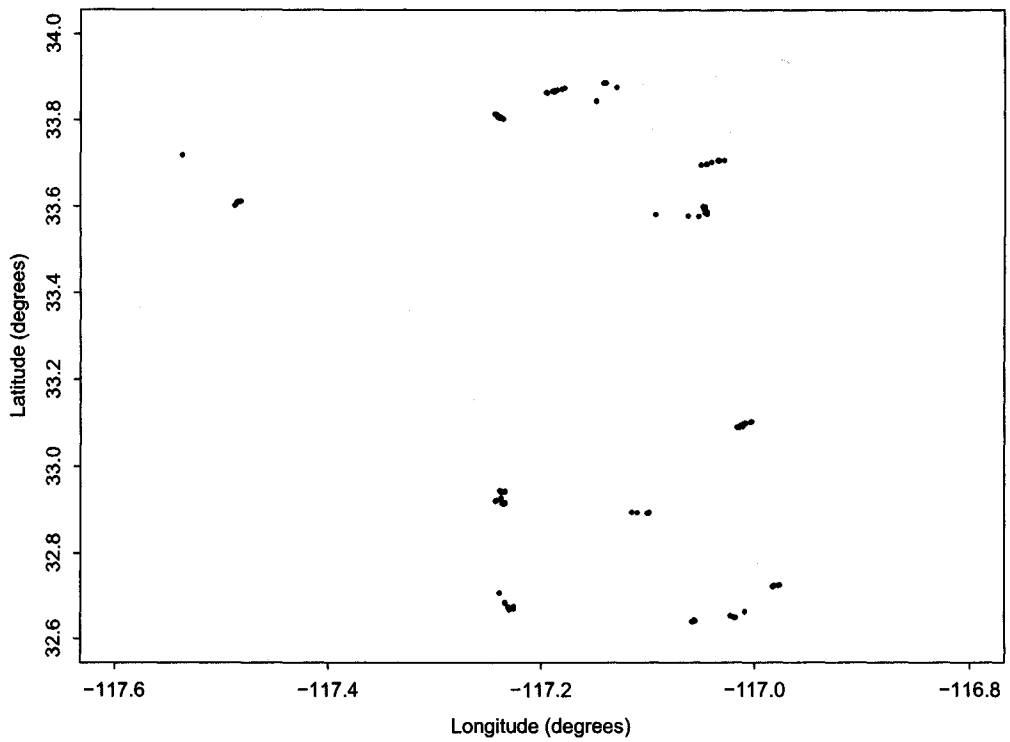

Figure 2.8: Locations in southern California where the whiptail lizard was observed: ($n = 148$).

Table 2.6: Model selection results for the whiptail lizard data. Listed are the explanatory variables selected using AIC as the selection criterion. The rank of the model (by AIC) is provided under both model selection strategies. Ant_1 corresponds to low abundance and Ant_2 corresponds to medium abundance.

Predictors	Spatial Rank	Independent Rank
Ant_1, % sand	1	66
Ant_1, Ant_2, % sand	2	56
Ant_1, % sand, % cover	3	59
Ant_1, Ant_2, % sand, %c over, elevation, barerock, % chaparral	41	1
Ant_1, Ant_2, % sand, elevation, barerock, % chaparral	33	2
Ant_1, % sand, % cover, elevation, barerock, % chaparral	38	3

In fact, the full model has the smallest AIC when the correlation structure is ignored during model selection. It should be noted that the top three models selected by the MDL method exactly matched those selected by spatial AIC.

2.5.1.1 Lizard Simulation Data

The highly clustered sample locations for the lizard data motivated the following simulation study (see Figure 2.8). First we define the "true" model as the model selected by spatial AIC, i.e., log(abundance) = $f(\text{Ant}_1, \%\text{sand})$ with Matérn autocorrelation function. The maximum likelihood estimates of $\Psi = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ were used to parameterize the model. A distinct advantage to this approach is that we need not assume anything about the nature of the explanatory variables. Each variable may or may not be autocorrelated and, perhaps, correlated with one another, thus providing a powerful paradigm to examine the impact of ignoring spatial correlation during model selection.

Using the selected model we generated 100 realizations of the Gaussian random field at the 148 locations used to fit the "true" model. For each realization we

Table 2.7: Model selection results for the simulated lizard data set. Listed is the percentage of time each model was selected out of 100 simulations using AIC. Only models with at least 5% support are presented.

Variables in Model	Spatial AIC	Independent AIC
$\boldsymbol{X}_1, \boldsymbol{X}_4$ (*True model*)	75	5
$\boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}_5$	8	5
$\boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}_5, \boldsymbol{X}_8$	1	10
$\boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}_5, \boldsymbol{X}_7$	0	6
$\boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}_6$	2	5
$\boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}_8$	3	5

performed model selection over the complete set of 160 models using both spatial and independent AIC. Table 2.7 lists the models most commonly selected. Clearly spatial AIC outperforms independent AIC, selecting the "true" model 75% of the time. Independent AIC once again generally selects models with additional explanatory variables (over fitting). Note that nearly 90% of the models selected using spatial AIC are listed while less than 40% of the selected models for independent AIC are identified. This implies that the variability associated with model selection is higher for independent AIC. Overall, the lizard simulation results are consistent with the simulation results presented in Section 2.4.1.2 for the highly clustered sample pattern.

2.5.2 Maryland Biological Stream Survey

The second example attempts to model water chemistry variables collected along streams throughout the state of Maryland. An important complication arose because the distance between two sites can be defined differently depending on whether or not one restricts movement to be *within* the stream network. As a consequence, there were competing distance measures to be used in the correlation function. It was decided that the best model be identified for each distance measure

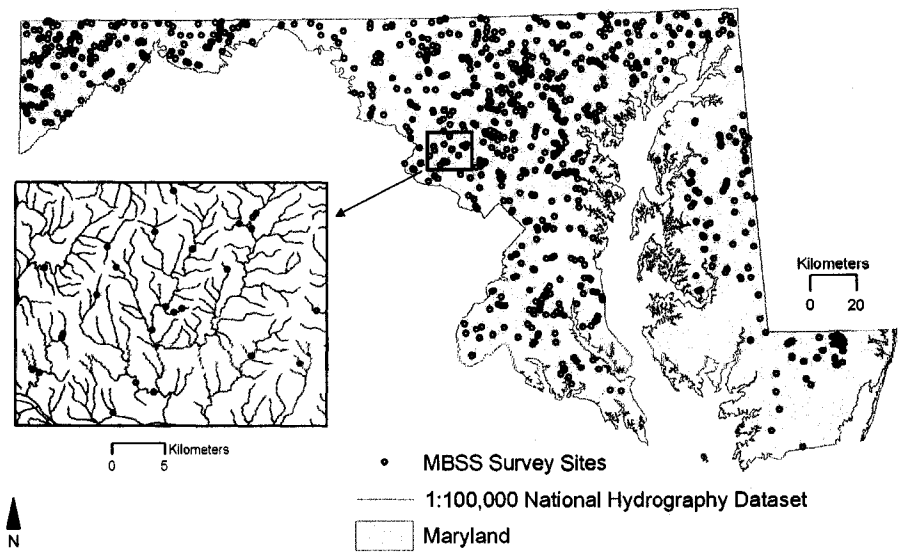

Figure 2.9: Spatial Distribution of the Maryland Biological Stream Survey (MBSS) Data.

separately. Note that the spatial model selection procedure can easily incorporate competing correlation functions but it is unclear if spatial AIC can properly differentiate between different distance measures used within the same correlation function.

The Maryland Biological Stream Survey (MBSS) was conducted by the Maryland Department of Natural Resources over the years 1995, 1996, and 1997. A stratified probability-based random survey design was employed collecting observations at over 880 locations spread throughout 17 interbasins. Figure 2.9 illustrates the spatial relationship of the sampling locations. Data collected included the acid neutralizing capacity, dissolved organic carbon, whether or not the surrounding landscape was urban or rural, etc. Peterson et al. (2006) provide a full description of the modeling process including a complete list of the response and explanatory variables. What follows is a brief accounting of the unique features of the stream network problem.

2.5.2.1 Hydrologic Distance and Flow Connectivity

Perhaps the greatest complication involved in this analysis is the nature of a stream networks. All of the simulations and examples presented thus far have been in two spatial dimensions and the distance between sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ has been defined by Euclidean distance, $\Delta_{ij} = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$. Conceptually this equates to "as the crow flies" across a two dimensional landscape and can be easily modified to accommodate one and three spatial dimensions. Indeed one can imagine a stream network as part of a two dimensional surface. If travel across land is allowed then the distance between two locations along the stream network is consistent with Euclidean distance. But what if travel between the two sites is restricted to the stream network within the two dimensional landscape?

Conceptually we are proposing a distance measure akin to "as the fish swims". Two locations that are geographically close may now be far apart hydrologically or, perhaps, not connected at all (such as the head waters on either side of the continental divide). We define *symmetric hydrologic distance* as the "in-stream" distance between locations. In recent years there has been significant development of Graphical Information System (GIS) tools that allow users to measure (approximate) the hydrologic distance between any two locations within a stream network (Theobald et al., 2005). To further complicate the issue we can also incorporate the direction of flow. For example, suppose a pollutant has been introduced to the stream network at a specific location (point source). If its transport mechanism is solely dependent on the direction of (water) flow, only locations downstream will be influenced by the presence of the pollutant. Thus downstream sites are flow connected but upstream sites are not. This relationship is referred to as *asymmetric flow distance.* Figure 2.10 demonstrates each distance measure used in the analysis.

Ver Hoef et al. (2006) have developed a series of correlation structures that can accommodate both symmetric and asymmetric hydrologic distance. They show

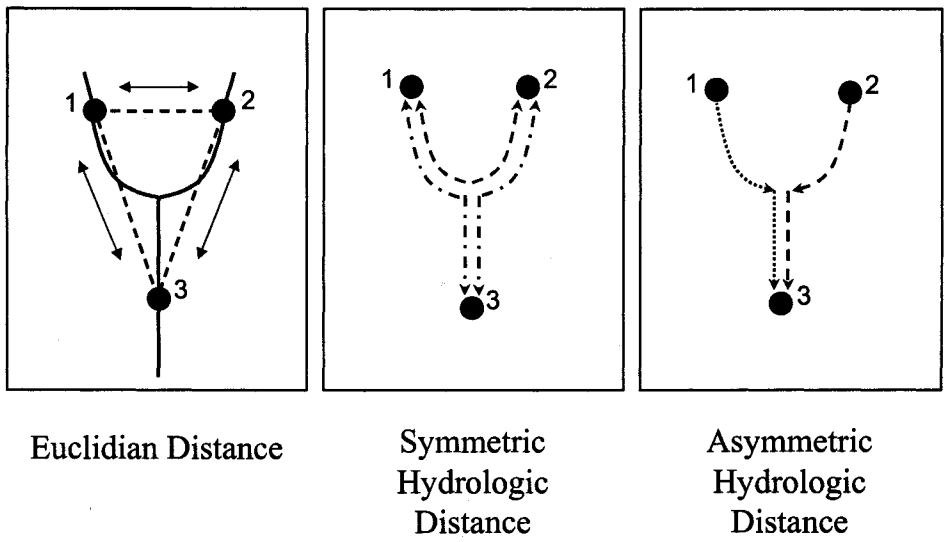

Figure 2.10: Distance measures for stream networks. The figure represents a confluence of two streams coming together. The are two upstream observation sites (1 and 2) on separate tributaries and a single downstream site at 3. The dashed lines of the first panel illustrate the Euclidean distance between sites. The second panel illustrates symmetric hydrologic distance. Note that travel is allowed to move both up and down stream. The third panel illustrates asymmetric hydrologic distance. Site 3 is flow connected to both sites 1 and 2 but sites 1 and 2 are not flow connected with one another.

that the exponential correlation function, among others, is well suited for working with hydrologic distance. The symmetric case is very similar to the traditional Euclidean distance case while the asymmetric case requires the introduction of a carefully constructed weight matrix to preserve positive definiteness of the variance-covariance matrix.

2.5.2.2 Summary of Key Analysis Results

Models were developed for eight different water chemistry variables. The Maryland Biological Stream Survey (MBSS) data contains many potential explanatory variables so the number of potential explanatory variables was restricted to five for each response. The set of competing models was limited to linear models (without interactions) with exponential spatial correlation. This set-up was chosen because the exponential correlation function is compatible with all of the proposed distance measures. It was assumed that measurement error was present and hence the correlation matrix of (2.3) was updated with

$$\Gamma(\boldsymbol{d};\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } d_{ij} = 0, \\ (1-\theta_1)\exp\{-d_{ij}/\theta_2\} & \text{otherwise,} \end{cases}$$

where d_{ij} is the distance between locations i and j (distance measure dependent), θ_1 is the proportion of variability due to measurement error, and θ_2 is the range parameter. A general rule associated with the exponential correlation function is that sites that are within $3\theta_2$ of one another are considered to be correlated.

For each response variable, model selection was performed over the set of all possible linear models by maximizing the profile likelihood (2.4) with respect to $\boldsymbol{\theta}$ for each distance measure. The selected models were then compared by evaluating MSPE over a fixed subset of data that was withheld from the model selection process as defined by (2.13).

Table 2.8, originally published by Peterson et al. (2006, Table 6), list the MSPE for model selected for each distance measure. For five of the eight responses, symmetric hydrologic distance provided a moderate gain over Euclidean distance ("straight-line distance") with respect to MSPE and to the coefficient of determination, r^2. Asymmetric hydrologic distance was only shown to be superior for the water temperature response. Unfortunately the utility of the models for water temperature were generally poor.

Table 2.8: Autocorrelation parameter estimates, mean square prediction error (MSPE), and sample coefficient of determination (r^2) for the general linear model (GLM), straight-line distance (SLD), symmetric hydrologic distance (SHD), and weighted asymmetric hydrologic distance (WAHD) model for each unscaled response variable. The MSPE and r^2 values were calculated using the observed and predicted values contained in the validation set.

Response Variable	Distance Measure	MSPE	r^2	Nugget %	Sill	Range (km)
ANC	GLM	293,000	0.411	NA	NA	NA
	SLD	50,600	0.899	1.90	0.388	26.8
	SHD	55,800	0.877	2.00	0.286	57.5
	WAHD	87,200	0.843	0.90	0.644	47.7
CONDLAB	GLM	34,900	0.712	NA	NA	NA
	SLD	11,600	0.921	0.90	0.961	12.4
	SHD	3,230	0.959	1.30	0.573	27.6
	WAHD	4,010	0.948	1.10	0.569	45.2
DOC	GLM	7.85	0.520	NA	NA	NA
	SLD	5.46	0.644	1.54	0.282	56.3
	SHD	5.37	0.656	2.89	0.693	180
	WAHD	5.47	0.649	1.99	0.734	82.3
DO	GLM	1.91	0.294	NA	NA	NA
	SLD	1.58	0.414	5.45	0.202	62.7
	SHD	1.64	0.392	7.04	0.283	301
	WAHD	1.74	0.355	3.98	0.263	82.3
NO^3	GLM	1.14	0.671	NA	NA	NA
	SLD	0.82	0.772	3.60	0.593	20.7
	SHD	0.75	0.783	7.40	0.957	45.1
	WAHD	0.95	0.725	6.50	0.937	73.3
PHLAB	GLM	0.16	0.504	NA	NA	NA
	SLD	0.11	0.663	4.70	0.647	16.3
	SHD	0.10	0.679	6.40	0.500	36.4
	WAHD	0.11	0.663	3.50	0.503	33.9
SO_4	GLM	363	0.190	NA	NA	NA
	SLD	210	0.400	1.81	0.271	23.4
	SHD	259	0.360	3.06	0.443	40.8
	WAHD	292	0.286	1.76	0.922	82.3
TEMP	GLM	8.81	0.712	NA	NA	NA
	SLD	7.72	0.278	1.25	0.310	6.90
	SHD	7.49	0.298	4.20	0.702	14.0
	WAHD	7.37	0.309	1.88	0.473	15.9

2.6 Conclusions

Our results demonstrate the problems that can be encountered in the selection of an appropriate set of explanatory variables when spatial correlation is ignored. Both spatial AIC and MDL based on the geostatistical model performed well in the selection of appropriate explanatory variables. Ignoring spatial correlation in the selection of explanatory variables and/or in the modeling of the data can lead to mis-specification of the model as well as higher prediction errors. In addition, we showed that for the sampling patterns considered here, it is advantageous to consider a clustered type of sampling design that offers observation pairs at both small and large distances. Other aspects of model mis-specification, such as the appropriateness of the adoption of a Gaussian random field and stationarity of the autocorrelation function, are also important. Cressie (1993, p. 289) and Smith (2000, pp. 94–96) summarize some of the research on these issues.

For this presentation, we have assumed that only the candidate explanatory variables enter the model as main effects with no interactions or higher order terms under consideration. Thus for a data set with 10 potential explanatory variables there are $2^{10} = 1024$ candidate models if only a single error structure is examined. But a data set with 20 potential explanatory variables leads to 1.04×10^6 candidate models under the same set up. This number will increase further if, for example, we allow interactions or higher-order polynomial fits. Thus there are practical limitations to defining the collection of candidate models. To overcome this limitation the researcher has many avenues open to her. She can perform exploratory data analyses to reduce the number of potential explanatory variables, limit the candidate models to a particular class (such as linear), restrict the error structure to a single form, e.g., Matérn without nugget, etc.. Often a researcher can rely on their expertise to further reduce the size of the family of candidate models. Both spatial

AIC and MDL allow the researcher to restrict the type and class of models that best suits her needs.

The derivation of the spatial AIC statistic in Section 2.3.1.1 required the introduction of standard asymptotic assumptions about the distribution of the maximum likelihood estimators of the model parameters. We need to verify that these assumptions hold in the spatial context. Chapter 3 investigates the asymptotic behavior of the maximum likelihood estimators in one dimension by supposing a underlying continuous first order autoregressive process. It will be proved that the assumptions required for the AIC derivation hold for the AR(1) process. Empirical results will demonstrate that the assumptions also appear to hold for the Matérn correlation function.

Chapter 3

ASYMPTOTIC ANALYSIS FOR THE ONE-DIMENSIONAL CASE

In Chapter 2 we demonstrated the importance of accounting for the residual dependence structure during model selection. We found that ignoring potential correlation in the residuals can lead to the selection of an inappropriate model. In the argument given for the spatial AIC statistic some assumptions about the asymptotic distribution of the maximum likelihood estimator of the model parameters were made. In this chapter we investigate the asymptotic distribution of the estimator for the spatial parameter vector $\boldsymbol{\theta}$ in the special case of the one-dimensional Gaussian process with an exponential covariance function. We begin by first examining the concept of increasing sample size in the spatial context. There are three general schemes for increasing the number of observations: collecting data between previous observation sites (infill), collecting data outside of the current domain (expanding domain), and a combination of both infill and expanding domain. We begin our analysis by first assuming observations are available at regular fixed intervals along a one-dimensional transect (or lattice). Next we allow the spacings between subsequent locations to be random yet prescribed by a known (continuous) probability distribution.

3.1 Infill and Expanding Domain Asymptotics

A major obstacle in the analysis of spatial data is that maximum likelihood estimators do not have an explicit form in terms of the data and this in turn precludes calculating its distribution (Cressie, 1993, p. 480). Instead, we typically have to rely

on approximating the finite sample distribution using the asymptotic distribution. Second, the spatial framework allows for the collection of additional observations of the response within the current area of study (referred to as infill). A natural question that arises is at what level of sampling effort, i.e., the number of sampling locations per unit length or area, should the investigator target to maximize the precision of the parameter estimates? In other words, is there a point of diminishing returns such that additional observations made at close proximity provide limited or minimal additional information about the model parameters?

Cressie and Zimmerman (1992) and Cressie (1993) outline the geostatistical model including the asymptotic distribution of model parameters with respect to expanding domain. Additional work by Ying (1993) demonstrates the consistency and distribution of the product of MLE of the range and variance parameters ($\theta\sigma^2$) along a regular lattice in two-dimensions. However, we are interested in the distribution of the individual parameters themselves. For example, in the ecological sciences the researcher is often interested in making inferences about the range parameter θ alone (Peterson et al., 2006).

In recent years there has been interest in identifying optimal sampling schemes. Anecdotal evidence over the past 10 to 15 years has suggested that for fitting geostatistical models clustering sampling locations throughout the region of study appears to be best. Formal arguments are put forth by Loh and Lam (2000), Zhu and Stein (2005), Xia et al. (2005), Zhang and Zimmerman (2005), and Irvine et al. (2006). Extensive simulation results can be found in Thompson (2001). Although not a primary objective of this dissertation, we do identify (heuristically) an optimal sampling pattern for the exponential correlation function in one-dimension.

Finally, the ease of collecting spatially referenced data has provided both the opportunity and dilemma of having large data sets available for fitting geostatistical models. When the sampling locations are fixed to a lattice and the correlation

function being considered is exponential there are numerical "shortcuts" to compute the likelihood function. This is especially helpful during model selection where the modeler may be trying to identify the best subset of covariates from a large collection. Unfortunately, most data sets do not restrict observations to a regular lattice. This greatly increases the computational time required to evaluate the log-likelihood. Many modelers approximate the likelihood, often working in the spectral domain, by "snapping" the data to a fine grid or approximating the covariance matrix; see Fuentes (2005) and Johns et al. (2003). Although this is extremely effective and cost efficient with respect to computational time, we have chosen to avoid the introduction of an error term due to approximation and work with the exact likelihood instead.

3.2 Processes with Exponential Covariance Function

A Gaussian process in one-dimension with an exponential covariance function is often referred to as a first order autoregressive process (AR(1)) in the time series literature. That is to say that the observed value of the process at the current time t is related to the observation at a time $s < t$ such that $Y_t = e^{-\tau(t-s)}Y_s + \sigma\sqrt{2\tau}\int_s^t e^{-\tau(t-u)}dB_u$ where $\tau > 0$. For example, an ecologist may be interested in the water temperature or flow volume at a particular site over time. Although conceptually she could sample the site continuously with respect to time, it is not uncommon to sample at regular fixed intervals such as hourly, daily, weekly, etc. Hence, observations are made a discrete points in time such that $t_1 < t_2 < \ldots < t_n$. The discrete autoregressive model, $Y_{t_i} = \phi Y_{t_{i-1}} + \varepsilon_{t_i}$ where $\phi > 0$ and $\varepsilon_{t_i} \sim$ iid $\mathcal{N}(0, \sigma^2)$, is used in place of the continuous model.

Often spatial data fit into this framework as well. Many investigators sample locations at regular fixed distances along a transect or over a lattice in two- or three-dimensions. For example, the pH level of water at 50 meter intervals along a

stream or the percent silt sampled at regularly spaced locations across a farmer's field. Typically, in the spatial context one assumes that the domain of the random field is continuous and one is able to sample at any location along a transect or across the domain of interest. For example, one common sampling approach is to take many observations around a central site close to one another, move to another central site a good distance away and repeat the process (this is known as cluster sampling). For the remainder of this chapter we will focus on the continuous time AR(1) process, also known as the Ornstein-Uhlenbeck process.

3.2.1 Ornstein-Uhlenbeck Process

The continuous AR(1) process is defined as the solution to the following stochastic differential equation (SDE):

$$dY_t = -\tau Y_t dt + \sigma\sqrt{2\tau}dB_t, \quad \tau > 0. \tag{3.1}$$

Here t represents time or space, τ is defined as the range parameter, and B_t is a standard Brownian motion, i.e., B_t is a Gaussian process such that

i. $B_0 = 0$ almost surely,

ii. B_t has independent increments, i.e., $B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_k} - B_{t_{k-1}}$ are independent for all $0 \leq t_1 < t_2 < \ldots < t_k$,

iii. $B_t - B_s \sim \mathcal{N}(0, t-s)$ for $s < t$,

iv. B_t has continuous sample paths almost surely.

The solution to equation (3.1), known as the *Ornstein-Uhlenbeck process*, is given by

$$Y_t = e^{-\tau(t-s)}Y_s + \sigma\sqrt{2\tau}\int_s^t e^{-\tau(t-u)}dB_u \tag{3.2}$$

for $s < t$ and $\tau > 0$. Therefore,

$$\begin{aligned} EY_t &= e^{-\tau(t-s)}EY_s, \\ EY_t^2 &= e^{-2\tau(t-s)}EY_s^2 + \sigma^2(1 - e^{-2\tau(t-s)}), \text{ and} \\ EY_tY_{t+h} &= e^{-\tau|h|}EY_t^2 \text{ where } h \in \mathbb{R}. \end{aligned}$$

If $Y_0 \sim \mathcal{N}(0, \sigma^2)$ then it follows that $\{Y_{t_i}\}$ is strictly stationary and that $EY_t = 0$, $EY_t^2 = \sigma^2$, and $\text{Cov}(Y_t, Y_{t+h}) = EY_tY_{t+h} = e^{-\tau|h|}\sigma^2$. Under the current parameterization we see that for small values of τ, locations t_i and t_j are more strongly correlated compared to when τ is large. In the spatial literature it is more common to parameterize the process with $\theta = \tau^{-1}$. Under this parameterization locations t_i and t_j are strongly correlated for large values of θ. Thus θ is often interpreted as the range parameter where locations that are separated by less than (approximately) 3θ are strongly to moderately correlated since $\exp(-3\theta/\theta) = 0.05$.

Finally, it should be noted that as the distance between sampling locations tends toward zero, the Y_{t_i}'s become very strongly correlated. In the lexicon of the time series literature we approach a unit root (or random walk) for autoregressive models (the exponential covariance model in one-dimension is equivalent to an AR(1) process structure). See, for example, Dickey and Fuller (1979), Dickey and Fuller (1981), Evans and Savin (1981), Ahtola and Tiao (1984), and Chan and Wei (1987). From the spatial statistical framework similar analyses have been explored using various correlation functions with and without measurement error. See, for example, Whittle (1954), Stein (1995), Bhattacharyya et al. (1997), Stein (1999b), Chen et al. (2000), Johns et al. (2003), Loh (2005), and Zhu and Zhang (2005).

3.2.1.1 Discrete case

Although we assume that the data are sampled from a continuous process such as defined above, it is infeasible to sample processes continuously, so in practice

one is limited to collecting a finite set of observations at discrete locations. For example, if we assume that subsequent sampling locations are separated by a fixed distance Δ and we define $X_k = Y(k\Delta), k = \{0, \pm 1, \ldots\}$, then the sequence $\{X_k\}$ is a discrete-time AR(1) process.

An important special case is to assume that the observations occur at regular intervals along a transect. Suppose that our domain is $[0, m]$ and each unit interval is subdivided into n equal intervals. Let t_i index each site within the first interval such that $\{t_i = i/n; i = 1, \ldots, n\}$. The corresponding index set for the mn sites is then $\{t_i = i/n; i = 1, \ldots, mn\}$. Hence, (3.2) is rewritten as

$$Y_{t_i} = e^{-1/\theta n} Y_{t_{i-1}} + \varepsilon_{t_i}, \quad t_i = \left\{ \frac{i}{n}; i = 2, \ldots, mn \right\}, \tag{3.3}$$

where $\{\varepsilon_{t_i}\} \sim$ iid $\mathcal{N}\left(0, \sigma^2(1 - e^{-2/n\theta})\right)$. The covariance between any two locations t_i and t_j is $\text{Cov}(Y_{t_i}, Y_{t_j}) = \sigma^2 e^{-|i-j|/\theta n}$. This will prove to be a convenient notation for investigating the asymptotic behavior of the estimator of θ as both $n \to \infty$ and $m \to \infty$. Infill will be modeled by increasing the number of observations n per unit length. As n increases the spacings between subsequent observations decreases and reaches zero in the limit. Similarly, expansion of the domain will be modeled by increasing the total number of unit lengths m.

3.3 Asymptotic Results for a Regular Lattice

We begin by supposing that the sampling design is a regular lattice such that there are n observations per unit length and there are a total of m contiguous units (see Section 3.2.1.1). To simplify notation let $\phi_n = e^{-1/(\theta n)}$ denote the correlation between subsequent observations. Thus as n increases ϕ_n increases to one. For a fixed $n \geq 1$ (3.3) can be written as $Y_{t_i} = \phi_n Y_{t_{i-1}} + \varepsilon_{t_i}$. If we assume that the mean is exactly zero, i.e., $\boldsymbol{\beta} = \mathbf{0}$, the general spatial model (1.2) simplifies to $\mathbf{Y} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{\Gamma})$ where the $(i,j)^{\text{th}}$ element of the correlation matrix $\boldsymbol{\Gamma}$ is $\phi_n^{|i-j|}$ for $i, j = 1, 2, \ldots, mn$.

To evaluate the profile log-likelihood (1.7), it will be convenient to condition on $Y_{t_0} = 0$ where $t_0 = 0$. Conditioning on the *unobserved* response Y_{t_0} we find that

$$\begin{aligned}\ell_{\text{profile}}(\phi_n; \mathbf{Y}|Y_{t_0}, \hat{\sigma}^2) &= -\frac{mn}{2}\log(2\pi) - \frac{mn}{2}\log\left(\frac{\sum_{i=1}^{mn}(Y_{t_i} - \phi_n Y_{t_{i-1}})^2}{mn(1-\phi_n^2)}\right) \\ &\quad - \frac{1}{2}\log(1-\phi_n^2)^{mn} - \frac{mn}{2} \\ &= -\frac{mn}{2}\log(2\pi) - \frac{mn}{2}\log\left(\frac{\sum_{i=1}^{mn}(Y_{t_i} - \phi_n Y_{t_{i-1}})^2}{mn}\right) - \frac{mn}{2}. \end{aligned} \tag{3.4}$$

Maximizing (3.4) with respect to ϕ_n is equivalent to minimizing the quantity $\sum_{i=1}^{mn}(Y_{t_i} - \phi_n Y_{t_{i-1}})^2$. Therefore,

$$\hat{\phi}_n = \frac{\sum_{i=1}^{mn} Y_{t_i} Y_{t_{i-1}}}{\sum_{i=1}^{mn} Y_{t_{i-1}}^2}. \tag{3.5}$$

Chan and Wei (1987) note that (3.5) is also the least squares estimator of ϕ_n conditioned on $Y_{t_0} = 0$. We now proceed to derive the asymptotic distribution of the range parameter θ.

THEOREM 1

Let $\mathbf{Y}$ *be an* mn*-vector of observations generated by the continuous AR(1) process as defined by equation (3.2) such that* $Y_i = Y_{t_i}$ *where* $t_i = \{i/n; i = 1, \ldots, mn\}$. *Let* $\hat{\theta}$ *be the maximum likelihood estimate of the range parameter* θ *conditional on* $Y_{t_0} = 0$. *Let* $\xrightarrow{\text{d}}$ *denote convergence in distribution. Then,*

(a) for fixed n *and* $m \to \infty$,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\text{d}} \mathcal{N}\left(0, \theta^4 n(e^{(2/\theta n)} - 1)\right),$$

(b) as $n \to \infty$ *and* $m \to \infty$,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\text{d}} \mathcal{N}(0, 2\theta^3).$$

Proof of Theorem 1

(a) For n fixed and $m \to \infty$, this corresponds to a discrete AR(1) process. While the derivation of the limit result is standard in time series, we provide a proof to illustrate key ideas that will be used in later arguments. Substituting $Y_{t_i} = \phi_n Y_{t_{i-1}} + \varepsilon_{t_i}$ into equation (3.5), we have

$$\begin{aligned}\hat{\phi}_n &= \frac{\sum_{i=1}^{mn}(\phi_n Y_{t_{i-1}} + \varepsilon_{t_i})Y_{t_{i-1}}}{\sum_{i=1}^{mn} Y_{t_{i-1}}^2} \\ &= \frac{\sum_{i=1}^{mn}\phi_n Y_{t_{i-1}^2}}{\sum_{i=1}^{mn} Y_{t_{i-1}}^2} + \frac{\sum_{i=1}^{mn}\varepsilon_{t_i} Y_{t_{i-1}}}{\sum_{i=1}^{mn} Y_{t_{i-1}}^2} \\ &= \phi_n + \frac{\sum_{i=1}^{mn}\varepsilon_{t_i} Y_{t_{i-1}}}{\sum_{i=1}^{mn} Y_{t_{i-1}}^2}.\end{aligned} \tag{3.6}$$

Rearranging and rescaling the estimator gives

$$\sqrt{m}n(\hat{\phi}_n - \phi_n) = \frac{\frac{1}{\sqrt{m}}\sum_{i=1}^{mn}\varepsilon_{t_i} Y_{t_{i-1}}}{\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}}^2}. \tag{3.7}$$

To establish the asymptotic distribution of $\sqrt{m}n(\hat{\phi}_n - \phi_n)$ we consider the terms in the numerator and denominator separately.

The expected value of the denominator is

$$\mathrm{E}\left(\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}}^2\right) = \sigma^2.$$

Noting that $EY_t^2Y_s^2 = (2e^{-2\theta(s-t)} + 1)\sigma^4$ for $s > t$, the variance of the denominator can be computed as

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}}^2\right) &= \mathrm{E}\left(\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}}^2\right)^2 - \left(\mathrm{E}\left(\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}}^2\right)\right)^2 \\
&= \frac{1}{m^2n^2}\sum_{i=1}^{mn}\sum_{j=1}^{mn}(2\phi_n^{2|i-j|} + 1)\sigma^4 - \sigma^4 \\
&= \frac{1}{m^2n^2}\sum_{i=1}^{mn}\sum_{j=1}^{mn} 2\phi_n^{2|i-j|}\sigma^4 \\
&\Rightarrow \mathrm{Var}\left(\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}}^2\right) \to 0 \text{ as } m \to \infty.
\end{aligned}$$

Hence,

$$\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}}^2 \xrightarrow{\mathrm{P}} \sigma^2.$$

Now we change the focus to the numerator. Since the ε_{t_i}'s are an independent sequence and ε_{t_i} is independent of $Y_{t_{i-1}}$ it follows that $Y_{t_{i-1}}\varepsilon_{t_i}$ and $Y_{t_{j-1}}\varepsilon_{t_j}$ are uncorrelated for every $i \neq j$. Hence, the expected value of the numerator is

$$\mathrm{E}\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{mn} Y_{t_{i-1}}\varepsilon_{t_i}\right) = 0,$$

and the variance is

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{mn} Y_{t_{i-1}}\varepsilon_{t_i}\right)^2 &= \frac{1}{m}\sum_{i=1}^{mn} EY_{t_{i-1}}^2\varepsilon_{t_i}^2 \\
&= n(1-\phi_n^2)\sigma^4.
\end{aligned}$$

Using a standard central limit theorem for martingales (Hall and Heyde, 1980), the numerator is asymptotically normal with mean zero and variance $n(1-\phi_n^2)\sigma^4$, i.e.,

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{mn} Y_{t_{i-1}}\varepsilon_{t_i} \xrightarrow{\mathrm{d}} \mathcal{N}(0, n(1-\phi_n^2)\sigma^4) \tag{3.8}$$

as $m \to \infty$.

Combining the results for the numerator and denominator and employing Slutsky's Theorem, the asymptotic distribution for equation (3.7) is

$$\sqrt{m}n(\hat{\phi}_n - \phi_n) \xrightarrow{\mathrm{d}} \mathcal{N}(0, n(1-\phi_n^2)) \tag{3.9}$$

as $m \to \infty$.

A second application of Slutsky's Theorem can be used to evaluate the asymptotic distribution of $\hat{\theta} - \theta$. Following Brockwell and Davis (1991, Ch. 6), let $g(x)$ be a real-valued function that is differentiable at a, and let $\{b_m\}$ be a sequence of positive constants satisfying $b_m \to \infty$ as $m \to \infty$. Then $b_m(X_m - a) \xrightarrow{\mathrm{d}} X$ implies $b_m(g(X_m) - g(a)) \xrightarrow{\mathrm{d}} g'(a)X$. Take $g(x) = -(n\log(x))^{-1}$, $b_m = \sqrt{m}$, and $a = e^{-1/(\theta n)}$. Therefore, $g'(x) = (nx\log^2(x))^{-1}$ and

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \theta^4 n e^{2/(\theta n)}(1-\phi_n^2)) = \mathcal{N}(0, \theta^4 n(e^{2/(\theta n)} - 1)),$$

as claimed.

(b) Here we assume that n is a function of m such that as $m \to \infty$, $n_m = n(m) \to \infty$ as well. Equation (3.7) can be rewritten as

$$\sqrt{m}n_m(\hat{\phi}_{n_m} - \phi_{n_m}) = \frac{\frac{1}{\sqrt{m}}\sum_{i=1}^{mn_m} \varepsilon_{t_i} Y_{t_{i-1}}}{\frac{1}{mn_m}\sum_{i=1}^{mn_m} Y_{t_{i-1}}^2}.$$

Using the exact same argument as before, the denominator can be shown to go to σ^2 in probability, i.e.,

$$\frac{1}{mn_m}\sum_{i=1}^{mn_m} Y_{t_{i-1}}^2 \xrightarrow{\mathrm{P}} \sigma^2.$$

The asymptotic distribution of the numerator remains nearly the same. The only change is that the variance term goes to $2\sigma^4\theta^{-1}$ as $n_m \to \infty$. Hence, equation (3.8) can be rewritten as

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{mn_m} Y_{t_{i-1}}\varepsilon_{t_i} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 2\sigma^4\theta^{-1}).$$

Thus,

$$\sqrt{m}n_m(\hat{\phi}_{n_m} - \phi_{n_m}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 2\theta^{-1}). \tag{3.10}$$

Using the mean value theorem,

$$e^{-1/(\hat{\theta}n_m)} = e^{-1/(\theta n_m)} + (\hat{\theta} - \theta)\frac{e^{-1/(\theta^\star n_m)}}{\theta^2 n_m}, \tag{3.11}$$

where $\theta^\star$ is between $\hat{\theta}$ and θ. We obtain the desired result by substituting (3.11) into (3.10) and noting that $\exp\{-1/(\theta^\star n_m)\} \to 1$ as $m \to \infty$. Hence,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \theta^2\mathcal{N}(0, 2\theta^{-1}) \sim \mathcal{N}(0, 2\theta^3).$$

□

REMARK 1 By holding n fixed for most of the proof, it is evident that expansion of the domain, i.e., increasing m, induces normality of the estimator.

REMARK 2 Increasing infill (n) reduces the overall variance of the estimator but with diminishing returns as $n \to \infty$. This is due to the fact that the variance of equation (3.9) can be written as

$$\begin{aligned} n(1 - \phi_n^2) &= n\left(1 - \left(1 - \frac{2}{\theta n} + \frac{4}{2\theta^2 n^2} - \dots\right)\right) \\ &= \frac{2}{\theta} + \mathcal{O}(\theta n^{-1}). \end{aligned}$$

For sufficiently large n, the error term is close to zero and little gain is made by further increasing n.

3.3.1 Simulation results over a regular lattice

We present a series of simulations to illustrate how the mean square error (MSE) varies relative to infill and expanding domain over a regular lattice. Figure 3.1 summarizes the results for $\theta = \{1, 2\}$. The top two panels are the observed MSE based on 1000 simulations. The middle and bottom panels decompose the MSE into

the contributions associated with the standard error and the bias, respectively. Note that the scales of the axes are $\log_2$ units such that $m, n = \{2^0, 2^1, \ldots, 2^8\}$. Expansion of the domain is represented by moving from left to right across a particular image plot and moving from bottom to top represents increased infill.

Evident in Figure 3.1 is that the standard error dominates the MSE. Furthermore, there appears to be little or no reduction of the MSE with respect to increased infill. Notice that the contours of the MSE images are parallel to the infill axis. In contrast, as the domain is expanded for any fixed level of infill the MSE tends toward zero. This illustrates the consistency of the estimator with respect to expanding domain (m).

Figure 3.2 illustrates similar results for the case $\theta = 4$. The top two panels plot the MSE and the bottom two panels plot the contributions attributable to the standard error and bias. Note that the MSE image is very similar to the previous two cases with respect to shape. However, the bias term contributes a noticeable portion to the overall MSE for small domains ($m \leq 2$). This is due in part to the fact that the individual observations are very strongly correlated over these small domains. Notice that as the domain is expanded beyond $1 \times \theta$ that the bias term tends toward zero at a faster rate than the standard error term, consistent with the previous results. This illustrates the need to have an adequately large domain for estimation.

Table 3.1 lists the observed and asymptotic variance for a fixed total sampling effort of $mn = 256$ for $\theta = \{1, 2, 4\}$. The reported observed variances, based on 1000 independent simulations, have been scaled by m to match the order of magnitude of the asymptotic variances, denoted by parentheses, which were computed using part (a) of Theorem 1, i.e., $\text{Var}(\hat{\theta}) = \theta^4 n(\exp\{2/(\theta n)\} - 1)$. In general the observed values agree with the asymptotic values.

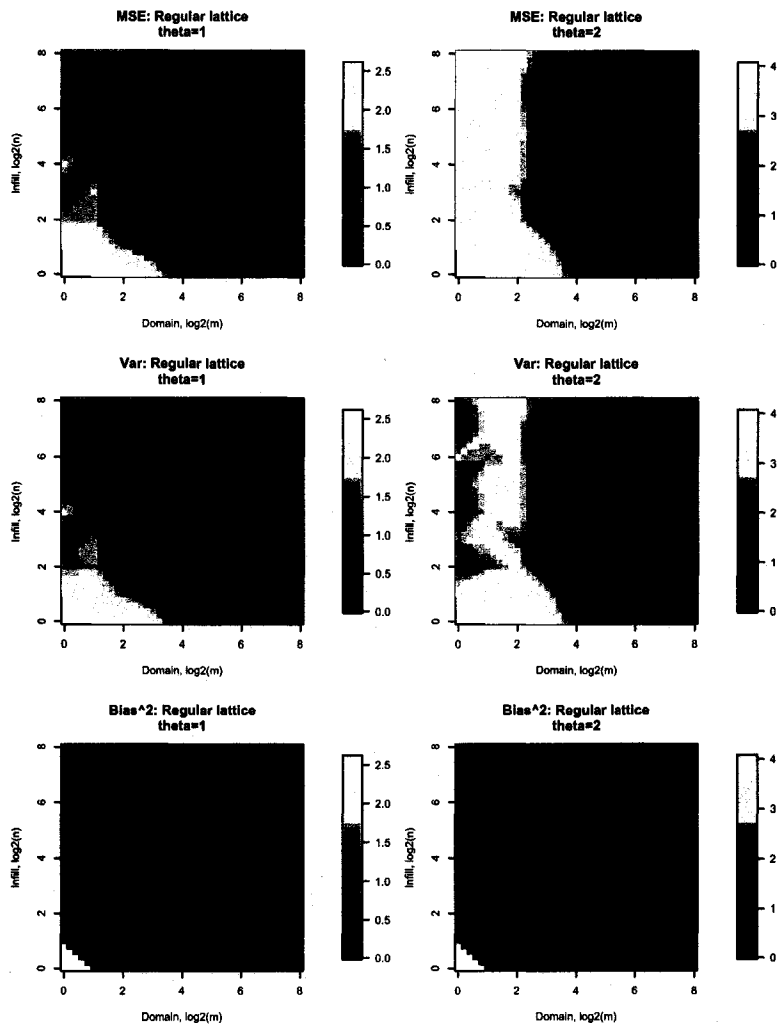

Figure 3.1: Simulated results for the AR(1) process over a regular lattice for $\theta = \{1, 2\}$. The top two panels are the observed MSE for 1000 simulations. The middle and bottom panels display the contributions associated with the standard error and bias, respectively. Note that the axes are $\log_2$ such that $m, n = \{2^0, 2^1, \ldots, 2^8\}$.

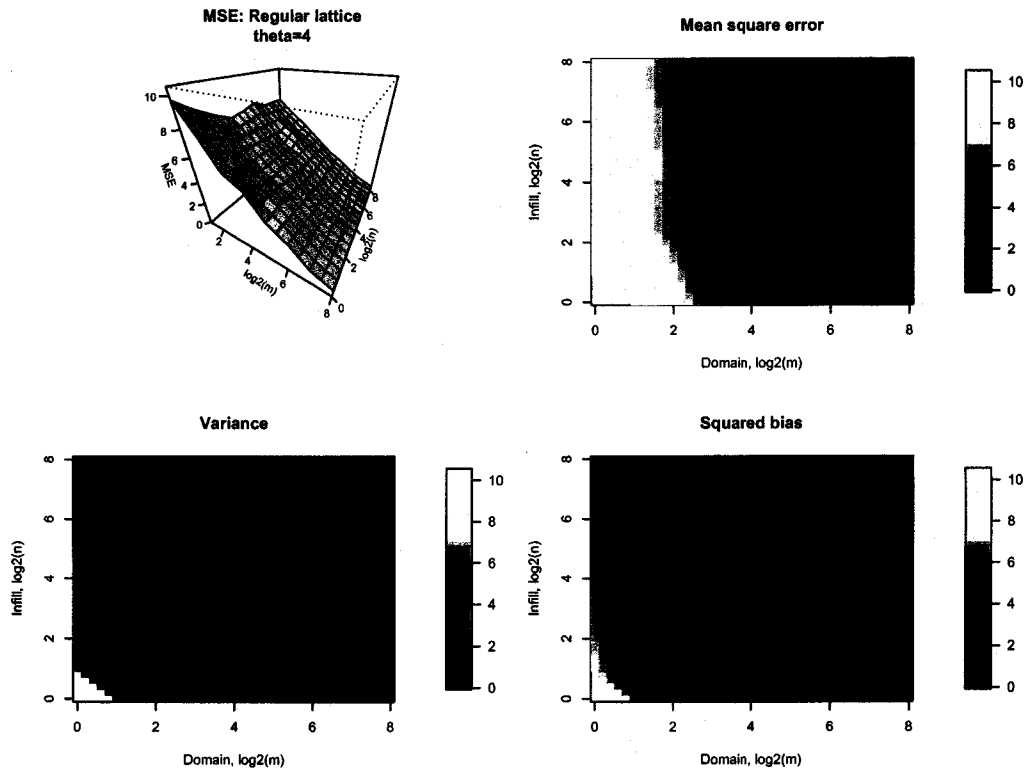

Figure 3.2: Simulated results for the AR(1) process over a regular lattice for $\theta = 4$. The upper panels are the observed MSE for 1000 simulations. The lower panels display the contributions associated with the standard error and bias, respectively. Note that the axes are $\log_2$ such that $m, n = \{2^0, 2^1, \ldots, 2^8\}$.

Table 3.1: Comparison of the observed and asymptotic variance for the estimator $\hat{\theta}$, scaled by m, for a fixed sampling effort of $mn = 256$. The asymptotic variances are delimited by parentheses and were calculated using part (a) of Theorem 1.

Parameter	(Domain, Infill) $= (m, n)$			
Value	$(256, 1)$	$(128, 2)$	$(64, 4)$	$(32, 8)$
$\theta = 1$	6.13	3.38	2.76	2.12
	(6.38)	(3.43)	(2.59)	(2.27)
$\theta = 2$	29.2	22.6	19.1	17.3
	(27.4)	(20.7)	(18.1)	(17.0)
$\theta = 4$	169	152	148	140
	(166)	(145)	(136)	(132)

3.3.2 Expected mean square error for the regular lattice

A benefit to having the estimator for θ in closed form is that we are able to approximate the mean square error (MSE) as a function of domain and level of infill using second-order Taylor expansion. Recall the estimator for $\phi_n = \exp\{-1/(\theta n)\}$ is

$$\hat{\phi}_n = \frac{\sum_{i=1}^{mn} Y_{t_{i-1}} Y_{t_i}}{\sum_{i=1}^{mn} Y_{t_{i-1}}^2}.$$

Let $U = (mn)^{-1} \sum_{i=1}^{mn} Y_{t_{i-1}}^2$ and $V = (mn)^{-1} \sum_{i=1}^{mn} Y_{t_{i-1}} Y_{t_i}$ for notational convenience. Therefore, the estimator for $\theta n = -(\log(\phi_n))^{-1}$ can be written as

$$\begin{aligned} \hat{\theta} n &= \left(\log \frac{1}{mn} \sum_{i=1}^{mn} Y_{t_{i-1}}^2 - \log \frac{1}{mn} \sum_{i=1}^{mn} Y_{t_{i-1}} Y_{t_i} \right)^{-1} \\ &= \frac{1}{(\log U - \log V)}. \end{aligned} \tag{3.12}$$

Approximating (3.12) using a second-order Taylor expansion about $\mathrm{E}U$ and $\mathrm{E}V$, we have

$$\begin{aligned} \hat{\theta} n \approx & \frac{1}{\log \mathrm{E}U - \log \mathrm{E}V} - \frac{1}{(\log \mathrm{E}U - \log \mathrm{E}V)^2} \left(\frac{U - \mathrm{E}U}{\mathrm{E}U} - \frac{V - \mathrm{E}V}{\mathrm{E}V} \right) \\ & + \frac{1}{2} (U - \mathrm{E}U)^2 \frac{2 + \log \mathrm{E}U - \log \mathrm{E}V}{(\mathrm{E}U)^2 (\log \mathrm{E}U - \log \mathrm{E}V)^3} \\ & + \frac{1}{2} (V - \mathrm{E}V)^2 \frac{2 - \log \mathrm{E}U + \log \mathrm{E}V}{(\mathrm{E}V)^2 (\log \mathrm{E}U - \log \mathrm{E}V)^3} \\ & - (U - \mathrm{E}U)(V - \mathrm{E}V) \frac{2}{\mathrm{E}U \mathrm{E}V (\log \mathrm{E}U - \log \mathrm{E}V)^3}. \end{aligned} \tag{3.13}$$

The expected value of U is evaluated as

$$\begin{aligned} \mathrm{E}U &= \mathrm{E} \frac{1}{mn} \sum_{i=1}^{mn} Y_{t_{i-1}}^2 \\ &= \frac{1}{mn} \sum_{i=1}^{mn} \mathrm{E} Y_{t_{i-1}}^2 \\ &= \sigma^2. \end{aligned}$$

Similarly, the expected value of V is

$$\begin{aligned} \mathrm{E}V &= \mathrm{E}\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}} Y_{t_i} \\ &= \frac{1}{mn}\sum_{i=1}^{mn} \mathrm{E}Y_{t_{i-1}}(\phi_n Y_{t_{i-1}} + \varepsilon_{t_i}) \\ &= \frac{1}{mn}\sum_{i=1}^{mn} \mathrm{E}(\phi_n Y_{t_{i-1}}^2 + Y_{t_{i-1}}\varepsilon_{t_i}) \\ &= \phi_n \sigma^2. \end{aligned}$$

Therefore $(\log \mathrm{E}U - \log \mathrm{E}V)^{-1} = \theta n$. Substituting this result into (3.13) and rearranging terms yields

$$\begin{aligned} \hat{\theta} - \theta \approx & -\theta^2 n \left(\frac{U - \mathrm{E}U}{\mathrm{E}U} - \frac{V - \mathrm{E}V}{\mathrm{E}V} \right) \\ & + \frac{1}{2}\frac{(U - \mathrm{E}U)^2}{(\mathrm{E}U)^2}(2\theta^3 n^2 + \theta^2 n) + \frac{1}{2}\frac{(V - \mathrm{E}V)^2}{(\mathrm{E}V)^2}(2\theta^3 n^2 - \theta^2 n) \\ & - \frac{2(U - \mathrm{E}U)(V - \mathrm{E}V)}{\mathrm{E}U\mathrm{E}V} 2\theta^3 n^2, \end{aligned} \tag{3.14}$$

where $\hat{\theta} - \theta$ is the bias. The approximate expected bias is

$$\begin{aligned} \mathrm{E}[\hat{\theta} - \theta] &\approx \theta^3 n^2 \left(\frac{\mathrm{Var}U}{(\mathrm{E}U)^2} + \frac{\mathrm{Var}V}{(\mathrm{E}V)^2} - \frac{2\mathrm{Cov}(U,V)}{\mathrm{E}U\mathrm{E}V} \right) + \frac{\theta^2 n}{2}\left(\frac{\mathrm{Var}U}{(\mathrm{E}U)^2} - \frac{\mathrm{Var}V}{(\mathrm{E}V)^2} \right) \\ &= \theta^3 n^2 \left(\frac{\mathrm{E}U^2}{(\mathrm{E}U)^2} + \frac{\mathrm{E}V^2}{(\mathrm{E}V)^2} - \frac{2\mathrm{E}UV}{\mathrm{E}U\mathrm{E}V} \right) + \frac{\theta^2 n}{2}\left(\frac{\mathrm{E}U^2}{(\mathrm{E}U)^2} - \frac{\mathrm{E}V^2}{(\mathrm{E}V)^2} \right). \end{aligned}$$

To approximate the variance of $\hat{\theta}$ we evaluate the variance of (3.14) using only the first-order terms. Thus,

$$\begin{aligned} \mathrm{Var}[\hat{\theta}] &\approx \theta^4 n^2 \left(\frac{\mathrm{Var}U}{(\mathrm{E}U)^2} + \frac{\mathrm{Var}V}{(\mathrm{E}V)^2} - \frac{2\mathrm{Cov}(U,V)}{\mathrm{E}U\mathrm{E}V} \right) \\ &= \theta^4 n^2 \left(\frac{\mathrm{E}U^2}{(\mathrm{E}U)^2} + \frac{\mathrm{E}V^2}{(\mathrm{E}V)^2} - \frac{2\mathrm{E}UV}{\mathrm{E}U\mathrm{E}V} \right). \end{aligned}$$

Note that both approximations require the evaluation of $\mathrm{E}U^2$, $\mathrm{E}V^2$, and $\mathrm{E}UV$.

