Abstract

Agarwal, Prasheen Kumar. Bootstrapping of Spatially Correlated Data. (Under the direction of Professor Montserrat Fuentes and Professor Margery Overton)

The application of the bootstrap to spatially correlated data has not been studied as widely as its application to time series data. This is a challenging problem since it is difficult to preserve the correlation structure of the data while implementing the bootstrap method. Kunsch (Kunsch 1989) , Politis and Romano (Politis & Romano 1993), Liu and Singh (Liu & Singh 1992) have suggested bootstrapping methods for higher dimensional data. We are proposing a new bootstrapping method for spatial data and are studying the properties of the estimators for the mean and the semi-variogram under our method. We demonstrate the performance and usefulness of this method by a simulation study. We will also show consistency and derive asymptotic distributional properties of the estimators. As an application we are studying the problem of modeling shoreline erosion along the coast of North Carolina and we apply our method in an effort to model the underlying covariance structure and build a complete model for the shoreline erosion process.

Bootstrapping of Spatially Correlated Data

by

Prasheen Agarwal

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

STATISTICS

in the

GRADUATE SCHOOL
at
NC STATE UNIVERSITY
2003

Professor Montserrat Fuentes
Chair of Advisory Committee

Professor Margery Overton
Co-Chair of Advisory Committee

Professor Dennis Boos

Professor B.B. Bhattacharya

Professor David Dickey

Biography

Prasheen K. Agarwal was born in Kharagpur, India on the 20th of September, 1974. He got his Masters of Science degree in Mathematics from the Indian Institute of Technology, Kharagpur in May, 1997. He was accepted for graduate studies in Statistics by the Department of Statistics at North Carolina State University in August, 1997. He finished his Masters in Statistics from North Carolina State University in May 1999. He met his wife, Parul Shah, at North Carolina State University while she was a student of the Masters program in the Department of Parks Recreation and Tourism Management. He got married to Parul on August 17, 2001. He is now finishing work on his dissertation for the doctoral degree in Statistics. He has been accepted by the Bristol-Myers Squibb Pharmaceutical Research Institute as a Research Biostatistician.

Acknowledgements

I would like to thank the members of my advisory committee for their kind support, encouragement and advice during the preparation of my thesis. I am extremely grateful to Dr. Montserrat Fuentes, my advisor, for her enthusiasm and her support and encouragement during every aspect of my education. She has taught me how to think independently. I am also very grateful to her for her advice and help on non thesis related issues. I am extremely grateful to Dr. Margery Overton for her patience and her encouragement during my work on my thesis. I am very thankful to her for her careful reading of my thesis and for her helpful comments and questions that have helped me understand the material better. Dr. Bhattacharya has been a father figure to me during my stay here and he has not only been very helpful in answering my many questions but has also helped me with his kindness and encouragement in overcoming the many personal challenges that I have faced. I am very grateful to Dr. Dickey for giving me his valuable time and help in answering my questions about the thesis and also for his advice and encouragement on other matters. His thorough reading of my thesis and his invaluable comments have helped me iron out the bumps in my thesis. I am very grateful for having had the opportunity to learn from Dr. Boos and interact with him. Dr. Pantula's support and caring attitude have helped me immensely all along, and I am very grateful to him. I am also very thankful to Dr. Swallow for all his help. I would like to thank all my professors here at NC State for all their help, encouragement and guidance over these past 5 years. I would not have achieved my goal without their constant encouragement.

My wife, Parul Agarwal, has been my ”happy haven” ever since I met her. She has encouraged and supported me unwaveringly in my goal to complete my graduate studies. I am very appreciative of her patience while I have worked on my dissertation.

Jimmy Doi, to whom I owe perhaps many a lunches, has given me some of the best advice that I have received. I am very grateful that I had the good fortune to have his friendship.

I am very grateful to my parents for having had the foresight to make the decisions they did, because of which I was able to attend this prestigious university.

My sister, Shamita Martin has always lent me her ear, and I am very grateful to her for listening attentively (I think) to my, sometimes, incessant groaning and complaining.