To evaluate $\mathrm{E}U^2$ first expand the summation such that

$$\begin{aligned} \mathrm{E}U^2 &= \mathrm{E}\left[\frac{1}{mn}\sum_{i=1}^{mn} Y_{t_{i-1}}^2 \right]^2 \\ &= \frac{1}{m^2 n^2}\sum_{i=1}^{mn}\sum_{j=1}^{mn} \mathrm{E}[Y_{t_{i-1}}^2 Y_{t_{j-1}}^2]. \end{aligned}$$

For all i and j we can evaluate $\mathrm{E}[Y^2_{t_{i-1}} Y^2_{t_{j-1}}]$. For example, for $i = j$, $\mathrm{E}[Y^4_{t_{i-1}}] = 3\sigma^4$. Similarly, for $j = i + 1$,

$$\begin{aligned}
\mathrm{E}[Y^2_{t_{i-1}} Y^2_{t_i}] &= \mathrm{E}[Y^2_{t_{i-1}}(\phi_n^2 Y^2_{t_{i-1}} + 2\phi_n Y_{t_{i-1}}\varepsilon_{t_i} + \varepsilon^2_{t_i})] \\
&= \phi_n^2 \mathrm{E}[Y^4_{t_{i-1}}] + 2\phi_n \mathrm{E}[Y^3_{t_{i-1}}\varepsilon_{t_i}] + \mathrm{E}[Y^2_{t_{i-1}}\varepsilon^2_{t_i}] \\
&= 3\phi_n^2\sigma^4 + (1 - \phi_n^2)\sigma^4 \\
&= (2\phi_n^2 + 1)\sigma^4,
\end{aligned}$$

and for $j = i + 2$,

$$\begin{aligned}
\mathrm{E}[Y^2_{t_{i-1}} Y^2_{t_{i+1}}] &= \mathrm{E}[Y^2_{t_{i-1}}(\phi_n^2 Y^2_{t_i} + 2\phi_n Y_{t_i}\varepsilon_{t_{i+1}} + \varepsilon^2_{t_{i+1}})] \\
&= \phi_n^2 \mathrm{E}[Y^2_{t_{i-1}} Y^2_{t_i}] + 2\phi_n \mathrm{E}[Y^3_{t_i}\varepsilon_{t_{i+1}}] + \mathrm{E}[Y^2_{t_i}\varepsilon^2_{t_{i+1}}] \\
&= \phi_n^2(2\phi_n^2 + 1)\sigma^4 + (1 - \phi_n^2)\sigma^4 \\
&= (2\phi_n^4 + 1)\sigma^4
\end{aligned}$$

By symmetry (recall that the process is isotropic) it is evident these results must hold when $j = i - 1$ and $j = i - 2$, respectively. Thus, using induction the above relations can be summarized by the expression

$$\mathrm{E}U^2 = \frac{\sigma^4}{m^2 n^2} \sum_{i=1}^{mn} \sum_{j=1}^{mn} (2\phi_n^{2|i-j|} + 1). \tag{3.15}$$

We expand $\mathrm{E}V^2$ in the same manner such that

$$\begin{aligned}
\mathrm{E}V^2 &= \mathrm{E}\left[\frac{1}{mn} \sum_{i=1}^{mn} Y_{t_{i-1}} Y_{t_i}\right]^2 \\
&= \frac{1}{m^2 n^2} \sum_{i=1}^{mn} \sum_{j=1}^{mn} \mathrm{E}[Y_{t_{i-1}} Y_{t_i} Y_{t_{j-1}} Y_{t_j}].
\end{aligned}$$

Once again we can evaluate each term for all i and j. When $j = i$ we have a previous result, i.e., $\mathrm{E}[Y_{t_{i-1}}^2 Y_{t_i}^2] = (2\phi_n^2 + 1)\sigma^4$. For $j = i + 1$,

$$\begin{aligned}
&Y_{t_{i-1}} Y_{t_i}^2 Y_{t_{i+1}} \\
&\quad = Y_{t_{i-1}} Y_{t_i}^2 (\phi_n Y_{t_i} + \varepsilon_{t_{i+1}}) \\
&\quad = \phi_n Y_{t_{i-1}} Y_{t_i}^3 + Y_{t_{i-1}} Y_{t_i}^2 \varepsilon_{t_{i+1}} \qquad (3.16) \\
&\quad = \phi_n Y_{t_{i-1}} (\phi_n^3 Y_{t_{i-1}}^3 + 3\phi_n^2 Y_{t_{i-1}}^2 \varepsilon_{t_i} + 3\phi_n Y_{t_{i-1}} \varepsilon_{t_i}^2 + \varepsilon_{t_i}^3) + Y_{t_{i-1}} Y_{t_i}^2 \varepsilon_{t_{i+1}} \\
&\quad = \phi_n^4 Y_{t_{i-1}}^4 + 3\phi_n^3 Y_{t_{i-1}}^3 \varepsilon_{t_i} + 3\phi_n^2 Y_{t_{i-1}}^2 \varepsilon_{t_i}^2 + \phi_n Y_{t_{i-1}} \varepsilon_{t_i}^3 + Y_{t_{i-1}} Y_{t_i}^2 \varepsilon_{t_{i+1}}.
\end{aligned}$$

Taking expectations on both sides yields

$$\begin{aligned}
\mathrm{E}[Y_{t_{i-1}} Y_{t_i}^2 Y_{t_{i+1}}] &= 3\phi_n^4 \sigma^4 + 3\phi_n^2 (1 - \phi_n^2)\sigma^4 \\
&= 3\phi_n^2 \sigma^4.
\end{aligned}$$

Setting $j = i - 1$ we have $Y_{t_{i-2}} Y_{t_{i-1}}^2 Y_{t_i}$ which is identical in form to (3.16). By symmetry $\mathrm{E}[Y_{t_{i-2}} Y_{t_{i-1}}^2 Y_{t_i}] = 3\phi_n^2 \sigma^4$. For $j = i + 2$,

$$\begin{aligned}
&Y_{t_{i-1}} Y_{t_i} Y_{t_{i+1}} Y_{t_{i+2}} \\
&\quad = \phi_n Y_{t_{i-1}} Y_{t_i} Y_{t_{i+1}}^2 + Y_{t_{i-1}} Y_{t_i} Y_{t_{i+1}} \varepsilon_{t_{i+2}} \\
&\quad = \phi_n^3 Y_{t_{i-1}} Y_{t_i}^3 + 2\phi_n^2 Y_{t_{i-1}} Y_{t_i}^2 \varepsilon_{t_{i+1}} + \phi_n Y_{t_{i-1}} Y_{t_i} \varepsilon_{t_{i+1}}^2 + Y_{t_{i-1}} Y_{t_i} Y_{t_{i+1}} \varepsilon_{t_{i+2}} \\
&\quad = \phi_n^3 Y_{t_{i-1}} Y_{t_i}^3 + 2\phi_n^2 Y_{t_{i-1}} Y_{t_i}^2 \varepsilon_{t_{i+1}} + \phi_n^2 Y_{t_{i-1}}^2 \varepsilon_{t_{i+1}}^2 + \phi_n Y_{t_{i-1}} \varepsilon_{t_i} \varepsilon_{t_{i+1}}^2 + Y_{t_{i-1}} Y_{t_i} Y_{t_{i+1}} \varepsilon_{t_{i+2}}.
\end{aligned}$$

Applying the expectation operator and using a previous result we find

$$\begin{aligned}
\mathrm{E}[Y_{t_{i-1}} Y_{t_i} Y_{t_{i+1}} Y_{t_{i+1}}] &= 3\phi_n^4 \sigma^4 + \phi_n^2 (1 - \phi_n^2)\sigma^4 \\
&= \phi_n^2 (2\phi_n^2 + 1)\sigma^4.
\end{aligned}$$

Once again, taking advantage of the symmetry and using induction we derive the relation

$$\mathrm{E}V^2 = \frac{\sigma^4}{m^2 n^2} \left(mn(2\phi_n^2 + 1) + \phi_n^2 \sum_{i=1}^{mn} \sum_{j \neq i} (2\phi_n^{2(|i-j|-1)} + 1) \right).$$

Finally, expand $\mathrm{E}UV$ such that

$$\begin{aligned}
\mathrm{E}UV &= \mathrm{E}\left[\frac{1}{mn} \sum_{i=1}^{mn} Y_{t_{i-1}}^2 \frac{1}{mn} \sum_{j=1}^{mn} Y_{t_{j-1}} Y_{t_j} \right] \\
&= \frac{1}{m^2 n^2} \sum_{i=1}^{mn} \sum_{j=1}^{mn} \mathrm{E}[Y_{t_{i-1}}^2 Y_{t_{j-1}} Y_{t_j}].
\end{aligned}$$

For $j = i$,

$$\begin{aligned} \mathrm{E}[Y_{t_{i-1}}^3 Y_{t_i}] &= \phi_n \mathrm{E}[Y_{t_{i-1}}^4] + \mathrm{E}[Y_{t_{i-1}}^3 \varepsilon_{t_i}] \\ &= 3\phi_n \sigma^4. \end{aligned}$$

Note that for $j = i - 1$ we have $Y_{t_{i-2}} Y_{t_{i-1}}^3$ which is identical in form. Thus the expected value is also $3\phi_n\sigma^4$. For $j = i + 1$,

$$\begin{aligned} \mathrm{E}[Y_{t_{i-1}}^2 Y_{t_i} Y_{t_{i+1}}] &= \phi_n \mathrm{E}[Y_{t_{i-1}}^2 Y_{t_i}^2] + \mathrm{E}[Y_{t_{i-1}}^2 Y_{t_i} \varepsilon_{t_{i+1}}] \\ &= \phi_n (2\phi_n^2 + 1)\sigma^4. \end{aligned}$$

The same form occurs when $j = i - 2$. By induction we derive the expression for $\mathrm{E}UV$ to be

$$\mathrm{E}UV = \frac{\phi_n \sigma^4}{m^2 n^2} \sum_{i=1}^{mn} \left(\sum_{j \geq i}^{mn} (2\phi_n^{2(j-i)} + 1) + \sum_{j<i} (2\phi_n^{2(i-j-1)} + 1) \right).$$

Thus, given m, n, and θ, the expected value and variance of the estimator $\hat{\theta}$ can be well approximated. Note that neither the expected value nor the variance depend on σ^2 as it cancels from both expressions.

Figures 3.3 and 3.4 illustrate the impact of domain expansion and infill on the MSE of $\hat{\theta}$ for $\theta = \{1, 2\}$. The left panels of Figure 3.3 demonstrate how the MSE converges quickly with respect to expansion of the domain for various levels of infill. Note that difference between the curves is small for $n = 4$ and $n = 8$ implying that additional infill does not significantly reduce the expected MSE for even moderate domains. The right panels further illustrate this observation by fixing the domain and infilling. Note that the scale of the x-axis is $\log_2$. Hence the reduction in the MSE is nearly zero for sampling efforts in excess of $2^3 = 8$ locations per unit length. Figure 3.3 demonstrates how the expected standard error begins to dominate the MSE for moderate domains. The left panels plot the expected MSE surface as a function of both expanding domain and level of infill. Note that the surfaces tend

toward zero as the domain increases but are approximately constant along contours of fixed domain. This observation coincides with the plots of Figure 3.3 and, in fact, the plots of Figure 3.3 are nothing more than slices of the surface at fixed levels of infill or domain. The panels on the right of Figure 3.4 illustrate the contribution to the MSE attributable to the standard error. Regions of red indicate that the bias term dominates the MSE whereas white regions indicate that the standard error term dominates. For moderate domain sizes the contribution due to bias is small. As a footnote, it should be noted that the magnitude of the expected MSE is suspect because the approximation using the second-order Taylor expansion will be poor for small m and n. Of principle interest are the trends in the standard error and bias terms and their relationship to one another.

For other correlation functions, e.g., Matérn, we will wish to estimate a "smoothness" parameter. One expects that it will be beneficial to make observations at small distances. Anecdotal evidence from our model selection analysis suggests that clustering the observations is superior for identifying the appropriate model when a smoothness parameter is to be specified. This type of analysis will be used to assess the impact of expansion of the domain, infill, and the sampling pattern on the MSE for θ. As we will see, for the case of the Matérn correlation function, we do not have the benefit of a closed form for the estimator of θ. Thus we will be forced to rely on simulations to provide insight into the relative importance of the aforementioned factors (see Chapter 5).

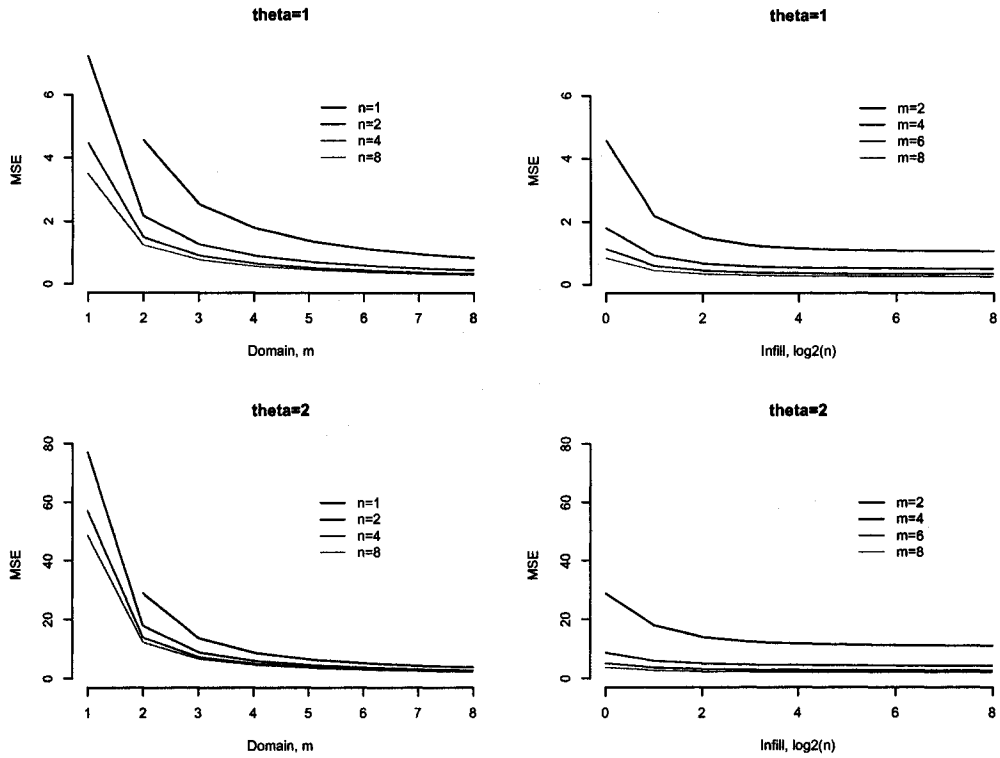

Figure 3.3: Expected mean square error for $\hat{\theta}$ as a function of m and n for $\theta = \{1, 2\}$. The left panels depict the MSE as a function of domain for four levels of infill where $n = \{1, 2, 4, 8\}$. The right hand panels illustrate the MSE as a function of increasing sampling effort per unit length (infill) for four fixed domains $m = \{2, 4, 6, 8\}$. Note that the scale for the level of infill is $\log_2$ such that $n = \{2^0, 2^1, \ldots, 2^8\}$.

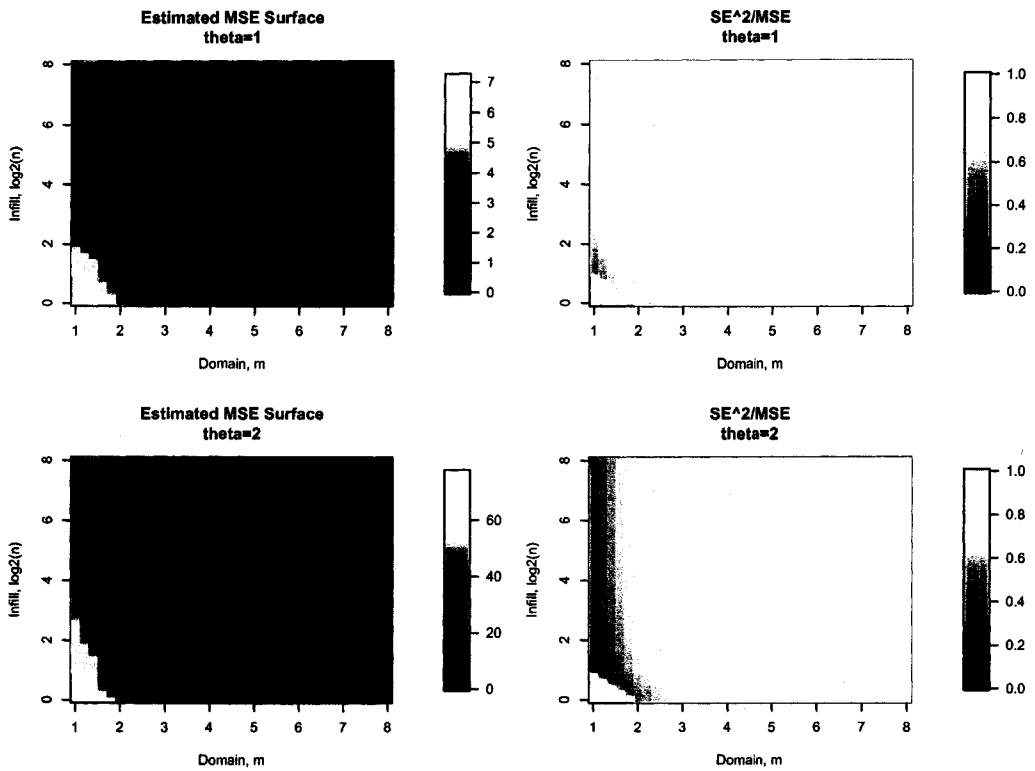

Figure 3.4: The left panels illustrate the expected mean square error for $\hat{\theta}$ as a function of m and n for $\theta = \{1, 2\}$. The right panels illustrate the proportion of the MSE attributable to the square of the standard error. Red regions indicate that the bias squared term dominates the MSE whereas white regions indicate that the standard error dominates. Note that the scale for the level of infill is $\log_2$ such that $n = \{2^0, 2^1, \ldots, 2^8\}$ in all four panels.

3.4 Asymptotic results for random locations along a transect

In the previous section, the results were restricted to data observed at regular intervals along a one-dimensional lattice. We would like to now present a similar analysis where we are no longer restricted to regular (fixed) intervals. Conceptually this is equivalent to making observations at random or irregularly spaced locations along a transect.

To begin we will assume that there exists a transect of finite length that can be divided into m unit lengths (or *blocks*). Examples include a strip of forest or open plain, or along a stream or river (assuming no confluences). We next assume that $\mathbf{t}_i$ is an $n-$vector of ordered locations within block i. To construct each $\mathbf{t}_i$ we first draw a simple random sample of size n from the sampling distribution $f_T(t)$ defined over the interval $[0,1]$. Define $\mathbf{t}_i = (i-1)\mathbf{1}+(t_{(1)}, t_{(2)}, \ldots, t_{(n)})' = (t_{i,1}, t_{i,2}, \ldots, t_{i,n})'$ where $t_{(j)}$ is the jth order statistic. We repeat this procedure for every $i = \{1, 2, \ldots, m\}$. It is not necessary for the number of sampling locations per block to be equal, only that the distribution of the sampling locations for each block remain identical. For example, one might consider letting the number of observations in block i to be a discrete random variable, e.g., negative binomial distribution, and then locate the n_i sites uniformly in that block. For the subsequent discussion and for notational convenience we will assume that n is constant relative to block. Therefore the total sampling effort is $N = mn$. We avoid resampling the same location twice by restricting the distribution of $f_{\mathbf{T}}(t)$ to be continuous. This assumption can be relaxed to include discrete probability distributions as long as the selection of sites is restricted to sampling without replacement. Finally, we assume that the selection of the sampling locations occurs independently from the process $Y(t)$.

The asymptotic behavior of parameter estimation can be easily explored using the current paradigm. For example, one can concatenate additional observations for blocks $m+1, \ldots, m+k$ to the current sample for expanding domain asymptotics.

The process of infill can modeled by generating additional observations for each block over the current domain.

Often we will require the distance between subsequent locations (spacings). Define δ as the $mn-$vector of spacings such that $\delta_{i,j} = t_{i,j} - t_{i,j-1}$ for $i = \{1, \ldots, m\}$ and $j = \{1, \ldots, n\}$. For all $i > 1$ we define $t_{i,0} = t_{i-1,n}$, i.e., use the last sampling site from the previous block. Hence $\delta_{i,1} = t_{i,1} - t_{i-1,n}$. Let $\delta_{1,1} = \infty$; this corresponds to not having a sampling location prior to $t_{1,1}$ and, consequently, implies that the first observation of the process is a single realization from a mean zero normal random variable with variance σ^2, i.e., $Y_{t_{1,1}} \sim \mathcal{N}(0, \sigma^2)$. By constructing the sampling design as outlined above, the realizations of $\boldsymbol{\delta}_i = (\delta_{i,1}, \delta_{i,2}, \ldots, \delta_{i,n})'$ for $i \geq 2$ are identically distributed. However the $\boldsymbol{\delta}_i$'s are not independent due to the presence $t_{i,n}$ in both $\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_{i-1}$. Figure (3.5) illustrates how the $\delta_{i,j}$'s vary for different sampling patterns. Generalize (3.3) with

$$Y_{t_{i,j}} = e^{-\delta_{i,j}/\theta} Y_{t_{i,j-1}} + \varepsilon_{t_{i,j}} \quad i = \{1, \ldots, m\}; j = \{1, \ldots, n\}, \tag{3.17}$$

where the $\varepsilon_{t_{i,j}}$ are independent, mean zero normal random variables with variance $\sigma^2(1 - e^{-2\delta_{i,j}/\theta})$. Note that for $i > 1$ and $j = 1$, $Y_{t_{i,j-1}} = Y_{t_{i-1,n}}$. The expected value and variance at each site remains unchanged, i.e., $\mathrm{E}Y_{t_{i,j}} = 0$ and $\mathrm{E}Y_{t_{i,j}}^2 = \sigma^2$. Similarly, the covariance between any two sites $l_{i,j}$ and $l_{i',j'}$, conditioned on the sampling locations $\mathbf{t}$, is only a function of the distance between them and can be written as $\mathrm{Cov}(Y_{t_{i,j}}, Y_{t_{i',j'}}) = \sigma^2 e^{-|t_{i',j'} - t_{i,j}|/\theta}$.

For notational convenience define $\phi = \exp(-1/\theta)$ and rewrite (3.17) as $Y_{t_{i,j}} = \phi^{\delta_{i,j}} Y_{t_{i,j-1}} + \varepsilon_{t_{i,j}}$. Note that as $\delta_{i,j}$ tends toward zero, $\phi^{\delta_{i,j}}$ tends toward one, which is consistent with near neighbors being more highly correlated with one another. Conversely, for (relatively) large $\delta_{i,j}$, subsequent locations are nearly independent since $\phi^{\delta_{i,j}} \to 0$.

A consequence of allowing the distance between observation sites to be random is that there no longer exists a closed form for the (conditional) maximum

likelihood estimator $\hat{\phi}$. Estimation requires maximization of the likelihood function with respect to $\phi = \exp\{-1/\theta\}$. We will show that the asymptotic distribution of the estimator of θ goes in distribution to a normal random variable. This will be established using a local asymptotic argument. Prior to working with the likelihood function we will first demonstrate the analysis by determining the asymptotic distribution of a weighted least squares (WLS) estimate of θ. Recall that the conditional MLE and the least squares estimate are identical when the sampling locations are equispaced. The general motivation for working with WLS first is that the objective function to be optimized is considerably simpler than the likelihood function and it will provide an opportunity to illustrate the solution strategy to be applied to the MLE. We will show that the WLS estimator is also asymptotically normal.

3.4.1 Weighted Least Squares Approach

We begin by defining the weighted least squares objective function to be minimized with respect to $\phi = \exp\{-1/\theta\}$:

$$S(\phi) = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(Y_{t_{i,j}} - \phi^{\delta_{i,j}} Y_{t_{i,j}-1}\right)^2 / \delta_{i,j}. \tag{3.18}$$

Let $\hat{\phi}$ be the value of ϕ that minimizes (3.18), i.e., $\hat{\phi} = \underset{\phi\in(0,1)}{\arg\min}\, S(\phi)$. The distance $\delta_{i,j}$ found in the denominator of the individual terms in the sum approximates the expression $1 - \phi^{2\delta_{i,j}}$ that appears in the quadratic term of likelihood function. For $\delta_{i,j}$ small, $1 - \phi^{2\delta_{i,j}} \sim 2\theta^{-1}\delta_{i,j}$. Hence, sampling locations that are relatively close to one another are weighted more heavily than locations that are further apart. Thus (3.18) is a weighted least squares estimate for ϕ. It should be noted that the contribution to $S(\phi)$ is zero when $i = 1$ and $j = 1$ since $\phi^{\infty} = 0$ for all $\phi \in (0, 1)$ which further implies $Y_{t_{1,1}}^2/\delta_{1,1} = 0$. The following theorem describes the asymptotic distribution of $\hat{\theta}$.

THEOREM 2

Let $\mathbf{t}$ *be an* mn*-vector of sampling locations along a transect of length* m *such that each unit length (block) contains* n *sampling locations constructed using* $f_T(t)$ *as outlined above. Let* $\boldsymbol{\delta}$ *be the* $mn-$*vector of spacings between subsequent sampling locations such that* $\delta_{i,j} = t_{i,j} - t_{i,j-1}$ *as also described above. For ease of notation we introduce the random variable* δ *which corresponds to a randomly selected spacing from block* $i \geq 2$ *such that* $\delta = \delta_{i,j}$ *with probability* $1/n$ *so that* $\mathrm{E}[f(\delta)] = n^{-1}\sum \mathrm{E}[f(\delta)]$. *Let* $\mathbf{Y}$ *be an* mn*-vector of observations corresponding to the sampling locations* $t_{i,j}$ *generated by the continuous AR(1) process as defined by (3.2) replacing* $Y_i = Y_{t_{i,j}}$. *Define* $\hat{\theta}$ *to be the weighted least squares estimator of the range parameter* θ *that minimizes (3.18). Then,*

(a) for fixed n *and* $m \to \infty$,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\theta^4}{n}\frac{\mathrm{E}\left[\phi_0^{2\delta}(1-\phi_0^{2\delta})\right]}{\left(\mathrm{E}\left[\phi_0^{2\delta}\delta\right]\right)^2}\right),$$

(b) as $n \to \infty$ *and* $m \to \infty$,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 2\theta^3).$$

Proof of Theorem 2:

(a) We will parallel the approach taken in Davis and Dunsimir (1997) by first reparameterizing the objective function $S(\phi)$ by setting $\phi = \phi_0 + \gamma/\sqrt{m}$ where $\phi_0 = \exp\{-1/\theta\}$. One can think of $\gamma/\sqrt{m}$ as a scaled deviation from the "true" value ϕ_0. Rewrite (3.18) as

$$T_m(\gamma) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{t_{i,j}} - (\phi_0 + \gamma/\sqrt{m})^{\delta_{i,j}} Y_{t_{i,j-1}}\right)^2/\delta_{i,j}. \qquad (3.19)$$

Next define the objective function $T_m^\star(\gamma) = T_m(\gamma) - T_m(0) = S(\phi) - S(\phi_0)$. Note that minimizing $T_m^\star(\gamma)$ with respect to γ is equivalent to minimizing

$S(\phi)$ with respect to ϕ since $S(\phi_0)$ is a constant. We intend to show that $T_m^\star(\gamma) \xrightarrow{\mathrm{d}} T(\gamma)$ on $C[-k,k]$ for every $k>0$ and furthermore, that the γ that minimizes $T(\gamma)$ is a mean zero normal random variable. We begin by showing that $T_m^\star(\gamma)$ converges pointwise to $T(\gamma)$. Hence,

$$\begin{aligned}
T_m^\star(\gamma) &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{t_{i,j}} - (\phi_0+\gamma/\sqrt{m})^{\delta_{i,j}}Y_{t_{i,j-1}}\right)^2/\delta_{i,j} \\
&\qquad - \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{t_{i,j}} - \phi_0^{\delta_{i,j}}Y_{t_{i,j-1}})^2/\delta_{i,j} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\left(\phi_0^{\delta_i}Y_{t_{i,j-1}} + \varepsilon_{t_{i,j}} - (\phi_0+\gamma/\sqrt{m})^{\delta_i}Y_{t_{i,j-1}}\right)^2 - \varepsilon_{t_{i,j}}^2\right)/\delta_{i,j} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\left(\varepsilon_{t_{i,j}} + \left[\phi_0^{\delta_{i,j}} - (\phi_0+\gamma/\sqrt{m})^{\delta_{i,j}}\right]Y_{t_{i,j-1}}\right)^2 - \varepsilon_{t_{i,j}}^2\right)/\delta_{i,j} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\phi_0^{\delta_{i,j}} - (\phi_0+\gamma/\sqrt{m})^{\delta_{i,j}}\right]^2 Y_{t_{i,j-1}}^2/\delta_{i,j} \\
&\qquad - 2\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\phi_0^{\delta_{i,j}} - (\phi_0+\gamma/\sqrt{m})^{\delta_{i,j}}\right]Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/\delta_{i,j}.
\end{aligned} \tag{3.20}$$

The second order Taylor expansion of $(\phi_0+\gamma/\sqrt{m})^{\delta_{i,j}}$ is

$$\begin{aligned}
(\phi_0+\gamma/\sqrt{m})^{\delta_{i,j}} &= \phi_0^{\delta_{i,j}} + \gamma\frac{\delta_{i,j}}{\sqrt{m}}\phi_0^{\delta_{i,j}-1} \\
&\qquad + \frac{\gamma^2}{2}\frac{\delta_{i,j}(\delta_{i,j}-1)}{m}(\phi_0+\gamma_{i,j}^*/\sqrt{m})^{\delta_{i,j}-2},
\end{aligned} \tag{3.21}$$

where $\gamma_{i,j}^*$ is an intermediate value such that $|\gamma_{i,j}^*| \in (0,|\gamma|)$. Similarly, the third order Taylor expansion of $(\phi_0+\gamma/\sqrt{m})^{\delta_{i,j}}$ is

$$\begin{aligned}
(\phi_0+\gamma/\sqrt{m})^{\delta_{i,j}} &= \phi_0^{\delta_{i,j}} + \gamma\frac{\delta_{i,j}}{\sqrt{m}}\phi_0^{\delta_{i,j}-1} + \frac{\gamma^2}{2}\frac{\delta_{i,j}(\delta_{i,j}-1)}{m}\phi_0^{\delta_{i,j}-2} \\
&\qquad + \frac{\gamma^3}{6}\frac{\delta_{i,j}(\delta_{i,j}-1)(\delta_{i,j}-2)}{m^{3/2}}(\phi_0+\gamma_{i,j}^{**}/\sqrt{m})^{\delta_{i,j}-3},
\end{aligned} \tag{3.22}$$

where $\gamma_{i,j}^{**}$ is an intermediate value such that $|\gamma_{i,j}^{**}| \in (0,|\gamma|)$.

Substituting expressions (3.21) and (3.22) into (3.20) and rearranging terms yields

$$\begin{aligned} T_m^*(\gamma) = \frac{\gamma^2}{m}\phi_0^{-2}\sum_{i=1}^{m}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}Y_{t_{i,j-1}}^2 - \frac{2\gamma}{\sqrt{m}}\phi_0^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}} \\ + \frac{\gamma^3}{m^{3/2}}\phi_0^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\delta_{i,j}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}(\phi_0+\gamma_{i,j}^*/\sqrt{m})^{\delta_{i,j}-2}Y_{t_{i,j-1}}^2 \\ + \frac{\gamma^4}{4m^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\delta_{i,j}(\delta_{i,j}-1)^2(\phi_0+\gamma_{i,j}^*/\sqrt{m})^{2\delta_{i,j}-4}Y_{t_{i,j-1}}^2 \\ - \frac{\gamma^2}{m}\phi_0^{-2}\sum_{i=1}^{m}\sum_{j=1}^{n}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}} \\ - \frac{\gamma^3}{3m^{3/2}}\sum_{i=1}^{m}\sum_{j=1}^{n}(\delta_{i,j}-1)(\delta_{i,j}-2)(\phi_0+\gamma_{i,j}^{**}/\sqrt{m})^{\delta_{i,j}-3}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}. \end{aligned} \tag{3.23}$$

The first term of (3.23) converges in probability to a constant, the second term converges in distribution to a mean zero normal random variable, and the remaining four terms (which are effectively error terms) converge to zero in probability.

Define $U_{n_i} = \sum_{j=1}^{n}\delta_{i,j}\phi_{0_{i,j}}^{2\delta_{i,j}}Y_{t_{i,j-1}}^2$ and let $\{\mathcal{F}_i, i = 2,3,\ldots\}$ be an increasing sequence of σ-fields. Note that the U_{n_i} are identically distributed by assumption but they are not independent due to the presence of $Y_{t_{i,j}}^2$ (which are correlated across blocks). In fact $\{U_{n_i}, i = 2,3,\ldots\}$ is a strictly stationary ergodic sequence. Thus a direct application of the ergodic theorem (Hall and Heyde, 1980) yields

$$\frac{1}{m}\sum_{i=1}^{m}\left[\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}Y_{t_{i,j}}^2\right] \xrightarrow{\mathrm{P}} \sigma^2 n\mathrm{E}\left[\phi_0^{2\delta}\delta\right] \tag{3.24}$$

and the desired result is obtained.

Now we turn our attention to the second term of (3.23). Let $V_{n_i} = \sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}$ such that $\{V_{n_i}, i = 2,3,\ldots\}$ is also a strictly stationary sequence of random variables on Ω.

Next we evaluate the mean and variance of V_{n_i}. We have

$$\begin{aligned} \mathrm{E}V_{n_i} &= \mathrm{E}\left[\mathrm{E}\left[\sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right]\right] \\ &= \mathrm{E}\left[\sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j}\ -1}\mathrm{E}\left[\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right]\right] \\ &= 0 \end{aligned} \tag{3.25}$$

since $Y_{t_{i,j-1}}$ and $\varepsilon_{t_{i,j}}$ are independent and $\mathrm{E}\left[\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right] = 0$. Moreover,

$$\begin{aligned} \mathrm{E}V_{n_i}^2 &= \mathrm{E}\left[\left(\sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}\right)^2\right] \\ &= \mathrm{E}\left[\sum_{j=1}^{n}\phi_0^{2\delta_{i,j}}Y_{t_{i,j-1}}^2\varepsilon_{t_{i,j}}^2\right] + \mathrm{E}\left[\sum_{j\neq k}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}\phi_0^{\delta_{i,k}}Y_{t_{i,k-1}}\varepsilon_{t_{i,k}}\right]. \end{aligned} \tag{3.26}$$

The second term of (3.26) is zero because $\varepsilon_{t_{i,j}}$ and $\varepsilon_{t_{i,k}}$ are uncorrelated for all $j \neq k$ and for $j < k$, $\varepsilon_{t_{i,k}}$ is independent of $Y_{t_{i,j}}$. Continuing,

$$\begin{aligned} \mathrm{E}\left[\sum_{j=1}^{n}\phi_0^{2\delta_{i,j}}Y_{t_{i,j-1}}^2\varepsilon_{t_{i,j}}^2\right] &= \mathrm{E}\left[\mathrm{E}\left[\sum_{j=1}^{n}\phi_0^{2\delta_{i,j}}Y_{t_{i,j-1}}^2\varepsilon_{t_{i,j}}^2|\boldsymbol{\delta}\right]\right] \\ &= \mathrm{E}\left[\sum_{j=1}^{n}\phi_0^{2\delta_{i,j}}Y_{t_{i,j-1}}^2\mathrm{E}\left[\varepsilon_{t_{i,j}}^2|\boldsymbol{\delta}\right]\right] \\ &= \sigma^4\mathrm{E}\left[\sum_{j=1}^{n}\phi_0^{2\delta_{i,j}}(1-\phi_0^{2\delta_{i,j}})\right] \\ &= \sigma^4 n\mathrm{E}\left[\phi_0^{2\delta}(1-\phi_0^{2\delta})\right]. \end{aligned} \tag{3.27}$$

Thus by direct application of the central limit theorem for a stationary martingale sequence (Hall and Heyde, 1980), $m^{-1/2}\sum_{i=2}^{m}V_{n_i}$ converges in distribution to a normal random variable with mean zero and variance $n\mathrm{E}\left[\phi_0^{2\delta}(1-\phi_0^{2\delta})\right]$, i.e.,

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\sum_{i=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}\right] \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, n\mathrm{E}\left[\phi_0^{2\delta}(1-\phi_0^{2\delta})\right]\right). \tag{3.28}$$

We now take each of the remaining four terms and show that each converges to zero in probability as $m \longrightarrow \infty$. First recall that $\phi_0 \in (0,1)$ and $\delta_{i,j} \in (0,2)$

for all i and j (since we are assuming there are $n \geq 1$ observations per block). Thus $\phi_0^{\delta_{i,j}} < 1$ and all polynomials of $\delta_{i,j}$ are bounded in absolute value by some constant, say C_1. Furthermore, since $|\gamma_{i,j}^*| \in (0, |\gamma|)$ there exists for every $\gamma > -\sqrt{m}\phi_0$ a constant, say C_2, such that $(\phi_0 + \gamma_{i,j}^*/\sqrt{m})^{\delta_{i,j}-2} < C_2$. Hence,

$$\begin{aligned}
&\mathrm{E}\left|\frac{\gamma^3}{m^{3/2}}\phi_0^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\delta_{i,j}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}(\phi_0+\gamma_{i,j}^*/\sqrt{m})^{\delta_{i,j}-2}Y_{t_{i,j-1}}^2\right| \\
&\quad\leq \frac{1}{\sqrt{m}}\left[\frac{|\gamma^3|}{m}\phi_0^{-1}\sum_{i=1}^{m}\left[C_1\cdot 1\cdot C_2\sum_{j=1}^{n}\mathrm{E}\left(Y_{t_{i,j-1}}^2\right)\right]\right] \\
&\quad= \frac{1}{\sqrt{m}}|\gamma^3|\phi_0^{-1}C_1C_2n\sigma^2 \\
&\quad\to 0 \text{ as } m\to\infty.
\end{aligned} \tag{3.29}$$

The exact same argument can be made for the next term in (3.23) replacing C_1 and C_2 with C_1' and C_2', respectively, i.e.,

$$\begin{aligned}
&\mathrm{E}\left|\frac{\gamma^3}{m^{3/2}}\sum_{i=1}^{m}\sum_{j=1}^{n}\delta_{i,j}(\delta_{i,j}-1)^2(\phi_0+\gamma_{i,j}^*/\sqrt{m})^{2\delta_{i,j}-4}Y_{t_{i,j-1}}^2\right| \\
&\quad\leq \frac{1}{\sqrt{m}}\left[\frac{|\gamma^3|}{m}\sum_{i=1}^{m}\left[C_1'\cdot 1\cdot C_2'\sum_{j=1}^{n}\mathrm{E}\left(Y_{t_{i,j-1}}^2\right)\right]\right] \\
&\quad= \frac{1}{\sqrt{m}}|\gamma^3|C_1'C_2'n\sigma^2 \\
&\quad\to 0 \text{ as } m\to\infty.
\end{aligned} \tag{3.30}$$

For the fifth term, define $W_{n_i} = \sum_{j=1}^{n}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}$ and note that $\{W_{n_i}, i = 2,3,\ldots\}$ is a strictly stationary sequence of correlated random variables on Ω. We evaluate the mean as follows.

$$\begin{aligned}
\mathrm{E}W_{n_i} &= \mathrm{E}\left[\mathrm{E}\left[\sum_{j=1}^{n}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right]\right] \\
&= \mathrm{E}\left[\sum_{j=1}^{n}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\mathrm{E}\left[\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right]\right] \\
&= 0
\end{aligned} \tag{3.31}$$

since $Y_{t_{i,j-1}}$ and $\varepsilon_{t_{i,j}}$ are independent and $\mathrm{E}\left[\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right] = 0$. Direct application of the ergodic theorem yields

$$\frac{\gamma^2}{\phi_0^2} \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} W_{n_i} \xrightarrow{\mathrm{P}} 0. \tag{3.32}$$

The limit of the last term can be determined in exactly the same manner as the previous term. That is to say, if one defines $X_{n_i} = \sum_{j=1}^{n}(\delta_{i,j}-1)(\delta_{i,j}-2)(\phi_0+\gamma_{i,j}^{**}/\sqrt{m})^{\delta_{i,j}-3}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}$, then $\{X_{n_i}, i = 2,3,\ldots\}$ is a strictly stationary sequence of random variables. Therefore,

$$\frac{\gamma^3}{3}\frac{1}{\sqrt{m}}\frac{1}{m}\sum_{i=1}^{m}|X_{n_i}| \to 0,$$

and hence,

$$\frac{\gamma^3}{3}\frac{1}{\sqrt{m}}\frac{1}{m}\sum_{i=1}^{m}X_{n_i} \xrightarrow{\mathrm{P}} 0.$$

Having determined the limiting distributions for each term in the objective function (3.20), we see that for fixed γ

$$T_m^{\star}(\gamma) \xrightarrow{\mathrm{d}} \gamma^2\phi_0^{-2}n\mathrm{E}\left[\phi_0^{2\delta}\delta\right] - 2\gamma\phi_0^{-1}N_n, \tag{3.33}$$

where $N_n \sim \mathcal{N}\left(0, n\mathrm{E}\left[\phi_0^{2\delta}(1-\phi_0^{2\delta})\right]\right)$.

We now extend the pointwise convergence to process convergence over a compact set of $\mathbb{R}$, i.e., $C[-k,k]$ for any $k > 0$. First rewrite (3.23) as

$$\begin{aligned} T_m^{\star}(\gamma) &= \frac{\gamma^2}{m}\phi_0^{-2}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_0^{2\delta_{i,j}}\delta_{i,j}Y_{t_{i,j-1}}^2 - \frac{2\gamma}{\sqrt{m}}\phi_0^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}} + o_{\mathrm{p}}(1) \\ &= T_{m_1}^{\star}(\gamma) - T_{m_2}^{\star}(\gamma) + o_{\mathrm{p}}(1), \end{aligned}$$

where $T_{m_1}^{\star}(\gamma) = \gamma^2\phi_0^{-2}m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_0^{2\delta_{i,j}}\delta_{i,j}Y_{t_{i,j-1}}^2$,
$T_{m_2}^{\star}(\gamma) = 2\gamma\phi_0^{-1}m^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}$, and for any constant k,

$$\sup_{|\gamma|\le k}|o_{\mathrm{p}}(1)| \xrightarrow{\mathrm{P}} 0.$$

First note that

$$\begin{aligned}
&\sup_{|\gamma|\leq k} \left|T^{\star}_{m_1}(\gamma) - \gamma^2\phi_0^{-2}n\mathrm{E}\left[\phi_0^{2\delta}\delta\right]\right| \\
&\qquad = \sup_{|\gamma|\leq k} \gamma^2 \left|\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_0^{2\delta_{i,j}}\delta_{i,j}Y^2_{t_{i,j-1}} - \phi_0^{-2}n\mathrm{E}\left[\phi_0^{2\delta}\delta\right]\right| \qquad (3.34)\\
&\qquad \leq \sup_{|\gamma|\leq k} k^2\left|o_{\mathrm{p}}(1)\right| \\
&\qquad = o_{\mathrm{p}}(1),
\end{aligned}$$

implying that $T^{\star}_{m_1}(\gamma)$ converges in probability to $\gamma^2\phi_0^{-2}n\mathrm{E}[\phi_0^{2\delta}\delta]$ on $C[-k,k]$.

Now consider $T^{\star}_{m}(\tilde{\gamma})$. Since $m^{-1/2}\sum_{i=1}^{m}\left[\sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}\right] \xrightarrow{\mathrm{d}} N_n$, then

$$\begin{aligned}
\left|T_{m_2}(\tilde{\gamma}) - T_{m_2}(\gamma)\right| &= |\tilde{\gamma}-\gamma|\left|2\phi_0^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}\right| \qquad (3.35)\\
&= |\tilde{\gamma}-\gamma||2\phi_0^{-1}O_{\mathrm{p}}(1)|.
\end{aligned}$$

Therefore for any $\eta > 0$ there exists $\epsilon > 0$ such that

$$\begin{aligned}
&\mathrm{P}\left(\lim_{m}\sup_{|\tilde{\gamma}-\gamma|\leq\epsilon}\left|T^{\star}_{m_2}(\tilde{\gamma}) - T^{\star}_{m_2}(\gamma)\right| > \eta\right) \\
&\qquad = \mathrm{P}\left(\lim_{m}\sup_{|\tilde{\gamma}-\gamma|\leq\epsilon}|\tilde{\gamma}-\gamma||2\phi_0^{-1}O_{\mathrm{p}}(1)| > \eta\right) \\
&\qquad \leq \mathrm{P}\left(\lim_{m}\sup_{|\tilde{\gamma}-\gamma|\leq\epsilon}2\epsilon\phi_0^{-1}|O_{\mathrm{p}}(1)| > \eta\right) \\
&\qquad = \mathrm{P}\left(\lim_{m}\sup_{|\tilde{\gamma}-\gamma|\leq\epsilon}|O_{\mathrm{p}}(1)| > \frac{\phi_0}{2}\frac{\eta}{\epsilon}\right) \xrightarrow{\mathrm{P}} 0 \text{ as } m\to\infty \text{ and } \epsilon\to 0.
\end{aligned} \qquad (3.36)$$

Therefore $T_{m_2}(\gamma)$ is tight. Thus if $T(\gamma)$ has a unique minimum, say $\gamma_{\min}$, then there exists a minimum $\gamma^{\star}_{\min} = \arg\min T^{\star}_{m}(\gamma)$ such that $\gamma^{\star}_{\min} \xrightarrow{\mathrm{d}} \gamma_{\min}$ (Billingsley, 1995).

Equation (3.33) can now be minimized by taking the derivative with respect to γ, setting the result to zero, and solving for γ. Therefore,

$$\hat{\gamma} = \sqrt{m}(\hat{\phi}-\phi_0) \xrightarrow{\mathrm{d}} \phi_0\left(n\mathrm{E}\left[\phi_0^{2\delta}\delta\right]\right)^{-1}N_n. \qquad (3.37)$$

Taking $g(x) = -(\log(x))^{-1}$ and applying Slutsky's Theorem (as we did for the regular lattice) yields

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\theta^4}{n} \frac{\mathrm{E}\left[\phi_0^{2\delta}(1 - \phi_0^{2\delta})\right]}{\left(\mathrm{E}\left[\phi_0^{2\delta}\delta\right]\right)^2}\right), \tag{3.38}$$

and we obtain the desired result.

(b) Now we turn to the case where $n \to \infty$. As for the regular lattice we assume that $n(m) = n_m \to \infty$ as $m \to \infty$. Next we assume that $f_T(t) > 0$ *a.e.* where $t \in (0, 1]$. This implies that as $m \to \infty$, $t_{(1)} \to 0$ and $t_{(n)} \to 1$. Moreover, since for each $i \in \{1, 2, \ldots, m\}$ $\delta_{i,j} = t_{i,j} - t_{i,j-1}$, then

$$\begin{aligned} \sum_{j=1}^{n_m} \delta_{i,j} &= \sum_{j=1}^{n_m} [t_{i,j} - t_{i,j-1}] \\ &= t_{i,0} - t_{i,n} \\ &= t_{i-1,n} - t_{i,n} \to 1 \text{ as } m \to \infty. \end{aligned} \tag{3.39}$$

Note that for $i = 1$ we redefine $t_{i,0} = 0$; that is to say the sampling location of the first observation approaches zero from the right and the sampling location of the prior *unobserved* response approaches zero from the left. Since the previous arguments for fixed n still hold, all that remains to show is the limiting values for the expressions $\sum_{j=1}^{n_m} \phi_0^{2\delta_{i,j}} \delta_{i,j}$ and $\sum_{j=1}^{n_m} \phi_0^{2\delta_{i,j}}(1 - \phi_0^{2\delta_{i,j}})$. Hence for every $i \in \{1, 2, \ldots\}$,

$$\begin{aligned} \lim_{m\to\infty} \sum_{j=1}^{n_m} \phi_0^{2\delta_{i,j}} \delta_{i,j} &= \lim_{m\to\infty} \sum_{j=1}^{n_m} \left[1 - \frac{2\delta_{i,j}}{\theta} + \frac{1}{2}\left(\frac{2\delta_{i,j}}{\theta}\right)^2 - \ldots\right] \delta_{i,j} \\ &= \lim_{m\to\infty} \sum_{j=1}^{n_m} \delta_{i,j} - \lim_{m\to\infty} \sum_{j=1}^{n_m} \left[2\theta^{-1}\delta_{i,j}^2 - 2\theta^{-2}\delta_{i,j}^3 + \ldots\right] \\ &= 1, \end{aligned}$$

and

$$\begin{aligned} \lim_{m\to\infty} \sum_{j=1}^{n_m} \phi_0^{2\delta_{i,j}}(1 - \phi_0^{2\delta_{i,j}}) &= \lim_{m\to\infty} \sum_{j=1}^{n_m} \phi_0^{2\delta_{i,j}} \left[\frac{2\delta_{i,j}}{\theta} - \frac{1}{2}\left(\frac{2\delta_{i,j}}{\theta}\right)^2 + \ldots\right] \\ &= \frac{2}{\theta} \lim_{m\to\infty} \sum_{j=1}^{n_m} \phi_0^{2\delta_{i,j}} \delta_{i,j} - \lim_{m\to\infty} \sum_{j=1}^{n_m} \left[\frac{2}{\theta^2}\phi_0^{2\delta_{i,j}}\delta_{i,j}^2 + \ldots\right] \\ &= 2\theta^{-1}. \end{aligned}$$

Substituting the above expressions into (3.38) we obtain the desired result, i.e.,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 2\theta^3).$$

□

REMARK 3 This result clearly demonstrates that the distribution of the sampling procedure has a direct impact on the asymptotic variance of the estimator. Note that if $\boldsymbol{\delta} \equiv n^{-1}\mathbf{1}$ we recover the same limiting distribution derived for the regular lattice. Explicitly, the variance reduces to

$$\begin{aligned} \frac{\theta^4 \mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}(1 - \phi_0^{2\delta_j})\right]}{\left(\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}\delta_j\right]\right)^2} &= \frac{\theta^4 n \phi_n^2 (1 - \phi_n^2)}{(n\phi_n^2 n^{-1})^2} \\ &= \theta^4 n \left(e^{2/n\theta} - 1\right). \end{aligned} \tag{3.40}$$

Furthermore, we can show heuristically that sampling along a regular lattice is most efficient (minimizes the asymptotic variance for a fixed sampling effort n per block). For small δ_j the asymptotic variance can be approximated by

$$\begin{aligned} \theta^4 \frac{\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}(1 - \phi_0^{2\delta_j})\right]}{\left(\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}\delta_j\right]\right)^2} &= \theta^4 \frac{\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}(2\delta_j\theta^{-1} + \mathcal{O}(\delta_j^2))\right]}{\left(\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}\delta_j\right]\right)^2} \\ &\approx 2\theta^3 \frac{\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}\delta_j\right]}{\left(\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}\delta_j\right]\right)^2} \\ &= 2\theta^3 \left(\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}\delta_j\right]\right)^{-1}. \end{aligned}$$

To minimize the asymptotic variance, we need to maximize $\mathrm{E}[f(\boldsymbol{\delta})]$, where $f(\boldsymbol{\delta}) = \sum_{j=1}^{n} \phi_0^{2\delta_j}\delta_j$ subject to the constraint $\sum_{j=1}^{n} \delta_j = 1$. Symmetry considerations suggest the maximum occurs when $\delta_j = n^{-1}$ for $j = 1, \ldots, n$. We explore this result through simulation in Section 3.4.1.1 where we compare several different sampling patterns.

3.4.1.1 Simulation Results

To demonstrate the impact of the sampling pattern on the asymptotic variance we generated a series of realizations using five different sampling patterns to compare against the expected asymptotic variance. The five sampling patterns include (1) making observations along a regular lattice, (2) generating n random locations uniformly per block, (3) sampling n random locations over each block using a beta(3,3) distribution, (4) sampling n random locations using the beta(10,10) distribution, and (5) sampling N_i uniformly distributed locations where N_i has a negative binomial distribution. Each of the first four sampling patterns has n observations per block. Patterns (3) and (4) tend to cluster the sampling locations toward the center of each block with the latter more tightly clustered. The final sampling pattern, referred to hereafter as the clustered pattern, varies the number of sampling locations per block thereby allowing some blocks to contain few or no observations and others to contain many observations. In order to compare the sampling patterns on the same footing, we set the expected number of sampling locations per block for the clustered pattern equal to the number of locations per block for the other four patterns. Figure (3.5) illustrates a single realization of each pattern over the first two blocks with $n = 4$ for patterns (1) through (4) and $\mathrm{E}N_i = 4$ for the clustered pattern.

Figures 3.6 and 3.7 illustrate the MSE surface as a function of domain and level of infill for the non-regular sampling patterns where $\theta = 1$ and $\theta = 2$, respectively. Note that the overall shape of the surfaces are somewhat different than those generated by restricting observations to a regular lattice(see Figure 3.1). For $\theta = 1$ the contours of the MSE are not as close to parallel to the infill axis as for the regular lattice sampling pattern. This is in part to the relatively large range of the temperature (color) scale. The scale for the WLS non-regular sampling images is roughly

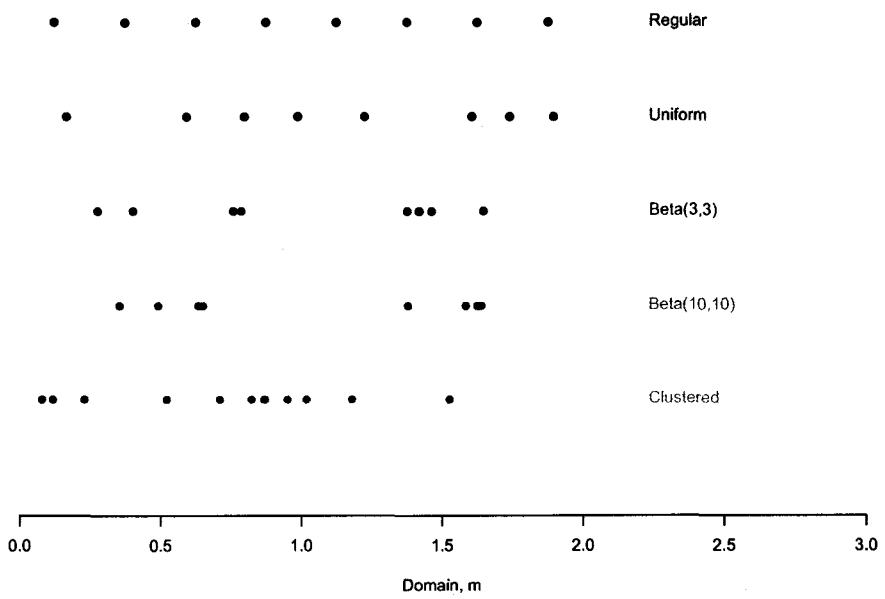

Figure 3.5: Partial realization of the sampling patterns used during simulation. The first four patterns (regular, uniform, beta(3,3), and beta(10,10)) all have four locations per block while the number of sites per block is random for the clustered pattern ($\mathrm{E}N = 4$).

twice that of the comparable regular pattern image due to large MSE for very meager sampling efforts, i.e., $(m, n) = (1, 2)$. For $\theta = 2$ the WLS results closely agree with the regular sampling pattern results. Once again the contours are nearly parallel to the infill axis and significant decreases in MSE are attributable to expansion of the domain. The uniform sampling pattern, for both $\theta = 1$ and $\theta = 2$, appears to be more efficient than the remaining three patterns. We illustrated earlier that the optimal sampling pattern for the AR(1) process is the regular pattern. The uniform pattern most resembles the regular lattice for moderate to high levels of infill unlike the beta patterns which tend to cluster observations toward the center of each unit length and the clustering pattern which, by design, varies the density of sampling from unit to unit. Thus it is not unexpected that the uniform sampling method is most efficient with respect to the standard error.

The asymptotic variance can be calculated exactly when sampling along a regular lattice. For the four remaining cases, the asymptotic variance was estimated using Monte Carlo simulation. We generated many realizations of each sampling pattern and empirically computed

$$\frac{\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}(1 - \phi_0^{2\delta_j})\right]}{\left(\mathrm{E}\left[\sum_{j=1}^{n} \phi_0^{2\delta_j}\delta_j\right]\right)^2}$$

with $\theta = 1$. Figure (3.8) illustrates how the asymptotic variance approaches $2\theta^3 = 2$ as the sampling effort per block increases. Note that the asymptotic variance for the regular pattern is uniformly best for all sampling efforts and coincides with the heuristic result presented in the first half of the proof. For the most meager of sampling efforts ($n = 1$) the beta distributions outperform the uniform distribution. This is because the beta distributions tend to place the sampling location toward the center of the block thereby creating a more regular pattern across blocks. However, as n increases the sampling locations tend to remain near the center of each block and the beta patterns diverge from the optimal design. Conversely, the uniform pattern

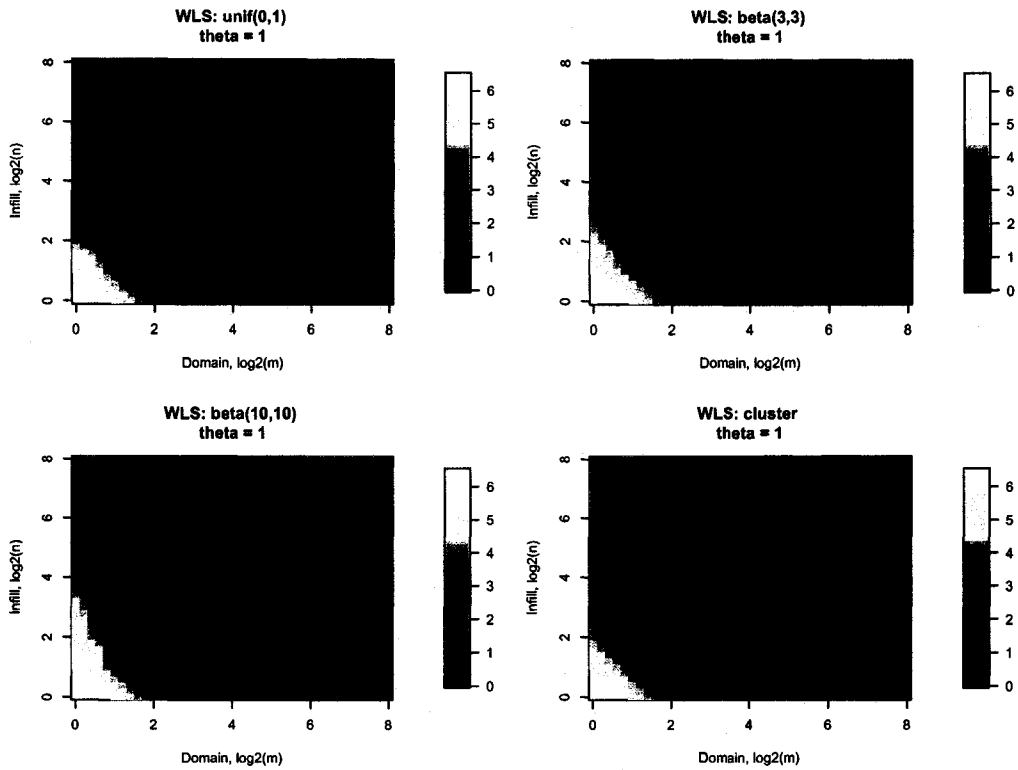

Figure 3.6: MSE for the weighted least squares estimator (WLS) of an AR(1) process for each non-regular sampling pattern where $\theta = 1$. The surfaces represent the results based on 1000 independent realizations for each combination of domain and level of infill. Note that the axes are $\log_2$ such that $m, n = \{2^0, 2^1, \ldots, 2^8\}$.

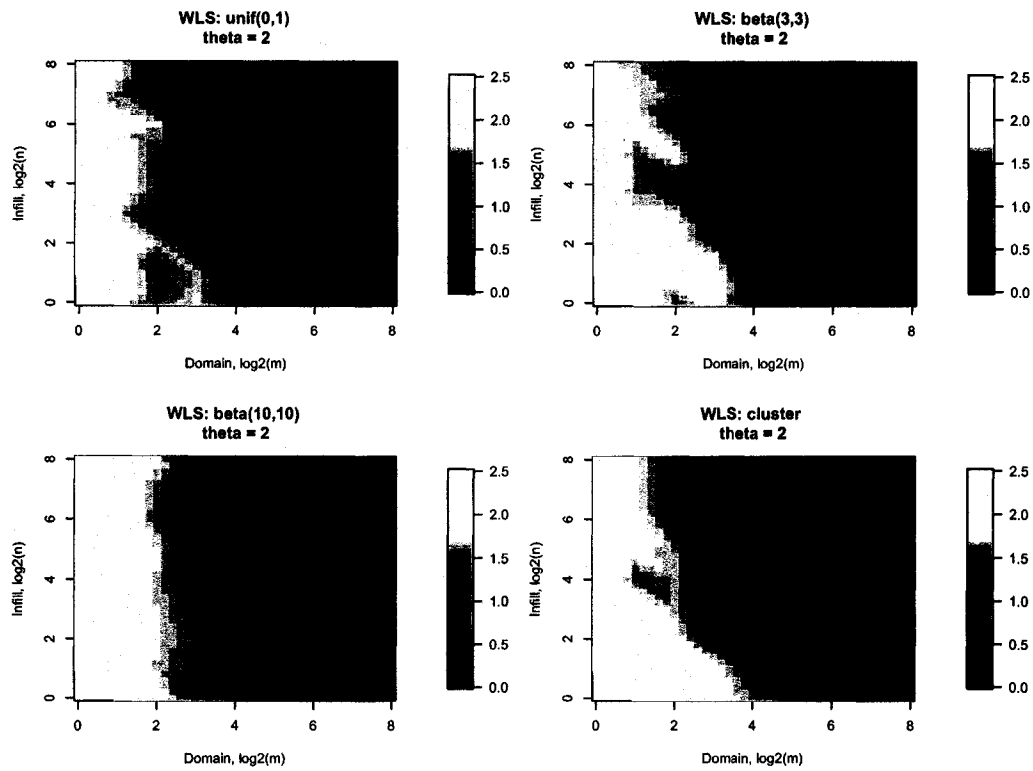

Figure 3.7: MSE for the weighted least squares estimator (WLS) of an AR(1) process for each non-regular sampling pattern where $\theta = 2$. The surfaces represent the results based on 1000 independent realizations for each combination of domain and level of infill. Note that the axes are $\log_2$ such that $m, n = \{2^0, 2^1, \ldots, 2^8\}$.

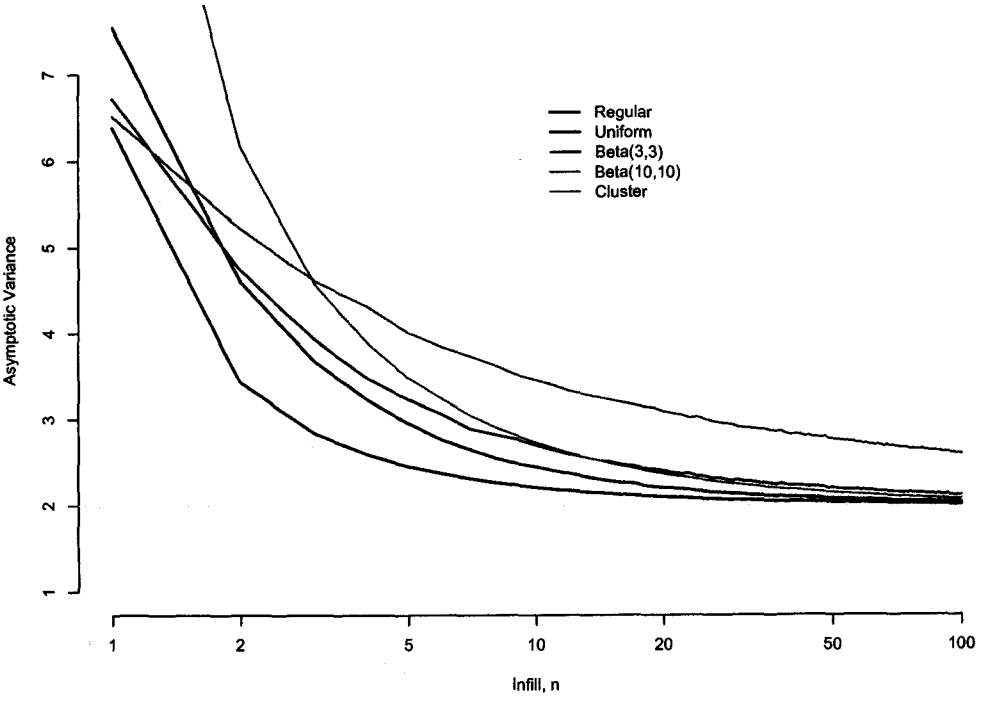

Figure 3.8: Asymptotic variance for the WLS estimator. The curve for the regular pattern is exact. The curves for the remaining four sample patterns are each based on 100 Monte Carlo simulations of $m = 1024$ blocks for all sampling efforts $n \in \{1, \ldots, 100\}$.

more evenly covers each block and, consequently, the overall pattern becomes more regular.

Table (3.2) lists the observed variance of the estimator $\hat{\theta}$. A total of 1000 simulations for each combination of sampling pattern and sampling effort were generated. For all simulations $\theta = 1$ and the domain was fixed at $m = 1024$. Thus the total sampling effort varies with the sampling effort. The estimate $\hat{\theta}$ is the optimal solution of the weighted least squares objective function (3.18) where $\phi = \exp\{-1/\theta\}$. Overall the observed variances very closely match the asymptotic variances, delim-

Table 3.2: Observed and expected variance of $\hat{\theta}$ for 1000 simulations with $\theta = 1$ as a function of sampling pattern and sampling effort. For the first four patterns the total sampling effort is $N = 1024n$ and for the clustered pattern the expected effort per block was set to $n = \{1, 2, 4, 8\}$. The expected asymptotic variances are delimited by parentheses.

Sampling Pattern	Sampling effort per block, n			
	$n = 1$	$n = 2$	$n = 4$	$n = 8$
Regular	6.71	3.26	2.39	2.28
	(6.38)	(3.43)	(2.59)	(2.27)
Uniform	7.94	4.28	3.23	2.58
	(7.55)	(4.62)	(3.23)	(2.55)
Beta(3,3)	6.89	4.80	3.36	3.01
	(6.72)	(4.75)	(3.48)	(2.83)
Beta(10,10)	6.12	5.09	4.24	3.63
	(6.52)	(5.23)	(4.31)	(3.63)
Clustered	11.6	5.86	3.73	2.97
	(11.9)	(6.17)	(3.89)	(2.91)

ited by parentheses. The largest relative difference is 8% and the mean difference is -0.1%.

3.4.2 Maximum Likelihood Approach

We now extend the estimation procedure to involve the exact likelihood function conditioned on the observed data $\mathbf{Y}$ at the sampling locations $\mathbf{t}$. In this case we will show that the asymptotic distribution of the MLE $\hat{\theta}$ is the same independent of sampling pattern. In other words, the likelihood function accounts for the distance between locations. The likelihood consists of two terms: a determinant component and a quadratic component. We will show that both components play an important role in the asymptotics.

We begin by assuming that the observed data are sampled from the Ornstein-Uhlenbeck process as defined in (3.2). We further assume that there are mn sampling locations, n per unit length, generated as described previously. The corresponding

observation vector $\mathbf{Y}$ is mn-dimensional with joint distribution, in matrix form, $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{\Gamma})$ where $\mathbf{0}$ is an mn-vector of zeros and $\mathbf{\Gamma}$ is an $mn \times mn$ correlation matrix such that the correlation between locations $t_{i,j}$ and $t_{i',j'}$ is $\exp\{-|t_{i,j} - t_{i',j'}|/\theta\}$. Since the sampling locations are random, we write the likelihood as

$$\mathrm{L}(\phi, \sigma^2; \mathbf{Y}, \mathbf{t}) \propto C \frac{1}{(2\pi)^{mn/2}} \frac{1}{|\sigma^2\Gamma|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{Y}'\Gamma^{-1}\mathbf{Y}\right\} \tag{3.41}$$

where $\phi = \exp\{-1/\theta\}$ and C is a proportionality constant resulting from $f(\mathbf{t}) = f(t_{1,1}, \ldots, t_{m,n})$.

We obtain the profile likelihood by taking the partial derivative of the log of (3.41) with respect to σ^2, setting the result to zero and solving for $\hat{\sigma}^2 = \sigma^2(\phi)$, and substituting $\hat{\sigma}^2$ in place of σ^2 into (3.41). Thus,

$$\ell_{\text{profile}}(\phi; \hat{\sigma}^2, \mathbf{Y}, \mathbf{t}) \propto \frac{mn}{2} \log \hat{\sigma}^2 - \frac{1}{2} \log |\mathbf{\Gamma}|, \tag{3.42}$$

where $\hat{\sigma}^2 = (mn)^{-1}\mathbf{Y}'\mathbf{\Gamma}^{-1}\mathbf{Y}$.

Defining $k = (i-1)n + j$, Brockwell and Davis (1996) show that (3.42) can be expressed in terms of the one-step prediction errors, $Y_{t_k} - \hat{Y}_{t_k}$, and their variances $\nu_{t_{k-1}}$, thus avoiding direct computation of Γ^{-1} and $\det\Gamma$. They show that $\mathbf{Y}'\Gamma^{-1}\mathbf{Y} = \sum_{k=1}^{mn}(Y_{t_k} - \hat{Y}_{t_k})^2/\nu_{t_{k-1}}$ and $\det\Gamma = \prod_{i=1}^{mn}\nu_{t_{k-1}}$. For this model, we find that $\hat{Y}_{t_k} = \phi^{\delta_{t_k}} Y_{t_{k-1}}$ and $\nu_{t_{k-1}} = 1 - \phi^{2\delta_{t_k}}$ where $\delta_{t_k} = t_k - t_{k-1}$. (We define $\delta_{t_1} = \infty$ which implies $\nu_{t_0} = 1$ and $\hat{Y}_{t_1} = 0$.) Returning to double indexing for the sampling locations, the reduced likelihood function is

$$\begin{aligned}\ell(\phi) = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\log(1-\phi^{2\delta_{i,j}})\right] \\ + \log\left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}(Y_{t_{i,j}} - \phi^{\delta_{i,j}}Y_{t_{i,j-1}})^2/(1-\phi^{2\delta_{i,j}})\right]\right).\end{aligned} \tag{3.43}$$

The first term arises from the determinant in the likelihood expression and the second term corresponds to the quadratic term. Minimizing (3.43) with respect to ϕ is equivalent to maximizing the conditional likelihood with respect to θ.

THEOREM 3

Let $\mathbf{t}$ *be an* mn*-vector of sampling locations along a transect of length* m *such that each unit length (block) contains* n *sampling locations. Define* $\delta_{i,j}$ *as the distance between subsequent sites such that* $\delta_{i,j} = t_{i,j} - t_{i,j-1}$ *where* $\delta_{i,1} = t_{i,1} - t_{i-1,n}$ *for* $i \geq 2$ *and* $\delta_{1,1} = \infty$. *Once again we introduce the random variable* δ *which corresponds to a randomly selected spacing from block* $i \geq 2$ *such that* $\delta = \delta_{i,j}$ *with probability* $1/n$ *so that* $\mathrm{E}[f(\delta)] = n^{-1} \sum \mathrm{E}[f(\delta)]$. *Let* $\mathbf{Y}$ *be an* mn*-vector of observations corresponding to the sampling locations* $\mathbf{t}$ *generated by the continuous AR(1) process as defined by (3.2) replacing* $Y_i = Y_{t_{i,j}}$. *Define* $\hat{\theta}$ *to be the estimator that minimizes (3.43). Then,*

(a) for fixed n *and* $m \to \infty$,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\theta^4}{n}\left(2\mathrm{Var}\left[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right] + \mathrm{E}\left[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right]\right)^{-1}\right),$$

(b) as $n \to \infty$ *and* $m \to \infty$,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 2\theta^3).$$

Proof of Theorem 3:

(a) As in the proof of Theorem 2, we begin by reparameterizing (3.43) using $\phi = \phi_0 + \gamma/\sqrt{m}$ where $\phi_0 = \exp\{-1/\theta\}$. The new parameter γ is interpreted as the deviation from the "true" value ϕ_0. Thus the reduced-likelihood can be written as

$$\ell(\gamma) = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} g_{i,j}(\gamma/\sqrt{m})\right] + \log\left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h_{i,j}(\gamma/\sqrt{m})\right]\right), \tag{3.44}$$

where

$$g_{i,j}(\gamma/\sqrt{m}) = \log\left(1 - (\phi_0 + \gamma/\sqrt{m})^{2\delta_{i,j}}\right),$$

$$h_{i,j}(\gamma/\sqrt{m}) = \left(Y_{t_{i,j}} - (\phi_0 + \gamma/\sqrt{m})^{\delta_{i,j}} Y_{t_{i,j-1}}\right)^2 / (1 - (\phi_0 + \gamma/\sqrt{m})^{2\delta_{i,j}}).$$

Next define $\ell_m^\star(\gamma) = m\left(\ell(\gamma) - \ell(0)\right)$. Note that minimizing $\ell_m^\star(\gamma)$ is equivalent to minimizing $\ell(\gamma)$ since m and $\ell(0)$ are constants. Therefore,

$$\begin{aligned}\ell_m^\star(\gamma) = &\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} g_{i,j}(\gamma/\sqrt{m})\right] - \sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\log\left(1-\phi_0^{2\delta_{i,j}}\right)\right] \\ &+ m\log\left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h_{i,j}(\gamma/\sqrt{m})\right]\right) \\ &- m\log\left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})\right]\right).\end{aligned} \tag{3.45}$$

We will show that $\ell_m^\star(\gamma) \to \ell(\gamma)$ on $C(\mathbb{R})$ and that... To evaluate the asymptotic distribution of $\hat{\gamma}$ with respect to m we need to expand (3.45) about $\gamma = 0$ using Taylor's theorem. We begin by looking at the terms contributed by the determinant. Define

$$\begin{aligned}D_m(\gamma) &= \sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} g_{i,j}(\gamma/\sqrt{m})\right] - \sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\log\left(1-\phi_0^{2\delta_{i,j}}\right)\right] \\ &= \sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\left(g_{i,j}(0) + \gamma g'_{i,j}(0) + \frac{\gamma^2}{2}g''_{i,j}(0) + R_1(i,j)\right)\right] \\ &\quad - \sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\log\left(1-\phi_0^{2\delta_{i,j}}\right)\right].\end{aligned} \tag{3.46}$$

where $R_1(i,j)$ is the remainder. Note that $R_1(i,j) \sim m^{-3/2}$ since

$$R_1(i,j) = \frac{\gamma^3}{6} g'''_{i,j}(\gamma^\star_{i,j})$$

where $|\gamma^\star_{i,j}| \in (0, |\gamma|)$ for all i and j. Define $\phi_\gamma = \phi_0 + \gamma/\sqrt{m}$. Thus

$$\begin{aligned}g_{i,j}(\gamma/\sqrt{m}) &= \log\left(1-\phi_\gamma^{2\delta_{i,j}}\right) \\ g'_{i,j}(\gamma/\sqrt{m}) &= -\frac{2\delta_{i,j}}{\sqrt{m}}\frac{\phi_\gamma^{2\delta_{i,j}-1}}{(1-\phi_\gamma^{2\delta_{i,j}})} \\ g''_{i,j}(\gamma/\sqrt{m}) &= -\frac{2\delta_{i,j}(2\delta_{i,j}-1)}{m}\frac{\phi_\gamma^{2\delta_{i,j}-2}}{1-\phi_\gamma^{2\delta_{i,j}})} - \frac{4\delta_{i,j}^2}{m}\frac{\phi_\gamma^{4\delta_{i,j}-2}}{\left(1-\phi_\gamma^{2\delta_{i,j}}\right)^2}.\end{aligned}$$

Therefore,

$$\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} g_{i,j}(0)\right] = \sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\log(1-\phi_0^{2\delta_{i,j}})\right], \tag{3.47}$$

and

$$\begin{aligned}\sum_{i=1}^{m}\left[\frac{1}{n}g'_{i,j}(0)\right] &= -2\phi_0^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= -2\phi_0^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}U_{n_i},\end{aligned} \tag{3.48}$$

where $U_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})$. The sequence $\{U_{n_i}, i = 2,3,\ldots\}$ is strictly stationary and 1-dependent; that is to say that U_{n_i} and U_{n_j} are independent for all $|i-j| > 1$. Application of the central limit theorem for strictly stationary m-dependent sequences (Brockwell and Davis, 1991) yields

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(U_{n_i} - \mathrm{E}[U_n]) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathrm{Var}[U_{n_i}]), \tag{3.49}$$

where

$$\mathrm{E}[U_n] = \mathrm{E}[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})], \tag{3.50}$$

and

$$\mathrm{Var}[U_n] = \mathrm{Var}\left[n^{-1}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \tag{3.51}$$

Continuing,

$$\begin{aligned}&\sum_{i=1}^{m}\left[\frac{1}{n}g''_{i,j}(0)\right] \\ &\quad= -\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\frac{2\delta_{i,j}(2\delta_{i,j}-1)\phi_0^{2\delta_{i,j}}}{1-\phi_0^{2\delta_{i,j}}} + \frac{1}{n}\sum_{j=1}^{n}\frac{4\delta_{i,j}^2\phi_0^{4\delta_{i,j}}}{\left(1-\phi_0^{2\delta_{i,j}}\right)^2}\right] \\ &\quad= -2\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}V_{n_i} - 4\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}W_{n_i},\end{aligned} \tag{3.52}$$

where $V_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}(2\delta_{i,j}-1)\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})$
and $W_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{4\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2$. Both sequences $\{V_{n_i}, i = 2,3,\ldots\}$ and $\{W_{n_i}, i = 2,3,\ldots\}$ are strictly stationary and 1-dependent, and hence

by direct application of the central limit theorem for strictly stationary m-dependent series we have

$$\frac{1}{m}\sum_{i=1}^{m} V_{n_i} \xrightarrow{\mathrm{P}} \mathrm{E}\left[V_n\right] = \mathrm{E}\left[\delta(2\delta-1)\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right], \tag{3.53}$$

and

$$\frac{1}{m}\sum_{i=1}^{m} W_{n_i} \xrightarrow{\mathrm{P}} \mathrm{E}\left[W_n\right] = \mathrm{E}\left[\delta^2\phi_0^{4\delta}/(1-\phi_0^{2\delta})^2\right]. \tag{3.54}$$

Since $R_1(i,j) \sim m^{-3/2}$, then $\sum_{i=1}^{m}\left[n^{-1}\sum_{j=1}^{n} R_1(i,j)\right] = o_p(m^{1/2}) \to 0$ as $m \to \infty$ for fixed n.

Substituting the above expressions into (3.46) yields

$$\begin{aligned} D_m(\gamma) = {} & -2\gamma\phi_0^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m} U_{n_i} \\ & - \gamma^2\phi_0^{-2}\left(\frac{1}{m}\sum_{i=1}^{m} V_{n_i} + \frac{1}{m}\sum_{i=1}^{m} 2W_{n_i}\right) + o_p(m^{1/2}). \end{aligned} \tag{3.55}$$

We now present a similar analysis for the components of (3.45) that arise from the quadratic term. Define

$$\begin{aligned} Q_m(\gamma) = {} & m\log\left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h_{i,j}(\gamma/\sqrt{m})\right]\right) \\ & - m\log\left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} \varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})\right]\right). \end{aligned} \tag{3.56}$$

We expand (3.56) about $\gamma = 0$ as follows:

$$\begin{aligned} Q_m(\gamma) = {} & m\log\left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h_{i,j}(0)\right]\right) + \gamma\frac{\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h'_{i,j}(0)\right]}{\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h_{i,j}(0)\right]} \\ & + \frac{\gamma^2}{2}\left(\frac{\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h''_{i,j}(0)\right]}{\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h_{i,j}(0)\right]} - \left(\frac{\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h'_{i,j}(0)\right]}{\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h_{i,j}(0)\right]}\right)^2\right) \\ & - m\log\left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} \varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})\right]\right) + R_2. \end{aligned} \tag{3.57}$$

Here R_2 is the remainder term, is a function of the $h_{i,j}(\gamma/\sqrt{m})$ and its first three derivatives, and is of order $\mathcal{O}(m^{1/2})$.

Recall $\phi_\gamma = \phi_0 + \gamma/\sqrt{m}$. Therefore,

$$
\begin{aligned}
h_{i,j}(\gamma/\sqrt{m}) &= \left(Y_{t_{i,j}} - \phi_\gamma^{\delta_{i,j}} Y_{t_{i,j-1}}\right)^2 / (1 - \phi_\gamma^{2\delta_{i,j}}) \\
h'_{i,j}(\gamma/\sqrt{m}) &= \frac{2\delta_{i,j}\phi_\gamma^{2\delta_{i,j}-1}}{\sqrt{m}} \frac{\left(Y_{t_{i,j}} - \phi_\gamma^{\delta_{i,j}} Y_{t_{i,j-1}}\right)^2}{\left(1 - \phi_\gamma^{2\delta_{i,j}}\right)^2} \\
&\quad - \frac{2\delta_{i,j}\phi_\gamma^{\delta_{i,j}-1}}{\sqrt{m}} \frac{\left(Y_{t_{i,j}} - \phi_\gamma^{\delta_{i,j}} Y_{t_{i,j-1}}\right) Y_{t_{i,j-1}}}{\left(1 - \phi_\gamma^{2\delta_{i,j}}\right)} \\
h''_{i,j}(\gamma/\sqrt{m}) &= \frac{2\delta_{i,j}(2\delta_{i,j}-1)}{m} \frac{\phi_\gamma^{2\delta_{i,j}-2}\left(Y_{t_{i,j}} - \phi_\gamma^{\delta_{i,j}} Y_{t_{i,j-1}}\right)^2}{\left(1 - \phi_\gamma^{2\delta_{i,j}}\right)^2} \\
&\quad - \frac{8\delta_{i,j}^2}{m} \frac{\phi_\gamma^{3\delta_{i,j}-2}\left(Y_{t_{i,j}} - \phi_\gamma^{\delta_{i,j}} Y_{t_{i,j-1}}\right) Y_{t_{i,j-1}}}{\left(1 - \phi_\gamma^{2\delta_{i,j}}\right)^2} \\
&\quad + \frac{8\delta_{i,j}^2}{m} \frac{\phi_\gamma^{4\delta_{i,j}-2}\left(Y_{t_{i,j}} - \phi_\gamma^{\delta_{i,j}} Y_{t_{i,j-1}}\right)^2}{\left(1 - \phi_\gamma^{2\delta_{i,j}}\right)^3} \\
&\quad - \frac{2\delta_{i,j}(\delta_{i,j}-1)}{m} \frac{\phi_\gamma^{\delta_{i,j}-2}\left(Y_{t_{i,j}} - \phi_\gamma^{\delta_{i,j}} Y_{t_{i,j-1}}\right) Y_{t_{i,j-1}}}{\left(1 - \phi_\gamma^{2\delta_{i,j}}\right)} \\
&\quad + \frac{2\delta_{i,j}^2}{m} \frac{\phi_\gamma^{2\delta_{i,j}-2} Y_{t_{i,j-1}}^2}{\left(1 - \phi_\gamma^{2\delta_{i,j}}\right)}.
\end{aligned}
$$

Therefore,

$$
\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} h_{i,j}(0)\right] = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n} \varepsilon_{t_{i,j}}^2/(1 - \phi_0^{2\delta_{i,j}})\right], \tag{3.58}
$$

since $Y_{t_{i,j}} - \phi_0^{\delta_{i,j}} Y_{t_{i,j-1}} = \varepsilon_{t_{i,j}}$. Define $S_{n_i} = n^{-1}\sum_{j=1}^{n} \varepsilon_{t_{i,j}}^2/(1 - \phi_0^{2\delta_{i,j}})$ and note that the sequence $\{S_{n_i}, i = 2, 3, \ldots\}$ is strictly stationary and ergodic, thus by the ergodic theorem

$$
\frac{1}{m}\sum_{i=1}^{m}[S_{n_i}] \xrightarrow{\mathrm{P}} \mathrm{E}[S_n], \tag{3.59}
$$

where

$$\begin{aligned} \mathrm{E}[S_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^2|\boldsymbol{\delta}\right]/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= \sigma^2. \end{aligned} \tag{3.60}$$

Furthermore,

$$\begin{aligned} \mathrm{Var}[S_n] &= \mathrm{E}[S_{n_i}^2] - (\mathrm{E}[S_n])^2 \\ &= \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\varepsilon_{t_{i,j}}^4/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &\quad + \mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n}\varepsilon_{t_{i,j}}^2\varepsilon_{t_{i,k}}^2/(1-\phi_0^{2\delta_{i,j}})(1-\phi_0^{2\delta_{i,k}})\right] - \sigma^4 \\ &= \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^4|\boldsymbol{\delta}\right]/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &\quad + \mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^2\varepsilon_{t_{i,k}}^2|\boldsymbol{\delta}\right]/(1-\phi_0^{2\delta_{i,j}})(1-\phi_0^{2\delta_{i,k}})\right] - \sigma^4 \\ &= 2\sigma^4 n^{-1}. \end{aligned} \tag{3.61}$$

Next,

$$\begin{aligned} \sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}h'_{i,j}(0)\right] &= 2\phi_0^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &\quad - 2\phi_0^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= 2\phi_0^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}A_{n_i} - 2\phi_0^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}B_{n_i}, \end{aligned} \tag{3.62}$$

where $A_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^2$ and $B_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})$ such that $\{A_{n_i}, i=2,3,\ldots\}$ and $\{B_{n_i}, i=2,3,\ldots\}$ are strictly stationary 1-dependent sequences. By applying the central limit theorem to the first sequence we have

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(A_{n_i}-\mathrm{E}[A_n]) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathrm{Var}[A_n]), \tag{3.63}$$

where

$$\begin{aligned}
\mathrm{E}[A_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^2|\boldsymbol{\delta}\right]\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\
&= \sigma^2\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\
&= \sigma^2\mathrm{E}\left[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right] \\
&= \sigma^2\mathrm{E}[U_n],
\end{aligned} \tag{3.64}$$

and

$$\begin{aligned}
\mathrm{Var}[A_n] &= \mathrm{E}[A_n^2] - (\mathrm{E}[A_n])^2 \\
&= \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{4\delta_{i,j}}\varepsilon_{t_{i,j}}^4/(1-\phi_0^{2\delta_{i,j}})^4\right] \\
&\quad + \mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n}\frac{\delta_{i,j}\delta_{i,k}\phi_0^{2(\delta_{i,j}+\delta_{i,k})}\varepsilon_{t_{i,j}}^2\varepsilon_{t_{i,k}}^2}{(1-\phi_0^{2\delta_{i,j}})^2(1-\phi_0^{2\delta_{i,k}})^2}\right] - \sigma^4\mathrm{E}[U_{n_i}]^2 \\
&= \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^4|\boldsymbol{\delta}\right]\delta_{i,j}^2\phi_0^{4\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^4\right] \\
&\quad + \mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^2\varepsilon_{t_{i,k}}^2|\boldsymbol{\delta}\right]\frac{\delta_{i,j}\delta_{i,k}\phi_0^{2\delta_{i,j}}\phi_0^{2\delta_{i,k}}}{(1-\phi_0^{2\delta_{i,j}})^2(1-\phi_0^{2\delta_{i,k}})^2}\right] - \sigma^4\mathrm{E}[U_{n_i}]^2 \\
&= 3\sigma^4\mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{4\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\
&\quad + \sigma^4\mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n}\frac{\delta_{i,j}\delta_{i,k}\phi_0^{2\delta_{i,j}}\phi_0^{2\delta_{i,k}}}{(1-\phi_0^{2\delta_{i,j}})(1-\phi_0^{2\delta_{i,k}})}\right] - \sigma^4\mathrm{E}[U_n]^2 \\
&= 2\sigma^4\mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{4\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] + \sigma^4\mathrm{Var}[U_n] \\
&= 2\sigma^4 n^{-1}\mathrm{E}\left[\delta^2\phi_0^{4\delta}/(1-\phi_0^{2\delta})^2\right] + \sigma^4\mathrm{Var}[U_n].
\end{aligned} \tag{3.65}$$

Furthermore, the sequence $\{A_{n_i}, i = 2, 3, \ldots\}$ is ergodic, hence by direct application of the ergodic theorem we have

$$\frac{1}{m}\sum_{i=1}^{m}A_{n_i} \xrightarrow{\mathrm{P}} \sigma^2\mathrm{E}[U_n]. \tag{3.66}$$

Application of the central limit theorem for strictly stationary m-dependent series to the second sequence $\{B_{n_i}, i = 2, 3, \ldots\}$ yields

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} B_{n_i} \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathrm{Var}[B_n]\right), \tag{3.67}$$

since

$$\begin{aligned} \mathrm{E}[B_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n} \delta_{i,j}\phi_0^{\delta_{i,j}} Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n} \mathrm{E}\left[\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right] \delta_{i,j}\phi_0^{\delta_{i,j}} Y_{t_{i,j-1}}/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= 0, \end{aligned} \tag{3.68}$$

and where

$$\begin{aligned} \mathrm{Var}[B_n] &= \mathrm{E}\left[B_n^2\right] \\ &= \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n} \delta_{i,j}^2\phi_0^{2\delta_{i,j}} Y_{t_{i,j-1}}^2\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &\quad + \mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n} \frac{\delta_{i,j}\delta_{i,k}\phi_0^{\delta_{i,j}}\phi_0^{\delta_{i,k}} Y_{t_{i,j-1}}Y_{t_{i,k-1}}\varepsilon_{t_{i,j}}\varepsilon_{t_{i,k}}}{(1-\phi_0^{2\delta_{i,j}})(1-\phi_0^{2\delta_{i,k}})}\right] \\ &= \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n} \mathrm{E}\left[Y_{t_{i,j-1}}^2\varepsilon_{t_{i,j}}^2|\boldsymbol{\delta}\right] \delta_{i,j}^2\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &\quad + \mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n} \frac{\mathrm{E}\left[Y_{t_{i,j-1}}Y_{t_{i,k-1}}\varepsilon_{t_{i,j}}\varepsilon_{t_{i,k}}|\boldsymbol{\delta}\right] \delta_{i,j}\delta_{i,k}\phi_0^{\delta_{i,j}}\phi_0^{\delta_{i,k}}}{(1-\phi_0^{2\delta_{i,j}})(1-\phi_0^{2\delta_{i,k}})}\right] \\ &= \sigma^4\mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n} \delta_{i,j}^2\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= \sigma^4 n^{-1}\mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})]. \end{aligned} \tag{3.69}$$

Continuing,

$$\begin{aligned}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}h''_{i,j}(0)\right] &= 2\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}(2\delta_{i,j}-1)\phi_0^{2\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &\quad - 8\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{3\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &\quad + 8\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{4\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^3\right] \\ &\quad - 2\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &\quad + 2\phi_0^{-2}\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{2\delta_{i,j}}Y_{t_{i,j-1}}^2/(1-\phi_0^{2\delta_{i,j}})\right].\end{aligned} \tag{3.70}$$

We next show that each term of (3.70) converges in probability to its mean. First define $C_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}(2\delta_{i,j}-1)\phi_0^{2\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^2$ and note that the sequence $\{C_{n_i}, i=2,3,\ldots\}$ is strictly stationary and 1-dependent. Hence,

$$\frac{1}{m}\sum_{i=1}^{m}C_{n_i} \xrightarrow{\mathrm{P}} \mathrm{E}[C_n], \tag{3.71}$$

where

$$\begin{aligned}\mathrm{E}[C_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}(2\delta_{i,j}-1)\phi_0^{2\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^2|\boldsymbol{\delta}\right]\delta_{i,j}(2\delta_{i,j}-1)\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &= \sigma^2\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}(2\delta_{i,j}-1)\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= \sigma^2\mathrm{E}\left[\delta(2\delta-1)\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right] \\ &= \sigma^2\mathrm{E}[V_n].\end{aligned} \tag{3.72}$$

Define $D_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{3\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2$. The sequence $\{D_{n_i}, i=2,3,\ldots\}$ is also strictly stationary and 1-dependent. Therefore,

$$\frac{1}{m}\sum_{i=1}^{m}D_{n_i} \xrightarrow{\mathrm{P}} \mathrm{E}[D_n] = 0, \tag{3.73}$$

since

$$\begin{aligned} \mathrm{E}[D_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{3\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right]\delta_{i,j}^2\phi_0^{3\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &= 0. \end{aligned} \tag{3.74}$$

Define $E_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{4\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^3$. The sequence $\{E_{n_i}, i = 2,3,\ldots\}$ is strictly stationary and 1-dependent. Therefore,

$$\frac{1}{m}\sum_{i=1}^{m}E_{n_i} \xrightarrow{\mathrm{P}} \mathrm{E}[E_n], \tag{3.75}$$

where

$$\begin{aligned} \mathrm{E}[E_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{4\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^3\right] \\ &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^2|\boldsymbol{\delta}\right]\delta_{i,j}^2\phi_0^{4\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^3\right] \\ &= \sigma^2\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{4\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &= \sigma^2\mathrm{E}\left[\delta^2\phi_0^{4\delta}/(1-\phi_0^{2\delta})^2\right] \\ &= \sigma^2\mathrm{E}[W_n]. \end{aligned} \tag{3.76}$$

Next define $F_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2$. The sequence $\{F_{n_i}, i = 2,3,\ldots\}$ is strictly stationary and 1-dependent, and

$$\frac{1}{m}\sum_{i=1}^{m}F_{n_i} \xrightarrow{\mathrm{P}} \mathrm{E}[F_n] = 0, \tag{3.77}$$

since

$$\begin{aligned} \mathrm{E}[F_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[Y_{t_{i,j-1}}\varepsilon_{t_{i,j}}|\boldsymbol{\delta}\right]\delta_{i,j}(\delta_{i,j}-1)\phi_0^{\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2\right] \\ &= 0. \end{aligned} \tag{3.78}$$

Finally, define $G_{n_i} = n^{-1}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{2\delta_{i,j}}Y_{t_{i,j-1}}^2/(1-\phi_0^{2\delta_{i,j}})$. The sequence $\{G_{n_i}, i = 2,3,\ldots\}$ is strictly stationary and 1-dependent, and

$$\frac{1}{m}\sum_{i=1}^{m} G_{n_i} \xrightarrow{\mathrm{P}} \mathrm{E}[G_n], \tag{3.79}$$

where

$$\begin{aligned}
\mathrm{E}[G_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{2\delta_{i,j}}Y_{t_{i,j-1}}^2/(1-\phi_0^{2\delta_{i,j}})\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[Y_{t_{i,j-1}}^2|\boldsymbol{\delta}\right]\delta_{i,j}^2\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\
&= \sigma^2\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}^2\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\
&= \sigma^2\mathrm{E}\left[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right].
\end{aligned} \tag{3.80}$$

Substituting the above relations into (3.57) yields

$$\begin{aligned}
Q_m(\gamma) = {} & 2\gamma\phi_0^{-1}\frac{1}{\sqrt{m}}\left(\frac{\sum_{i=1}^{m}[A_{n_i}-B_{n_i}]}{m^{-1}\sum_{i=1}^{m}S_{n_i}}\right) \\
& + \gamma^2\phi_0^{-2}\frac{1}{m}\left(\frac{\sum_{i=1}^{m}[C_{n_i}-4D_{n_i}+4E_{n_i}-F_{n_i}+G_{n_i}]}{m^{-1}\sum_{i=1}^{m}S_{n_i}}\right) \\
& - 2\gamma^2\phi_0^{-2}\frac{1}{m}\left(\frac{m^{-1/2}\sum_{i=1}^{m}[A_{n_i}-B_{n_i}]}{m^{-1}\sum_{i=1}^{m}S_{n_i}}\right)^2 \\
& + o_p(m^{1/2}).
\end{aligned} \tag{3.81}$$

Finally, by substituting (3.55) and (3.81) into the reduced likelihood (3.44) we obtain

$$
\begin{aligned}
\ell_m^\star(\gamma) &= D_m(\gamma) + Q_m(\gamma) \\
&= -2\gamma\phi_0^{-1}\frac{\frac{1}{\sqrt{m}}\sum_{i=1}^m U_{n_i}\frac{1}{m}\sum_{i=1}^m S_{n_i} - \frac{1}{\sqrt{m}}\sum_{i=1}^m A_{n_i} + \frac{1}{\sqrt{m}}\sum_{i=1}^m B_{n_i}}{m^{-1}\sum_{i=1}^m S_{n_i}} \\
&\quad + \gamma^2\phi_0^{-2}\frac{m^{-1}\sum_{i=1}^m [C_{n_i} - 4D_{n_i} + 4E_{n_i} - F_{n_i} + G_{n_i}]}{m^{-1}\sum_{i=1}^m S_{n_i}} \\
&\quad - \gamma^2\phi_0^{-2}\frac{1}{m}\sum_{i=1}^m [V_{n_i} + 2W_{n_i}] \\
&\quad - 2\gamma^2\phi_0^{-2}\frac{(m^{-1}\sum_{i=1}^m [A_{n_i} - B_{n_i}])^2}{(m^{-1}\sum_{i=1}^m S_{n_i})^2} \\
&\quad + o_p(m^{1/2}) \\
&= -2\gamma\phi_0^{-1}\frac{\frac{1}{\sqrt{m}}\sum_{i=1}^m (U_{n_i} - \mathrm{E}[U_n])\frac{1}{m}\sum_{i=1}^m S_{n_i}}{m^{-1}\sum_{i=1}^m S_{n_i}} \\
&\quad + 2\gamma\phi_0^{-1}\frac{\frac{1}{\sqrt{m}}\sum_{i=1}^m (A_{n_i} - \sigma^2\mathrm{E}[U_n])}{m^{-1}\sum_{i=1}^m S_{n_i}} \\
&\quad - 2\gamma\phi_0^{-1}\frac{\frac{1}{\sqrt{m}}\sum_{i=1}^m (S_{n_i} - \sigma^2)\,\mathrm{E}[U_n] + \frac{1}{\sqrt{m}}\sum_{i=1}^m B_{n_i}}{m^{-1}\sum_{i=1}^m S_{n_i}} \\
&\quad + \gamma^2\phi_0^{-2}\frac{m^{-1}\sum_{i=1}^m [C_{n_i} - 4D_{n_i} + 4E_{n_i} - F_{n_i} + G_{n_i}]}{m^{-1}\sum_{i=1}^m S_{n_i}} \\
&\quad - \gamma^2\phi_0^{-2}\frac{1}{m}\sum_{i=1}^m [V_{n_i} + 2W_{n_i}] \\
&\quad - 2\gamma^2\phi_0^{-2}\frac{(m^{-1}\sum_{i=1}^m [A_{n_i} - B_{n_i}])^2}{(m^{-1}\sum_{i=1}^m S_{n_i})^2} \\
&\quad + o_p(m^{1/2})
\end{aligned}
\tag{3.82}
$$

Taking the limit with respect to m and employing Slutsky's theorem we find

$$\begin{aligned}
\lim_{m\to\infty} \ell_m^\star(\gamma) &= -2\gamma\phi_0^{-1}\frac{1}{\sigma^2}\left(N_U\cdot\sigma^2 - N_A + N_S\cdot \mathrm{E}[U_n] + N_B\right) \\
&\quad + \gamma^2\phi_0^{-2}\frac{1}{\sigma^2}\left(\mathrm{E}[C_n] - 4\mathrm{E}[D_n] + 4\mathrm{E}[E_n] - \mathrm{E}[F_n] + \mathrm{E}[G_n]\right) \\
&\quad - \gamma^2\phi_0^{-2}\left(\mathrm{E}[V_n] + 2\mathrm{E}[W_n]\right) \\
&\quad - 2\gamma^2\phi_0^{-2}\frac{1}{\sigma^4}\left(\mathrm{E}[A_n]\right)^2 \\
&= -2\gamma\phi_0^{-1}\frac{1}{\sigma^2}\left(N_U\cdot\sigma^2 - N_A + N_S\cdot \mathrm{E}[U_n] + N_B\right) \\
&\quad + \gamma^2\phi_0^{-2}\frac{1}{\sigma^2}\left(\sigma^2\mathrm{E}[V_n] + 4\sigma^2\mathrm{E}[W_n] + \sigma^2\mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})]\right) \\
&\quad - \gamma^2\phi_0^{-2}\left(\mathrm{E}[V_n] + 2\mathrm{E}[W_n]\right) \\
&\quad - 2\gamma^2\phi_0^{-2}\frac{1}{\sigma^4}\left(\sigma^2\mathrm{E}[U_n]\right)^2 \\
&= -2\gamma\phi_0^{-1}\frac{1}{\sigma^2}\left(N_U\cdot\sigma^2 - N_A + N_S\cdot \mathrm{E}[U_n] + N_B\right) \\
&\quad + \gamma^2\phi_0^{-2}\left(2\mathrm{E}[\delta^2\phi_0^{4\delta}/(1-\phi_0^{2\delta})] - 2\left(\mathrm{E}[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})]\right)^2\right) \\
&\quad + \gamma^2\phi_0^{-2}\mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})] \\
&= -2\gamma\phi_0^{-1}\frac{1}{\sigma^2}\left(N_U\cdot\sigma^2 - N_A + N_S\cdot \mathrm{E}[U_n] + N_B\right) \\
&\quad + \gamma^2\phi_0^{-2}\left(2\mathrm{Var}[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})] + \mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})]\right),
\end{aligned} \tag{3.83}$$

where N_U, N_A, N_S, and N_B have joint multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix such that

$$\begin{aligned}
\mathrm{Var}(N_U) &= \mathrm{Var}[U_n], \\
\mathrm{Var}(N_A) &= 2\sigma^4 n^{-1}\mathrm{E}[\delta^2\phi_0^{4\delta}/(1-\phi_0^{2\delta})^2] + \sigma^4\mathrm{Var}[U_n], \\
\mathrm{Var}(N_S) &= 2\sigma^4 n^{-1}, \\
\mathrm{Var}(N_B) &= \sigma^4 n^{-1}\mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})], \\
\mathrm{Cov}(N_U, N_A) &= \sigma^2\mathrm{Var}[U_n], \text{ and} \\
\mathrm{Cov}(N_A, N_S) &= 2\sigma^4 n^{-1}\mathrm{E}[U_n].
\end{aligned} \tag{3.84}$$

To compute the covariances we have

$$\begin{aligned}
\mathrm{E}[U_n A_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\frac{1}{n}\sum_{k=1}^{n}\delta_{i,k}\phi_0^{2\delta_{i,k}}\varepsilon_{t_{i,k}}^2/(1-\phi_0^{2\delta_{i,k}})^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\left[\varepsilon_{t_{i,k}}^2|\boldsymbol{\delta}\right]\delta_{i,k}\phi_0^{2\delta_{i,k}}/(1-\phi_0^{2\delta_{i,k}})^2\right] \\
&= \sigma^2\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\frac{1}{n}\sum_{k=1}^{n}\delta_{i,k}\phi_0^{2\delta_{i,k}}/(1-\phi_0^{2\delta_{i,k}})\right] \\
&= \sigma^2\mathrm{E}[U_n^2],
\end{aligned} \tag{3.85}$$

which implies

$$\mathrm{Cov}(N_U, N_A) = \sigma^2\mathrm{Var}[U_n]. \tag{3.86}$$

Next,

$$\begin{aligned}
\mathrm{E}[A_n S_n] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})^2\frac{1}{n}\sum_{k=1}^{n}\varepsilon_{t_{i,j}}^2/(1-\phi_0^{2\delta_{i,j}})\right] \\
&= \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^4|\boldsymbol{\delta}\right]\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^3\right] \\
&\quad + \mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n}\mathrm{E}\left[\varepsilon_{t_{i,j}}^2\varepsilon_{t_{i,k}}^2|\boldsymbol{\delta}\right]\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})^2(1-\phi_0^{2\delta_{i,k}})\right] \\
&= 3\sigma^4\mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\
&\quad + \sigma^4\mathrm{E}\left[\frac{1}{n^2}\sum_{j\neq k}^{n}\delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\
&= 2\sigma^4 n^{-1}\mathrm{E}[U_n] + \sigma^4\mathrm{E}[U_n],
\end{aligned} \tag{3.87}$$

which implies

$$\mathrm{Cov}(N_A, N_S) = 2\sigma^4 n^{-1}\mathrm{E}[U_n]. \tag{3.88}$$

It is straight forward to show that each of the remaining covariances are zero. Thus,

$$\begin{aligned}
&\mathrm{Var}\left[\frac{1}{\sigma^2}\left(N_U\cdot\sigma^2 - N_A + N_S\cdot \mathrm{E}[U_{n_i}] + N_B\right)\right] \\
&\quad = \mathrm{Var}(N_U) + \sigma^{-4}\mathrm{Var}(N_A) + \sigma^{-4}\left(\mathrm{E}[U_n]\right)^2\mathrm{Var}(N_S) + \sigma^{-4}\mathrm{Var}(N_B) \\
&\qquad - 2\sigma^{-2}\mathrm{Cov}(N_U, N_A) - 2\sigma^{-4}\mathrm{E}[U_n]\mathrm{Cov}[(N_A, N_S) \\
&\quad = \mathrm{Var}[U_n] + 2n^{-1}\mathrm{E}[\delta^2\phi_0^{4\delta}/(1-\phi_0^{2\delta})^2] + \mathrm{Var}[U_n] + 2n^{-1}\left(\mathrm{E}[U_n]\right)^2 \\
&\qquad + n^{-1}\mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})] - 2\mathrm{Var}[U_n] - 4n^{-1}\left(\mathrm{E}[U_n]\right)^2 \\
&\quad = 2n^{-1}\left(\mathrm{E}[\delta^2\phi_0^{4\delta}/(1-\phi_0^{2\delta})^2] - \left(\mathrm{E}[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})]\right)^2\right) \\
&\qquad + n^{-1}\mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})] \\
&\quad = 2n^{-1}\mathrm{Var}[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})] + n^{-1}\mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})].
\end{aligned} \tag{3.89}$$

Therefore, after combining all of the results thus far, we have

$$\lim_{m\to\infty}\ell_m^\star(\gamma) = -2\gamma\phi_0^{-1}N_n + \gamma^2\phi_0^{-2}K_n, \tag{3.90}$$

where $N_n \sim \mathcal{N}(0, n^{-1}K_n)$ and $K_n = 2\mathrm{Var}[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})] + \mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})]$. We can show that the reduced likelihood converges over a compact set on $\mathbb{R}$ by following the methodology first presented for the weighted least squares approach. That is to say we can rewrite (3.82) as

$$\ell_m^\star(\gamma) = \ell_{m_1}^\star(\gamma) + \ell_{m_2}^\star(\gamma) + o_\mathrm{p}(1), \tag{3.91}$$

where $\ell_{m_1}^\star(\gamma)$ collects the quadratic terms, $\ell_{m_2}^\star(\gamma)$ collects the linear terms, and for any constant k,

$$\sup_{|\gamma|\le k}|o_\mathrm{p}(1)| \xrightarrow{\mathrm{P}} 0.$$

It is straightforward to show that

$$\sup_{|\gamma|\le k}\left|\ell_1(\gamma) - \gamma^2\phi_0^{-2}K_n\right| = o_\mathrm{p}(1),$$

implying that $\ell_1(\gamma)$ converges in probability to K_n on $C[-k,k]$. Moreover,

$$\ell^\star_m(\gamma) = \gamma^2\phi_0^{-2}K_n + \ell^\star_{m_2}(\gamma) + o_p(1).$$

Second, if we consider $\ell^\star_m(\tilde{\gamma})$ and note that $\ell^\star_{m_2}(\gamma) \xrightarrow{\mathrm{d}} N_n$ where N_{K_n} is defined as above, then

$$|\ell_{m_2}(\tilde{\gamma}) - \ell_{m_2}(\gamma)| = |\tilde{\gamma} - \gamma||\phi_0^{-1}O_{\mathrm{p}}(1)|,$$

where

$$\sup_{|\gamma|\leq k} |O_p(1)| = o_p(1).$$

Thus for any $\eta > 0$ there exists $\epsilon > 0$ such that

$$\mathrm{P}\left(\lim_m \sup_{|\tilde{\gamma}-\gamma|\leq\epsilon} |\ell_2(\tilde{\gamma}) - \ell_2(\gamma)|\right) \xrightarrow{\mathrm{P}} 0 \text{ as } m \to \infty \text{ and } \epsilon \to 0,$$

implying that $\ell_2(\gamma)$ is tight. It follows that $\ell^\star_m(\gamma) \xrightarrow{\mathrm{d}} \ell(\gamma)$ on $C(\mathbb{R})$.

If $\ell(\gamma)$ has a unique minimum, say $\gamma_{\min}$, then there exists a minimum $\gamma^\star_{\min} = \arg\min \ell^\star(\gamma)$ such that $\gamma^\star_{\min} \xrightarrow{\mathrm{d}} \gamma_{\min}$. Equation (3.90) can now be minimized by taking the derivative with respect to γ, setting the result to zero, and solving for γ. Hence,

$$\hat{\gamma} = \sqrt{m}(\hat{\phi} - \phi_0) \xrightarrow{\mathrm{d}} \phi_0\mathcal{N}\left(0, (nK_n)^{-1}\right).$$

taking $g(x) = -(\log(x))^{-1}$ and applying Slutsky's theorem yields the desired result,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\theta^4}{n}\frac{1}{2\mathrm{Var}[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})] + \mathrm{E}[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})]}\right). \quad (3.92)$$

(b) Now consider the case where $n \to \infty$. Once again we assume that $n(m) = n_m \to \infty$ as $m \to \infty$ and that $f_T(t) > 0$ *a.e.* where $t \in (0,1]$. Recall that as $m \to \infty$, $t_{(1)} \to 0$ and $t_{(n)} \to 1$ which in turn implies that $\sum_{j=1}^{n_m} \delta_{i,j} \to 1$ for every $i \in \{1, 2, \ldots, m\}$. Furthermore,

$$\max_{1\leq j\leq n_m} \delta_{i,j} \xrightarrow{\mathrm{P}} 0.$$

The previous arguments for fixed n still hold; therefore all that remains to show are the limiting values for $n_m \cdot 2\text{Var}[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})]$ and $n_m \cdot \text{E}[\delta\phi_0^{2\delta}(1-\phi_0^{2\delta})]$. By the mean value theorem we have for every $i \in \{1,2,\ldots\}$,

$$\begin{aligned} n_m \text{E}\left[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right] &= \text{E}\left[\sum_{j=1}^{n_m} \delta_{i,j}^2 \phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right] \\ &= \text{E}\left[\sum_{j=1}^{n_m} \delta_{i,j}^2 \phi_0^{2\delta_{i,j}}/\left(2\theta^{-1}\delta_{i,j}\phi_0^{2\tilde{\delta}_{i,j}}\right)\right] \\ &= \frac{\theta}{2}\text{E}\left[\sum_{j=1}^{n_m} \delta_{i,j}\phi_0^{2(\delta_{i,j}-\tilde{\delta}_{i,j})}\right], \end{aligned}$$

where $\tilde{\delta}_{i,j} \in (0, \delta_{i,j})$ for all j. Note that $\phi_0^{2\delta_{i,j}} \leq \phi_0^{2(\delta_{i,j}-\tilde{\delta}_{i,j})} \leq 1$. For the proof of Theorem 2 we showed

$$\lim_{m\to\infty} \sum_{j=1}^{n_m} \delta_{i,j}\phi_0^{2\delta_{i,j}} = 1.$$

Therefore as $m \to \infty$, $n_m \text{E}\left[\delta^2\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right] \to 2\theta^{-1}$.

In a similar manner we can show that the variance terms goes to zero. A second application of the mean value theorem yields

$$\begin{aligned} &n_m \text{Var}\left[\delta\phi_0^{2\delta}/(1-\phi_0^{2\delta})\right] \\ &= \text{E}\left[\sum_{j=1}^{n_m} \delta_{i,j}^2 \phi_0^{4\delta_{i,j}}/\left(1-\phi_0^{2\delta_{i,j}}\right)^2\right] - \frac{1}{n_m}\left(\text{E}\left[\sum_{j=1}^{n_m} \delta_{i,j}\phi_0^{2\delta_{i,j}}/(1-\phi_0^{2\delta_{i,j}})\right]\right)^2 \\ &= \text{E}\left[\sum_{j=1}^{n_m} \delta_{i,j}^2 \phi_0^{4\delta_{i,j}}/\left(2\theta^{-1}\delta_{i,j}\phi_0^{2\tilde{\delta}_{i,j}}\right)^2\right] - \frac{1}{n_m}\left(\text{E}\left[\sum_{j=1}^{n_m} \delta_{i,j}\phi_0^{2\delta_{i,j}}/(2\theta^{-1}\delta_{i,j}\phi_0^{2\tilde{\delta}_{i,j}})\right]\right)^2 \\ &= \frac{\theta^2}{4}\left\{\text{E}\left[\sum_{j=1}^{n_m} \phi_0^{4(\delta_{i,j}-\tilde{\delta}_{i,j})}\right] - \frac{1}{n_m}\left(\text{E}\left[\sum_{j=1}^{n_m} \phi_0^{2(\delta_{i,j}-\tilde{\delta}_{i,j})}\right]\right)^2\right\} \end{aligned}$$

Using the same arguments as before it is straightforward to show that the expression inside the braces goes to zero as $m \to \infty$. Substituting this result along with the previous result into (3.92) we obtain

$$\sqrt{m}(\hat{\theta}-\theta) \xrightarrow{\text{d}} \mathcal{N}(0, 2\theta^3). \tag{3.93}$$

□

REMARK 4 Once again the result for fixed n incorporates the sampling pattern into the calculation of the overall efficiency of the maximum likelihood estimator. However, unlike the weighted least squares approach, it is not immediately clear if there is an optimal design. To minimize the asymptotic variance we need to maximize nK_n. For a small sampling effort per block the variance term may be relatively large and perhaps dominate the expected value term. However, as n increases the variance will converge to zero and the expected value term will dominate. We recover the regular lattice result setting $\boldsymbol{\delta} \equiv n^{-1}\mathbf{1}$; $\mathrm{E}_n\left[\phi_0^{2\delta}\delta^2/(1-\phi_0^{2\delta})\right] = n^{-2}\phi_n^2/(1-\phi_n^2)$ and the asymptotic variance reduces to $\theta^4 n\phi_n^{-2}(1-\phi_n^2) = \theta^4 n(e^{2/\theta n}-1)$. We will confirm through simulation that the efficiency of the maximum likelihood estimator appears to be nearly invariant relative to the the sampling pattern provided that the effort per block remains constant, even for small n.

REMARK 5 Unlike many likelihood estimation procedures the determinant term plays an important role in the asymptotics; see for example Brockwell and Davis (1996). Recall the coefficient of γ in (3.90) is given by

$$-2\phi_0^{-1}\frac{\frac{1}{\sqrt{m}}\sum_{i=1}^m U_{n_i}\frac{1}{m}\sum_{j=1}^m S_{n_j} - \frac{1}{\sqrt{m}}\sum_{i=1}^m A_{n_i} + \frac{1}{\sqrt{m}}\sum_{i=1}^m B_{n_i}}{m^{-1}\sum_{i=1}^m S_{n_i}},$$

includes the random variable $\sum_{i=1}^m U_{n_i}$ that comes from the determinant term. The product $\sum_{i=1}^m U_{n_i}\cdot m^{-1}\sum_{j=1}^m S_{n_j}$ acts as a mean correction for $\sum_{i=1}^m A_{n_i}$, i.e., $\mathrm{E}[\sum_{i=1}^m U_{n_i} m^{-1}\cdot\sum_{j=1}^m S_{n_j}] = \mathrm{E}[\sum_{i=1}^m A_{n_i}]$. Thus $\sum_{i=1}^m U_{n_i}$ centers the distribution of $\hat{\gamma}$ at zero (since $\mathrm{E}[\sum_{i=1}^m B_{n_i}] = 0$).

Figure 3.9 presents a comparison of the weighted least squares and the maximum likelihood approaches. For this particular realization $\theta = 1$, $m = 1024$, $n = 4$, and the sampling distribution is uniform. The upper panels plot the objective functions as defined by (3.18) and (3.43), respectively. The lower two panels decompose (3.43) into the determinant and quadratic terms. Notice that the quadratic term has a distinct global minimum below $\theta = 0.5$. Although not readily apparent at the

current scale, the determinant effectively helps to better identify the global minimum, that is to say that the right tail of the WLS objective function increases at a slower rate than the right tail of the MLE objective function.

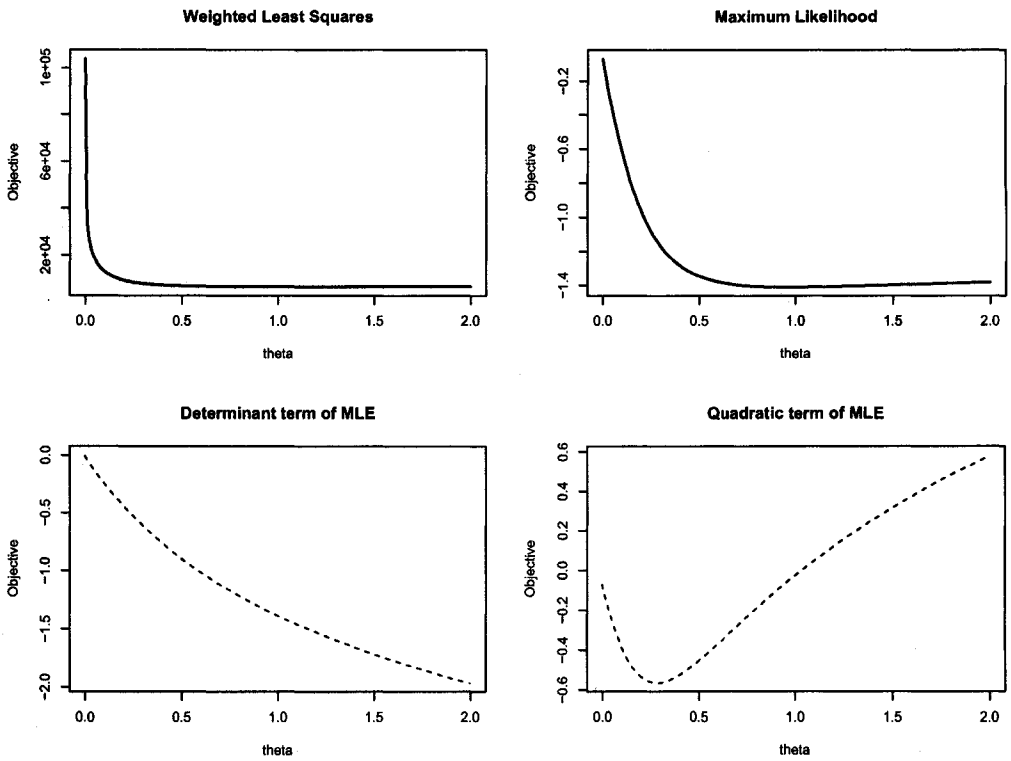

Figure 3.9: A comparison of the weighted least squares approach and the maximum likelihood approach. The upper two panels plot the respective objective functions. The lower two panels decompose the (reduced) likelihood function into the determinant term (left) and the quadratic term (right). For this particular realization $\theta = 1$, $m = 1024$, $n = 4$, and the sampling distribution is uniform. $\hat{\theta}_{\text{WLS}} = 0.983$ and $\hat{\theta}_{\text{MLE}} = 0.972$.

3.4.2.1 Simulation Results

Figures 3.10 and 3.11 illustrate the MSE surface as a function of domain and level of infill for the non-regular sampling patterns where $\theta = 1$ and $\theta = 2$, respectively. Once again the same general pattern is forthcoming: the contours of the surfaces run parallel to the infill axis and the MSE tends toward zero with respect to expanding domain. Further note that for a fixed θ the four surfaces are very similar to one another. This suggests that the likelihood function is better able to compensate for non-regularly sampling locations than the weighted least squares function. Resolution of the MSE images where $\theta = 1$ is diminished to account for the large MSE values at very meager sample sizes. In fact the maximum MSE is more than $2\times$ that for the sampling effort over a regular lattice.

We now proceed to quantifying the asymptotic variance of the maximum likelihood estimator $\hat{\theta}$ with respect to the five sampling patterns first presented in Section (3.4.1.1) at four different sampling efforts, $n = \{1, 2, 4, 8\}$. Figure 3.12 illustrates the expected asymptotic variance as a function of both sampling pattern and effort when $\theta = 1$. The plot for the regular lattice is the same exact calculation for the weighted least squares approach. The remaining four "curves" are plots of the sample variance at each level of sampling effort (n) computed from an extensive series of Monte Carlo simulations. Notice how all five plots follow a very similar trajectory. This is a very different result compared to the weighted least squares case where there was a pronounced influence due to sampling pattern (see Figure (3.8)). It appears that the (reduced) likelihood function is much more capable of compensating for unequal spacings in the sampling locations than the weighted least squares approach.

A comparison between the observed and theoretical variances is presented in Tables 3.3 and 3.4. We generated two sets of simulations where the total sampling effort $N = mn$ was fixed at 256 and 1024, respectively. This required that the

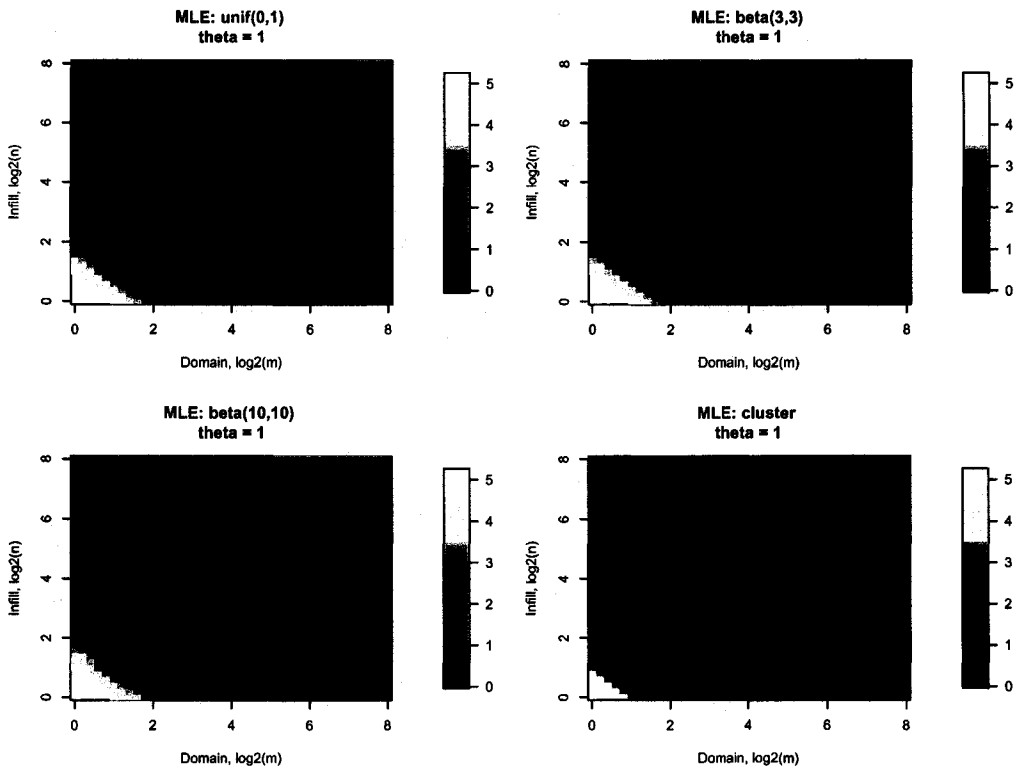

Figure 3.10: MSE for the maximum likelihood estimator (MLE) of an AR(1) process for each non-regular sampling pattern where $\theta = 1$. The surfaces represent the results based on 1000 independent realizations for each combination of domain and level of infill. Note that the axes are $\log_2$ such that $m, n = \{2^0, 2^1, \ldots, 2^8\}$.

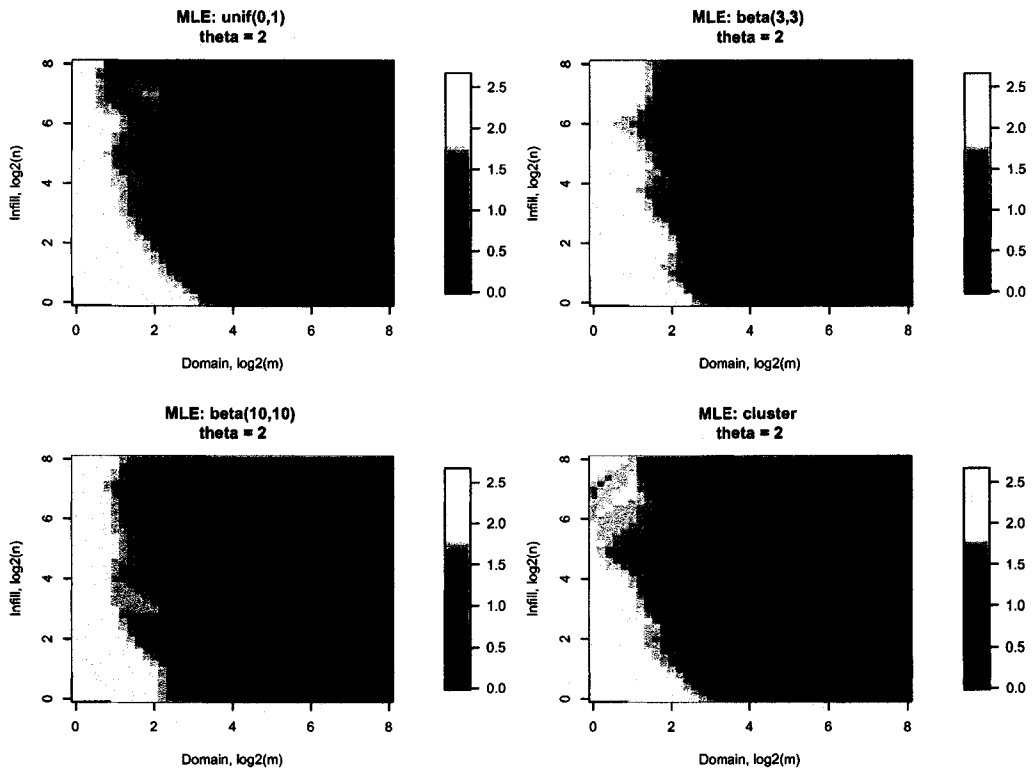

Figure 3.11: MSE for the maximum likelihood estimator (MLE) of an AR(1) process for each non-regular sampling pattern where $\theta = 2$. The surfaces represent the results based on 1000 independent realizations for each combination of domain and level of infill. Note that the axes are $\log_2$ such that $m, n = \{2^0, 2^1, \ldots, 2^8\}$.

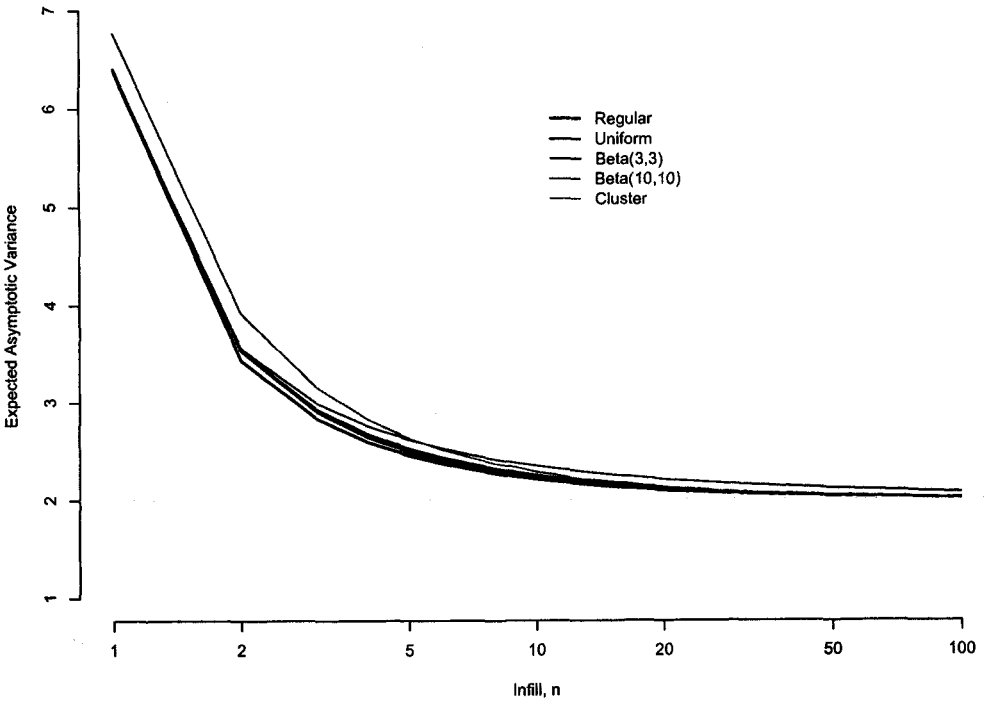

Figure 3.12: Expected asymptotic variance for the maximum likelihood estimator. The curve for the regular pattern is exact. The curves for the remaining four sample patterns are each based on 100 Monte Carlo simulations of $m = 1024$ blocks for all sampling efforts $n \in \{1, \ldots, 100\}$.

Table 3.3: Observed asymptotic variance as a function of sampling pattern and sampling effort per block. For the first four patterns the total sampling effort is $N = 256$ and for the clustered pattern the expected effort per block was set to $n = \{1, 2, 4, 8\}$. The results listed are the means of 1000 simulations with $\theta = 1$. The expected variances are delimited by parentheses.

Sampling Pattern	(Domain, Infill) $= (m, n)$			
	$(256, 1)$	$(128, 2)$	$(64, 4)$	$(32, 8)$
Regular	6.51	3.56	2.45	2.32
	(6.38)	(3.43)	(2.59)	(2.27)
Uniform	6.48	3.46	2.59	2.35
	(6.41)	(3.54)	(2.64)	(2.29)
Beta(3,3)	6.79	3.54	2.58	2.26
	(6.40)	(3.53)	(2.67)	(2.32)
Beta(10,10)	6.47	3.54	2.60	2.40
	(6.39)	(3.65)	(2.76)	(2.42)
Clustered	6.80	4.36	2.80	2.40
	(6.76)	(3.93)	(2.84)	(2.38)

domain be adjusted appropriately for each level of sampling effort. For the clustered sampling pattern we set the expected number of sampling locations per block to $n = \{1, 2, 4, 8\}$ and fixed the domains to match those of the other four patterns. Tables 3.3 and 3.4 report the variance of the MLE for θ. Overall the agreement between the observed and expected variances are excellent. For $N = 256$ the relative difference between the simulated variances and the expected variance never exceeds 11% with a mean difference of 0.4%. For $N = 1024$ the largest relative difference is 13% with a mean difference of just over 1%.

3.5 Conclusions

In this chapter we derived the asymptotic distribution of the MLE for the range parameter θ for a mean-zero Gaussian process with exponential covariance function in one-dimension. We began by deriving the asymptotics for equispaced sampling locations along a transect. Next we derived the asymptotics for a weighted least

Table 3.4: Observed asymptotic variance as a function of sampling pattern and sampling effort per block. For the first four patterns the total sampling effort is $N = 1024$ and for the clustered pattern the expected effort per block was set to $n = \{1, 2, 4, 8\}$. The results listed are the means of 1000 simulations with $\theta = 2$. The expected variances are delimited by parentheses.

Sampling Pattern	(Domain, Infill) $= (m, n)$			
	$(256, 1)$	$(128, 2)$	$(64, 4)$	$(32, 8)$
Regular	6.68	3.87	2.48	2.24
	(6.38)	(3.43)	(2.59)	(2.27)
Uniform	6.51	3.38	2.56	2.41
	(6.41)	(3.54)	(2.64)	(2.29)
Beta(3,3)	6.60	3.40	2.80	2.28
	(6.40)	(3.53)	(2.67)	(2.32)
Beta(10,10)	6.67	3.53	2.76	2.46
	(6.39)	(3.65)	(2.76)	(2.42)
Clustered	6.99	4.17	2.65	2.37
	(6.76)	(3.93)	(2.84)	(2.38)

squares approach which was constructed to mimic the likelihood function with a simpler mathematical form. We introduced several techniques that were later applied to the likelihood solution including reparameterization of the objective function and Taylor expansion. For each method we provide simulation studies that corroborate the theoretical results.

Chapter 4

EXPONENTIAL CORRELATION IN TWO-DIMENSIONS

In Chapter 3 we explored the limiting behavior of $\hat{\theta}$, the MLE of the range parameter, for the Ornstein-Uhlenbeck process in one-dimension. We examined the impact of the sampling pattern on the weighted least squares estimate as well as the maximum likelihood estimate. We were able to quantify the asymptotic variance for estimators for increasing levels of infill and an expanding domain including the limiting case, i.e., $n \to \infty$. Simulation results support the theoretical results.

The one-dimensional case provided insight into the behavior of the estimator as a function of increased sampling effort, both infill and expanding domain. Now we would like to broaden the scope to higher dimensions. While most ecological studies collect observations over a two-dimensional area such as a forest or on the surface of a body of water; atmospheric and oceanographic studies often require analysis in three spatial dimensions. The question of interest is "do the same asymptotic properties hold?" Many new obstacles are encountered when analyzing data in higher dimensions. For example, the concept of ordering sampling locations becomes much less intuitive when compared to the one-dimensional case. In addition, the order of the sample locations impacts the overall structure of the correlation matrix $\mathbf{\Gamma}$. Thus there are many factors that may impact the asymptotic behavior of $\hat{\theta}$.

This chapter examines the exponential covariance function model in two-dimensions. We empirically estimate θ by maximizing the likelihood function for various combinations of sampling patterns and the total sampling effort. The intent

is to gain an understanding of the behavior of $\hat{\theta}$ as a function of infill and expanding domain and compare it to the one-dimensional case.

4.1 Sampling patterns in two-dimensions

For the one-dimensional Gaussian process with exponential correlation function, the paraemter set is continuous and one is able to sample the process at any location t. Hence the investigator could conceivably make observations of the process using any of numerous sampling patterns, e.g., regularly spaced observations or randomly selected locations. Furthermore, the investigator can return to the process and make additional observations at new locations within the current domain (infill), outside the current domain (expansion of the domain), or some combination of the two. To extend this concept to two (spatial) dimensions we need to introduce the concept of a random field. We assume that the field can be observed at any spatially referenced location. Examples include observing the percent silt in the soil in a farmers field or recording the surface temperature at distinct locations across a body of water such as a lake. For both of these examples any location in the domain of interest can be referenced by two spatial coordinates such as latitude and longitude. Most importantly, the assumption of a continuous random field preserves the concepts of infill and expansion of the domain with respect to collecting additional observations. Hence we need only to generate sensible sampling schemes.

To mirror the general sampling patterns used for the one dimensional case we first consider two of the patterns found in previous chapters: the regular lattice and the uniform pattern. Figure 4.1 illustrates a regular lattice and a single realization of the uniform (random) pattern, also known as a homogeneous Poisson process, where $N = 121$ and the domain is 4×4 square. Accompanying each sampling pattern is a plot of the empirical (Ripley's) K-function. The K-function is the cumulative distribution function of the distances between all pairs of sampling

locations. Roughly speaking the K-function is a powerful diagnostic tool typically used to classify point processes as uniform, regularly spaced, or clustered. The K-function for the homogeneous Poisson process is $K(d) = \pi d^2$ where $d > 0$. Since the locations of the sampling points are assumed to be independent, the proportion of individuals within d units is proportional to the surrounding area. The plot of $L(d) = (K(d)/\pi)^{1/2}$ should be a line for a Poisson process. Hence we have a null hypothesis with which we can test the distribution of points. For the uniform pattern the empirical K-function, $\hat{L}(d)$, follows the trajectory of the theoretical $L(d)$ very closely. However, for the regular lattice $\hat{L}(d)$ is remarkably different. Since the minimum distance between any two locations is fixed by the lattice itself, $\hat{L}(d) = 0$ for all distances less than this minimum value and the empirical K-function will fall below the hypothesized line. Furthermore, the set of all possible distances between any two points is finite leading to the typical "stair-step" pattern as illustrated in the upper right panel.

To simulate clustering we departed from the original formulation as presented in Chapter 2. The cluster centers were generated by first uniformly selecting n sites within a 10×10 square, referred to as the parent locations. Centered at each parent location is a (radially symmetric) mean-zero bivariate normal distribution with covariance function $\Sigma = \tau^2\mathbf{I}$. To prevent points from being too close to one another we generated a very fine lattice onto which all sampling locations were restricted. Thus if two or more parents are close to one another and their "circles" of influence overlap, then the affected lattice locations are weighted by the sum of the densities. Thus higher weight is placed at locations that are at or near one or more parent locations while outlying locations far from the parents receive little or no weight. Figure 4.2 illustrates the construction of a clustered sampling pattern for $N = 121$ over a 4×4 square. The upper left panel is a temperature plot where white (think "white" hot) indicates locations with relatively large weights and red

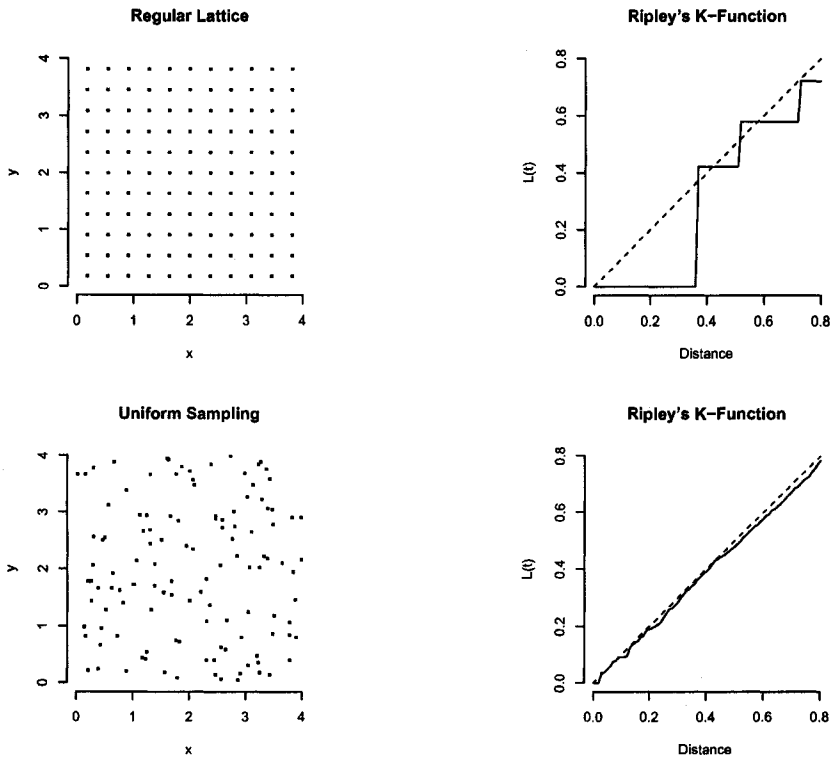

Figure 4.1: Illustration of two of four sampling patterns used in the two-dimensional studies. The upper panels demonstrate the regular lattice and its corresponding Ripley's K-function. The lower two panels present a single realization of the random pattern and the corresponding Ripley's K-function. For both sampling schemes $N = 121$ and the domain is $m \times m = 4 \times 4$.

corresponds to small (or zero) weights. The upper right panel is a perspective plot of the same weight surface. Large peaks correspond to regions with several parents in close proximity. The bottom two panels illustrate a single realization and the corresponding K-function. Note that the K-function indicates that there is a minimum distance between points ($\hat{L}(d) = 0$ for $d \leq 0.02$) and then quickly rises above the hypothesized line. This is because there are many pairs of points that are close to one another (within clusters) and many pairs that are far apart (across clusters) giving the K-function its distinct shape. Finally, note that the total

sampling effort for the two-dimensional clustered patterns is not random. For the one-dimensional "cluster" pattern we allowed the number of observations per unit block to be random.

A total of four sampling schemes were used to explore the behavior of the spatial parameter estimators: regular, uniform, and two clustering patterns ($\tau = 0.4$ and $\tau = 0.2$), hereafter referred to as light and heavy clustering, respectively. For both the exponential and Matérn correlation simulations in two-dimensions the sampling locations were constrained to a fine lattice for all three non-regular sampling patterns. For the non-regular sampling patterns a new realization of the sampling locations was generated for each simulation. Thus for the cluster patterns a new weight surface was generated first and then a single realization of the sampling locations was generated using the updated weight surface.

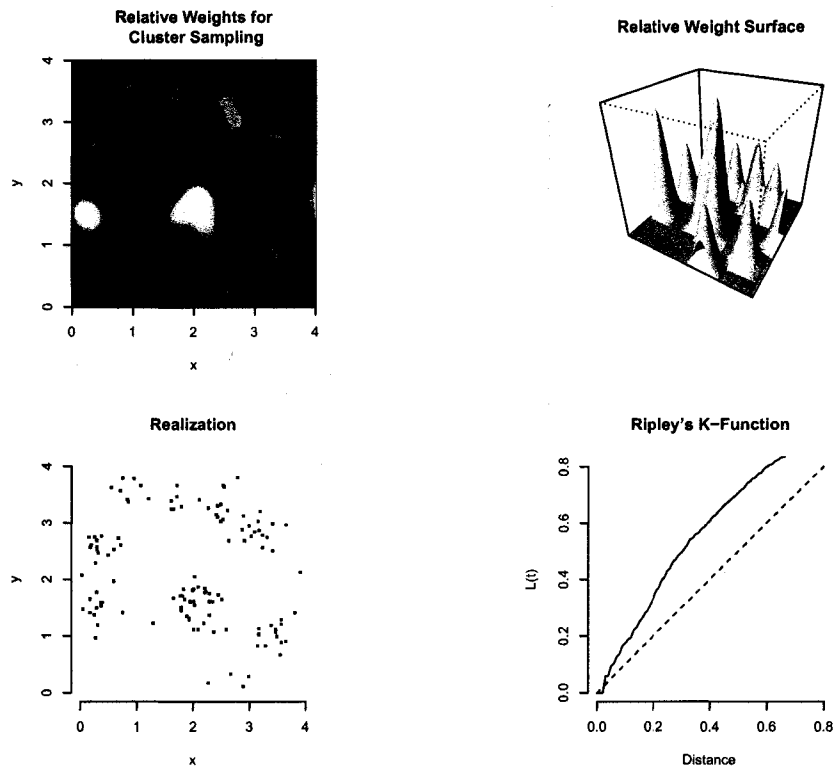

Figure 4.2: Illustration of the construction of a clustered sampling pattern. The upper left panel is a image plot of the relative weights assigned to each node of a fine grid. The upper right panel is the corresponding perspective plot of the same weight surface. The N sample locations are randomly selected over the $m \times m$ square according to this surfaces. The lower panels illustrate a single realization and the corresponding Ripley's K-function. Here $N = 121$, $m = 4$, and the individual weight surface for each parent location is the product of two independent mean-zero normal densities with $\sigma = 0.2$ (heavy clustering).

4.2 Exponential correlation function

We begin by first reviewing the derivation of the profile likelihood function. Let $\mathbf{Y} = \{Y(s_1), \ldots, Y(s_N)\}$ be a partial realization of a mean-zero Gaussian random field over some domain $\mathcal{D}$. The corresponding model for the N-vector $\mathbf{Y}$ is

$$\mathbf{Y} \sim \mathcal{N}(0, \sigma^2\mathbf{\Gamma}), \tag{4.1}$$

where N is the total sampling effort, σ^2 is the variance of the random field, and $\mathbf{\Gamma}$ is an $N \times N$ correlation matrix. If the correlation function for the Gaussian process is parameterized by θ, then after concentrating out σ^2, the profile likelihood for θ based on the obcerved data $\boldsymbol{Y}$ is given by

$$\ell_{profile}\left(\theta; \hat{\sigma}^2, \mathbf{Y}\right) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\hat{\sigma}^2 - \frac{1}{2}\log|\mathbf{\Gamma}| - \frac{N}{2}, \tag{4.2}$$

where $\hat{\sigma}^2 = N^{-1}\mathbf{Y}'\mathbf{\Gamma}^{-1}\mathbf{Y}$. For the exponential correlation function the $(i,j)^{th}$ element of $\mathbf{\Gamma}$ is $\exp\{-|\mathbf{s}_i - \mathbf{s}_j|/\theta\}$ where $|\mathbf{s}_i - \mathbf{s}_j|$ is the Euclidean distance between locations i and j and θ is the range parameter.

In general evaluating (4.2) requires calculation of $\det(\mathbf{\Gamma})$ and the quadratic form $\boldsymbol{Y}'\mathbf{\Gamma}^{-1}\boldsymbol{Y}$ which can be computationally intensive. Brockwell and Davis (1996) show that for any (one-dimensional) AR(p) process (4.2) can be written in terms of the one-step ahead predictions and their corresponding variances. By employing the Innovations algorithm, a recursive algorithm for estimating model parameters for series with finite second moments (Brockwell and Davis, 1996), one can greatly reduced the computational burden and evaluate the likelihood quickly. Hence, for the exponential correlation function in one-dimension, which we showed is exactly equivalent to an AR(1) process, evaluating the (reduced) likelihood is relatively inexpensive. Unfortunately, for higher dimensions no such algorithm currently exists, thus (4.2) must be evaluated by computing $\det\mathbf{\Gamma}$ and $\boldsymbol{Y}'\mathbf{\Gamma}\boldsymbol{Y}$ explicitly For the exponential correlation function there is only a single parameter over which to optimize

and hence the time required to find a solution is generally not too long. Optimization over a higher dimensional parameter can take much longer. For example, the profile likelihood function for the Matérn correlation function is a function of two spatial parameters. If one assumes the presence of measurement error the dimension of the parameter space increases to three.

It should be noted that there do exist methods that approximate the likelihood function. For example, Whittle's algorithm uses the spectral density and effectively eliminates the need to invert the correlation matrix (Whittle, 1954). For the current treatise we are interested in the asymptotic behavior of the maximum likelihood estimate of the range parameter θ, thus we have chosen not to use approximation methods to evaluate the likelihood function. In short we do not wish to introduce additional sources of error by using estimates for the likelihood when one can compute it exactly. The trade-off is that we are restricted to moderate sample sizes (especially when working with the Matérn correlation function).

4.2.1 2D simulation setup for the exponetial correlation case

We conducted a series of simulations using three different values of the range parameter ($\theta = \{1/2, 1, 2\}$) by simulating realizations of (4.1). Since the optimization requires inversion of the correlation matrix $\mathbf{\Gamma}$ we restricted the maximum sampling effort to $N = 256$. We generated realizations for all sampling efforts equal to the square numbers between $6^2 = 36$ and $16^2 = 256$ inclusive, i.e., $N = \{36, 49, \ldots, 225, 256\}$. The simulated domains include all square areas with sides of length $m = \{1, 2, \ldots, 8\}$, i.e, $m \times m = \{1 \times 1, 2 \times 2, \ldots, 8 \times 8\}$. Recall that the effective range (3θ) for the three parameter values of θ are 3/2, 3, and 6, respectively. Therefore only the large domains will contain numerous pairwise (approximately) independent points. This is especially true for $\theta = 6$. One hundred realizations for each combination of θ, $m \times m$, N, and sampling pattern were gen-

erated and the maximum likelihood estimate of θ computed and recorded. For all simulations the variance parameter was fixed at $\sigma^2 = 1$.

4.2.2 Simulation results and discussion

Figures 4.3, 4.4, and 4.5 illustrate how the mean square error (MSE) changes with respect to the domain and the sampling effort for each combination of θ and sampling pattern. Expansion of the domain is represented by moving from left to right across a particular image. Moving from the bottom to the top of the image represents increasing the level of infill. The highest density of sampling locations occurs when $m \times m = 1 \times 1$ and $N = 256$ (upper left corner) and the smallest density when $m \times m = 8 \times 8$ and $N = 36$ (lower right corner). Note that the x- and y-axes are linear with respect to m and $\sqrt{N}$, respectively. Each of the image plots have been smoothed using simple linear interpolation.

Of primary interest is the shape of the MSE surfaces and how they do (or do not) differ across sampling pattern and range parameter θ. Clearly evident from each image is the reduction of the MSE with respect to increasing the domain for a fixed sampling effort N. Each image has a large (relative) MSE for small domains regardless of sampling effort. This is perhaps best illustrated by Figure 4.4 where $\theta = 1$. Similar to the one-dimensional case, for a fixed domain, increasing the the level of infill does not appear to dramatically reduce the MSE. This trend is most evident for $\theta = 1$ but it does appear to be present for the remaining two cases. The "patchiness" of the images, i.e., less well defined contours compared to the one-dimensional simulation studies, is likely a result of the relative small number of simulations. It is presumed that the accumulation of additional realizations will temper the local variability of the MSE surfaces.

The relative rate at which the MSE decreases with respect to expansion of the domain at first glance appears to differ for each of the three values of θ. However,

closer inspection reveals that each MSE surface approximately reaches the "blue" at approximately 3θ. For example, for $\theta = 1$ the MSE surface is small along the transect of $m \times m = 3 \times 3$. Although not as well defined for $\theta = 2$, the same trend is evident in Figure 4.5. Similarly, for $\theta = 1/2$ the MSE surface is small along the $m \times m = 2 \times 2$ transect. (Currently no simulations have been produced for a domain of size 1.5×1.5.) This observation suggests the importance of collecting data over a domain that encompasses the effective range.

It is not surprising that, in general, for smaller sampling efforts the MSE is large. However, a peculiar feature of the upper left panel of Figure 4.3, where the $\theta = 1/2$ and the sampling pattern is regular, is the distinguishable plateau (or mesa if you prefer) for small N and large $m \times m$. This is not an artifact due to simulation. Note that when $N = 36$ and $m \times m = 8 \times 8$, the distance between adjacent points is $8/6 = 1.25$ for the regular sampling pattern. Hence there is enough separation between the sample locations that the observations are nearly independent, i.e., $\rho(-1.25/0.50) = 0.082$. This phenomenon was not encountered in the one-dimensional case due to the design of the experiment. Recall that the first four of the five sampling methods (regular, uniform, beta(3,3), beta(10,10)) used for the one-dimensional analysis enforced that at least one observation was made per unit length. The fifth pattern, referenced as cluster sampling, assigned on average at least one observation per unit length. Since none of the simulated processes had an effective range of less than one unit, subsequent sampling locations were (almost) guaranteed be within the effective range. For the present study this is not the case. For a fixed sampling effort, the mean distance between sampling sites necessarily increases as the domain increase. For the exponential correlation function sampling locations separated by more than 3θ are essentially independent. As the mean distance between locations approaches, and subsequently exceeds, 3θ the set of observations essentially becomes iid. Note that the regular sampling pattern is most

susceptible to this phenomenon of high MSE for large $m \times m$. By design the distance between adjacent sampling locations must increase as the domain is expanded for a fixed sampling effort. The remaining three patterns are less susceptible because of the manner in which they assign locations. For the uniform pattern the distribution of the pairwise distances is uniform. Until the domain is very large there is non-zero probability that two or more locations will be near one another. The cluster patterns virtually guarantee that there will be sample locations within close proximity of one another. Hence to properly examine the asymptotic behavior of the MLE for θ one must account for the density of the sampling locations.

The experimental design for the one-dimensional simulations prevented the possibility of having too few sample locations within the domain of interest. This is because the (expected) number of sampling locations per unit length was always held constant. Therefore as the domain was increased the total number of sampling locations also increased and the (expected) density remained constant. The equivalent idea in two-dimensions is to hold the mean number of observations per unit area constant as the domain is expanded. Referring to the image plots of the MSE, the contours of constant density are straight lines with positive slope such that the greater the density the steeper the slope.

Asymptotic analysis of the MLE for θ is only meaningful with respect to constant or increasing sampling densities. We define sampling density as the mean number of sampling locations per unit area. Tables 4.1 and 4.2 decompose the MSE into the standard error and bias. The tables illustrate how each varies with respect to expansion of the domain for fixed sampling densities of 4 and 9. The bias is negative for all θ and sampling patterns. Recall that the theoretical and observed bias was positive for all combinations of θ and sampling pattern in one-dimension. Thus in two-dimensions the range parameter is, on average, being underestimated. The magnitude of the bias is generally smaller than the corresponding standard error.

This is especially true when the range parameter is small, e.g., $\theta = 1/2$. This implies that the order of magnitude of the standard error dominates the calculation of the MSE. Furthermore the bias appears to go to zero more rapidly than the standard error. For example, for a sampling density of 4 locations per unit area and a range parameter of $\theta = 2$ (bottom panel of Table 4.1) the magnitudes of the standard error and bias are the same for the smallest domain of $(m \times m) = (3 \times 3)$. However, as the domain is expanded, the absolute value of the ratio of the standard error to bias increases from approximately 1 to somewhere between 2 and 3 (dependent on sampling pattern).

Figure 4.6 shows the observed MSE for constant sampling densities for each θ and sampling pattern. The left column of panels correspond to a densities of 4 sampling locations per unit area and the right column summarizes the MSE for a density of 9 locations per unit area. Immediately apparent is that the MSE decreases as the domain increases implying that the MLEs are consistent with respect to expanding domain. Since the bias term approaches zero, the MSE approximates the asymptotic squared standard error for the relatively large domains. Each plot also indicates that for moderate and large domains, the regular (——) and uniform (——) sampling patterns perform best with respect to the MSE. Furthermore, the heavy clustering (——) pattern appears to be the poorest performer for all but one case. The overall shape of the plots closely matches the expected asymptotic variance plots presented in Section 3.4.2.1 (see Figure 3.12).

Figure 4.7 plots the histogram of the MLE estimator $\hat{\theta}$ for each sampling method where $\theta = 1/2$ and $(N, m \times m) = (256, 8 \times 8)$. These plots suggest that the distributions are approximately normal. We employed the Anderson-Darling test to formally test the normality of $\hat{\theta}$ for each combination of θ, sampling effort $(N, m \times m)$, and sampling pattern. The results for $\theta = 1/2$ and $\theta = 1$ are summarized by Figures 4.8 and 4.9. The null hypothesis is that the distribution is normal. The blue squares

indicate that the null hypothesis was rejected at the 0.05 significance level. Red squares indicate that there was insufficient evidence to conclude that the distribution was not normal. For the case where $\theta = 1/2$ the regular sampling pattern is most successful at inducing normality for the distribution of $\hat{\theta}$, typically for domains of size 6×6 or larger and for moderate to large sample sizes. The remaining three sampling patterns are less likely to induce normality over the domains and sampling efforts studied. Notice that for $\theta = 1$ many fewer distributions are designated as approximately normal. (For the case where $\theta = 2$ only one distribution was designated as normal, results not included here.) There observations suggest that the simulated domains are not large enough to observe approximately normally distributed range parameter estimates for the cases $\theta = 1$ and $\theta = 2$ and sample sizes up to $N = 256$. Recall that the effective ranges are approximately 3.0 and 6.0, respectively. It is expected that if the simulated domains are increased that the resulting distributions for $\hat{\theta}$ will be approximately normal.

The simulation results in two-dimensions provide evidence to support the hypothesis that regularly spacing the sampling locations is optimal with respect to MSE with the caveat that if the domain of interest is large with respect to the accompanying sampling effort N, then it may be fruitful to employ the uniform pattern. For the researcher in the field this has important applications. For example, if it is known a priori that the correlation structure is exponential then a regular pattern is likely to be best. However, if there is no a priori knowledge about the effective range then the researcher risks locating the sampling locations too far apart if the true range parameter is small (relative to the domain). In this case, she may be better served to hedge her bet and employ a uniform pattern or conduct a preliminary study to better evaluate the optimal sampling pattern. Typically, however, the structure and range of the correlation function is unknown a priori. Thus the modeler often requires a more flexible family of correlation functions. The Matérn

correlation function contains as special cases both the exponential and Gaussian correlation functions making it a strong candidate for modeling spatially correlated isotropic data. It is precisely for this reason that we devote the next chapter to exploring the behavior of the MLE of the spatial parameter vector for both the one- and two-dimensional Gaussian processes with a Matérn covariance function.

4.3 Conclusions

One of the key results illustrated by the simulation study is the importance that the domain of the sampling locations encompasses the effective range (3θ). This ensures that there are sampling locations that are effectively independent of one another. However one must take care not to allow the mean distance between sampling locations to be too great or correlation may be undetectable. Furthermore, it is important to compare sampling designs with respect to the mean density of sampling locations per unit area. Empirical results confirm that for a constant density the MSE decreases with expanding domain. Finally, to induce normality of the MLE for the range parameter one requires sampling over a relatively large domain with respect to the true effective range. Additionally, the regular and uniform sampling designs appear to be most effective at obtaining this result.

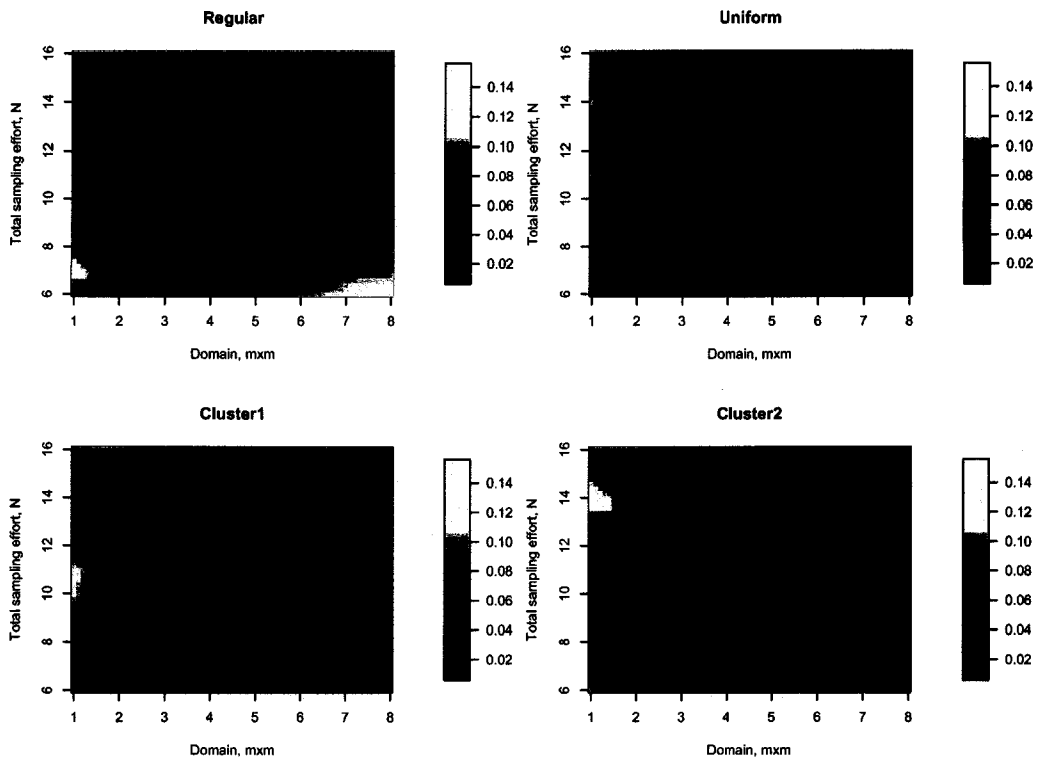

Figure 4.3: Image plots of the MSE for $\hat{\theta}$ sampled from a two-dimensional mean-zero exponentially correlated Gaussian random field with known range parameter of $\theta = 1/2$ as a function of four sampling pattern: regular, uniform, light and heavy clustering. Each panel plots the observed MSE for sampling effort versus domain for a particular sampling pattern. Note the color scale for MSE is the same for all four panels and that the scales for x- and y-axes are linear in $\sqrt{N}$ and m, respectively. High density sampling, i.e, large N and small $m \times m$, are found in the upper left hand corner of each image plot and, conversely, low density sampling falls to the lower right of each image.

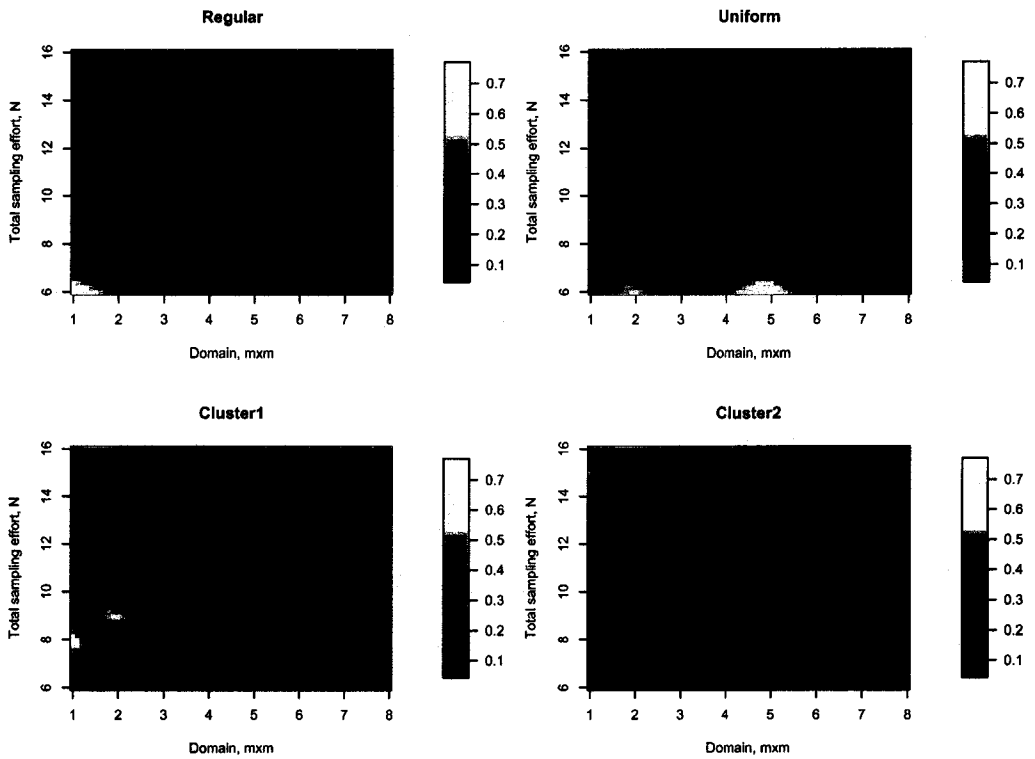

Figure 4.4: Image plots of the MSE for $\hat{\theta}$ sampled from a two-dimensional mean-zero exponentially correlated Gaussian random field with known range parameter of $\theta = 1$ as a function of four sampling pattern: regular, uniform, light and heavy clustering. Each panel plots the observed MSE for sampling effort versus domain for a particular sampling pattern. Note the color scale for MSE is the same for all four panels and that the scales for x- and y-axes are linear in $\sqrt{N}$ and m, respectively. High density sampling, i.e, large N and small $m \times m$, are found in the upper left hand corner of each image plot and, conversely, low density sampling falls to the lower right of each image.

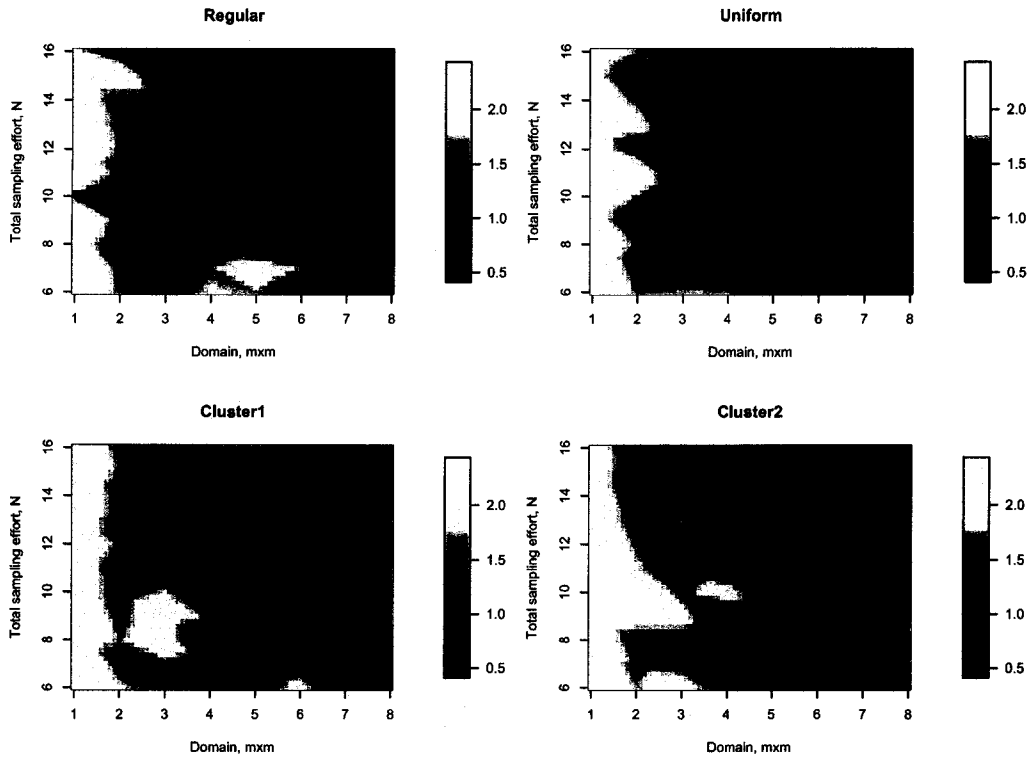

Figure 4.5: Image plots of the MSE for $\hat{\theta}$ sampled from a two-dimensional mean-zero exponentially correlated Gaussian random field with known range parameter of $\theta = 2$ as a function of four sampling pattern: regular, uniform, light and heavy clustering. Each panel plots the observed MSE for sampling effort versus domain for a particular sampling pattern. Note the color scale for MSE is the same for all four panels and that the scales for x- and y-axes are linear in $\sqrt{N}$ and m, respectively. High density sampling, i.e, large N and small $m \times m$, are found in the upper left hand corner of each image plot and, conversely, low density sampling falls to the lower right of each image.

Table 4.1: Observed standard error (se) and bias of the MLE $\hat{\theta}$ for a constant sampling density of $N/(m \times m) = 4$. The tabulated values are based on 100 independent simulations.

Range Parameter $\theta = 1/2$

Sampling Pattern		(Sample Size, Domain) $= (N, m^2)$ $(36, 3^2)$	$(64, 4^2)$	$(100, 5^2)$	$(144, 6^2)$	$(196, 7^2)$	$(256, 8^2)$
Regular	se	0.224	0.158	0.131	0.096	0.100	0.083
	bias	−0.116	−0.040	−0.018	−0.038	−0.021	−0.006
Uniform	se	0.190	0.176	0.113	0.108	0.089	0.091
	bias	−0.083	−0.042	−0.042	−0.029	−0.025	−0.021
Light Clusters	se	0.184	0.160	0.124	0.106	0.104	0.100
	bias	−0.102	−0.061	−0.046	−0.027	−0.011	−0.014
Heavy Clusters	se	0.271	0.168	0.188	0.118	0.113	0.120
	bias	−0.057	−0.063	−0.024	−0.047	−0.032	−0.021

Range Parameter $\theta = 1$

Sampling Pattern		(Sample Size, Domain) $= (N, m^2)$ $(36, 3^2)$	$(64, 4^2)$	$(100, 5^2)$	$(144, 6^2)$	$(196, 7^2)$	$(256, 8^2)$
Regular	se	0.593	0.370	0.332	0.305	0.235	0.247
	bias	−0.264	−0.103	−0.085	−0.054	−0.099	−0.027
Uniform	se	0.460	0.383	0.446	0.265	0.217	0.252
	bias	−0.230	−0.155	−0.057	−0.093	−0.079	−0.081
Light Clusters	se	0.380	0.411	0.338	0.320	0.254	0.253
	bias	−0.264	−0.186	−0.104	−0.054	−0.117	−0.046
Heavy Clusters	se	0.536	0.487	0.381	0.371	0.320	0.269
	bias	−0.294	−0.155	−0.152	−0.074	−0.088	−0.105

Range Parameter $\theta = 2$

Sampling Pattern		(Sample Size, Domain) $= (N, m^2)$ $(36, 3^2)$	$(64, 4^2)$	$(100, 5^2)$	$(144, 6^2)$	$(196, 7^2)$	$(256, 8^2)$
Regular	se	0.820	0.867	0.802	0.686	0.562	0.542
	bias	−0.808	−0.468	−0.429	−0.477	−0.353	−0.379
Uniform	se	1.12	1.17	0.752	0.773	0.754	0.620
	bias	−0.686	−0.503	−0.426	−0.308	−0.233	−0.197
Light Clusters	se	0.928	0.567	0.800	0.714	0.829	0.628
	bias	−0.845	−0.662	−0.546	−0.374	−0.255	−0.276
Heavy Clusters	se	1.12	0.805	0.850	0.850	0.782	0.856
	bias	−0.863	−0.736	−0.533	−0.329	−0.430	−0.387

Table 4.2: Observed standard error (se) and bias of the MLE $\hat{\theta}$ for a constant sampling density of $N/(m \times m) = 9$. The tabulated values are based on 100 independent simulations.

Range Parameter $\theta = 1/2$

Sampling Pattern		(Sample Size, Domain) $= (N, m^2)$ $(36, 2^2)$	$(81, 3^2)$	$(144, 4^2)$	$(225, 5^2)$
Regular	se	0.175	0.167	0.115	0.118
	bias	−0.128	−0.034	−0.038	−0.015
Uniform	se	0.180	0.165	0.129	0.096
	bias	−0.122	−0.054	−0.033	−0.026
Light Clusters	se	0.225	0.163	0.138	0.126
	bias	−0.082	−0.071	−0.032	−0.007
Heavy Clusters	se	0.208	0.217	0.131	0.161
	bias	−0.161	−0.041	−0.031	−0.008

Range Parameter $\theta = 1$

Sampling Pattern		(Sample Size, Domain) $= (N, m^2)$ $(36, 2^2)$	$(81, 3^2)$	$(144, 4^2)$	$(225, 5^2)$
Regular	se	0.599	0.371	0.335	0.329
	bias	−0.256	−0.187	−0.079	−0.053
Uniform	se	0.700	0.366	0.350	0.266
	bias	−0.253	−0.216	−0.145	−0.128
Light Clusters	se	0.401	0.323	0.334	0.313
	bias	−0.412	−0.246	−0.198	−0.049
Heavy Clusters	se	0.409	0.404	0.405	0.369
	bias	−0.370	−0.160	−0.136	−0.107

Range Parameter $\theta = 2$

Sampling Pattern		(Sample Size, Domain) $= (N, m^2)$ $(36, 2^2)$	$(81, 3^2)$	$(144, 4^2)$	$(225, 5^2)$
Regular	se	0.686	1.03	0.711	0.901
	bias	−1.07	−0.597	−0.610	−0.393
Uniform	se	0.822	0.866	0.748	0.822
	bias	−0.985	−0.621	−0.507	−0.452
Light Clusters	se	0.664	1.32	0.805	0.787
	bias	−1.15	−0.621	−0.555	−0.450
Heavy Clusters	se	0.638	1.27	1.03	0.674
	bias	−1.15	−0.566	−0.553	−0.553

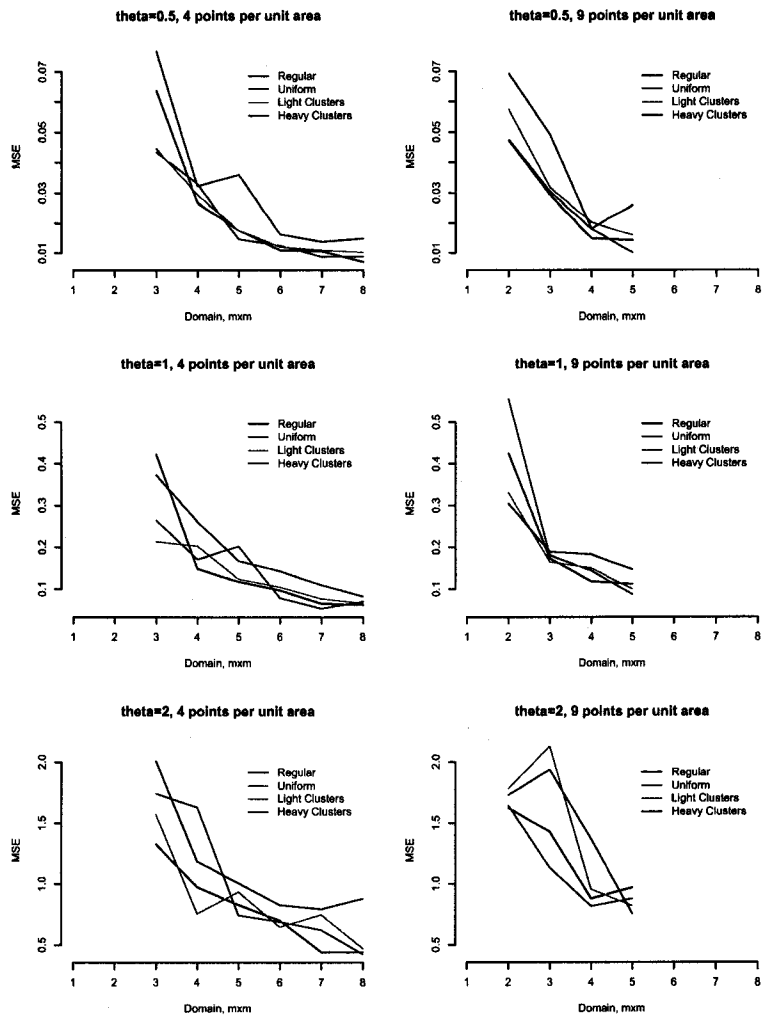

Figure 4.6: Observed MSE when sampling at a constant density, i.e., $N/(m \times m)$ where $(m \times m)$ is a constant. The left panels correspond to a constant density of 4 sampling locations per unit area and the right panels to a density of 9 locations per unit area.

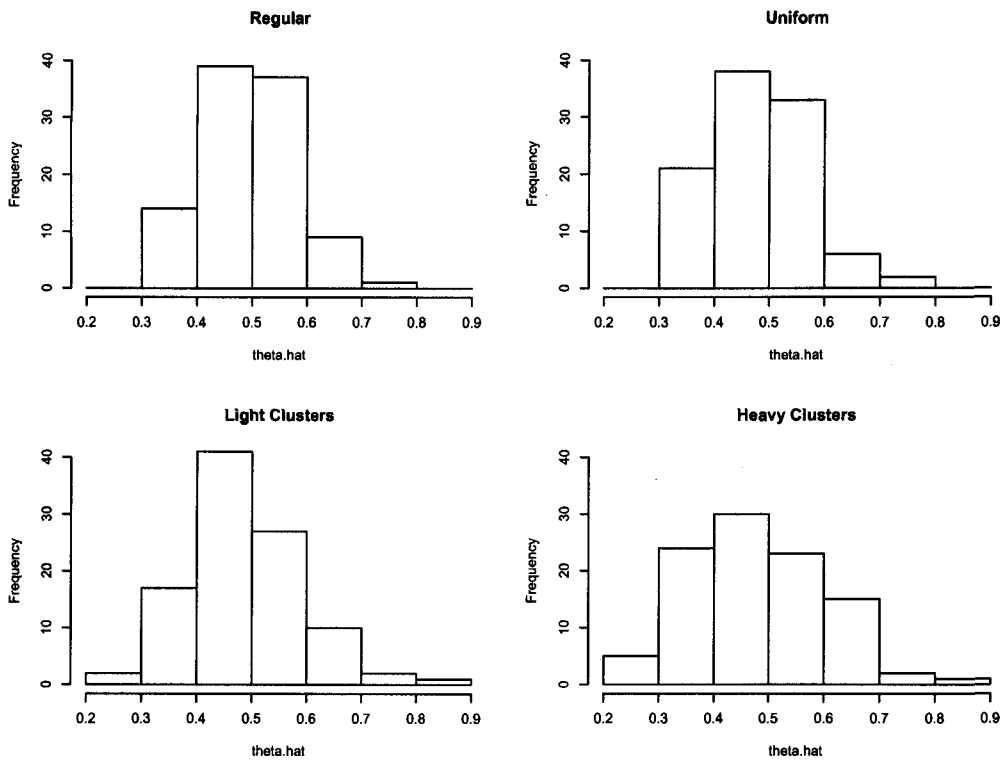

Figure 4.7: Distribution of the parameter estimator $\hat{\theta}$ for $\theta = 1/2$ and $(N, m \times m) = (256, 8 \times 8)$.

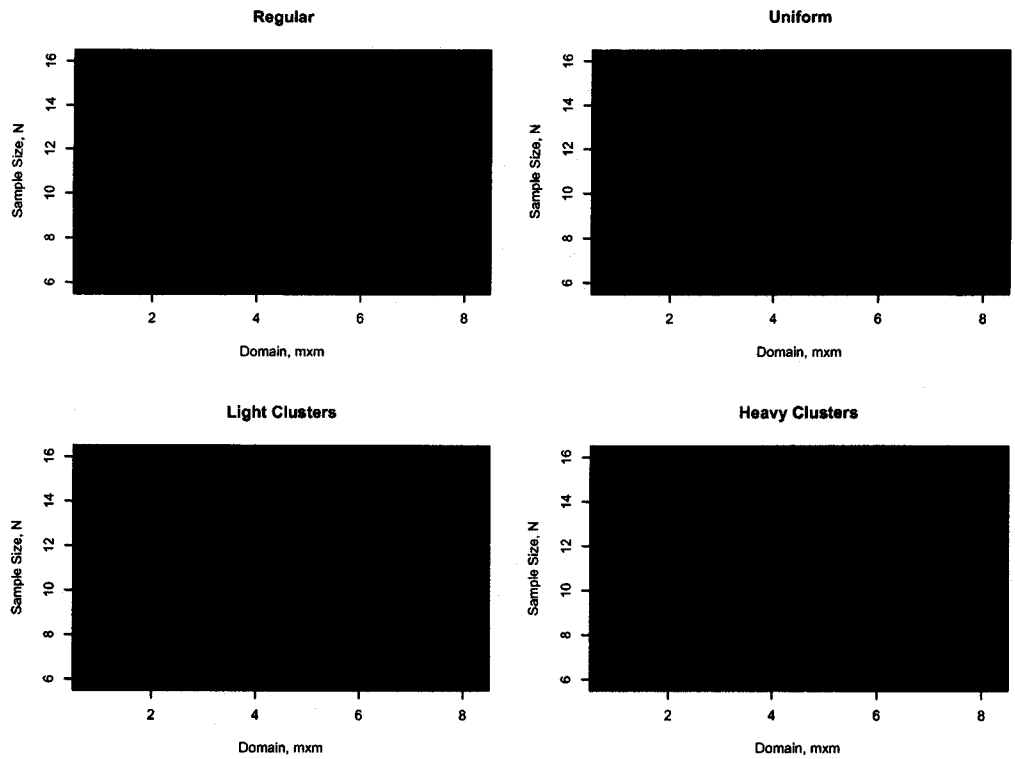

Figure 4.8: Normality testing results where $\mathbf{\Gamma}$ = exponential(1/2). The Anderson-Darling test was performed on the observed distribution of $\hat{\theta}$ at a significance level of 0.05 for all combinations of $(N, m \times m)$ and sampling pattern. The null hypothesis is that the distribution is normal. Blue squares correspond to non-normal distributions and red squares to normal distributions.

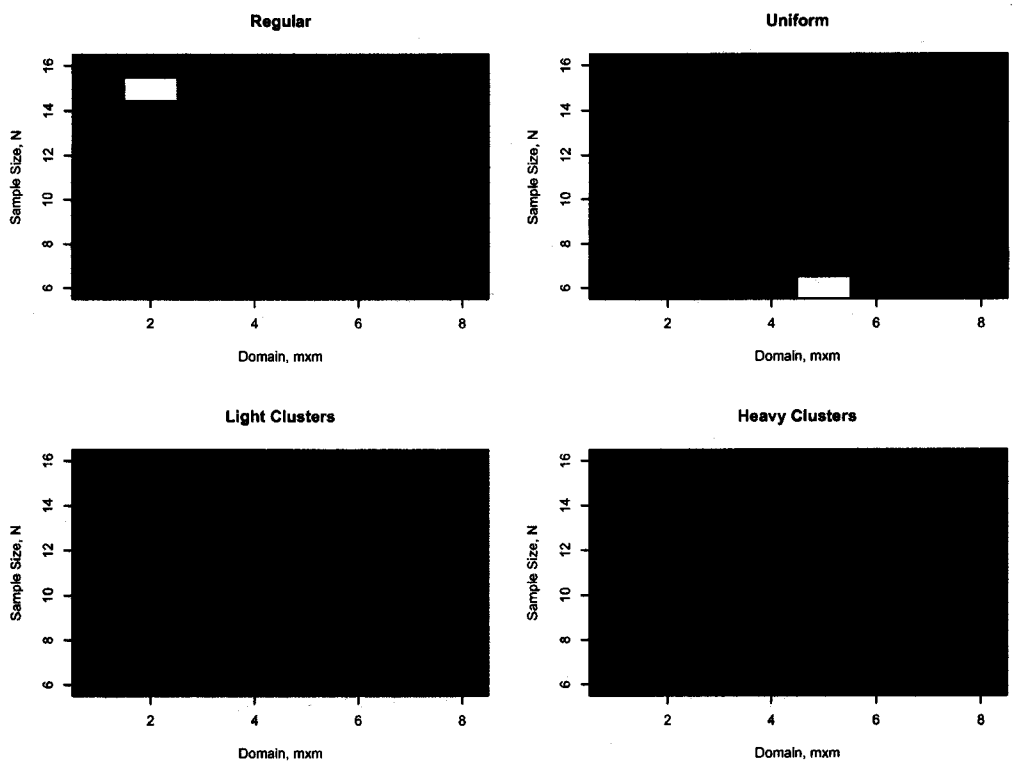

Figure 4.9: Normality testing results where Γ = exponential(1). The Anderson-Darling test was performed on the observed distribution of $\hat{\theta}$ at a significance level of 0.05 for all combinations of $(N, m \times m)$ and sampling pattern. The null hypothesis is that the distribution is normal. Blue squares correspond to non-normal distributions and red squares to normal distributions.

Chapter 5

MATÉRN CORRELATION FUNCTION

Up to this point we have focused on the exponential correlation function in one- and two-dimensions. The chief motivating factors for working with the exponential correlation function are its simple mathematical form, that it is characterized by a single parameter (when assuming no measurement error is present), and that the likelihood can be easily computed without the need to invert covariance matrices explicitly in the one-dimensional case. By exploiting each of these factors, we were able to develop the asymptotic distribution of the MLE for the range parameter $\hat{\theta}$ in one-dimension with respect to expanding domain and infill asymptotics. Simulations conducted in two-dimensions provided insight into the behavior of the MLE with respect to the expanding domain and infill and suggest that the form of the MLEs asymptotic distribution will be similar to the one-dimensional case. Now we would like to broaden the scope of the analysis and incorporate a larger class of correlation functions, namely the Matérn class.

5.1 Matérn correlation function

Over the past 10 to 20 years there has been increased interest in using the Matérn correlation function in data analysis (Peterson et al., 2006). Part of this stems from its flexibility; recall that both the exponential and Gaussian correlation functions are subclasses of the Matérn function (see Section 2.2). The Matérn function (assuming no measurement error) is characterized by the two-dimensional

vector $\boldsymbol{\theta} = (\theta_1, \theta_2)'$. Recall that the functional form is

$$\rho(d;\boldsymbol{\theta}) = \frac{1}{2^{\theta_2-1}\Gamma(\theta_2)}\left(\frac{2d\sqrt{\theta_2}}{\theta_1}\right)^{\theta_2}\mathcal{K}_{\theta_2}\left(\frac{2d\sqrt{\theta_2}}{\theta_1}\right), \qquad \theta_1 > 0,\ \theta_2 > 0, \tag{5.1}$$

where $\mathcal{K}_{\theta_2}(\cdot)$ is the modified Bessel function of order θ_2 (Abramowitz and Stegun, 1965). The "range" parameter, θ_1, controls the rate of decay of the correlation between observations as distance increases. Large values of θ_1 indicate that sites that are relatively far from one another are moderately (positively) correlated. The parameter θ_2, typically referred to as the "smoothness" parameter, can be described as controlling behavior of the correlation function for observations that are separated by small distances. We recover the exponential correlation function when $\theta_2 = 1/2$ which, in turn, implies $\theta = \theta_1/\sqrt{2}$. The Gaussian autocorrelation function is the limiting case when $\theta_2 \to \infty$. The Matérn class is very flexible, being able to to strike a balance between these two extremes. Figures 2.1 and 2.2 illustrate the flexibility of the Matérn correlation function. Notice that for small distances the correlation between sites is large and decreases as distance increases.

In Chapter 2 we demonstrated that for a fixed sampling effort we were more likely to identify the correct model form of the correlation structure when the spatial AIC statistic was used during model selection. We did not, however, investigate how well the maximum likelihood estimates compared with the known true parameter values for the simulated data. This is the focus of the subsequent material. Specifically we wish to determine whether the standard asymptotics assumed in the heuristic proof of the spatial AIC statistic hold. For example, is the distribution of maximum likelihood estimate $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})'$ approximately normal with mean $(\boldsymbol{\beta}, \boldsymbol{\theta})'$ and asymptotic covariance matrix given by the inverse of the Fisher information, I_n? We also hope to glean information regarding whether or not there is an optimal sampling design. Through simulation we hope to explore all of these aspects and lay the ground work for future theoretical work.

5.1.1 Simulation constraints for the Matérn class

The Matérn correlation function presents some difficulties not encountered with the exponential correlation function. Two primary concerns are (1) there are two parameters that must be estimated simultaneously, and (2) the presence of a Bessel function prevents easy calculation of the likelihood function.For this presentation we restrict attention to estimating the spatial parameter vector θ for a mean-zero Gaussian random field with Matérn correlation structure. Let $\mathbf{Y}$ denote the N-vector of observations from the random field $\{Y(s), s \in D\}$. Therefore $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \sigma^2\Gamma)$ where $\mathbf{\Gamma}$ is Matérn(θ_1, θ_2). Hence the profile log-likelihood function retains the same form with the only difference in the functional form of the correlation structure (see (1.7)). For large sampling efforts N optimization can be computationally expensive in terms of the number of operations required. Consequently we restricted the maximum data set size to be $N \leq 256$ (with a minor exception made for cluster sampling in one-dimension).

The correlation matrix $\mathbf{\Gamma}$ is positive definite for all $\theta_1 > 0$ and $\theta_2 > 0$ (Abramowitz and Stegun, 1965). Hence $\mathbf{\Gamma}^{-1}$ exists for all strictly positive values of θ_1 and θ_2. Computationally one is restricted to the numerical precision of the platform upon which one is working. For the Matérn correlation function, large range parameters generally imply that neighboring observations at close proximity are very strongly correlated, i.e., $\rho(|u - v|; \theta) \approx 1$. This leads to a nearly singular correlation matrix that may not be numerically invertible. The problem is further exacerbated when the smoothness parameter is also large. (See Figures 2.1 and 2.2.) Preliminary analysis demonstrated that the minimum distance between subsequent locations when the maxima for the range and smoothness parameters are constrained to 6 and 3, respectively, is 0.02 units. This number is somewhat arbitrary because it depends on the user defined maxima of "allowable" values of the spatial vector $\boldsymbol{\theta}$. Consequently, we implemented the following two constraints for

all simulations involving the Matérn correlation function: (1) the parameter space was constrained to $(\theta_1, \theta_2) \in \{[0.02, 6.00] \times [0.10, 3.00]\}$, and (2) all of the Matérn observations were restricted to a fine lattice.

Constraining the parameter space prevents the optimizer from straying too far from the "truth" and potentially generating a nearly singular correlation matrix. However, many realizations yielded optimal solutions that fell on the boundary of the parameter space, e.g., $\hat{\theta}_2 = 3.00$. This was especially true for the smaller sampling efforts, i.e., $N \leq 16$. We decided to include solutions on the boundary when computing summary statistics and creating maps of the mean square error (MSE) so long as the optimizer reported that the convergence criteria was met. (All likelihood functions in this study were optimized using the R function `optim()` developed by R Development Core Team (2006)). The optimization procedure `L-BFGS-B` is a quasi-Newton method that constrains each variable with an upper and lower bound defined by the user. The second constraint, placing all of the observation locations onto a fine lattice, eliminated the possibility of randomly assigned sampling locations being too close to one another. Section 4.1 describes the procedure(s) used for both the exponential and Matérn classes in two-dimensions. A similar discussion is presented in Section 5.1.2 for the Matérn class in one-dimension.

5.1.2 One-dimensional analysis

We begin with the exploration of the Matérn correlation function in one-dimension. As was performed for the exponential correlation function, we investigated the parameter estimates using five different sampling patterns first presented in Section 3.4.1.1: regular, uniform, beta(3,3) and beta(10,10) sampling patterns, and the clustered pattern. Recall that the first four patterns assign a fixed number of observations per block while the clustered pattern randomizes the number of observations for each block (following a negative binomial distribution). We restricted

the observation locations to a fine lattice to prevent neighboring observations from being too close. This in turn removed the potential for the correlation matrix $\mathbf{\Gamma}$ to be nearly singular as the optimizer searched over the parameter space (see Section 4.1). The minimum distance between any two sampling locations was 0.02 units.

Four parameter vectors $\boldsymbol{\theta}$ were used for simulation: (1) $\boldsymbol{\theta}_1 = (\sqrt{2}, 1/2)'$, (2) $\boldsymbol{\theta}_2 = (2, 1)'$, (3) $\boldsymbol{\theta}_3 = (2\sqrt{2}, 2)'$, and (4) $\boldsymbol{\theta}_4 = (4, 2)'$. The corresponding effective ranges are approximately 3.0, 4.0, 5.4, and 7.6, respectively. Note that the first case is exactly equivalent to the exponentially correlated model with $\theta = 1$. The domain m and level of infill n were incremented by powers of two such that $mn \in \{4, 8, \ldots, 256\}$. For each combination of $\boldsymbol{\theta}_i$, m, n, and sampling pattern 100 realizations were generated and the MLE for $\boldsymbol{\theta}$ computed and recorded.

5.1.2.1 Mean square error

Figures 5.1 and 5.2 illustrate how the MSE for each parameter varies with respect to the domain size and level of infill for sampling on the regular lattice. In these plots the scales of the x- and y-axes are $\log_2$. Note that the shape of the MSE surface is very similar for all four known spatial parameter vectors; the images differ in magnitude only. As both the range and smoothness parameters are increased, the MSE of the range parameter estimator increases and the MSE of the smoothness parameter estimator decreases. The general shape of the MSE surface for the range parameter closely resembles the observed surfaces for the exponentially correlated simulations in one-dimension. As the domain increases the MSE decreases but an increase in the level of infill (for a fixed domain) does not significantly reduce the MSE. This result closely mirrors the theoretical and observed behavior of $\hat{\theta}$ for the exponential correlation function in one-dimension. Note that θ is the "range" parameter for the exponential correlation function and it should not be surprising that the MSE surface of $\hat{\theta}_1$ is so similar in shape (if not magnitude). The MSE surface for the smoothness parameter estimator, $\hat{\theta}_2$, is decidedly different. For the

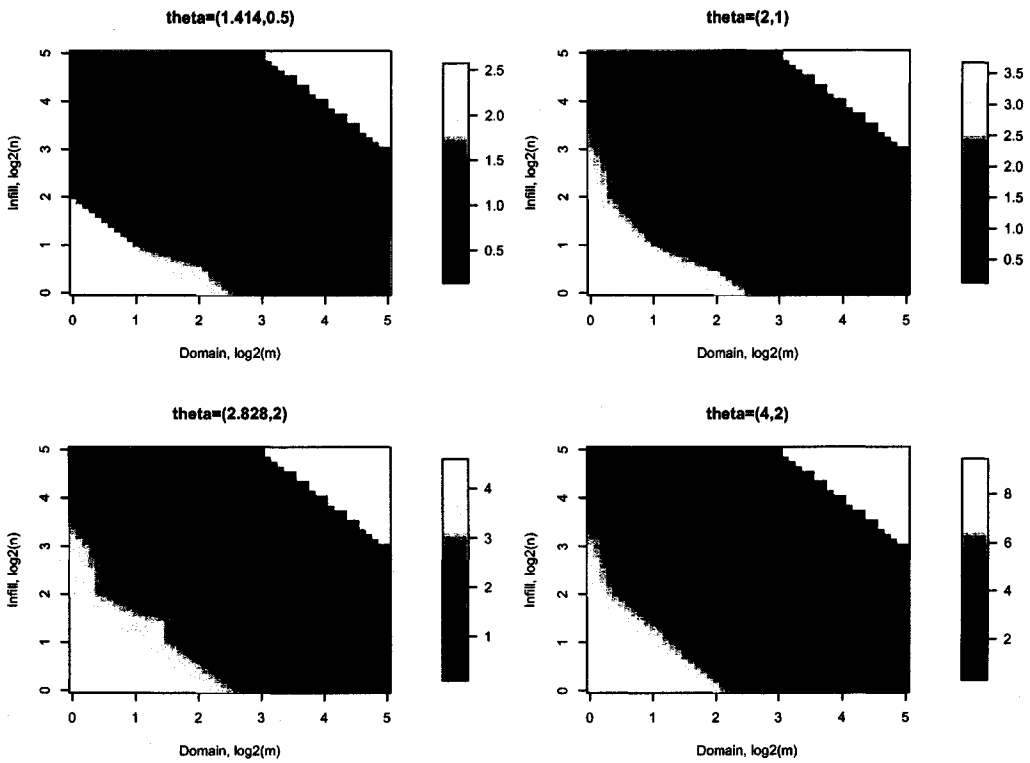

Figure 5.1: MSE of the range parameter estimator, $\hat{\theta}_1$, for the Matérn correlation function when sampling is restricted to a regular lattice in one-dimension. Note that the scales of the x- and y-axes are $\log_2$.

smoothness parameter there is a distinct reduction in the MSE as either the domain or level of infill is increased. Notice that the surface is not symmetric about the $n = m$ line. The MSE tends toward zero faster when the level of infill is increased. This observation matches what one would intuitively suspect. Since θ_2 controls the behavior of the correlation function for short distances it is not unexpected that making observations at close proximity should improve the overall estimate with respect to MSE.

Figures 5.3 and 5.4 represent the MSE surfaces for the four non-regular sampling patterns. Note that the temperature (color) scales have been fixed to match the

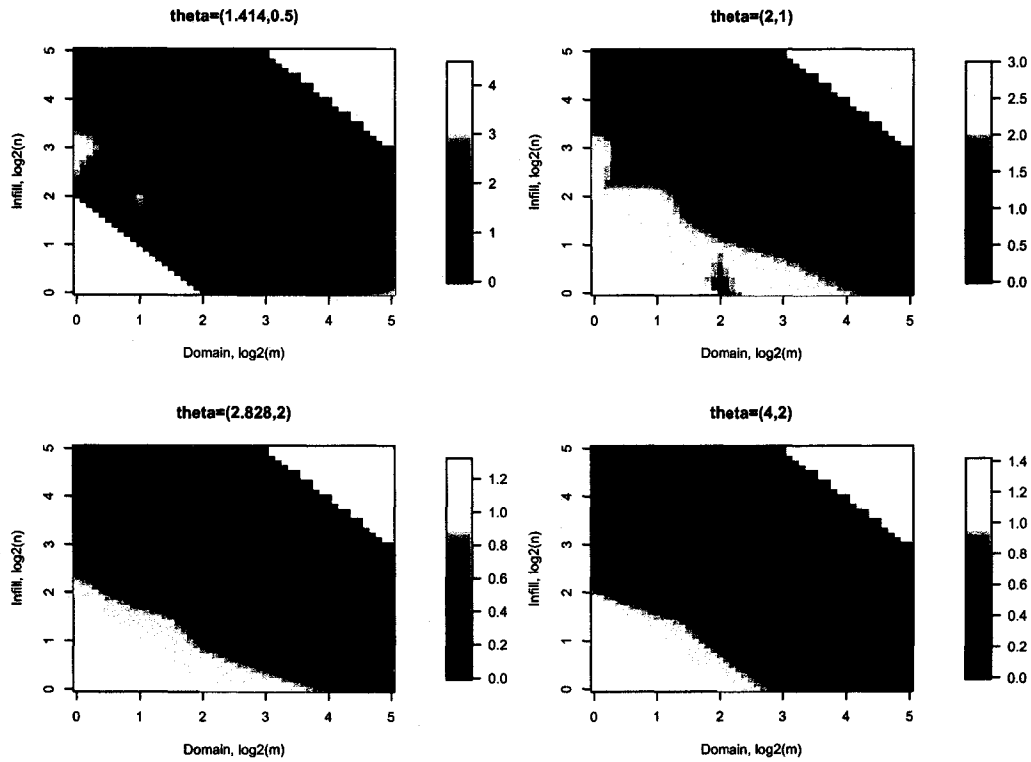

Figure 5.2: MSE of the smoothness parameter estimator, $\hat{\theta}_2$, for the Matérn correlation function when sampling is restricted to a regular lattice in one-dimension. Note that the scales of the x- and y-axes are $\log_2$.

corresponding scale for the regular lattice allowing for direct comparisons. Overall the MSE surfaces for each parameter are very similar to the corresponding surfaces for the regular lattice. The MSE for $\hat{\theta}_1$ decreases at a fast rate with respect to expansion of the domain. There is evidence that increasing the level of infill for a fixed domain does reduce the MSE for the range parameter estimate but at a decidedly slower rate, i.e., the asymmetry of the MSE surface about the one-to-one line is very pronounced. Similar to the regular pattern, the MSE for $\hat{\theta}_2$ decreases with respect to both the domain and level of infill. The asymmetry about the one-

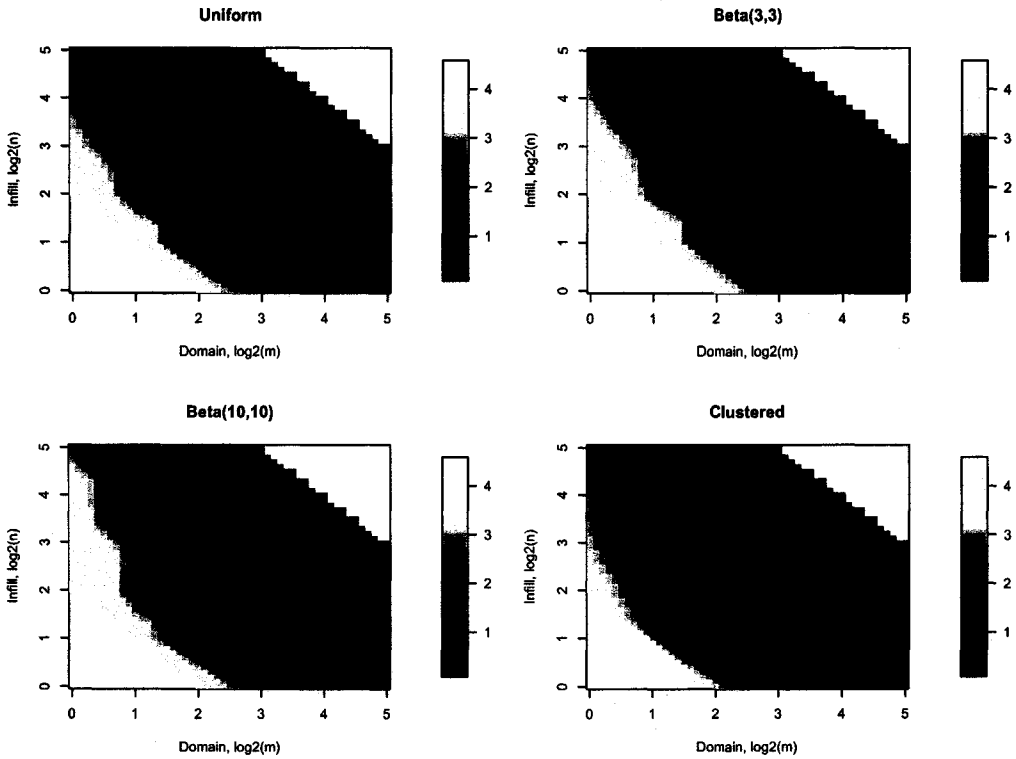

Figure 5.3: MSE of the range parameter estimator, θ_1, for Matérn$(2\sqrt{2}, 2)$ correlation when sampling is non-regular. Each panel illustrates how the MSE varies for a particular sampling pattern. The temperature (color) scales have been fixed to match the corresponding scale for the regular lattice. Note that the scales of the x- and y-axes are $\log_2$.

to-one line indicates that increasing the level of infill has stronger impact on the MSE.

The MSE surfaces indicate that there appear to be two competing strategies for placing sampling locations: (1) spreading the sampling locations over the entire domain of interest which results in small MSE for $\hat{\theta}_1$, and (2) locating sampling locations at close proximity which results in small MSE for $\hat{\theta}_2$. Of course one can combine both of these strategies into a third that assures some sampling locations are at close proximity while others are spread over the remainder of the domain. Cluster

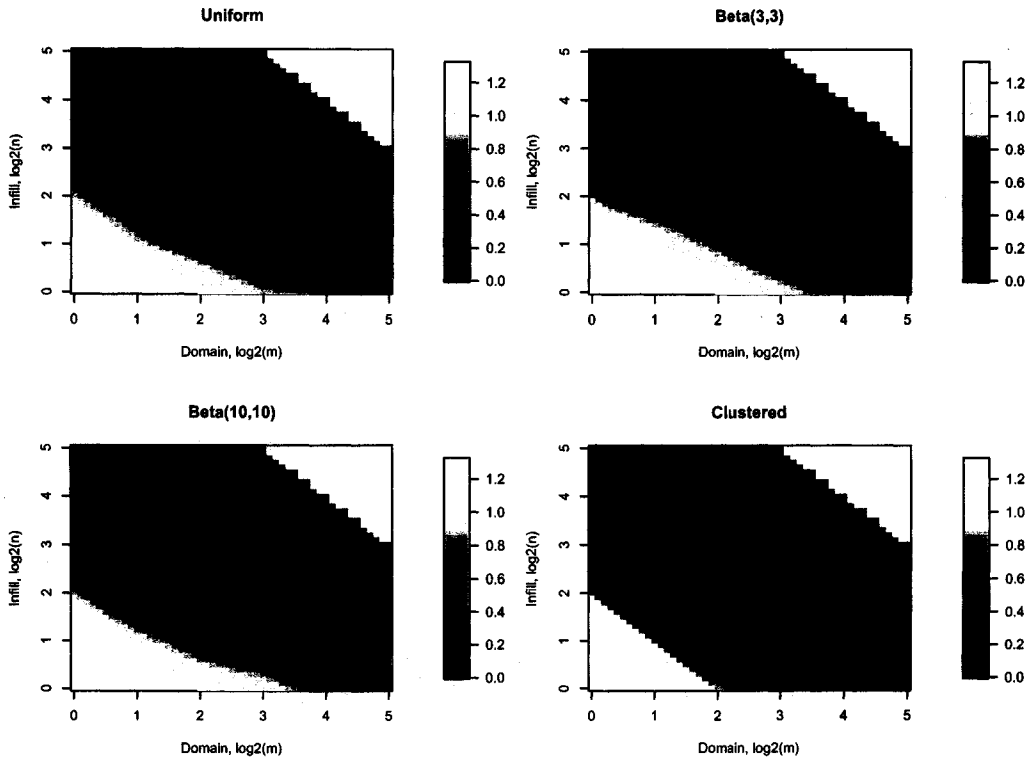

Figure 5.4: MSE of the smoothness parameter estimator, θ_2, for Matérn$(2\sqrt{2}, 2)$ correlation when sampling is non-regular. Each panel illustrates how the MSE varies for a particular sampling pattern. The temperature (color) scales have been fixed to match the corresponding scale for the regular lattice. Note that the scales of the x- and y-axes are $\log_2$.

sampling is one such sampling strategy (Zhang and Zimmerman (2005)). But which strategy to use is critical and may depend on the goal of the investigation. Typically a researcher's resources are finite and she desires to invest her sampling efforts in a manner that will glean the most precise estimate for θ. A second researcher, however, may be more interested in a subset of the spatial parameter vector and yet a third researcher may be interested in prediction.Since the scope of these goals is very broad, we will restrict attention to the asymptotic behavior of the MLE

of θ with respect to expanding domain and infill asymptotics and comment where appropriate about the impact of the sampling pattern on MLE.

Table 5.1 decomposes the MSE into the standard error (se) and bias terms for $\Gamma = \text{Matérn}(\sqrt{2}, 1/2)$ and $\Gamma = \text{Matérn}(4, 2)$ where the sampling effort is fixed at $mn = 256$. Empirical results are listed for the regular and three of the non-regular sampling patterns. The table illustrates that as the domain increases the standard error and the magnitude of the bias of $\hat{\theta}_1$ decrease. The same cannot be said for the estimator of the smoothness parameter $\hat{\theta}_2$. For some sampling patterns the standard error decreases with increasing domain and for others it increases. Furthermore there is no discernible general trend for the bias term. Recall for exponentially correlated data in one-dimension we were able to develop expressions for the expected variance and bias for any combination of sampling pattern and sampling effort (m, n). Currently we do not have that luxury for the Matérn class. Although the sampling effort is held constant, the density of observations is variable. Recall that the contours of constant sampling effort ($N = mn$) for Figures 5.1 through 5.4 are lines with a slope of -1. Since the MSE surfaces are asymmetric about the one-to-one line and the MSE is large when either m or n is small, then the MSE along the constant sampling effort contour must be bowl shape.

Tables 5.2 and 5.3 decompose the MSE of $\hat{\theta}_1$ and $\hat{\theta}_2$ with respect to a constant sampling density of $n = 4$. Listed are the empirical results of 100 independent realizations for each combination of sampling pattern and spatial parameter vector. Also tabulated is Pearson's sample correlation coefficient for $\hat{\theta}_1$ and $\hat{\theta}_2$. As expected there is a decrease in both the standard error and in the magnitude of the bias for both estimators as the domain is increased. The bias for the range parameter estimate is always negative implying that θ_1 is, on average, being underestimated. Recall for the exponentially correlated data in one-dimension that the bias was positive. The bias of the smoothness parameter estimate is positive. Furthermore, the

estimators are moderately (negatively) correlated over all simulations. Of concern is how these biases and correlation impact the overall estimation of the correlation function. (See Section 5.1.2.3 for further discussion.)

In general, the standard error and bias for the range parameter estimates are similar with respect to the sampling pattern. However, the standard error and bias are typically smaller for the beta(10,10) and cluster patterns than for either the regular or uniform patterns. In fact, the standard error is always largest for the regular pattern and the bias is nearly always largest. This is consistent with the observation that increasing the level of infill is more efficient at reducing the MSE of $\hat{\theta}_2$. By design the beta(10,10) sampling pattern generates clusters of size n centered in each unit length thereby being more capable of estimating θ_2 well. The cluster pattern and the uniform pattern also have built in mechanisms that virtually ensure that some if not many of the sampling locations will be at close proximity. In contrast, the design of the regular pattern spaces the sampling locations as far apart as possible for a given sampling effort. Hence it is not unexpected that the standard error for $\hat{\theta}_2$ is poorest for the regular pattern.

Another important observation illustrated by Tables 5.2 and 5.3 is that the standard error term dominates the bias term with respect to the MSE. In other words, as the domain increases the bias quickly approaches zero (in magnitude). For example, for $\mathbf{\Gamma} = \text{Matérn}(\sqrt{2}, 1)$ the ratio of the magnitudes of the standard error to the bias is approximately two for $(m, n) = (8, 4)$ and increases to approximately four when $(m, n) = (32, 4)$. Similarly, for $\mathbf{\Gamma} = \text{Matérn}(4, 2)$ the ratio is approximately one for a domain of length $m = 8$ and increases to about two for $m = 32$.Recall for the one-dimensional exponential correlation function that the expected variance for a fixed sampling effort (m, n) was a function of θ, m, and n. Since the exponential function is a subclass of the Matérn there is ample reason to believe that a similar asymptotic limits exist for $\hat{\theta}_1$ and $\hat{\theta}_2$.

The MSE surface plots clearly demonstrate that the MSE decrease with respect to expansion of the domain. Figures 5.5 and 5.6 are profile images of the MSE plot along the contour of a constant sampling density equal to four observations per unit length. The left panels are the MSE of the range parameter estimator and the right panels are for the smoothness parameter estimator. Overall, each sampling method performs nearly the same for each simulated spatial parameter vector $\boldsymbol{\theta}$. For the range parameter estimate there is some empirical evidence that the regular and uniform patterns out perform the other "clustered" patterns for very small domains. The story is reversed for the smoothness parameter estimate; the best performers in a MSE sense are the "clustered" patterns. These observations are consistent with intuition, i.e., sampling patterns that spread the data over a larger area will estimate the range parameter well whereas sampling patterns that locate observation locations at close proximity will, in general, estimate the smoothness parameter well. Finally, the MSE of $\hat{\theta}_1$ for each sampling pattern is very similar for the large domains. This is reminiscent of the derived MLE result for the AR(1) process, i.e., for medium to large domains the standard error of $\hat{\theta}$ is nearly indistinguishable across sampling patterns.

Table 5.1: Standard error (se), bias, and correlation for Matérn($\sqrt{2}, 1/2$) and Matérn$(4, 2)$. Estimates are based on 100 independent realizations for a fixed sampling effort of $N = 256$. $\hat{\rho}$ is Pearson's sample correlation coefficient for $\hat{\theta}_1$ and $\hat{\theta}_2$.

Spatial Parameter $(\sqrt{2}, 1/2)$

Sampling Pattern		(Domain, Infill) $= (m, n)$ $(m,n) = (8,32)$ se	bias	$(m,n) = (16,16)$ se	bias	$(m,n) = (32,8)$ se	bias
Regular	$\hat{\theta}_1$	0.850	−0.220	0.518	−0.118	0.431	−0.066
	$\hat{\theta}_2$	0.052	0.013	0.062	0.014	0.069	0.010
	$\hat{\rho}$	-0.421		-0.453		-0.570	
Uniform	$\hat{\theta}_1$	0.434	−0.454	0.531	−0.190	0.393	−0.130
	$\hat{\theta}_2$	0.047	0.009	0.046	0.009	0.042	0.010
	$\hat{\rho}$	-0.275		-0.451		-0.538	
Beta(10,10)	$\hat{\theta}_1$	0.692	−0.284	0.398	−0.258	0.380	−0.194
	$\hat{\theta}_2$	0.050	0.014	0.035	0.004	0.042	0.015
	$\hat{\rho}$	-0.322		-0.328		-0.494	
Clustered	$\hat{\theta}_1$	0.658	−0.279	0.530	−0.210	0.392	−0.056
	$\hat{\theta}_2$	0.053	0.006	0.049	0.013	0.043	0.006
	$\hat{\rho}$	-0.430		-0.416		-0.465	

Spatial Parameter $(4, 2)$

Sampling Pattern		(Domain, Infill) $= (m, n)$ $(m,n) = (8,32)$ se	bias	$(m,n) = (16,16)$ se	bias	$(m,n) = (32,8)$ se	bias
Regular	$\hat{\theta}_1$	0.872	−0.645	0.712	−0.334	0.627	−0.139
	$\hat{\theta}_2$	0.055	0.002	0.057	0.014	0.065	0.010
	$\hat{\rho}$	-0.621		-0.617		-0.639	
Uniform	$\hat{\theta}_1$	0.885	−0.851	0.668	−0.497	0.551	−0.279
	$\hat{\theta}_2$	0.052	0.010	0.054	0.015	0.057	0.014
	$\hat{\rho}$	-0.498		-0.351		-0.617	
Beta(10,10)	$\hat{\theta}_1$	0.867	−0.732	0.712	−0.442	0.646	−0.143
	$\hat{\theta}_2$	0.045	0.009	0.042	0.016	0.050	0.002
	$\hat{\rho}$	-0.538		-0.558		-0.583	
Clustered	$\hat{\theta}_1$	0.917	−0.640	0.689	−0.444	0.611	−0.209
	$\hat{\theta}_2$	0.059	0.004	0.045	0.004	0.054	0.008
	$\hat{\rho}$	-0.478		-0.582		-0.649	

Table 5.2: Standard error (se), bias, and correlation for Matérn($\sqrt{2}, 1/2$) and Matérn$(2, 1)$. Estimates are based on 100 independent realizations for a fixed sampling effort per block of $n = 4$. $\hat{\rho}$ is Pearson's sample correlation coefficient for $\hat{\theta}_1$ and $\hat{\theta}_2$.

Spatial Parameter $(\sqrt{2}, 1/2)$

Sampling Pattern		(Domain, Infill) $= (m, n)$ $(m,n) = (8,4)$ se	bias	$(m,n) = (16,4)$ se	bias	$(m,n) = (32,4)$ se	bias
Regular	$\hat{\theta}_1$	0.788	−0.337	0.756	−0.124	0.588	−0.102
	$\hat{\theta}_2$	0.784	0.418	0.362	0.170	0.211	0.088
	$\hat{\rho}$	-0.476		-0.537		-0.624	
Uniform	$\hat{\theta}_1$	0.812	−0.325	0.567	−0.267	0.502	−0.067
	$\hat{\theta}_2$	0.442	0.201	0.148	0.044	0.081	0.021
	$\hat{\rho}$	-0.455		-0.393		-0.584	
Beta(10,10)	$\hat{\theta}_1$	0.956	−0.212	0.847	−0.106	0.643	−0.069
	$\hat{\theta}_2$	0.265	0.073	0.125	0.035	0.069	0.021
	$\hat{\rho}$	-0.369		-0.591		-0.602	
Clustered	$\hat{\theta}_1$	0.868	−0.402	0.648	−0.191	0.473	−0.142
	$\hat{\theta}_2$	0.590	0.208	0.113	0.038	0.079	0.022
	$\hat{\rho}$	-0.295		-0.530		-0.487	

Spatial Parameter $(2, 1)$

Sampling Pattern		(Domain, Infill) $= (m, n)$ $(m,n) = (8,4)$ se	bias	$(m,n) = (16,4)$ se	bias	$(m,n) = (32,4)$ se	bias
Regular	$\hat{\theta}_1$	0.894	−0.365	0.768	−0.271	0.472	−0.127
	$\hat{\theta}_2$	0.469	0.237	0.313	0.137	0.142	0.034
	$\hat{\rho}$	-0.578		-0.570		-0.702	
Uniform	$\hat{\theta}_1$	0.932	−0.483	0.646	−0.265	0.428	−0.136
	$\hat{\theta}_2$	0.434	0.231	0.164	0.064	0.098	0.026
	$\hat{\rho}$	-0.509		-0.555		-0.663	
Beta(10,10)	$\hat{\theta}_1$	0.828	−0.365	0.685	−0.223	0.470	−0.103
	$\hat{\theta}_2$	0.198	0.111	0.119	0.028	0.071	0.013
	$\hat{\rho}$	-0.576		-0.669		-0.608	
Clustered	$\hat{\theta}_1$	0.755	−0.436	0.682	−0.198	0.516	−0.077
	$\hat{\theta}_2$	0.329	0.133	0.115	0.043	0.092	0.013
	$\hat{\rho}$	-0.587		-0.524		-0.659	

Table 5.3: Standard error (se), bias, and correlation for Matérn($2\sqrt{2}, 2$) and Matérn$(4, 2)$. Estimates are based on 100 independent realizations for a fixed sampling effort per block of $n = 4$. $\hat{\rho}$ is Pearson's sample correlation coefficient for $\hat{\theta}_1$ and $\hat{\theta}_2$.

Spatial Parameter $(2\sqrt{2}, 2)$

Sampling Pattern		(Domain, Infill) $= (m, n)$ $(m,n) = (8,4)$ se	bias	$(m,n) = (16,4)$ se	bias	$(m,n) = (32,4)$ se	bias
Regular	$\hat{\theta}_1$	0.946	−0.311	0.523	−0.346	0.455	−0.055
	$\hat{\theta}_2$	0.378	0.127	0.230	0.098	0.166	0.018
	$\hat{\rho}$	-0.758		-0.753		-0.802	
Uniform	$\hat{\theta}_1$	0.830	−0.405	0.599	−0.225	0.447	−0.138
	$\hat{\theta}_2$	0.306	0.103	0.183	0.045	0.129	0.013
	$\hat{\rho}$	-0.655		-0.652		-0.737	
Beta(10,10)	$\hat{\theta}_1$	0.782	−0.339	0.546	−0.306	0.404	−0.115
	$\hat{\theta}_2$	0.236	0.061	0.140	0.074	0.101	0.007
	$\hat{\rho}$	-0.705		-0.642		-0.755	
Clustered	$\hat{\theta}_1$	0.892	−0.346	0.581	−0.288	0.430	−0.108
	$\hat{\theta}_2$	0.270	0.091	0.146	0.054	0.088	0.014
	$\hat{\rho}$	-0.605		-0.604		-0.695	

Spatial Parameter $(4, 2)$

Sampling Pattern		(Domain, Infill) $= (m, n)$ $(m,n) = (8,4)$ se	bias	$(m,n) = (16,4)$ se	bias	$(m,n) = (32,4)$ se	bias
Regular	$\hat{\theta}_1$	1.12	−0.935	1.04	−0.378	0.686	−0.229
	$\hat{\theta}_2$	0.327	0.192	0.264	0.076	0.139	0.022
	$\hat{\rho}$	-0.716		-0.791		-0.667	
Uniform	$\hat{\theta}_1$	1.22	−0.764	0.863	−0.428	0.750	−0.240
	$\hat{\theta}_2$	0.275	0.108	0.175	0.049	0.119	0.031
	$\hat{\rho}$	-0.683		-0.704		-0.777	
Beta(10,10)	$\hat{\theta}_1$	0.949	−0.940	0.747	−0.582	0.609	−0.257
	$\hat{\theta}_2$	0.195	0.093	0.124	0.048	0.079	0.023
	$\hat{\rho}$	-0.566		-0.574		-0.601	
Clustered	$\hat{\theta}_1$	1.12	−0.694	0.887	−0.595	0.669	−0.112
	$\hat{\theta}_2$	0.242	0.085	0.140	0.043	0.080	0.003
	$\hat{\rho}$	-0.585		-0.711		-0.658	

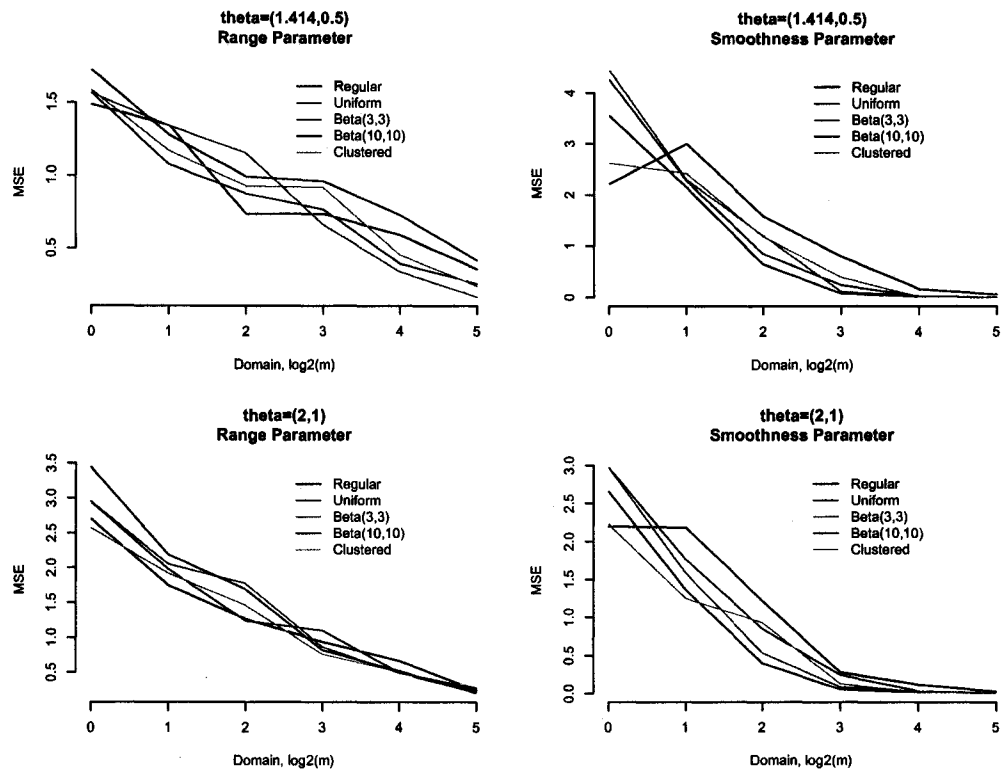

Figure 5.5: MSE of $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of domain for a constant sampling effort per block, $n = 4$, for $\Gamma = \text{Matérn}(\sqrt{2}, 1/2)$ and $\Gamma = \text{Matérn}(2, 1)$. The left panels plot the MSE of the range parameter estimator and the right panels plot the MSE of the smoothness parameter estimator.

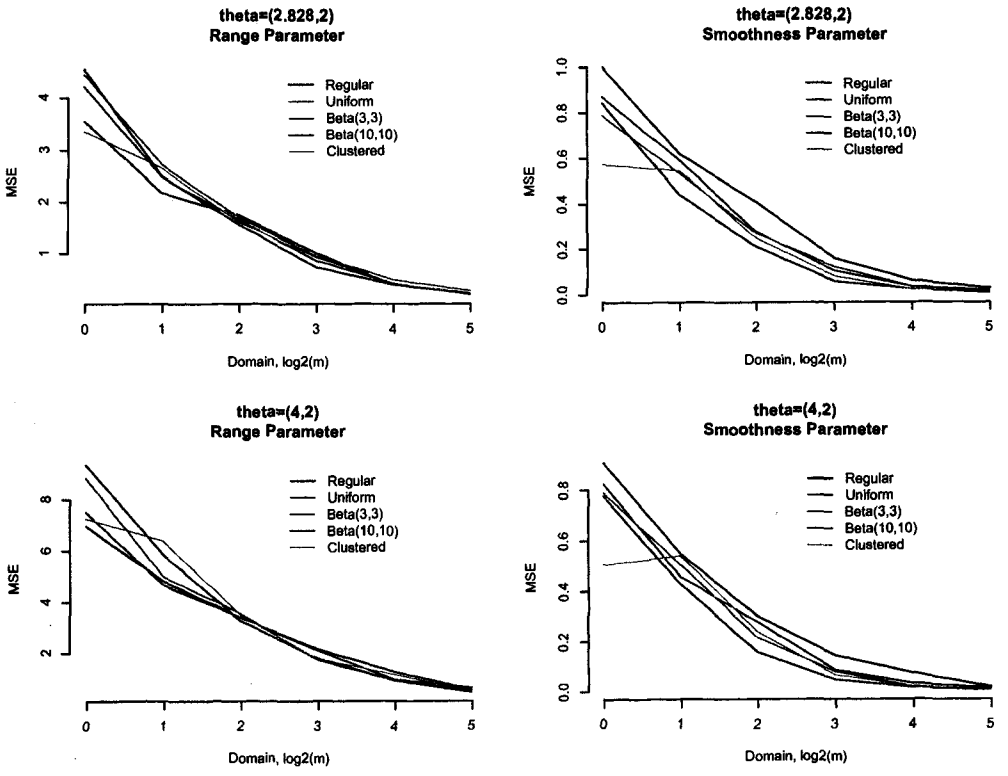

Figure 5.6: MSE of $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of domain for a constant sampling effort per block, $n = 4$, for $\mathbf{\Gamma} = \text{Matérn}(2\sqrt{2}, 2)$ and $\mathbf{\Gamma} = \text{Matérn}(4, 2)$. The left panels plot the MSE of the range parameter estimator and the right panels plot the MSE of the smoothness parameter estimator.

5.1.2.2 Distribution of the parameter estimates

In Section 5.1.2.1, we illustrated how MSE of the MLE varies with respect to the domain, level of infill, and sampling pattern. One important observation was the correlation between the parameter estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. The correlation was moderate for nearly all cases (typically of the order of 0.60) and always negative. Furthermore, the empirical biases, consistent with the sample correlation coefficient, indicate that the range parameter is typically underestimated and the smoothness parameter overestimated. So what impact do these results have on the estimation of the correlation function?

Note that the sample correlations reported in Tables 5.2 and 5.3 underestimate the true correlation. Pearson's correlation coefficient is a measure of linear correlation. Figures 5.7 through 5.10 plot the MLEs for $\boldsymbol{\theta}$ for all 100 realizations for $\Gamma = \text{Matérn}(2, 1)$ and a fixed sampling density of four observations per unit length. For the smaller domains, $m = 4$ and $m = 8$, the overall shape of the MLEs is that of a severely bent "banana", independent of sampling pattern. Thus, the correlation estimates for the small domains will, on average, be low. As the domain is expanded the cluster "straightens out" to an oblong shape with a principal (major) axis nearly parallel to the range parameter estimate axis. Note how the variability of the MLEs with respect to the smoothness parameter estimate is reduced more quickly for the "cluster" patterns. This is consistent with the observed standard errors listed in Tables 5.2 and 5.3. Although the variability of the MLEs with respect to $\hat{\theta}_1$ also decrease as the domain is expanded, no particular sampling pattern appears to be more efficient. The negative bias associated with $\hat{\theta}_1$ is clearly visible for the moderate and large domains as the center of the distribution of the MLEs is clearly to the left (less than) of the true parameter value.

Clearly the joint distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ is not bivariate normal for small domains. However as the domain is expanded the distribution of the MLEs appears

to form a more regular, albeit oblong, cluster. The heuristic proof of the spatial AIC statistic assumed that the parameter estimates are asymptotically normal with mean zero and covariance matrix equivalent to the Fisher Information. Referring to Figure 5.10, the distributions of the MLEs are nearly symmetric about the true parameter value. The beta(10,10) pattern appear to have a larger right tail (positively skewed) with respect to $\hat{\theta}_1$. We performed the Anderson-Darling test [*** Lookup AD reference ***] on the observed distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ for each combination of $\boldsymbol{\theta}$, sampling effort (m, n), and sampling pattern. The null hypothesis associated with the Anderson-Darling test is that the distribution is normal.Figures 5.11 through 5.14 summarize the results where blue indicates that there is sufficient evidence (at the 0.05 significance level) to conclude that the distribution is not normal. Red squares indicate that the null hypothesis was not rejected. The left columns correspond to the range parameter estimates and the right to the smoothness parameter estimates.

For the range parameter estimate, typically the domain must be large and the true parameter value, θ_1, must be relatively large (> 2) for the approximate distribution of $\hat{\theta}_1$ to approach normality. In contrast, the domain need not be large for the smoothness parameter estimate to approach normality but the level of infill must be reasonably large (generally at least 4 observations per unit length). No sampling pattern appears to reach normality more quickly than any other although there is mild evidence suggesting that the clustered pattern induces normality jointly more often.

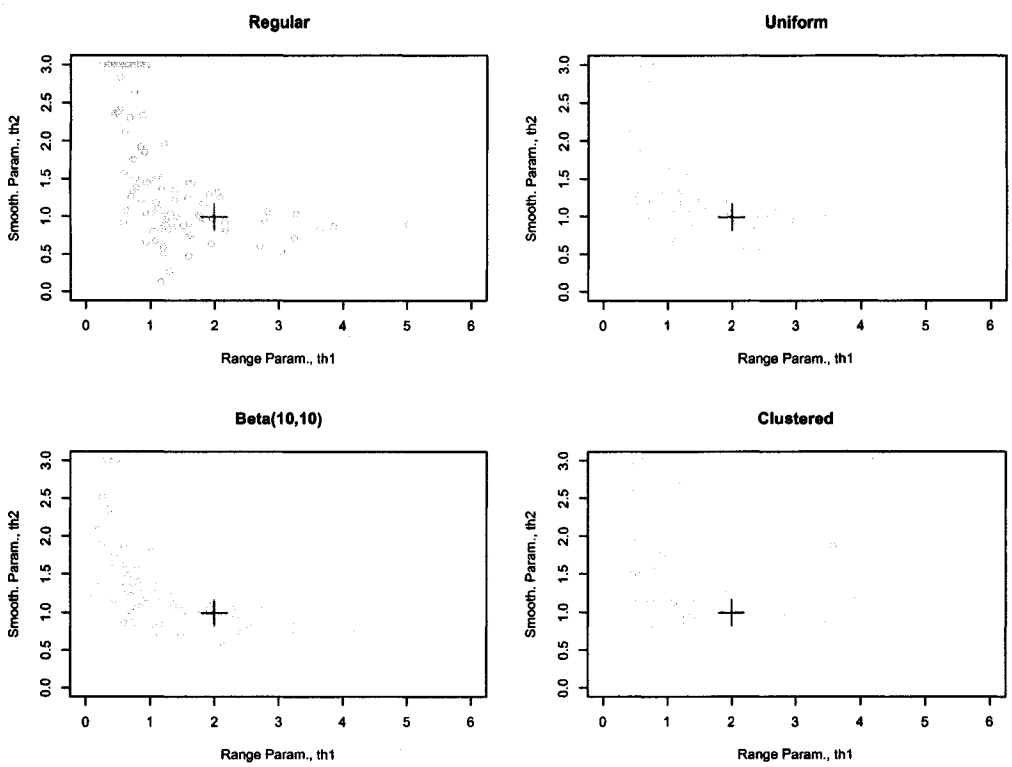

Figure 5.7: Plot of the MLE $(\hat{\theta}_1, \hat{\theta}_2)$ as a function of sampling method for $(m, n) = (4, 4)$ where $\boldsymbol{\Gamma} = \text{Matérn}(2, 1)$. The true value of $\boldsymbol{\theta}$ is indicated by $(+)$.

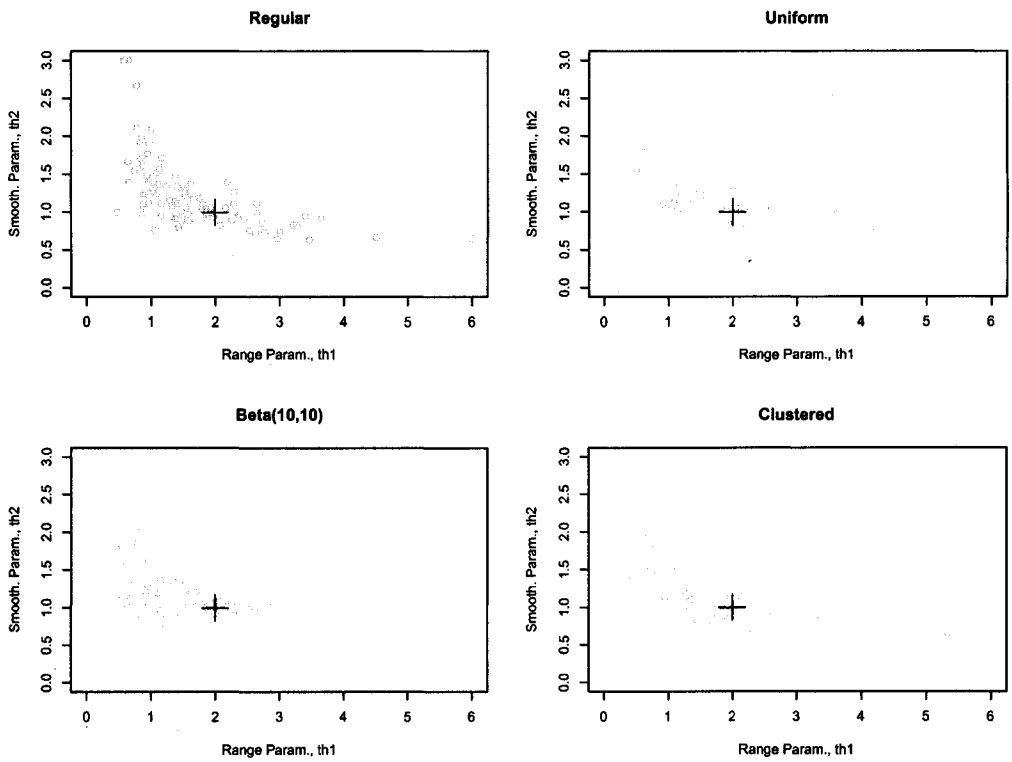

Figure 5.8: Plot of the MLE $(\hat{\theta}_1, \hat{\theta}_2)$ as a function of sampling method for $(m, n) = (8, 4)$ where $\mathbf{\Gamma} = \text{Matérn}(2, 1)$. The true value of $\boldsymbol{\theta}$ is indicated by (+).

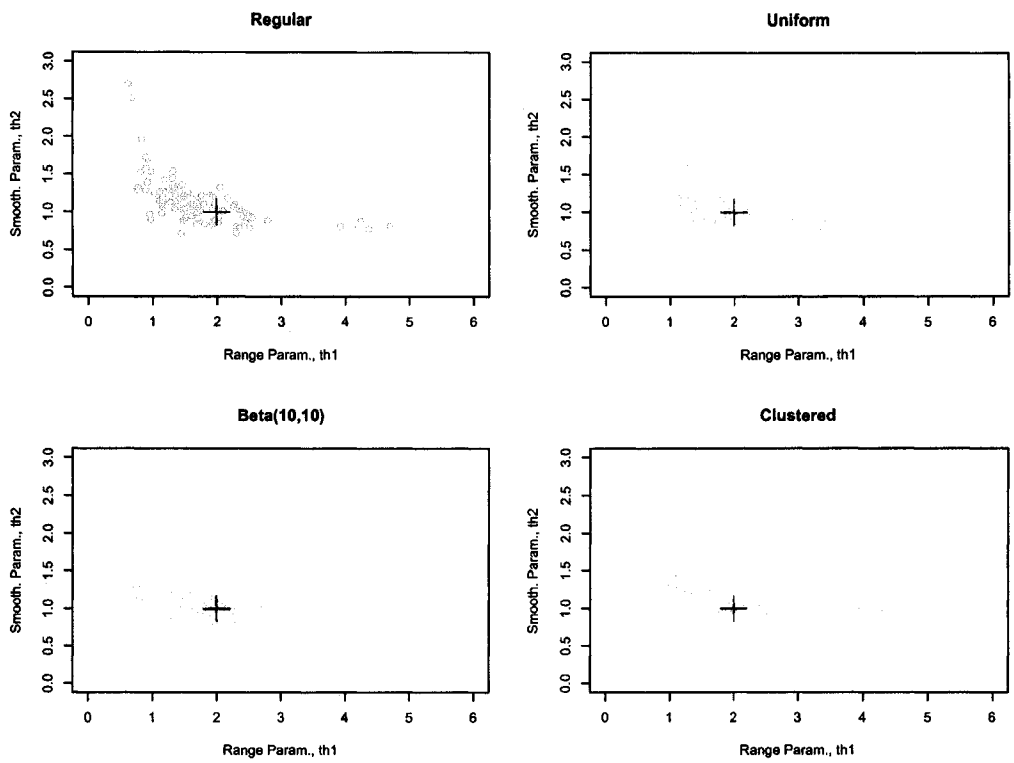

Figure 5.9: Plot of the MLE $(\hat{\theta}_1, \hat{\theta}_2)$ as a function of sampling method for $(m, n) = (16, 4)$ where $\mathbf{\Gamma} = \text{Matérn}(2, 1)$. The true value of $\boldsymbol{\theta}$ is indicated by $(+)$.

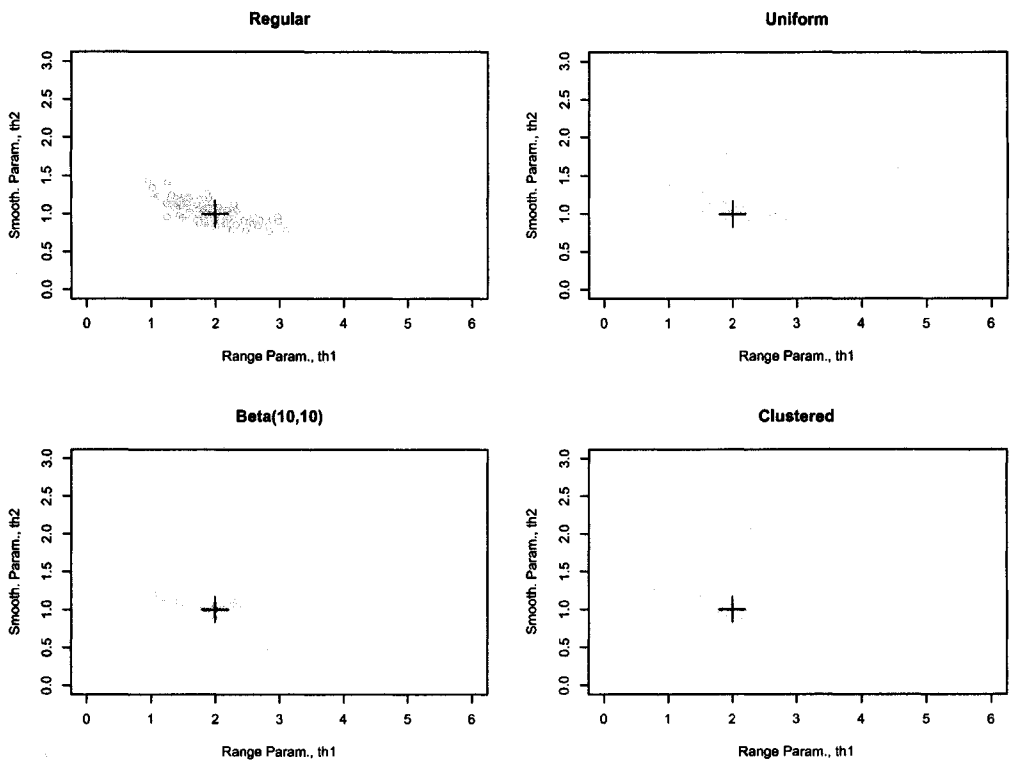

Figure 5.10: Plot of the MLE $(\hat{\theta}_1, \hat{\theta}_2)$ as a function of sampling method for $(m, n) = (32, 4)$ where $\mathbf{\Gamma} = \text{Matérn}(2, 1)$. The true value of $\boldsymbol{\theta}$ is indicated by $(+)$.

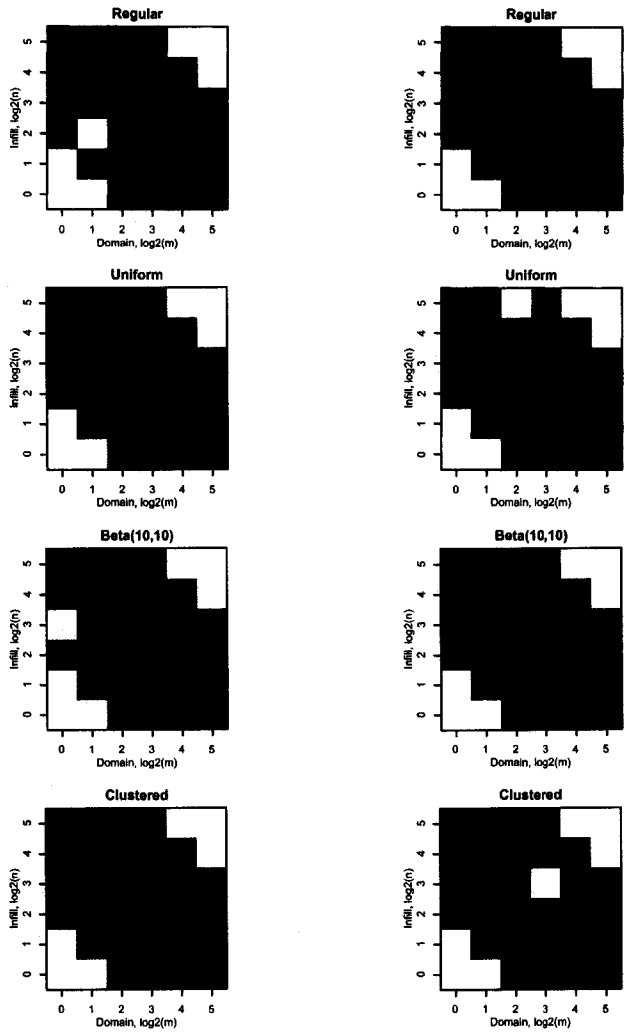

Figure 5.11: Normality testing results where $\Gamma = \text{Matérn}(\sqrt{2}, 1)$. The Anderson-Darling test was performed individually on the observed distribution of each parameter estimate for each combination of sampling effort (m, n) and sampling pattern. The null hypothesis is that the distribution is normal. Blue squares correspond to non-normal distributions and red squares to normal distributions.

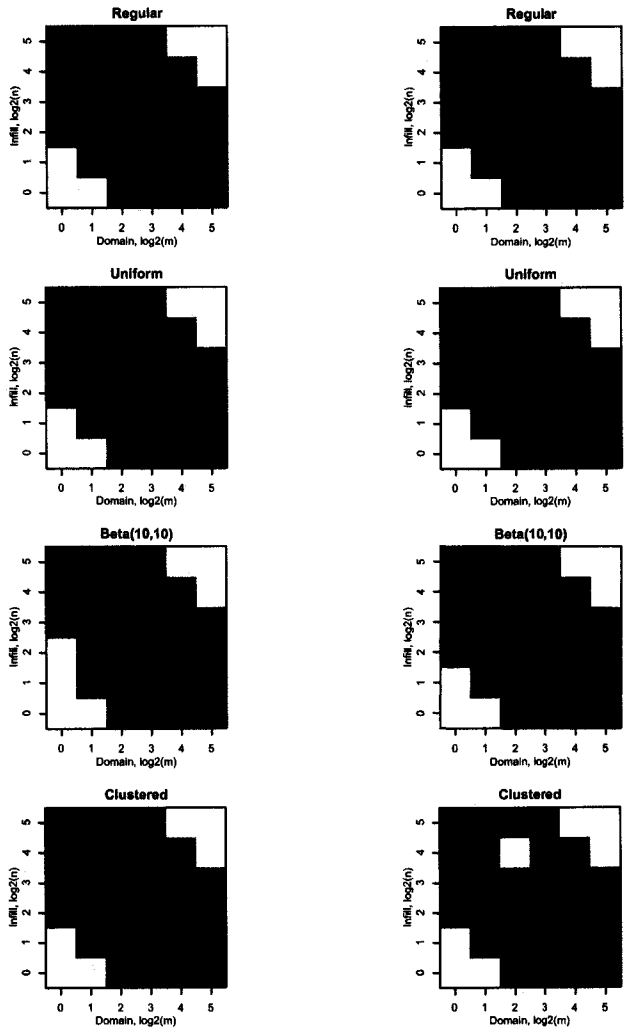

Figure 5.12: Normality testing results where $\mathbf{\Gamma} = \text{Matérn}(2, 1)$. The Anderson-Darling test was performed individually on the observed distribution of each parameter estimate for each combination of sampling effort (m, n) and sampling pattern. The null hypothesis is that the distribution is normal. Blue squares correspond to non-normal distributions and red squares to normal distributions.

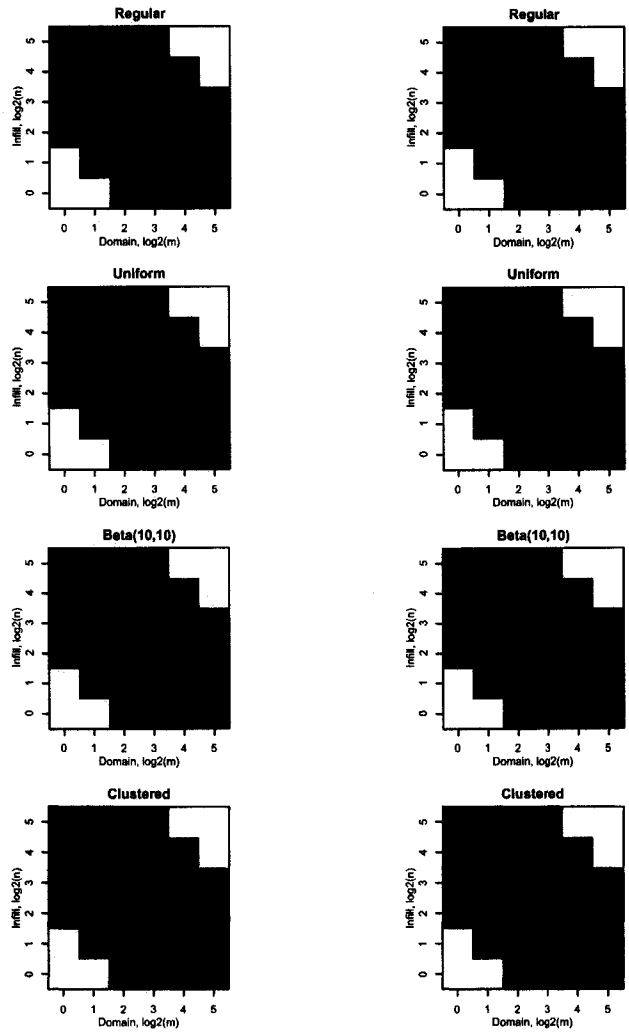

Figure 5.13: Normality testing results where $\Gamma = \text{Matérn}(2\sqrt{2}, 2)$. The Anderson-Darling test was performed individually on the observed distribution of each parameter estimate for each combination of sampling effort (m, n) and sampling pattern. The null hypothesis is that the distribution is normal. Blue squares correspond to non-normal distributions and red squares to normal distributions.

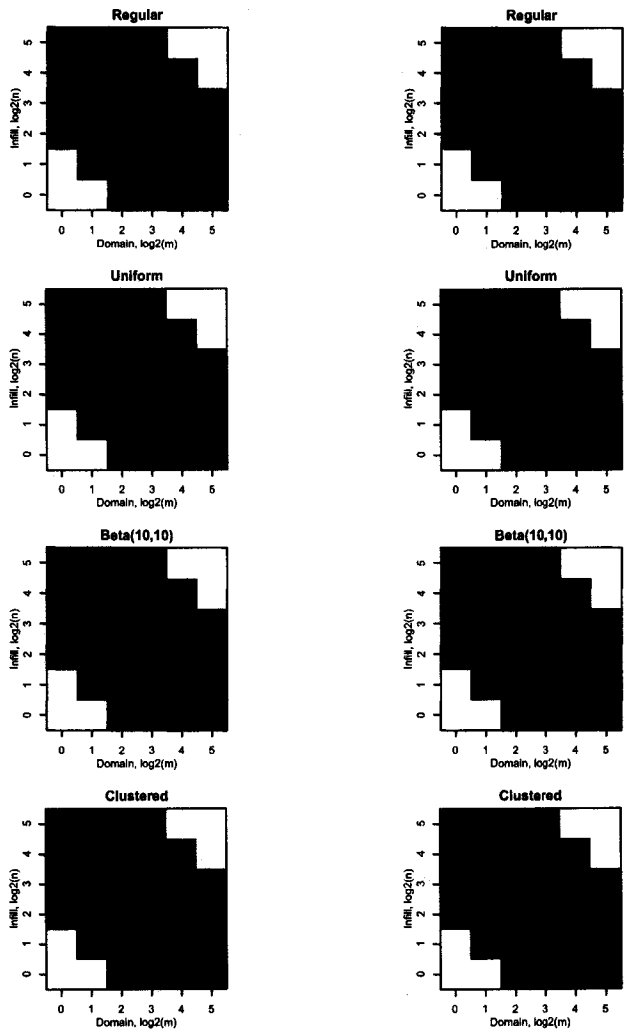

Figure 5.14: Normality testing results where $\mathbf{\Gamma}$ = Matérn$(4, 2)$. The Anderson-Darling test was performed individually on the observed distribution of each parameter estimate for each combination of sampling effort (m, n) and sampling pattern. The null hypothesis is that the distribution is normal. Blue squares correspond to non-normal distributions and red squares to normal distributions.

5.1.2.3 Estimated correlation function

Figures 5.15 through 5.18 plot the estimated correlation functions (light blue) using the MLEs of (θ_1, θ_2) for $(m, n) = (16, 16)$. Superimposed on the plots are the true correlation function (black)and the empirical 5th, 50th (median), and 95th percentiles (blue)Note that the plotted percentiles do not necessarily correspond to the 5th, 50th, and 95th percentiles of the parameter estimates. Instead the percentiles correspond to the values of the estimated correlation functions at each distance. For example, for the distance of 2 units, the value of each estimated correlation function was determined and the percentiles computed. This was done for each distance along a fine grid and the results plotted.

For all cases the median correlation function underestimates the true correlation. Recall that the bias for the range parameter is negative and the bias for the smoothness parameter is positive. Although these two biases appear to work against one another the overall effect is to underestimate the correlation function at all distances. In other words, underestimation of the range parameter cannot be fully overcome by overestimating the smoothness parameter. Furthermore, the positive bias associated with the smoothness parameter estimate implies that sampling locations at close proximity are estimated to be more strongly correlated than the "true" model dictates.

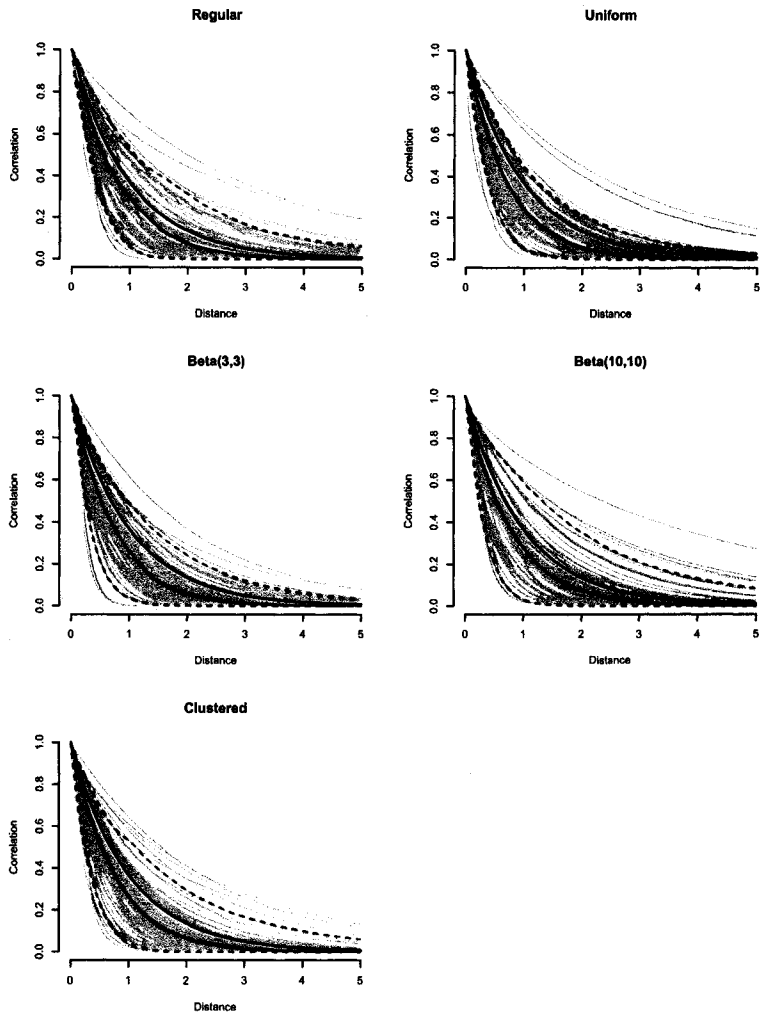

Figure 5.15: Fitted correlation functions for the Matérn($\sqrt{2}, 1/2$) where $(m, n) = (16, 4)$. Each light blue line corresponds to a single realization. The dark blue lines are the median (solid) and the 5th and 95th percentiles (dashed). The solid black line is the true correlation function.

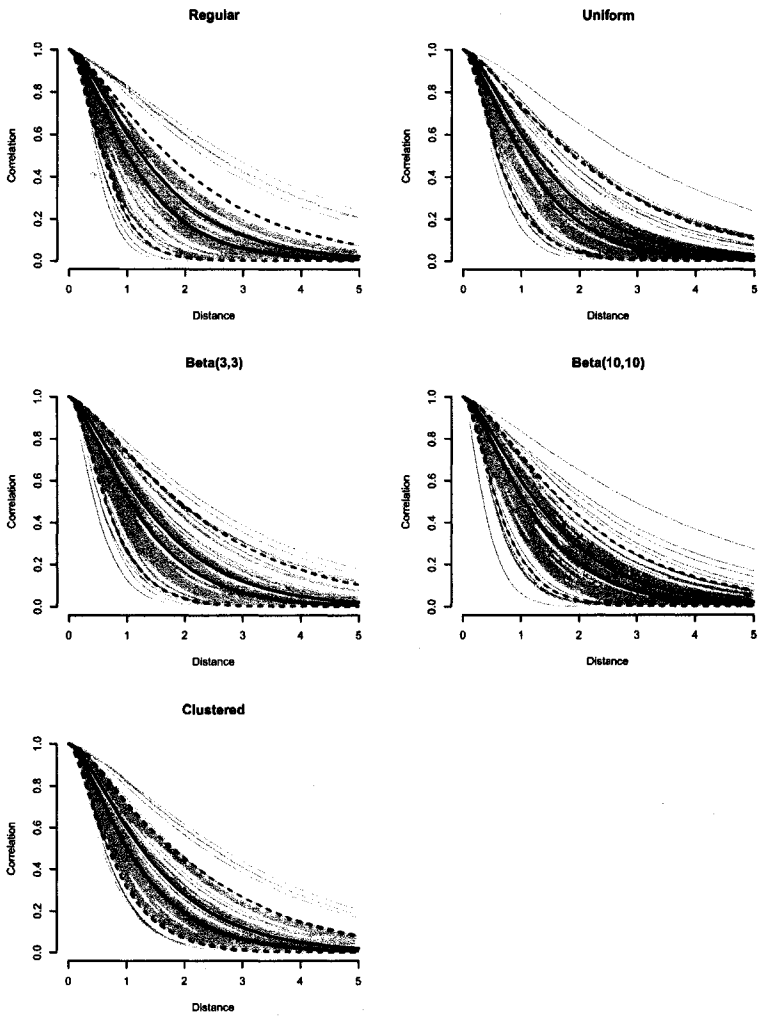

Figure 5.16: Fitted correlation functions for the Matérn$(2,1)$ where $(m,n) = (16,4)$. Each light blue line corresponds to a single realization. The dark blue lines are the median (solid) and the 5th and 95th percentiles (dashed). The solid black line is the true correlation function.

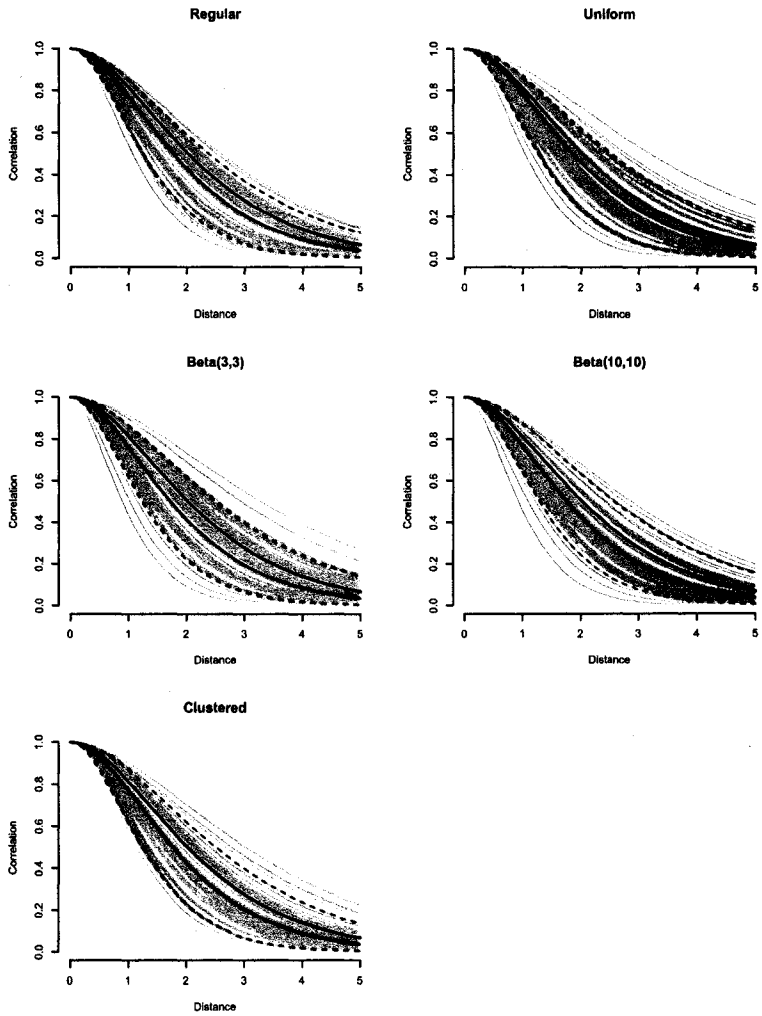

Figure 5.17: Fitted correlation functions for the Matérn$(2\sqrt{2}, 2)$ where $(m, n) = (16, 4)$. Each light blue line corresponds to a single realization. The dark blue lines are the median (solid) and the 5th and 95th percentiles (dashed). The solid black line is the true correlation function.

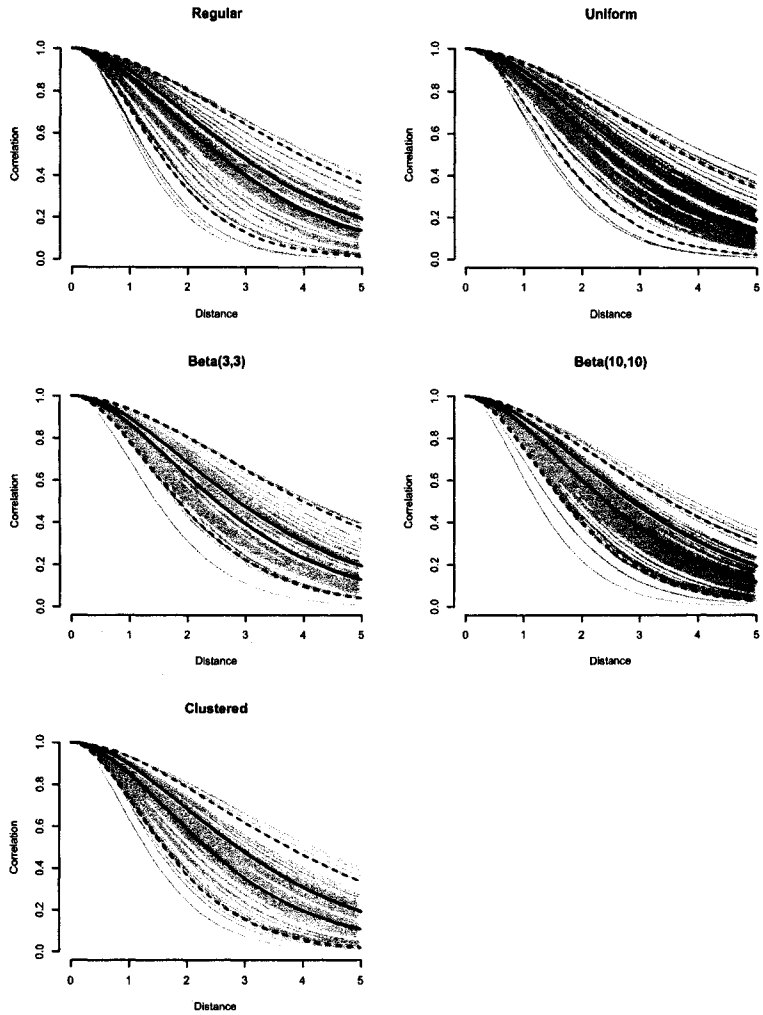

Figure 5.18: Fitted correlation functions for the Matérn$(4, 2)$ where $(m, n) = (16, 4)$. Each light blue line corresponds to a single realization. The dark blue lines are the median (solid) and the 5th and 95th percentiles (dashed). The solid black line is the true correlation function.

5.1.2.4 Mean integrated square error

To compare performance amongst the different sampling designs we adopt the mean integrated square error (MISE). Recall that the function $\rho(\cdot;\boldsymbol{\theta})$ defines the correlation between locations u and v as a function of the Euclidean distance between them, i.e., $d = |u - v|$. Substituting the MLE for $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$ one can plot the estimated correlation function with respect to distance. Integrating the square error between the estimated and known correlation functions over the positive real line yields a positive measure that will be zero if the estimated correlation function exactly coincides with the "truth". We refer to this statistic as the integrated square error (ISE) (Givens and Hoeting, 2005) written as

$$\text{ISE} = \int_0^\infty \left(\hat{\rho}(t;\hat{\boldsymbol{\theta}}) - \rho(t;\boldsymbol{\theta})\right)^2 dt, \tag{5.2}$$

where $\rho(\cdot;\theta) = \text{Matérn}(\theta_1, \theta_2)$. We define the mean integrated square error as the empirical mean of the ISE statistic for r realizations such that

$$\text{MISE} = \frac{1}{r}\sum_{i=1}^{r} \text{ISE}_i.$$

Figures 5.19 through 5.22 illustrate the observed MISE as a function of sampling pattern. Superimposed on the image plots are the approximated contours to better highlight the gradients. Note that the MISE surfaces are very similar to the MSE surfaces for the exponential correlation function simulations of Chapter 3. For example, compare the upper left panels of Figures 5.1 and 5.19. Although the magnitudes differ the direction of the contours are similar. Subtle differences are visible for meager levels of infill. Implied is that the (relatively) large MSE of $\hat{\theta}_2$ for small n has a modest impact on the estimated correlation function. This is consistent with the observations first presented in Section 5.1.2.3. The direction of the contours suggest that the first priority is to sample over a large domain and then incorporate infill second. There is modest evidence indicating that the clustered sampling pattern outperforms the remaining sampling patterns, especially for small sampling efforts.

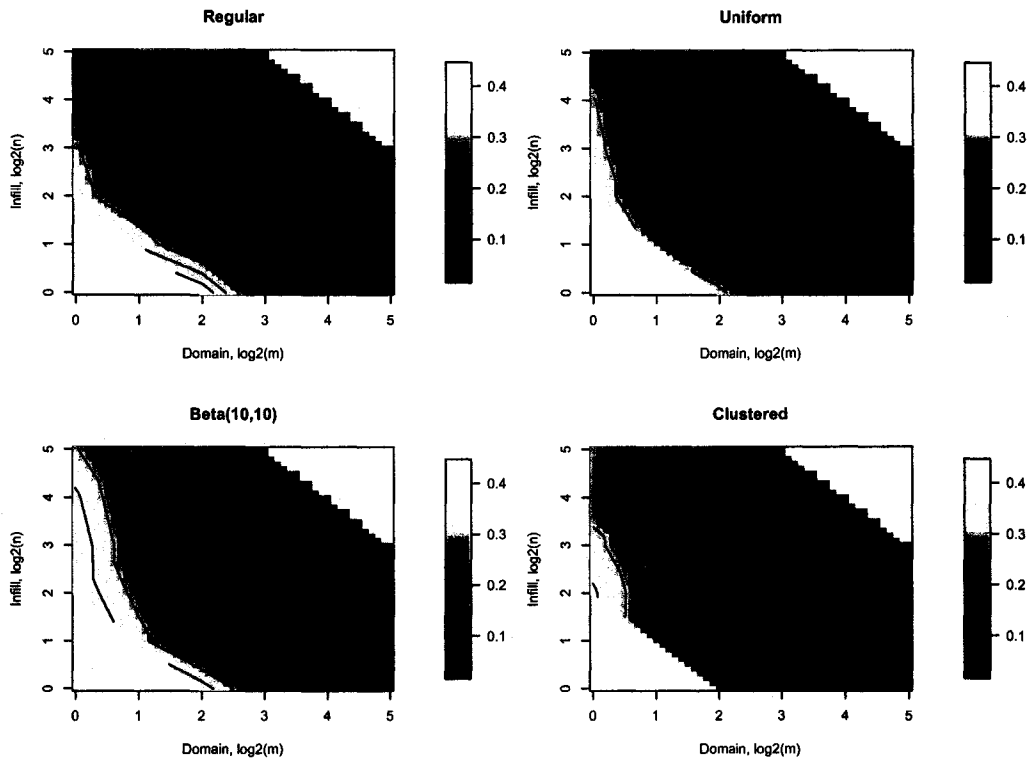

Figure 5.19: Mean integrated square error (MISE) maps for $\Gamma = \text{Matérn}(\sqrt{2}, 1/2)$ for each sampling pattern. Note the temperature (color) scale for MISE is the same for all four panels and that the scales for the x- and y-axes are $\log_2$.

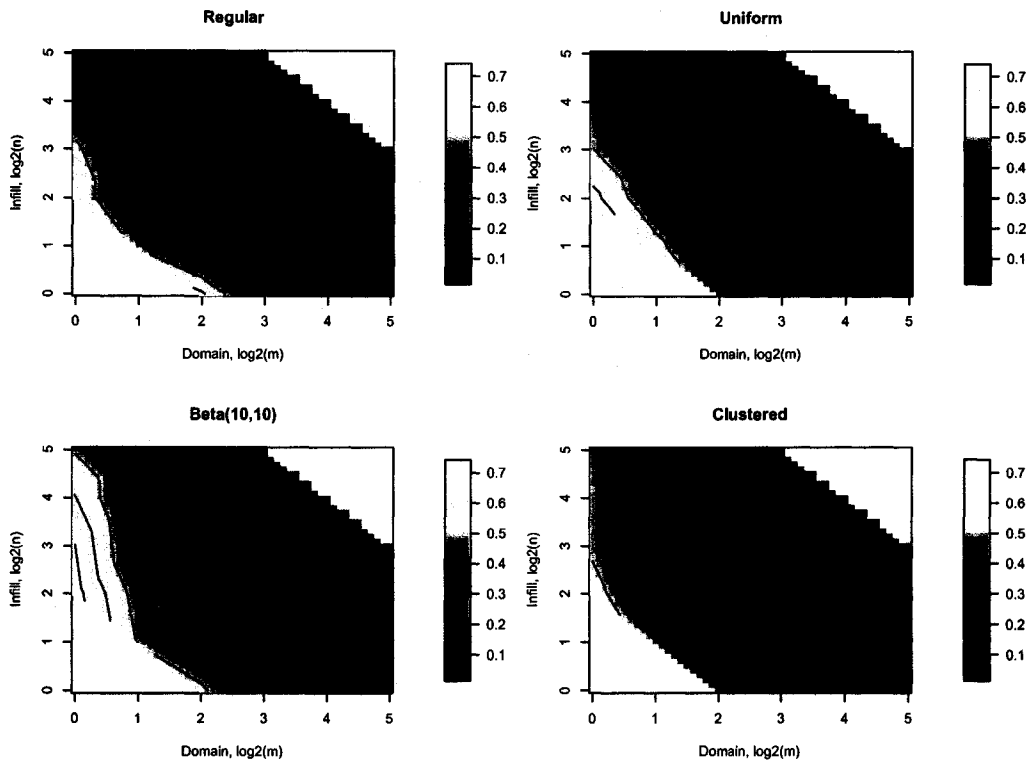

Figure 5.20: Mean integrated square error (MISE) maps for $\Gamma = \text{Matérn}(2, 1)$ for each sampling pattern. Note the temperature (color) scale for MISE is the same for all four panels and that the scales for the x- and y-axes are linear in m and $\sqrt{N}$, respectively.

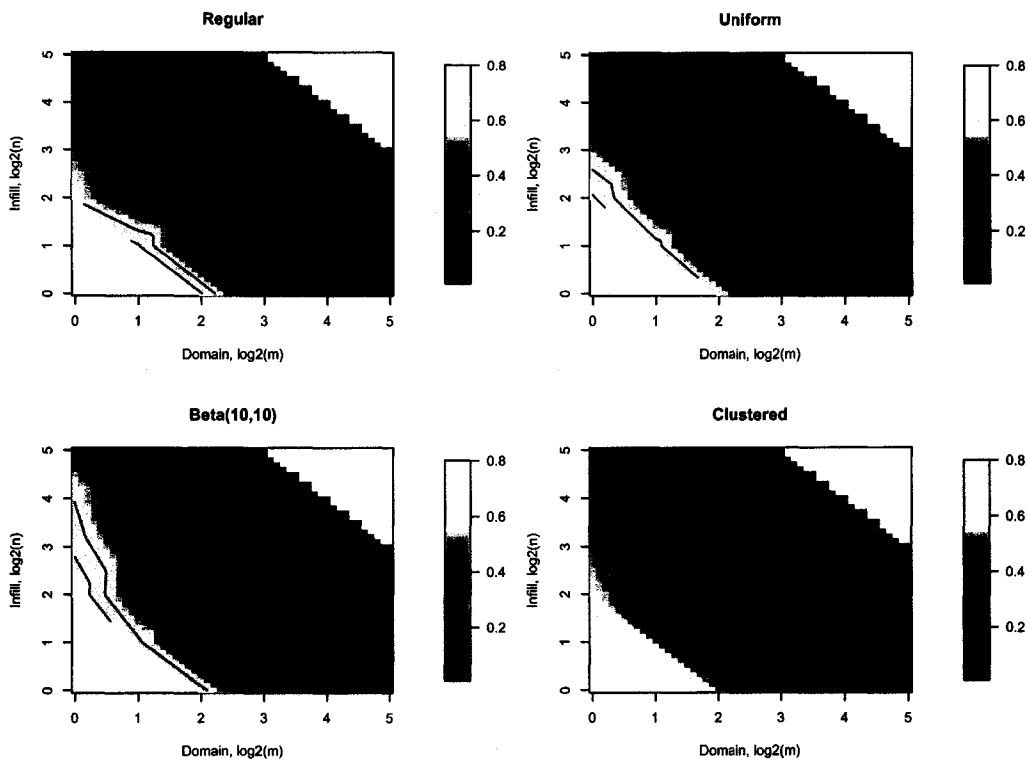

Figure 5.21: Mean integrated square error (MISE) maps for $\mathbf{\Gamma} = \text{Matérn}(2\sqrt{2}, 2)$ for each sampling pattern. Note the temperature (color) scale for MISE is the same for all four panels and that the scales for the x- and y-axes are linear in m and $\sqrt{N}$, respectively.

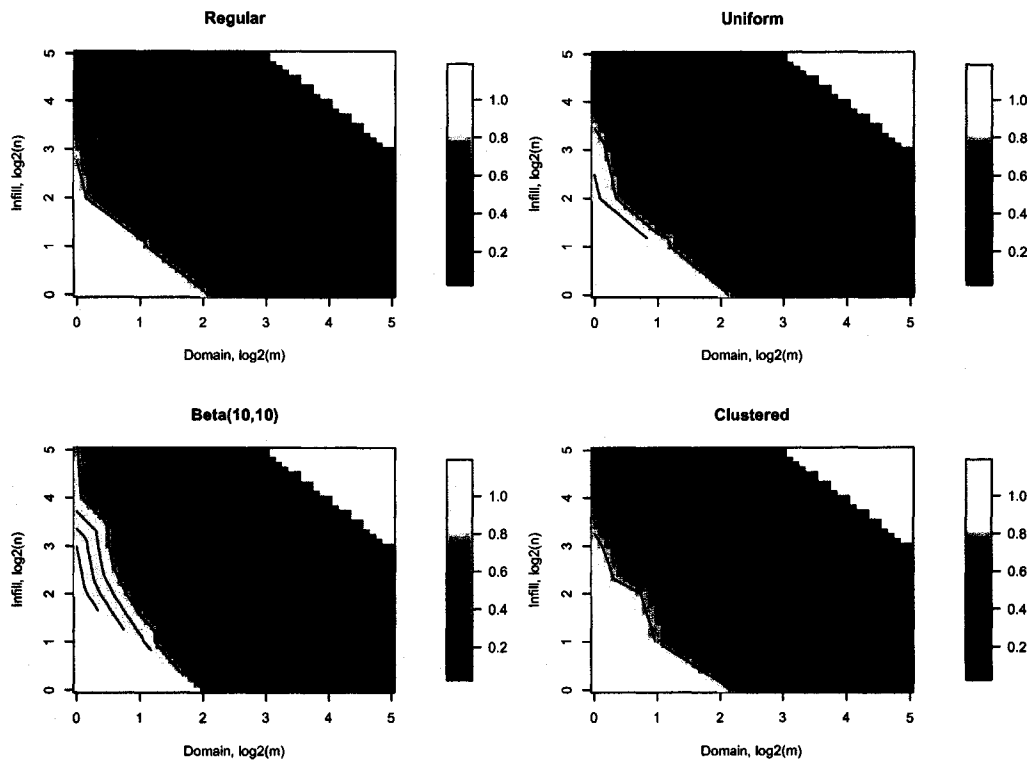

Figure 5.22: Mean integrated square error (MISE) maps for $\Gamma = \text{Matérn}(4, 2)$ for each sampling pattern. Note the temperature (color) scale for MISE is the same for all four panels and that the scales for the x- and y-axes are linear in m and $\sqrt{N}$, respectively.

5.1.3 Two-dimensional analysis

We extended the simulation study for the Matérn class to two-dimensions in the same manner as was done for the exponential correlation function. As described in Section 4.1, we generated sampling locations over square domains with sides of length $m = \{1, 2, \ldots, 8\}$. All sampling locations were constrained to a fine lattice to maintain numerical stability for the optimizer. The same four sampling patterns used for the exponential correlation function in two-dimensions were used for the Matérn class: regular, uniform, light and heavy clusters. Four spatial parameter vectors were employed to generate the realizations. The first three are identical to the ones used for the one-dimensional analysis, i.e., $\boldsymbol{\theta}_1 = (\sqrt{2}, 1/2)'$, $\boldsymbol{\theta}_2 = (2, 1)'$, and $\boldsymbol{\theta}_3 = (2\sqrt{2}, 2)'$. Recall that the effective range of each is 3.0, 4.0, and 5.4, respectively. Since the maximum area under study is 8×8 it was concluded that the fourth parameter vector, $\boldsymbol{\theta}_4 = (4, 2)$, which has an effective range of approximately 7.4, would not allow for proper expanding domain analysis, i.e., very few locations would be pairwise (approximately) independent, we replaced $\boldsymbol{\theta}_4$ with $\boldsymbol{\theta}_5 = (1/\sqrt{2}, 1)$ which has an effective range of approximately 1.5. What follows are selected results that mirror the presentation of Section 5.1.2 for the exponential covariance function.

5.1.3.1 Mean square error

Figures 5.25 through 5.30 illustrate the observed MSE for both the range and smoothness parameter estimates for each of the examined cases. Note that the axes follow the same format presented for the exponential class in two-dimensions, i.e., the x-axis is the domain $m \times m$ and is linear with respect to m and the y-axis lists the total sampling effort N and is linear with respect to $\sqrt{N}$. The first three cases have similar MSE surfaces for both the range and smoothness parameter estimates. For small domains the MSE is (relatively) large for the range parameter as well as for small sampling efforts. As the density of the sampling locations decreases, the MSE

of $\hat{\theta}_2$ increases. It is most dramatic for the regular sampling pattern where sampling locations are separated by design. Hence estimates for the smoothness parameter suffer. For the remaining case, $\boldsymbol{\theta}_5$, the MSE surfaces are quite different. Recall that the effective range for this case is approximately 1.5 units, thus moderate sized domains will contain numerous sampling locations that are pairwise (approximately) independent. For the cluster patterns, light and heavy, the MSE of $\hat{\theta}_1$ is small even for very small sampling densities. In contrast, for the regular pattern the MSE of $\hat{\theta}_1$ is large for moderate sized domains and small sampling efforts. Note that for a domain of 6×6 and sampling effort of $N = 36$ the sampling locations are no less than one unit length apart. Consequently the MSE for the smoothness parameter estimate is large for all small sampling densities. This effect is diminished for the uniform and clustered patterns because each of these sampling methods have a mechanism to place some or many sampling locations near one another. It should be noted that for the regular pattern there appears to be a ridge on the MSE surface with respect to $\hat{\theta}_2$. This suggests that there may be a least optimal sampling strategy.

The MSE is decomposed in Tables 5.4 and 5.5 for a sampling density of $N/(m \times m) = 4$ sampling locations per unit area. Once again the standard error of the estimators dominates the magnitude of the bias as the domain is expanded, i.e., the magnitude of the bias term decreases more rapidly. Also, the estimators for the range and smoothness parameters are moderately (negatively) correlated. These results parallel the one-dimensional case.

Figures 5.31 and 5.32 plot the profile of the MSE surfaces along the contour of a constant sampling density of 4. (Recall that the contours of constant sampling densities are straight lines with positive slope equal to the square root of the density.) For $\hat{\theta}_1$ the four sampling patterns perform approximately equally well with respect to MSE. However the regular pattern performs poorly over small domains for the smoothness parameter estimator, $\hat{\theta}_2$. Unsurprisingly, having no observations at close

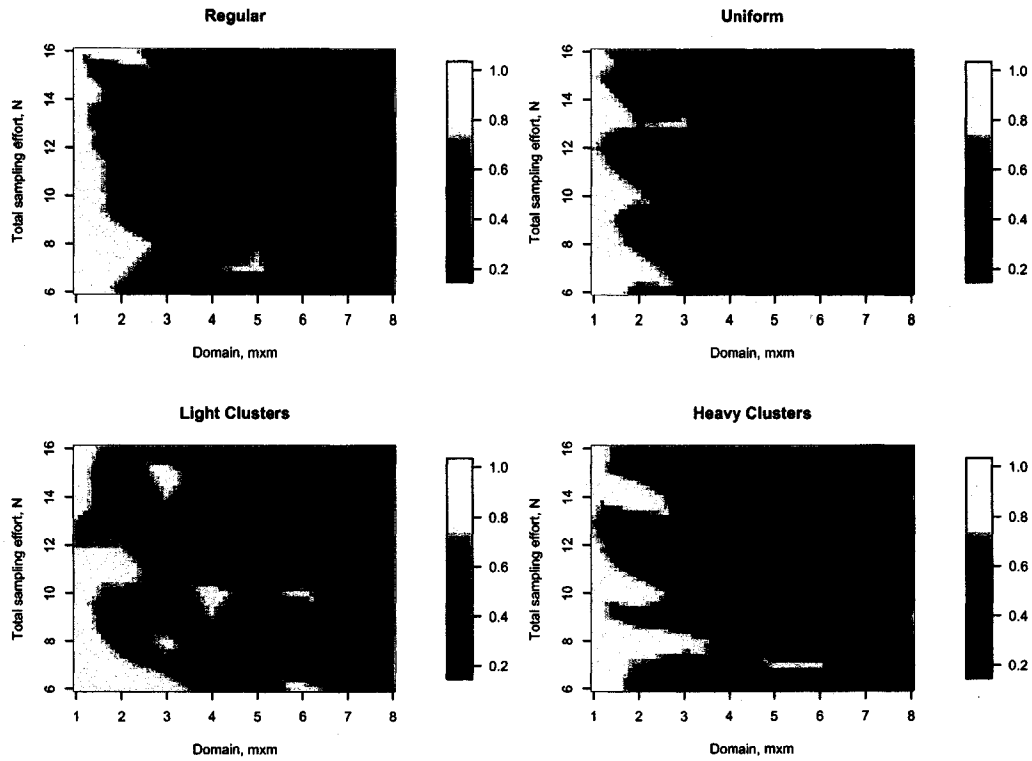

Figure 5.23: MSE of $\hat{\theta}_1$ where $\Gamma = \text{Matérn}(\sqrt{2}, 1/2)$ for each sampling pattern.

proximity the regular pattern must contain a large sample to significantly impact the MSE.

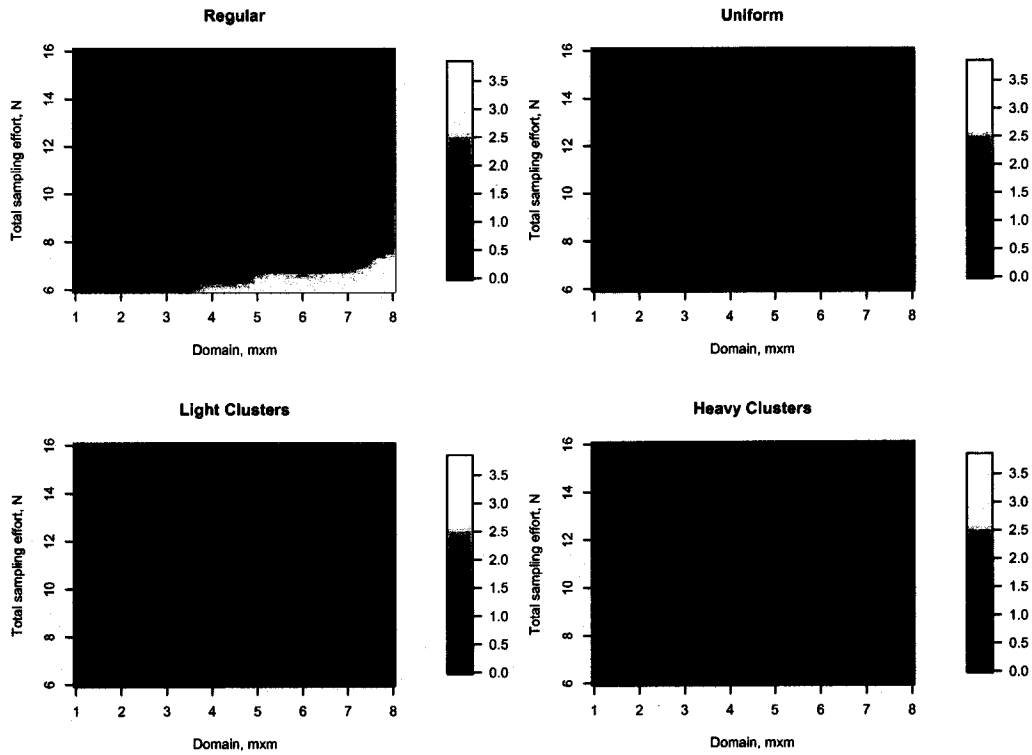

Figure 5.24: MSE of $\hat{\theta}_2$ where $\mathbf{\Gamma} = \text{Matérn}(\sqrt{2}, 1/2)$ for each sampling pattern.

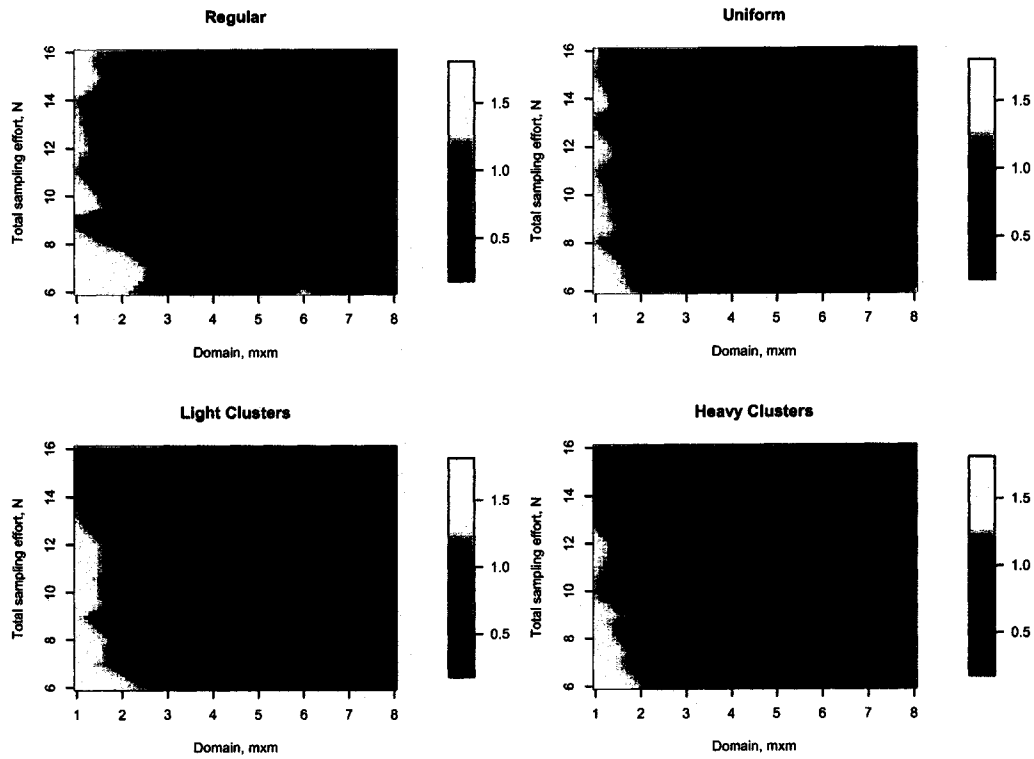

Figure 5.25: MSE for $\hat{\theta}_1$ where Γ = Matérn$(2, 1)$ for each sampling pattern. Note that the scales of the x- and y-axes are linear with respect to m and $\sqrt{N}$, respectively.

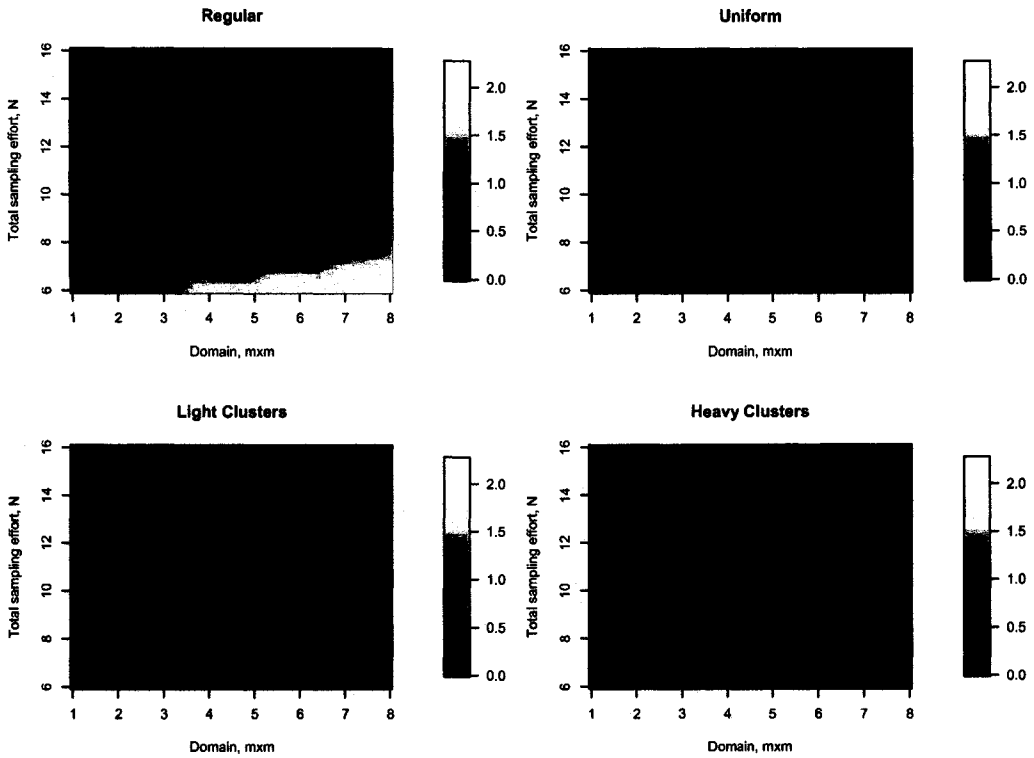

Figure 5.26: MSE of $\hat{\theta}_2$ where $\Gamma = \text{Matérn}(2, 1)$ for each sampling pattern. Note that the scales of the x- and y-axes are linear with respect to m and $\sqrt{N}$, respectively.

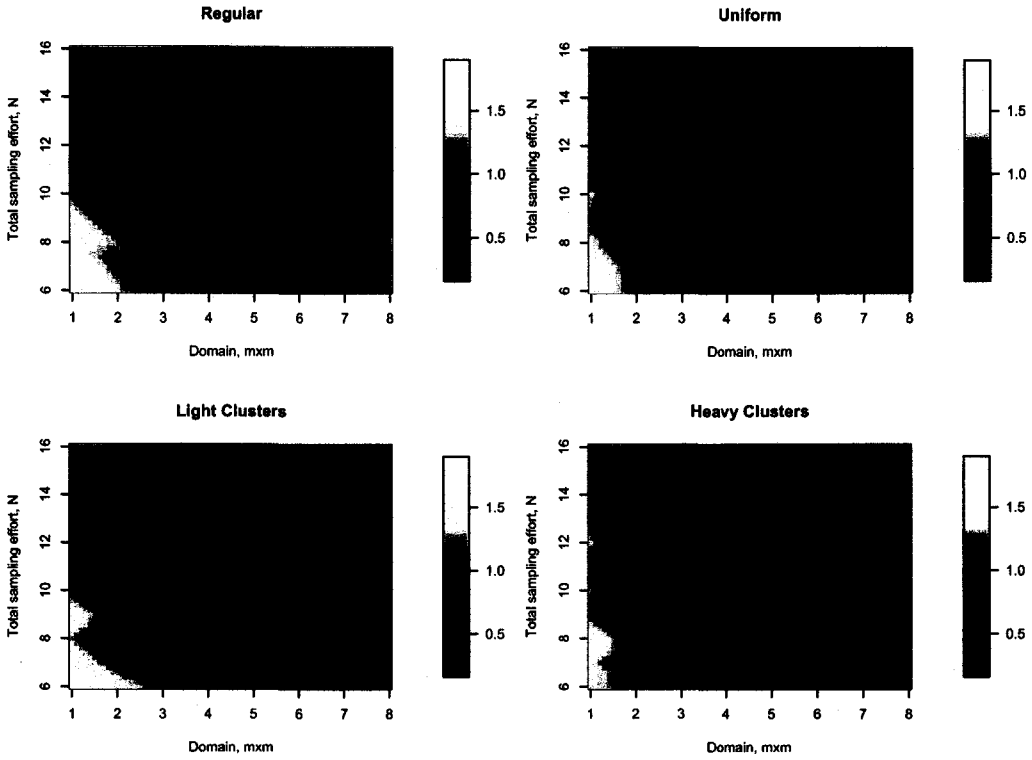

Figure 5.27: MSE of $\hat{\theta}_1$ where $\Gamma = \text{Matérn}(2\sqrt{2}, 2)$ for each sampling pattern. Note that the scales of the x- and y-axes are linear with respect to m and $\sqrt{N}$, respectively.

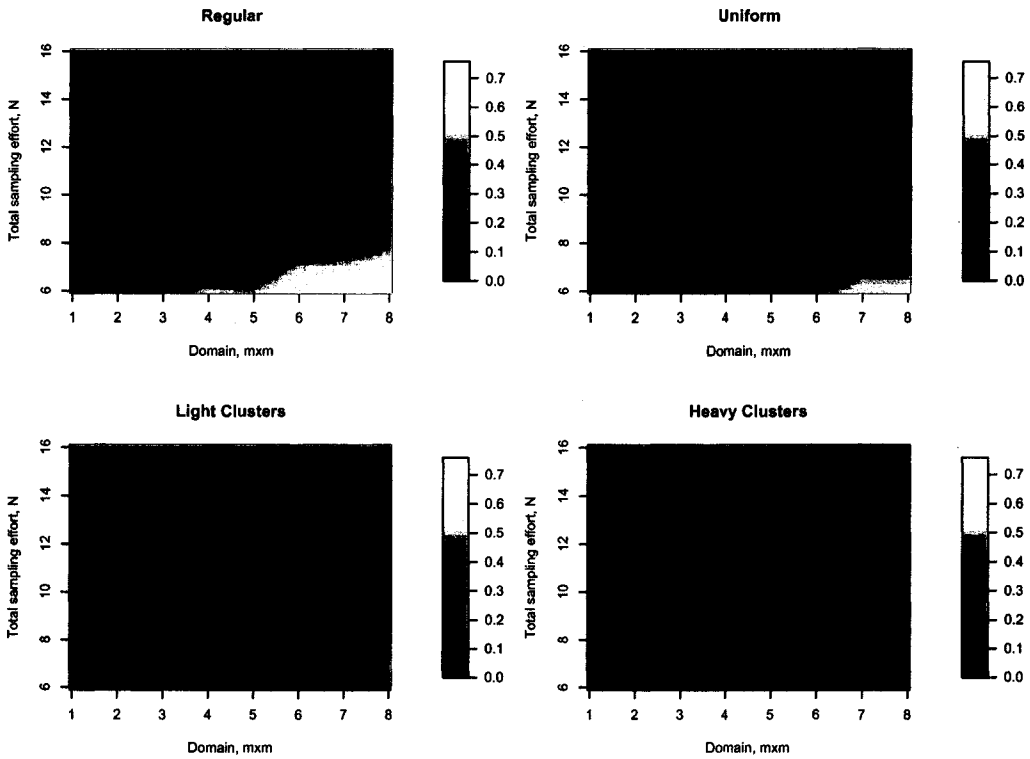

Figure 5.28: MSE of $\hat{\theta}_2$ where $\mathbf{\Gamma} = \text{Matérn}(2\sqrt{2}, 2)$ for each sampling pattern. Note that the scales of the x- and y-axes are linear with respect to m and $\sqrt{N}$, respectively.

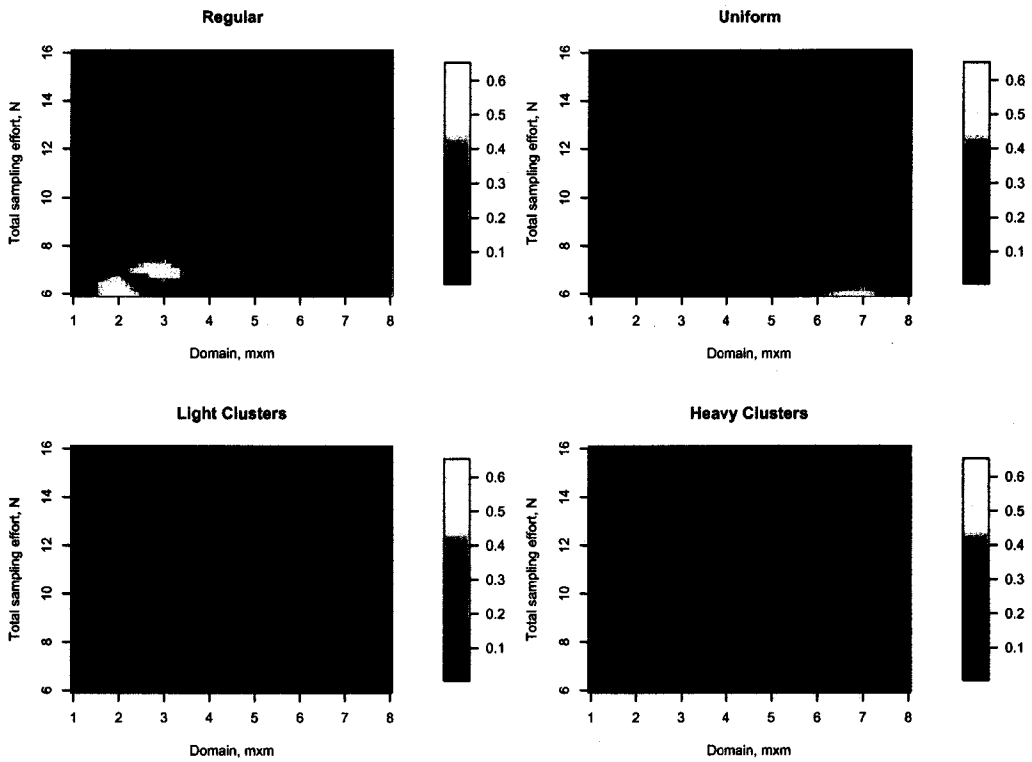

Figure 5.29: MSE of $\hat{\theta}_1$ where $\Gamma = \text{Matérn}(1/\sqrt{2}, 1)$ for each sampling pattern. Note that the scales of the x- and y-axes are linear with respect to m and $\sqrt{N}$, respectively.

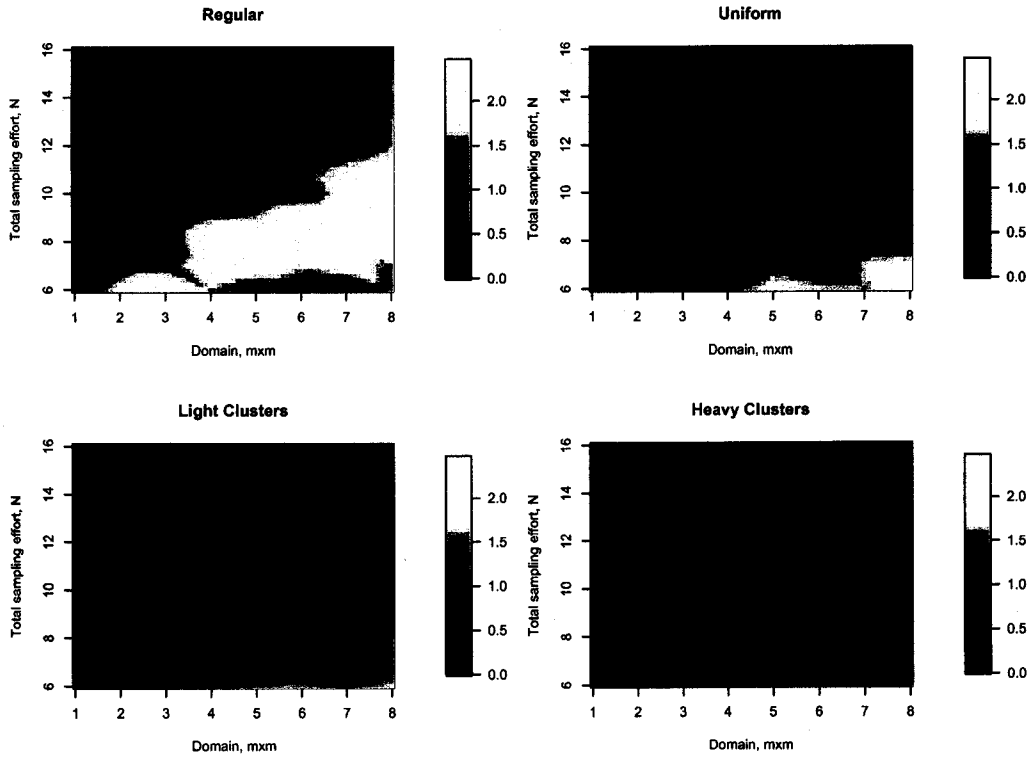

Figure 5.30: MSE of $\hat{\theta}_2$ where $\Gamma = \text{Matérn}(1/\sqrt{2}, 1)$ for each sampling pattern. Note that the scales of the x- and y-axes are linear with respect to m and $\sqrt{N}$, respectively.

Table 5.4: Standard error (se) and bias of the MLE for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ for a constant sampling density of $N/(m \times m) = 4$. Also listed is Pearson's sample correlation coefficient for $\hat{\theta}_1$ and $\hat{\theta}_2$. The tabulated values are based on 100 independent simulations.

Spatial Parameter $\boldsymbol{\theta}_1 = (\sqrt{2}, 1/2)'$

Sampling Pattern		(Sample Size, Domain) = $(N, m \times m)$ $(64, 4 \times 4)$ se	bias	$(144, 6 \times 6)$ se	bias	$(256, 8 \times 8)$ se	bias
Regular	$\hat{\theta}_1$	0.594	−0.373	0.516	−0.217	0.362	−0.149
	$\hat{\theta}_2$	0.824	0.586	0.434	0.196	0.156	0.083
	$\hat{\rho}$	-0.466		-0.491		-0.607	
Uniform	$\hat{\theta}_1$	0.681	−0.274	0.573	−0.098	0.411	−0.103
	$\hat{\theta}_2$	0.379	0.156	0.120	0.049	0.083	0.018
	$\hat{\rho}$	-0.399		-0.496		-0.531	
Light Clusters	$\hat{\theta}_1$	0.582	−0.292	0.470	−0.228	0.398	−0.191
	$\hat{\theta}_2$	0.192	0.102	0.089	0.052	0.065	0.020
	$\hat{\rho}$	-0.549		-0.521		-0.487	
Heavy Clusters	$\hat{\theta}_1$	0.725	−0.300	0.546	−0.219	0.526	−0.100
	$\hat{\theta}_2$	0.215	0.089	0.081	0.035	0.063	0.012
	$\hat{\rho}$	-0.478		-0.504		-0.590	

Spatial Parameter $\boldsymbol{\theta}_2 = (2, 1)'$

Sampling Pattern		(Sample Size, Domain) = $(N, m \times m)$ $(64, 4 \times 4)$ se	bias	$(144, 6 \times 6)$ se	bias	$(256, 8 \times 8)$ se	bias
Regular	$\hat{\theta}_1$	1.084	−0.124	0.687	−0.177	0.590	−0.029
	$\hat{\theta}_2$	0.478	0.199	0.307	0.146	0.149	0.039
	$\hat{\rho}$	-0.531		-0.661		-0.577	
Uniform	$\hat{\theta}_1$	0.832	−0.286	0.512	−0.211	0.464	−0.095
	$\hat{\theta}_2$	0.363	0.174	0.172	0.062	0.103	0.019
	$\hat{\rho}$	-0.626		-0.608		-0.624	
Light Clusters	$\hat{\theta}_1$	0.543	−0.458	0.590	−0.164	0.432	−0.139
	$\hat{\theta}_2$	0.303	0.156	0.117	0.043	0.081	0.030
	$\hat{\rho}$	-0.522		-0.568		-0.586	
Heavy Clusters	$\hat{\theta}_1$	0.704	−0.353	0.550	−0.222	0.498	−0.209
	$\hat{\theta}_2$	0.233	0.078	0.107	0.050	0.070	0.024
	$\hat{\rho}$	-0.526		-0.579		-0.641	

Table 5.5: Standard error (se) and bias of the MLE for $\boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_5$ for a constant sampling density of $N/(m \times m) = 4$. Also listed is Pearson's sample correlation coefficient for $\hat{\theta}_1$ and $\hat{\theta}_2$. The tabulated values are based on 100 independent simulations.

Spatial Parameter $\boldsymbol{\theta}_3 = (2\sqrt{2}, 2)'$

Sampling Pattern		(Sample Size, Domain) $= (N, m \times m)$ $(64, 4 \times 4)$ se	bias	$(144, 6 \times 6)$ se	bias	$(256, 8 \times 8)$ se	bias
Regular	$\hat{\theta}_1$	0.853	−0.343	0.573	−0.254	0.474	−0.147
	$\hat{\theta}_2$	0.441	0.247	0.289	0.154	0.192	0.080
	$\hat{\rho}$	-0.750		-0.799		-0.741	
Uniform	$\hat{\theta}_1$	0.738	−0.293	0.457	−0.275	0.429	−0.120
	$\hat{\theta}_2$	0.329	0.143	0.196	0.103	0.155	0.042
	$\hat{\rho}$	-0.707		-0.760		-0.754	
Light Clusters	$\hat{\theta}_1$	0.721	−0.296	0.492	−0.196	0.498	−0.089
	$\hat{\theta}_2$	0.270	0.130	0.172	0.041	0.103	0.028
	$\hat{\rho}$	-0.714		-0.783		-0.754	
Heavy Clusters	$\hat{\theta}_1$	0.769	−0.238	0.487	−0.166	0.422	−0.147
	$\hat{\theta}_2$	0.238	0.096	0.097	0.023	0.075	0.018
	$\hat{\rho}$	-0.701		-0.692		-0.664	

Spatial Parameter $\boldsymbol{\theta}_5 = (1/\sqrt{2}, 1)'$

Sampling Pattern		(Sample Size, Domain) $= (N, m \times m)$ $(64, 4 \times 4)$ se	bias	$(144, 6 \times 6)$ se	bias	$(256, 8 \times 8)$ se	bias
Regular	$\hat{\theta}_1$	0.253	−0.024	0.166	0.010	0.126	−0.004
	$\hat{\theta}_2$	1.078	1.023	0.819	0.389	0.699	0.334
	$\hat{\rho}$	-0.698		-0.694		-0.743	
Uniform	$\hat{\theta}_1$	0.256	−0.038	0.132	−0.038	0.101	−0.015
	$\hat{\theta}_2$	0.758	0.440	0.444	0.184	0.199	0.060
	$\hat{\rho}$	-0.642		-0.641		-0.690	
Light Clusters	$\hat{\theta}_1$	0.202	−0.037	0.147	−0.037	0.126	−0.013
	$\hat{\theta}_2$	0.550	0.269	0.268	0.103	0.129	0.055
	$\hat{\rho}$	-0.619		-0.601		-0.738	
Heavy Clusters	$\hat{\theta}_1$	0.215	−0.055	0.189	−0.024	0.143	−0.018
	$\hat{\theta}_2$	0.463	0.193	0.225	0.079	0.103	0.034
	$\hat{\rho}$	-0.582		-0.766		-0.706	

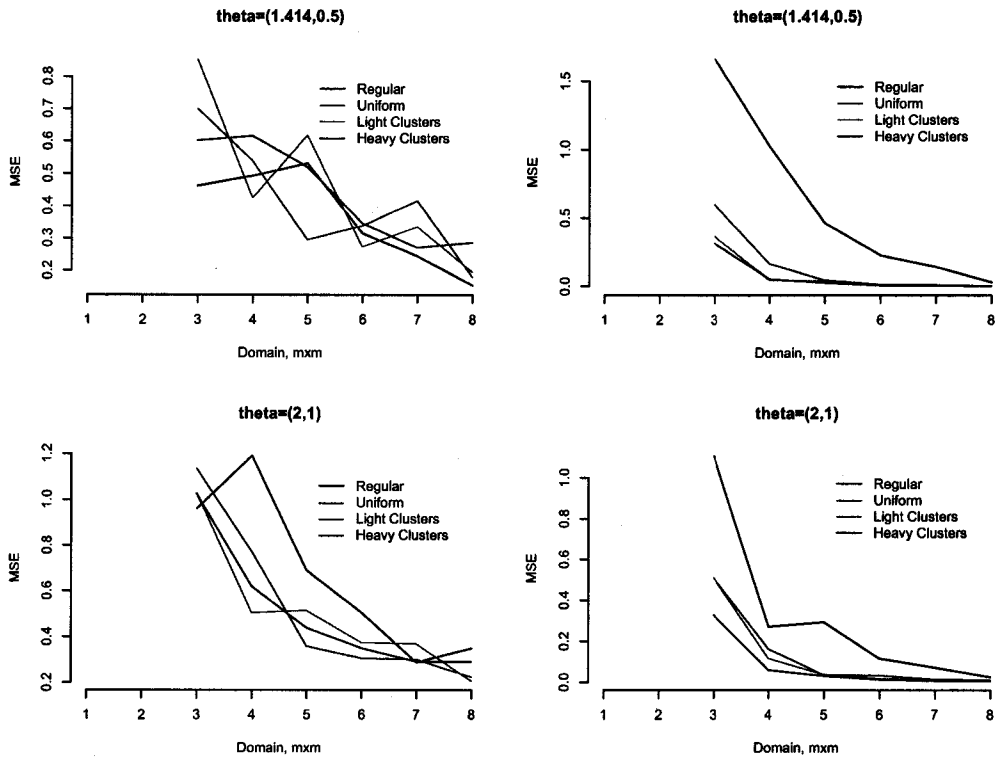

Figure 5.31: MSE of $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of domain for a constant sampling density of $N/(m \times m) = 4$ for $\mathbf{\Gamma} = \text{Matérn}(\sqrt{2}, 1/2)$ and $\mathbf{\Gamma} = \text{Matérn}(2, 1)$. The left panels plot the MSE for the range parameter estimator and the right panels plot the MSE of the smoothness parameter estimator.

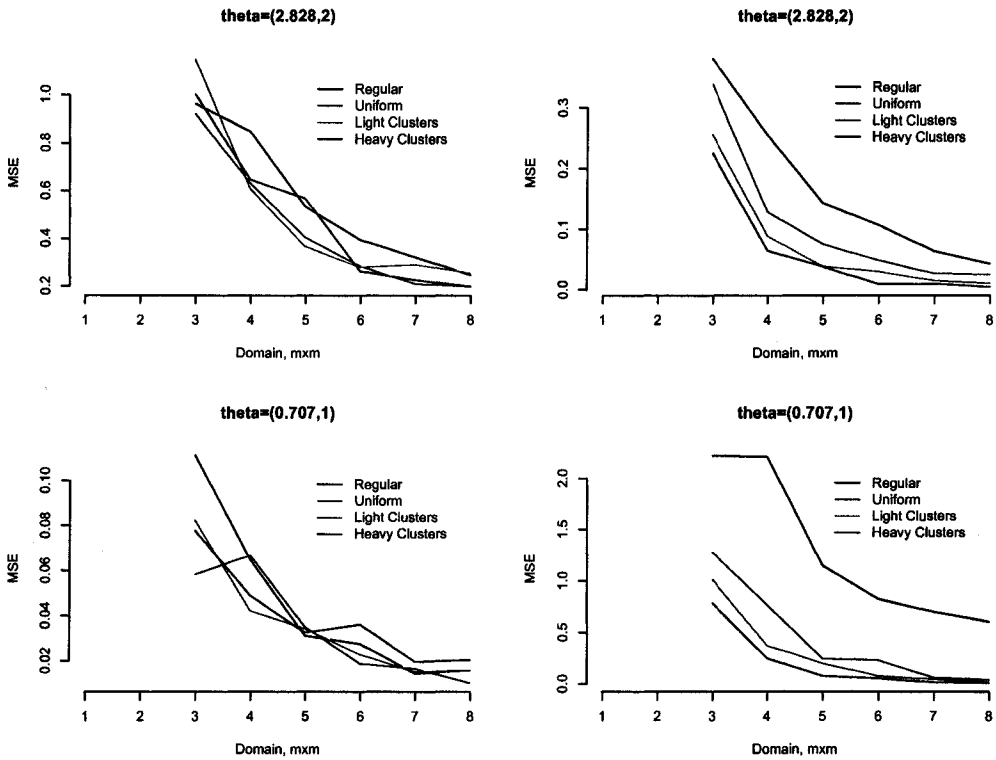

Figure 5.32: MSE of $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of domain for a constant sampling density of $N/m \times m = 4$ for $\Gamma = \text{Matérn}(2\sqrt{2}, 2)$ and $\Gamma = \text{Matérn}(1/\sqrt{2}, 1)$. The left panels plot the MSE for the range parameter estimator and the right panels plot the MSE of the smoothness parameter estimator.

5.1.3.2 Distribution of the parameter estimates

The joint distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ for a constant sampling density of 4 locations per unit area are illustrated by Figures 5.33 through 5.35. Note that the "banana" shape observed in the one dimensional case is present for a domain of 4×4. As the domain is expanded the MLEs form tighter oblong clusters and appear to be distributed as a bivariate normal. For each combination of $\boldsymbol{\theta}$, $m \times m$, N, and sampling pattern we tested each parameter estimate, individually, for normality using the Anderson-Darling test at a significance level of 0.05. Figures 5.36 and 5.37 illustrate the results for $\boldsymbol{\theta}_2 = (2, 1)$ and $\boldsymbol{\theta}_5 = (1/\sqrt{2}, 1)$, respectively. In general the distribution of the smoothness parameter estimate reaches normality more quickly than the range parameter estimator. In addition the range parameter estimator appears to require large domains to induce normality compared to the smoothness parameter estimator that can be approximately normal for all domains for moderate to large samples. For the two cases not presented, the results are similar and mirror Figure 5.36.

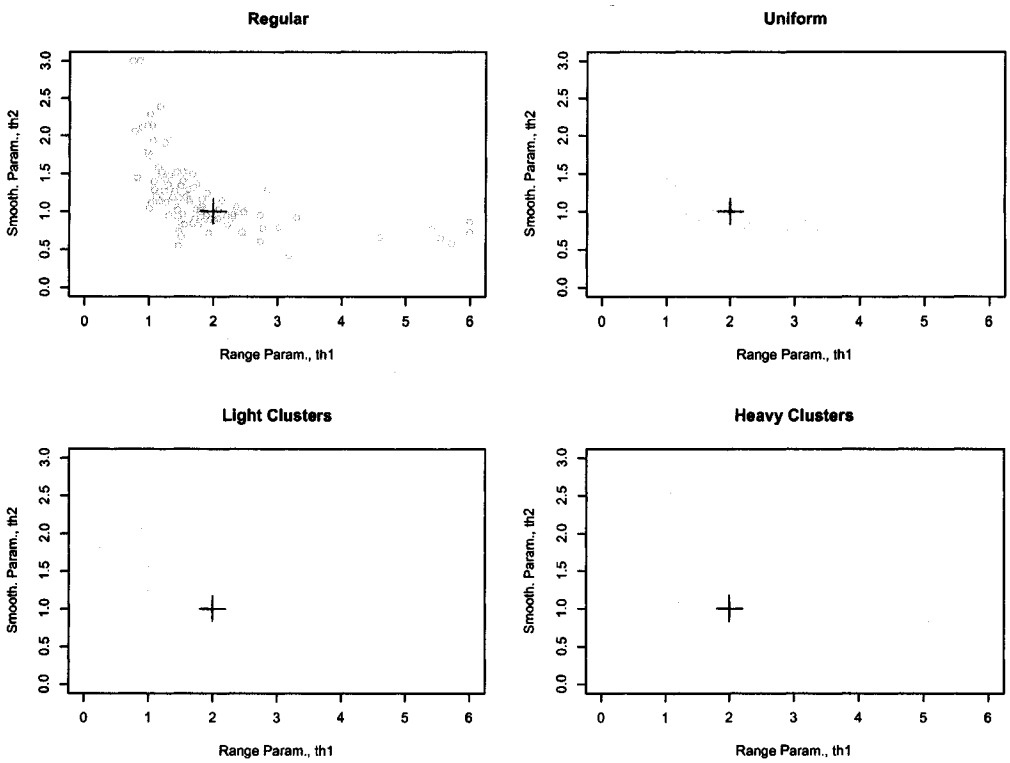

Figure 5.33: Plot of the MLE $(\hat{\theta}_1, \hat{\theta}_2)$ as a function of sampling method for $N, m \times m) = (64, 4 \times 4)$ where $\boldsymbol{\Gamma} = \text{Matérn}(2, 1)$. The sampling density, $N/(m \times m)$, is 4 observations per unit area. The true value of $\boldsymbol{\theta}$ is indicated by (+).

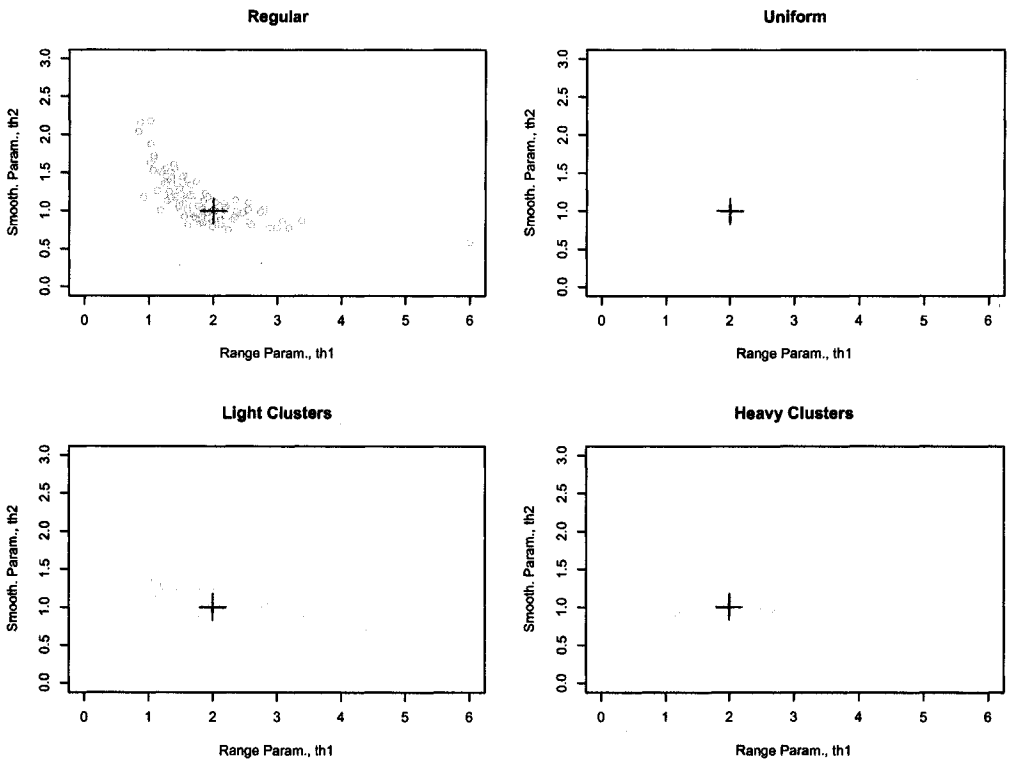

Figure 5.34: Plot of the MLE $(\hat{\theta}_1, \hat{\theta}_2)$ as a function of sampling method for $(N, m \times m) = (144, 6 \times 6)$ where $\mathbf{\Gamma} = \text{Matérn}(2, 1)$. The sampling density, $N/(m \times m)$, is 4 observations per unit area. The true value of $\boldsymbol{\theta}$ is indicated by $(+)$.

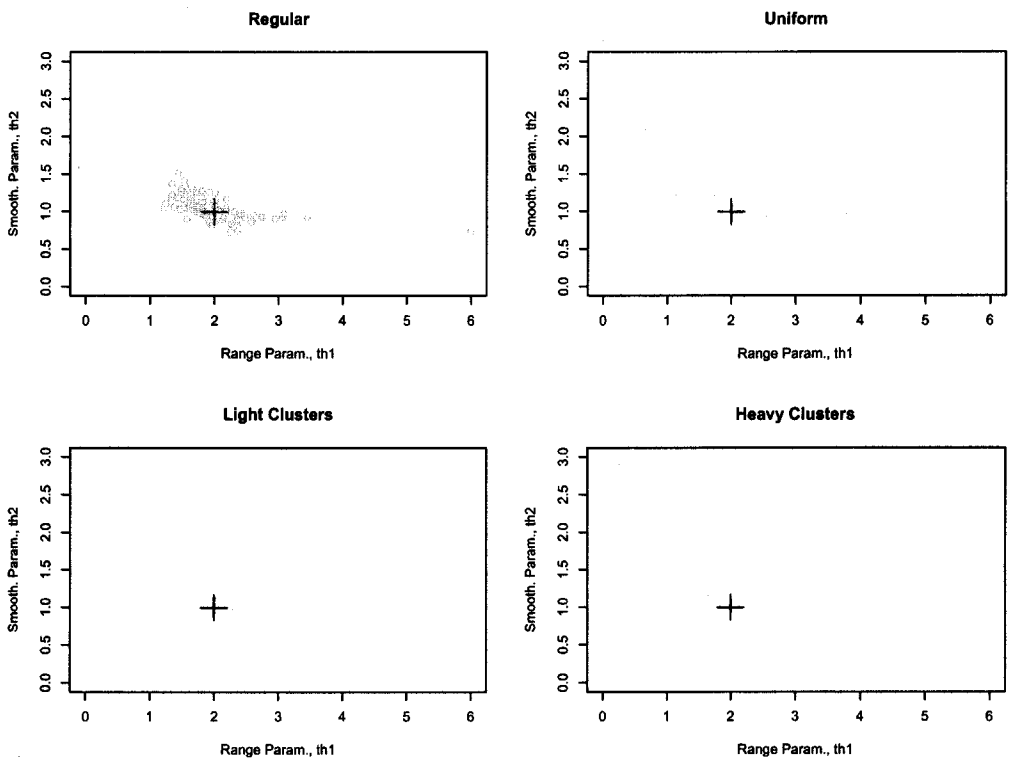

Figure 5.35: Plot of the MLE $(\hat{\theta}_1, \hat{\theta}_2)$ as a function of sampling method for $(N, m \times m) = (256, 8 \times 8)$ where $\mathbf{\Gamma} = \text{Matérn}(2, 1)$. The sampling density, $N/(m \times m)$, is 4 observations per unit area. The true value of $\boldsymbol{\theta}$ is indicated by (+).

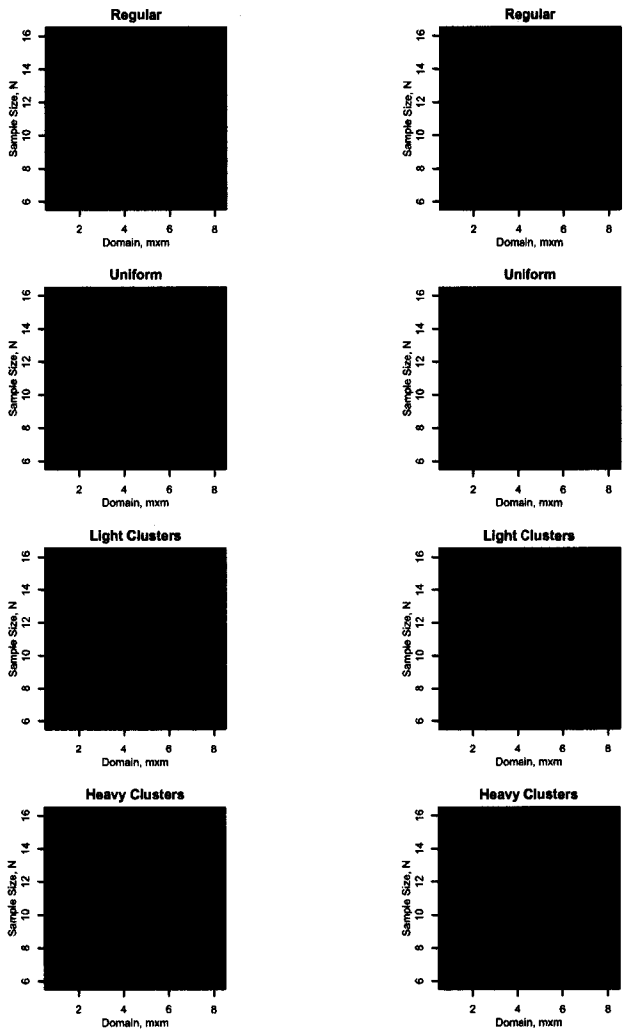

Figure 5.36: Normality testing results where $\mathbf{\Gamma} = \text{Matérn}(2, 1)$. The Anderson-Darling test was performed on the observed distribution of each parameter estimate individually at a significance level of 0.05 for all combinations of domain ($m \times m$), sampling effort N and sampling pattern. The null hypothesis is that the distribution is normal. Blue squares correspond to non-normal distributions and red squares to normal distributions.

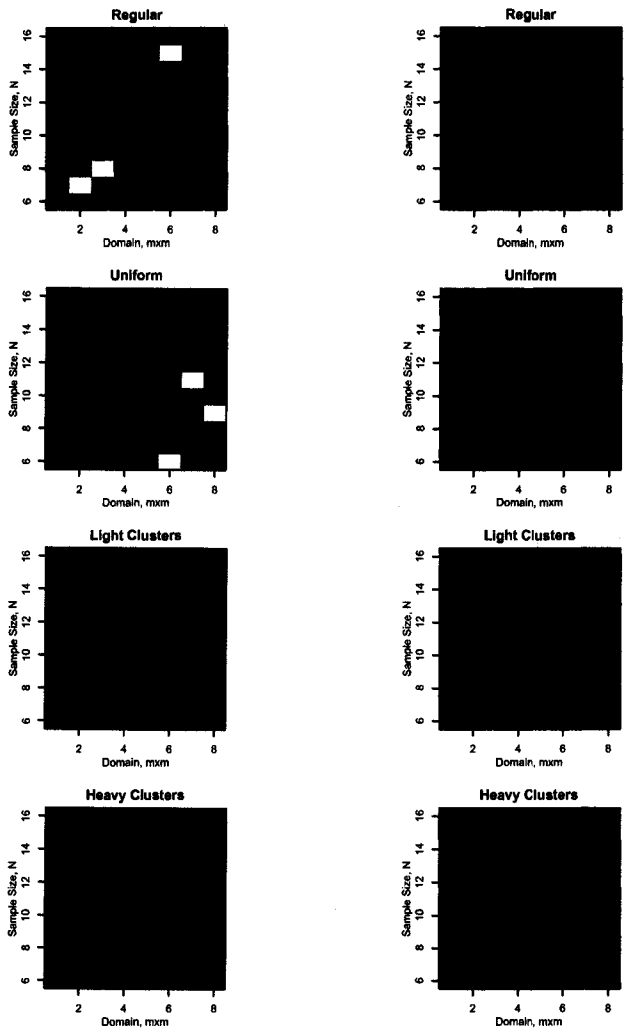

Figure 5.37: Normality testing results where $\Gamma = \text{Matérn}(1/\sqrt{2}, 1)$. The Anderson-Darling test was performed on the observed distribution of each parameter estimate individually at a significance level of 0.05 for all combinations of domain ($m \times m$), sampling effort N and sampling pattern. The null hypothesis is that the distribution is normal. Blue squares correspond to non-normal distributions and red squares to normal distributions.

5.1.3.3 Estimated correlation function

The plots of the estimated correlation functions for two of the cases are found in Figures 5.38 and 5.39. The plots correspond to a sampling density of 4 where $(N, m \times m) = (144, 6 \times 6)$. Each light blue line corresponds to a single realization, the dark blue lines identify the 5th, 50th, and 95th percentiles, and the solid black line is the true correlation. Similar to the one-dimensional case, the bias of $\hat{\theta}_1$ is negative and the bias of $\hat{\theta}_2$ is positive. Consequently the estimated correlation is typically below the true correlation. It should be noted that the variability of the estimated correlation functions is much larger for $\theta_1 = 2$ compared to $\theta_1 = 1/\sqrt{2}$. This stems from the fact that the estimator of the range parameter θ_1 is more variable for the former case and, as we have seen previously, the range parameter typically dominates the overall shape of the correlation function whereas the smoothness parameter θ_2 influences the shape near zero (distance).

5.1.3.4 Mean integrated square error

We illustrate the mean integrated square error (MISE) maps for each case with Figures 5.40 through 5.43. Superimposed on each plot are approximate contours to better differentiate gradients. For each spatial parameter vector $\boldsymbol{\theta}$ the four panels, one each for the sampling methods, are very similar in shape. For the first three cases the largest MISEs occur when the domain is small independent of sampling method. The contours suggest that the higher priority is to cover the domain of interest with fewer sampling locations but with sufficient density. Note that for smaller domains a greater sampling is required, hence denser sampling on average, in order to reach the region with the smallest MISE. For the remaining case, $\boldsymbol{\theta} = (1/\sqrt{2}, 1)'$, there is a significant increase in the MISE for moderate to large domains and small sampling efforts (low sampling densities). The inability to approximate the smoothness parameter in this region (in an MSE sense) is the source of the

error. This particular example demonstrates that the contribution to the MISE by the smoothness parameter estimator can be the dominant term. Note that the MISE surface for the previous three cases much more closely resemble the MSE surfaces of $\hat{\theta}_1$ (by shape, not magnitude). Conversely, the MISE surface for the last case resembles the MSE surface of $\hat{\theta}_2$.

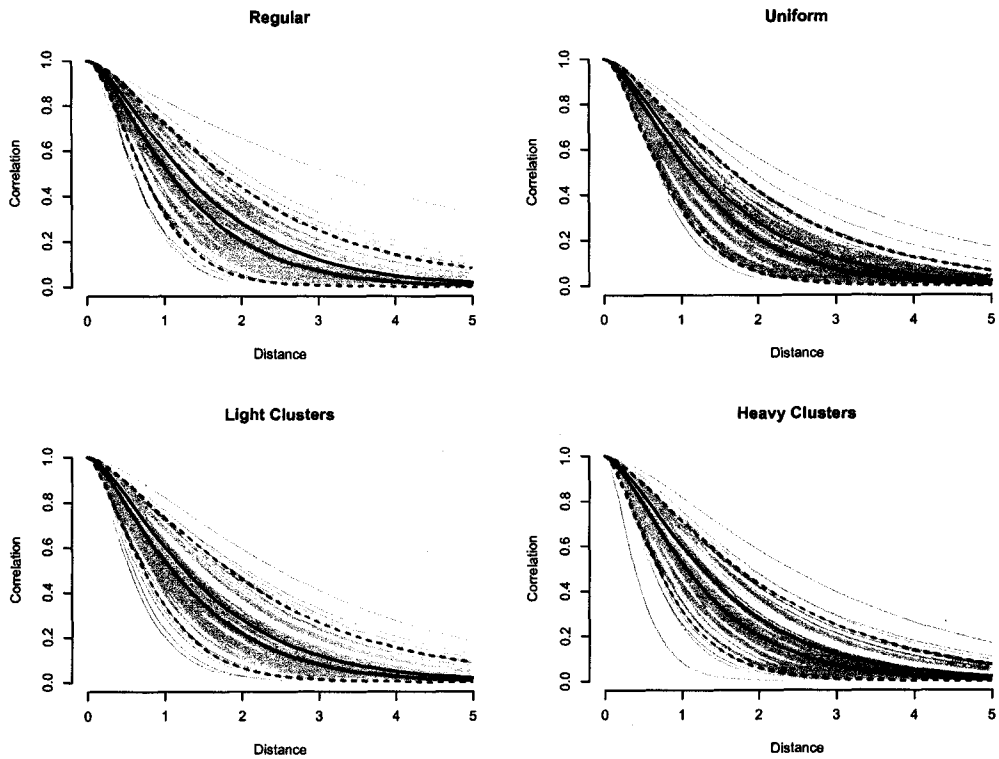

Figure 5.38: Fitted correlation functions for Matérn(2,1) where $(N, m \times m) = (144, 6 \times 6)$. Each light blue line corresponds to a single realization. The dark blue lines are the median (solid) and the 5th and 95th percentiles (dashed). The solid black line is the true correlation function.

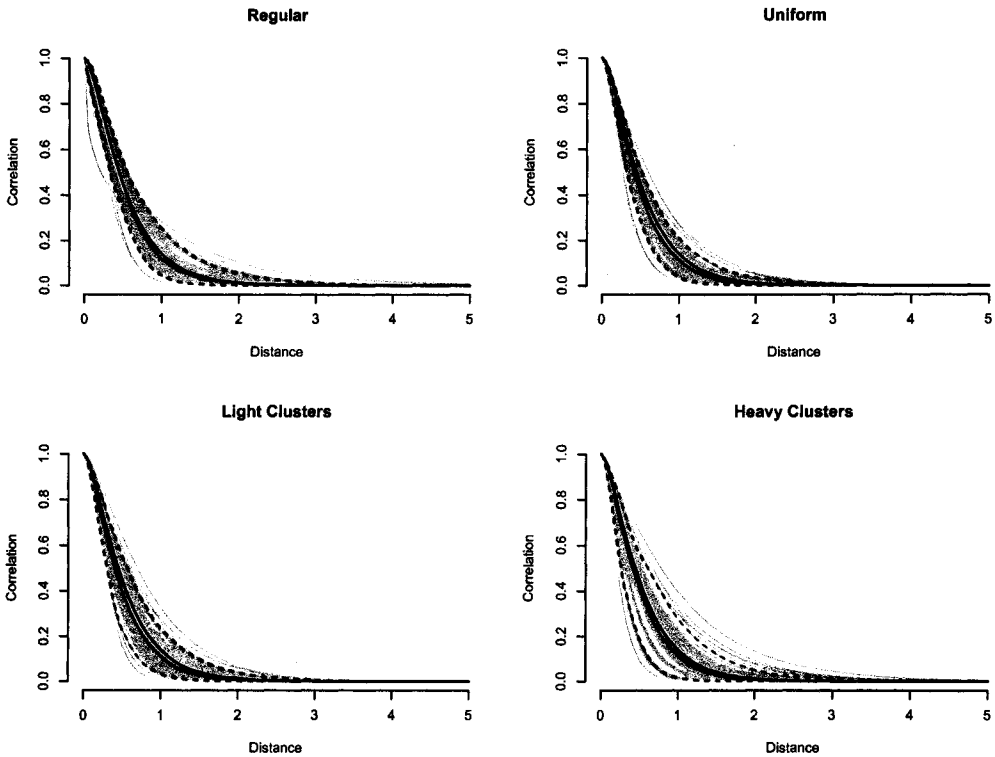

Figure 5.39: Fitted correlation functions for Matérn$(1/\sqrt{2}, 1)$ where $(N, m \times m) = (144, 6 \times 6)$. Each light blue line corresponds to a single realization. The dark blue lines are the median (solid) and the 5th and 95th percentiles (dashed). The solid black line is the true correlation function.

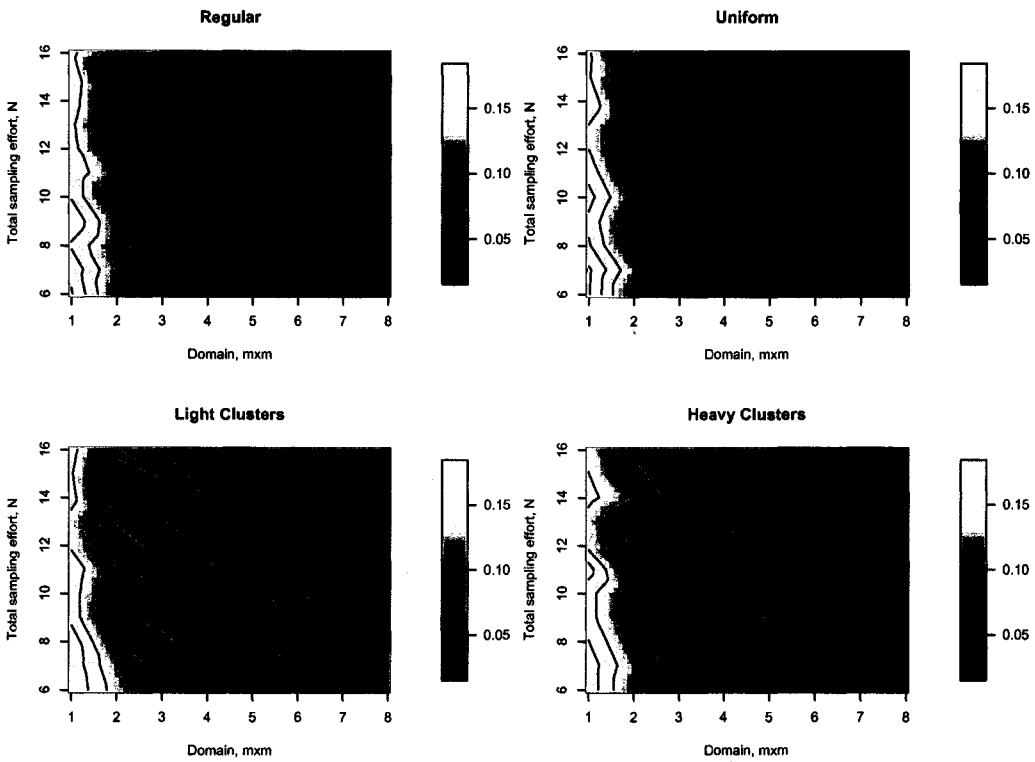

Figure 5.40: Mean integrated square error (MISE) maps for $\Gamma = \text{Matérn}(\sqrt{2}, 1/2)$ for each sampling pattern. Note the temperature (color) scale for MISE is the same for all four panels and that the scales for the x- and y-axes are linear in m and $\sqrt{N}$, respectively.

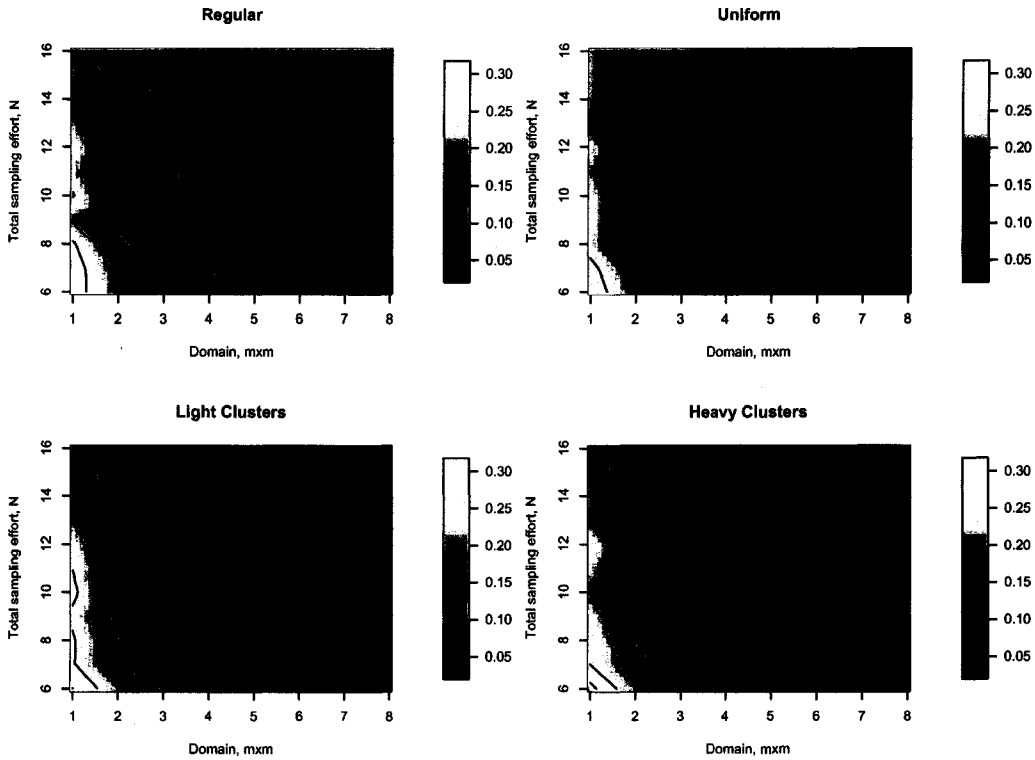

Figure 5.41: Mean integrated square error (MISE) maps for $\mathbf{\Gamma} = \text{Matérn}(2, 1)$ for each sampling pattern. Note the temperature (color) scale for MISE is the same for all four panels and that the scales for the x- and y-axes are linear in m and $\sqrt{N}$, respectively.

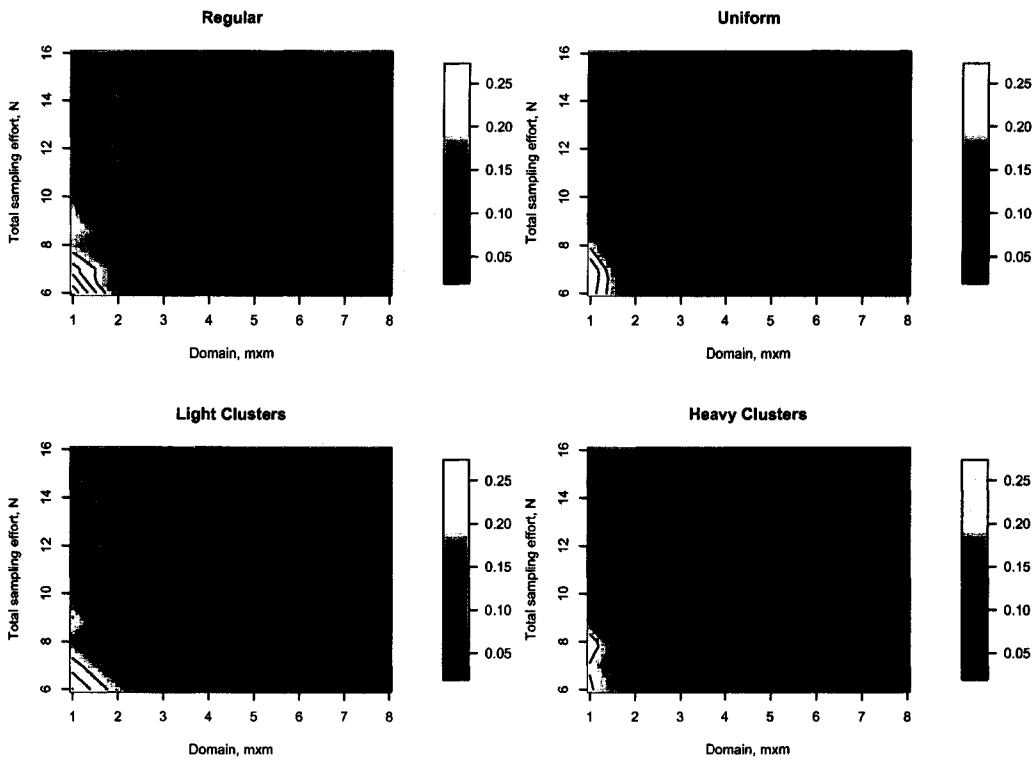

Figure 5.42: Mean integrated square error (MISE) maps for $\mathbf{\Gamma} = \text{Matérn}(2\sqrt{2}, 2)$ for each sampling pattern. Note the temperature (color) scale for MISE is the same for all four panels and that the scales for the x- and y-axes are linear in m and $\sqrt{N}$, respectively.

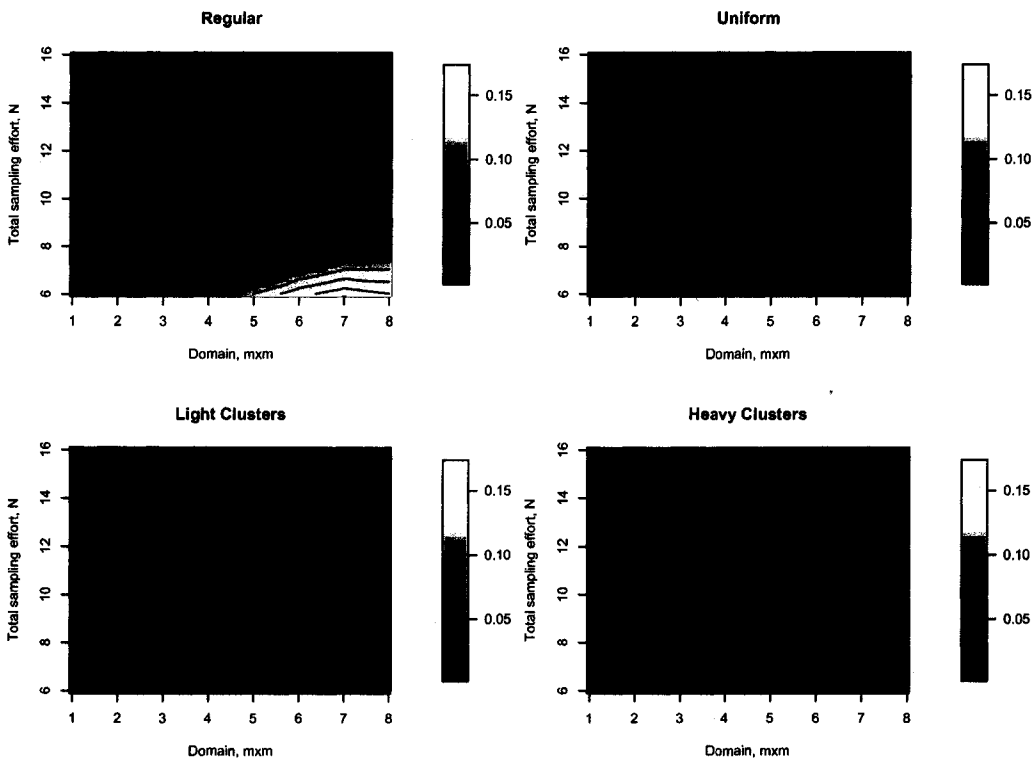

Figure 5.43: Mean integrated square error (MISE) maps for $\Gamma = \text{Matérn}(1/\sqrt{2}, 1)$ for each sampling pattern. Note the temperature (color) scale for MISE is the same for all four panels and that the scales for the x- and y-axes are linear in m and $\sqrt{N}$, respectively.

5.2 Conclusions

In this chapter we have performed an extensive number of simulation studies to characterize empirically the behavior of the MLE of the spatial parameter vector θ with respect to expanding domain and infill for the Matérn correlation function. The Matérn class is characterized by two parameters (in the absence of measurement error): a range parameter θ_1 and a smoothness parameter θ_2. The Matérn class includes as a special case the exponential class. A drawback in using the Matérn correlation function is the presence of two parameters over which to optimize increases the time required to find optima. Simulations indicate that the smoothness parameter estimator was approximately normal in distribution for sufficiently large domains and levels of infill. The normality of the range parameter estimator proved to be more elusive and mimicked the behavior observed for the exponential correlation function in two-dimensions. We identified that the bias for the range parameter estimator was negative and the bias for the smoothness parameter estimator was positive. Overall the effect on the estimated correlation function was underestimation of the true correlation for most distances. Using the correlation function we introduced an ad hoc measure for goodness-of-fit called the integrated square error (ISE). Observing that the parameter estimates are moderately (negatively) correlated we suggested integrating the squared difference between the estimated and true correlation function. ISE values close to zero indicate that the correlation function using the MLE $\hat{\boldsymbol{\theta}}$ is close to the true correlation function. The resulting image plots appeared to be composites of the MSE images for the individual MLEs $\hat{\theta}_1$ and $\hat{\theta}_2$. These clearly showed that the estimate of the range parameter dominates the overall mean ISE (MISE) in that the resulting MISE image is very similar to the MSE for $\hat{\theta}_1$ image plot. This suggests that the sampling design should place more emphasis no estimating the range parameter well compared to the smoothness parameter.

We extended the Matérn class simulations to two-dimension following the procedure first outlined in the two-dimensional simulation study for the exponential correlation function. Many of the observations of the two-dimensional study coincided with the the results of the one-dimensional study. Both studies suggested that the regular pattern is less than optimal for minimizing the overall MSE and that clustering the sampling locations was very beneficial provided that the the clusters were spread across the domain of interest.

Chapter 6

CONCLUSIONS AND FUTURE WORK

With the advent of new technologies such as Global Positioning Systems (GPS) and Graphical Information Systems (GIS), the ability to collect and analyze geo-referenced data has become increasingly easy. One positive consequence is researchers are now more apt to analyze the data in a spatial or spatial-temporal framework. In addition, the increased growth of computational capability allows the researcher to freely explore potentially huge model spaces with large numbers of covariates and many error processes from which she selects the "best" model. Some researchers are interested in identifying the underlying process while others may wish to focus on developing maps for prediction. This interest in fitting spatial models and selecting the "best" model from a collection of competing models is the motivation for the subsequent work found in Chapter 2. We began by first looking at these issues concurrently, spatial modeling and model selection, by developing the spatial AIC statistic heuristically. Spatial AIC provides a means to evaluate the goodness-of-fit and complexity of each candidate model and allows the researcher to make direct comparisons. Furthermore, by optimizing the profile likelihood all of the model parameters are fit simultaneously eliminating the need to iterate between fitting the mean surface and correlation parameters. We apply the spatial AIC statistic to both simulated and real world data sets showing it superior to more classic modeling approaches with respect to selecting the "true" model from a collection of candidates. We also demonstrate the statistic's utility by applying it to a stream network problem where movement between locations is restricted to "within-stream",

i.e., as the fish swims. Hence the spatial dimensionality is somewhere between one- (a stream without confluences) and two- (open water) dimensions.

The development of spatial AIC requires standard asymptotic assumptions. Chief among them is that the estimates of the model parameters, particularly the spatial parameters, are asymptotically unbiased and normally distributed with covariance structure equal to the inverse of the Fisher information. We set out in Chapter 3 to develop the asymptotic distribution of the MLE for the range parameter θ for the exponential correlation function in one-dimension. We assume the underlying process is continuous and can be modeled as the Uhlenbeck-Ornstein process (a stochastic differential equation). Furthermore, we assume that we observe the process without measurement error at a finite set of geo-referenced locations. We further assume that one can observe the process anywhere, hence there are two methods to increase sample size. The first is to extend the current domain and continue to collect information, and the second is to make additional observations inside the current domain at previously unobserved locations. The former is known as expanding the domain and the latter as infilling. Within this framework we develop the asymptotic distribution of the maximum likelihood estimator (MLE) of θ. We illustrate that the asymptotic results are good approximations for finite sample sizes through a series of simulation studies. Coupled with the asymptotic analysis is an analysis of the impact of the sampling design, i.e., the procedure for choosing sampling locations. We show theoretically and empirically that for a fixed domain and sampling effort, regular equispaced sampling is preferable to any other design with respect to minimizing the standard error of the estimator.

Due to complexities in the theory, the scope of Chapter 3 is rather restrictive: a Gaussian process in one-dimension with exponential correlation structure. Hence we seek to broaden our understanding of the behavior of the MLE in two-dimensions for both the exponential class and for the larger Matérn class in Chapter 4. The Matérn

class includes as a special case the exponential class and as the limiting case and Gaussian class. The difficulty of the problem forces us to first restrict our attention to empirically analyze the situation and defer development of the theoretical results for now. Nonetheless, through simulation we are able to develop distributions of the MLEs and observe that the standard assumptions appear to be satisfied. we present these results as a departure point for developing the asymptotic distribution for the Matérn class and to further our understanding of the impact of the sampling design on estimation of the spatial parameters.

6.1 Future work

The importance of these efforts is demonstrated by the continued interest in applying spatial models to larger and more complex data sets. It will only become easier with time to collect and analyze data in a spatial setting. Computational limitations are continuously being overcome. We suggest several avenues for continued work in this area.

6.1.1 Spatial AIC, model selection, and model misspecification

The development of the spatial AIC presented in Chapter 2 was heuristic in nature. The next step is to formally develop spatial AIC in a more rigorous manner. Furthermore, it is important to verify that the MLEs for each of model parameters meet the assumptions for the more general case. The work presented here focused on the correlation parameters; one can explore the distribution of the estimated regression coefficients $\hat{\beta}$ and estimated variance $\hat{\sigma}^2$ using the realizations from the current study.

A drawback to the AIC approach is the assumption that there exists an underlying "true model". Regardless of one's philosophical position, that either there is a true model of finite dimension or that there exists a best approximating model

to the truth, one is still required to assume that a "best" underlying model exists. A promising alternative approach is minimum description length (MDL) (see Section 2.3.3). MDL does not require the existence of a "true model". Developed through information theory and computer science, MDL seeks to describe the data in the fewest "bits". Much like AIC it consists of a likelihood term and a penalty term. In this context one can think of the likelihood term as the residuals and the penalty term as the model itself where less parsimonious models are more heavily penalized. One of the core principles of statistical modeling is identifying significant relationships in data and, barring evidence to the contrary, assuming simpler models whenever possible. But how does one interpret the fitted model parameters? For example, is it still appropriate to refer to θ as the range parameter when one is simply trying to minimize the information lost by compressing the data? and what can we say about inferences concerning θ?

6.1.1.1 Model misspecification

An important avenue that requires further investigation is the impact of misspecifying the model form during the model selection process. In this scenario one is assuming that there is a "true" underlying model that governs the process. In Section 2.3 we introduced the spatial AIC statistic for use during model selection which consists of two terms: the first term, $-2\times$log-likelihood, represents the quality of fit and the second term acts as a penalty that increases with increased model complexity. The AIC statistic was developed as an estimator of the Kullback-Leibler information criterion and roughly represents the loss of information incurred by fitting an incorrect model to the data.

The following two examples were generated to illustrate the impact of model misspecification. For the first example, the underlying model is $\mathbf{Y} = \sigma^2\mathbf{\Gamma}$ where $\sigma^2 = 1$ and $\mathbf{\Gamma} =$ exponential(1). For each sampling design ($N = mn \geq 8$) we generated 100 realizations. For each realization we fit both the exponential and

Matérn correlation functions and used spatial AIC to select the "better" model. Figure 6.1 illustrates the percentage of the time that the exponential model was selected in favor of the Matérn model. Red indicates that the exponential model was selected more often than the Matérn model. Regions of yellow and orange indicate that the Matérn model was selected nearly one quarter of the time; this implies that by allowing the smoothness parameter to vary a significant reduction of the likelihood function was achieved to compensate for the additional penalty. Recall that the exponential class is a subset of the the Matérn class ($\theta_2 = 1/2$). Therefore, for particular realizations of the random field the data may indeed be better represented by the Matérn model.

For the second example the true correlation matrix is $\Gamma = \text{Matérn}(2, 1)$. Once again, 100 realizations for each sampling design were generated, both model forms fit, and spatial AIC used to select the favored model. Figure 6.2 illustrates the results where blue regions indicate that the Matérn model was selected more often. The Matérn model was selected for nearly all realizations of size 64 or more independent of sampling design while for more modest sample sizes the exponential model was favored.

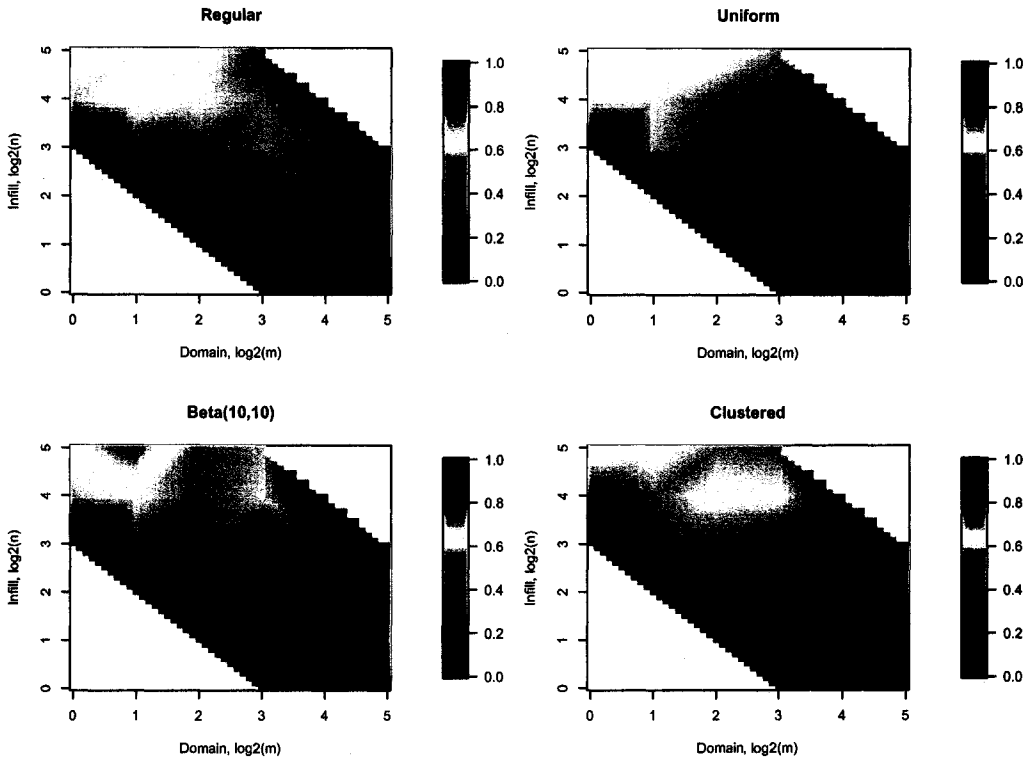

Figure 6.1: Summary of the model selection results for the Matérn($\sqrt{2}, 1/2$). The image plot illustrates the percentage of times that the exponential model was selected in favor of the Matérn model. The selected model was determined by the minimum spatial AIC. Red indicates that the exponential model was selected for a large majority of the realizations where as blue indicates that the Matérn model was selected for a large majority. Only realizations with a sample size of at least 8 were considered (due to the nature of the penalty term of the spatial AIC).

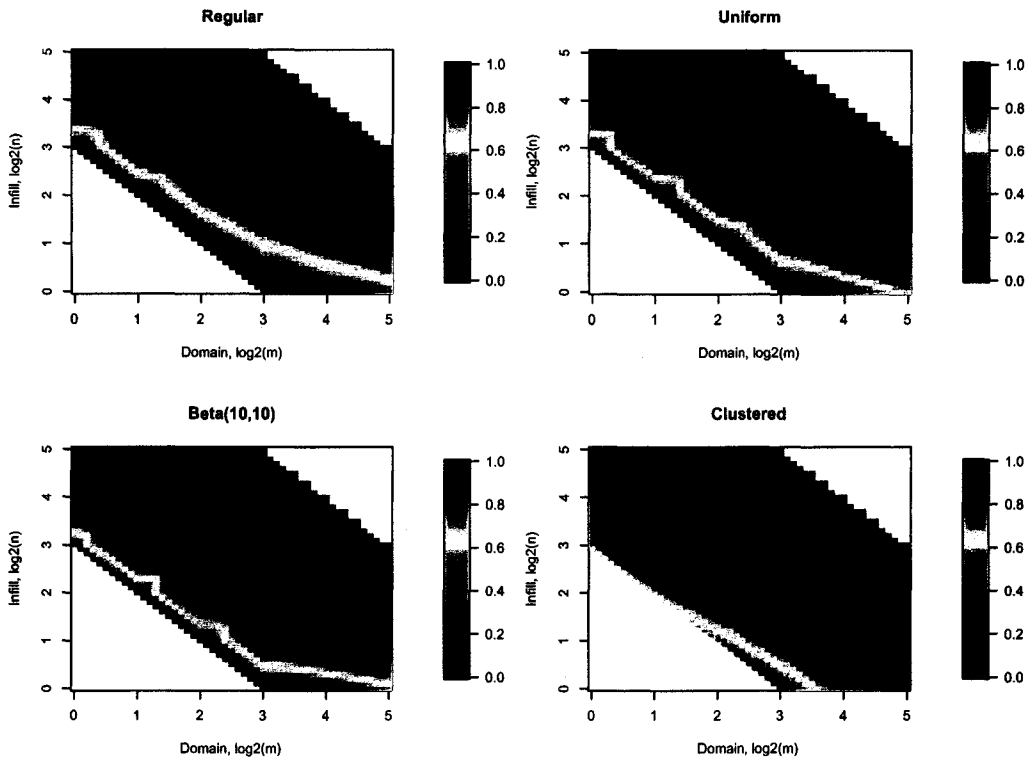

Figure 6.2: Summary of the model selection results for the Matérn($\sqrt{2}, 1/2$). The image plot illustrates the percentage of times that the exponential model was selected in favor of the Matérn model. The selected model was determined by the minimum spatial AIC. Red indicates that the exponential model was selected for a large majority of the realizations where as blue indicates that the Matérn model was selected for a large majority. Only realizations with a sample size of at least 8 were considered (due to the nature of the penalty term of the spatial AIC).

6.1.2 Asymptotic distribution of spatial parameter estimates

Standard asymptotic assumptions were required for the development of the spatial AIC statistic. In Chapter 3 we formally developed the asymptotic distribution of the MLE $\hat{\theta}$ for the exponential correlation function in one-dimension and much of Chapter 4 was devoted to exploring the behavior of the MLE with respect to expanding domain and infill. Empirical evidence developed from the simulations suggests that the derived asymptotics in one-dimension hold for the exponential class in two-dimension as well as for the Matérn class in both one- and two-dimensions. We now need to progress and formally develop the asymptotic distribution for the MLE for the exponential class in two-, and perhaps more, dimensions. Ideally one would next proceed to the Matérn class but it may prove beneficial to first examine the Gaussian class (the limiting case of the Matérn class). The simulation studies do provide some insight into the distributional properties of the MLEs including the rate and direction from which the bias goes to zero as well as the normality of the MLEs.

6.1.3 Measurement error

All of the work presented thus far assumes that the process has been observed without error. In part this was to maintain a reasonable scope for the study. However, the inclusion of additional noise in the signal can greatly complicate the covariance structure and precludes the possibility of closed form solutions (as was used for the exponential class in one-dimension). A typical model that incorporates measurement error into the covariance structure is $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Gamma} + \tau^2\mathbf{I}$ where σ^2 and $\boldsymbol{\Gamma}$ are defined as before, τ^2 is the variance of a white noise process, and $\mathbf{I}$ is an $n \times n$ identity matrix. In this modeling framework, the researcher does not observe the underlying process of interest $\boldsymbol{Z}$. Instead she observes the process plus white noise, i.e., she observes $\boldsymbol{Y} = \boldsymbol{Z} + \eta$ where $\{\eta_i\} \sim \text{WN}(0, \tau^2)$. The parameter τ^2

is often referred to as the "nugget" in the spatial context and typically represents measurement error (squared) due to instrumentation or observation.

The presence of white noise in the process signal greatly impacts optimization. For example, the Matérn class which is characterized by two parameters must now be extended to include a third parameter α. The parameter α is the proportion of the total variance $(\sigma^2+\tau^2)$ of the observed process that is attributable to white noise, i.e., $\alpha = \tau^2/(\sigma^2 + \tau^2)$. The same is true for the exponential class. Thus for both cases the presence of white noise increases the dimension of the parameter space by one and thus increases the difficulty to find numerical solutions. Furthermore, preliminary studies suggest that the behavior of the parameter estimate $\hat{\alpha}$ is poor, even for moderate values of α (0.10 or 0.20). Additionally, the sampling pattern may play a more crucial role with respect to expanding domain and infill asymptotics. For example, it may be "best" to include observations at (extremely) close proximity to better estimate the nugget.

One promising avenue for research is considering the exponential correlation function in one-dimension observed with noise. If the sampling locations are equispaced, then we can show that the resulting structure is equivalent to the ARMA(1,1) model from time series. Define Z_t to be the underlying process of interest, Y_t to be the observed process with noise, and η_t to be a Gaussian noise process such that $\{\eta_t, t = 1, 2, \ldots\} \sim$ iid $\mathcal{N}(0, \tau^2)$. Recall $Z_t = \phi Z_{t-1} + \varepsilon_t$ where $\{\varepsilon_t, t = 1, 2, \ldots\} \sim \mathcal{N}(0, \sigma^2(1 - \phi^2))$. Hence,

$$\begin{aligned} Y_t &= Z_t + \eta_t \\ &= (\phi Z_{t-1} + \varepsilon_t) + \eta_t \\ &= \phi (Z_{t-1} + \eta_{t-1}) + \varepsilon_t + \eta_t - \phi\eta_{t-1} \\ &= \phi Y_{t-1} + (\varepsilon_t + \eta_t) - \phi\eta_{t-1} \\ &= \phi Y_{t-1} + \eta_t^\star + \theta\eta_{t-1}^\star, \end{aligned}$$

where $\{\eta_t^\star\} \sim \mathcal{N}(0, \sigma^2(1-\phi^2)+\tau^2)$ and $\theta = -\phi\tau/\sqrt{\sigma^2(1-\phi^2)+\tau^2}$. Therefore $Y_t - \phi Y_{t-1} = \eta_t^\star + \theta\eta_{t-1}^\star$ which is precisely the ARMA(1,1) model (Brockwell and Davis, 1996). Note that we recover the underlying process $\{Z_t\}$ when $\tau = 0$. The hope is that a procedure similar to that employed to derive the asymptotic distribution for the exponential class in one-dimension without measurement error can be recycled for the current case.

Bibliography

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

Ahtola, J. and Tiao, G. C. (1984). Parameter inference for a nearly nonstationary first-oder autoregressive model. *Biometrika*, 71:263–272.

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.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Bhattacharyya, B. B., Richardson, G. D., and Franklin, L. A. (1997). Asymptotic inference for near unit roots in spatial autoregression. *The Annals of Statistics*, 25(4):1709–1724.

Billingsley, P. (1995). *Probability and measure.* Wiley Interscience [John Wiley & Sons], New York, 3rd edition.

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

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting.* Springer-Verlag.

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

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Inference A Practical Information Theoretic Approach.* Springer: New York, 2nd edition.

Chan, N. H. and Wei, C. Z. (1987). Asymptotic inference for nearly nonstationary ar(1) processes. *The Annals of Statistics*, 15:1050–1063.

Chen, H.-S., Simpson, D. G., and Ying, Z. (2000). Infill asymptotics for a stochastic process model with measurement error. *Statististica Sinica*, 10:141–156.

Cressie, N. and Zimmerman, D. L. (1992). On the stability if the geostatistical model. *Mathematical Geology*, 24(1).

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

Davis, R. A. and Dunsimir, W. T. M. (1997). Least absolute deviation estimation for regression with arma errors. *Journal of Theoretical Probability*, 10(2):481–497.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49.

Evans, G. B. and Savin, N. E. (1981). The calculation of the limiting distribution of the least squares estimator of the parameter in a random walk model. *The Annals of Statistics*, 9(5):1114–1118.

Fuentes, M. (2005). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association.*

Givens, G. H. and Hoeting, J. A. (2005). *Computational statistics.* Wiley Interscience [John Wiley & Sons], Hoboken NJ, 1st edition.

Haining, R. (1990). *Spatial Data Analysis in the Social and Enviromental Sciences.* Cambridge University Press: Cambridge.

Hall, P. and Heyde, C. C., editors (1980). *Martingale Limit Theory and Its Application.* Academic Press: New York.

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

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 sighting data. In Miller, R. I., editor, *Mapping the diversity of Nature.* Chapman and Hall: London.

Hurvich, C. M. and ling Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Irvine, K. M., Gitelman, A. I., and Hoeting, J. (2006). Spatial designs and strength of spatial signal: effects on covariance estimation. In press with ...

Johns, C. J., Nychka, D., Kittel, T. G. F., and Daly, C. (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association*, 98:796–806.

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

Lee, T. C. (2001). An introduction to coding theory and the two-part minimum description length principle. *International Statistical Review*, 69:169–183.

Loh, W.-L. (2005). Fixed-domain asymptotics for a subclass of matérn-type gaussian random fields. *The Annals of Statistics*, 33:2344–2394.

Loh, W.-L. and Lam, T.-K. (2000). Estimating structured correlated matrices in smooth gaussian random field models. *The Annals of Statistics*, 28:880–904.

McQuarrie, A. D. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific: New Jersey.

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

Peterson, E. E., Merton, A. A., Theobald, D. M., and Urquhart, N. S. (2006). Patterns of spatial autocorrelation in stream water chemistry. *Environmental Monitoring and Assessment*.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 14:1080–1100.

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

Stein, M. L. (1995). Fixed-domain asymptotics for spatial periodograms. *Journal of the American Statistical Association*, 90(432).

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

Stein, M. L. (1999b). Predicting random fields with increasing dense observations. *The Annals of Applied Probability*, 9:242–273.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Part A – Theory and Methods*, 7:13–26.

Theobald, D. M., Norman, J., Peterson, E. E., and Ferraz, S. (2005). Functional linkage of watersheds and streams (flows): Network-based arcgis tools to analyze freshwater ecosystems. ESRI – GIS Mapping Software. Online Manual for ArcGIS Software.

Thompson, S. E. (2001). *Bayesian Model Averaging and Spatial Prediction.* PhD thesis, Colorado State University.

Venables, W. N. and Ripley, B. D. (1999). *Statistics and Computing.* Springer: New York, third edition.

Ver Hoef, J. M., Cressie, N., Fisher, R. N., and Case, T. J. (2001). Uncertainty and spatial linear models for ecological data. In Hunsaker, C., Goodchild, M., Friedl, M., and Case, T., editors, *Spatial Uncertainty for Ecology: Implications for Remote Sensing and GIS Applications*, pages 214–237. Springer-Verlag, New York, NY.

Ver Hoef, J. M., Peterson, E. E., and Theobald, D. M. (2006). Some new spatial statistical models for stream networks. *Environmental and Ecological Statistics*, 13.

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

Xia, G., Miranda, M. L., and Gelfand, A. E. (2005). Approximately optimal spatial design approaches for environmental health data. In press with Environmetrics.

Ying, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *The Annals of Statistics*, 21(3):1567–1590.

Zhang, H. and Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92:921–936.

Zhu, Z. and Stein, M. (2005). Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference*, 134:583–603.

Zhu, Z. and Zhang, H. (2005). Spatial sampling design under the infill asymptotic framework. In press with Environmetrics.