I am very thankful to Janice Gaddy for never saying “yes” in answer to my question “am I bothering you?”, when I have gone to her for help, and for all her help throughout these 5 years. I am also grateful to Terry Byron for never poking fun of me and for always helping me with all my, though sometimes inane, computer requests! I am also thankful to him and the Statistics Department for making available to us such excellent computing facilities and resources.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The motivation for this research stems from the study of the changes in the shoreline along the coast of Pea Island, North Carolina. Development along the coastal United States has increased over the last few decades. Most of the coastal states have initiated programs to control coastal development. This usually takes the form of coastal construction setbacks. The setbacks are usually based on the amount of erosion anticipated to occur several years into the future. Therefore, accurate determination of erosion is a critical factor in determining the setbacks. Accurate forecasts of coastal erosion are critical for coastal managers and agencies responsible for protecting the coast. Crowell et al. (Crowell, Douglas & Leatherman 1997) compare the "end-point-rate method", linear regression and the MDL criterion method to evaluate the shoreline prediction. The most common approach taken by researchers to calculate the rate of erosion is the end-point-rate method. The rate is calculated as the shoreline movement between the earliest and the most recent positions of the shoreline divided by the time elapsed between the two measurements. This technique is simple to implement and only two shoreline positions suffice for the calculation of the rate. However, this could result in misleading results as there could be aberrations in the measurements of the two end points and this is a big disadvantage of this method.

Another technique used is the simple linear regression on time where a best fitting straight line is fitted to the observed data. This technique is simple to implement and has the advantage over the end-point-rate method of using all the available data and not just the end points.

The technique known as the minimum description length(MDL) criterion was proposed by Fenster (Fenster, Dolan & Elder 1993). This method fits a polynomial in time up to the third degree, by determining whether an increase in the complexity of the model is justified by an increase in the model precision. The results of a study conducted by Crowell et al. indicate that the linear regression technique performs as well as or better than the MDL criterion technique when the forecast date lies well beyond the temporal bounds of the data set. The MDL criterion involves a penalty term that depends upon the number of parameters of the model, that is added to the MSE to give the value for the criterion.

A further generalization of this technique was given by Douglas and Crowell (Douglas & Crowell 2000). They propose fitting a nth degree polynomial in time instead of stopping at the cubic polynomial.

There is a long standing and ongoing debate on the pros and cons of using storm affected erosion data for modeling purposes. Stated in other words: Should storm influenced shorelines be considered as a outliers in shoreline change data sets or do these data contribute to understanding the long term shoreline migration process? Fenster, Dolan and Morton (Fenster, Dolan & Morton 2001) show that storm influenced data points are not outliers and thus they oppose the idea proposed by Honeycutt, Crowell and Douglas (Honeycutt, Crowell & Douglas 2001), that the exclusion of storm influenced data points reduces prediction variability. Honeycutt et al. conclude that under the constraints of the limited data available, the best erosion forecasts will be those derived from linear regression rates using non-storm data. As pointed out by Douglas et al. (Douglas, Crowell & Honeycutt 2002) there are some pertinent issues that render the use of the simple linear regression on time a little tenuous, like the presence of seasonal cycles along with an underlying long term trend, and, the higher variability of winter shoreline positions compared to summer shoreline positions. They further conclude that shoreline positions affected by great storms are very inconsistent with a linear trend model of shoreline retreat for an extended time but, the linear trend model may hold over the long term in many cases, even though it is seriously violated in the short term following a storm.

A significant disadvantage of the simple linear regression technique is that it assumes the rate of erosion to be constant. Large scale changes result from severe tropical and extra tropical storms. During these storms, beach width changes within a short time interval(hours to days) can be much larger than the accumulated erosion over many previous decades. Following a severe storm of long duration, beach recovery can go on for many years. Douglas and Crowell (Douglas & Crowell 2000) warn that it is very important not to consider intervals of a few decades as adequate to characterize shoreline behavior at a site except during that time and perhaps a few years into the future. Long term forecasts in which the underlying trend will dominate must be made using as much high quality and long term data as possible to obtain the best possible estimate of the the underlying trend. Most of the examples in the literature are for very sparse data sets.

Most of the current methods for forecasting and modeling erosion rates use time as the only covariate. None of them use any information available on storms and storm related parameters to model the erosion. In this work we try to incorporate the knowledge of storm related parameters to build a sufficiently detailed model to explain the behaviour of the erosion and further to use this model with other techniques to forecast the erosion. We primarily focus on the coastline along Pea Island, however, this method can be easily applied to other shorelines and other data sets.

Shore line constructions like jetties and groins can fundamentally alter the flow of water and sediment supply and thus affect the erosion/accretion process of the shoreline. A groin was built on the Southern side of Oregon Inlet, NC to stabilize the Northern end of Pea Island and as a direct consequence also protect the Southern end of Herbert C. Bonner bridge.

A groin is a dam like structure, usually a few feet high and about a hundred feet long that is constructed perpendicular to the shoreline. Its objective is to slow down the loss of a beach, widening it by trapping the passing sand. Groins may be made of timber, sheet-steel pilings, stone or concrete and they may be built solid or they could be permeable to sand flow.

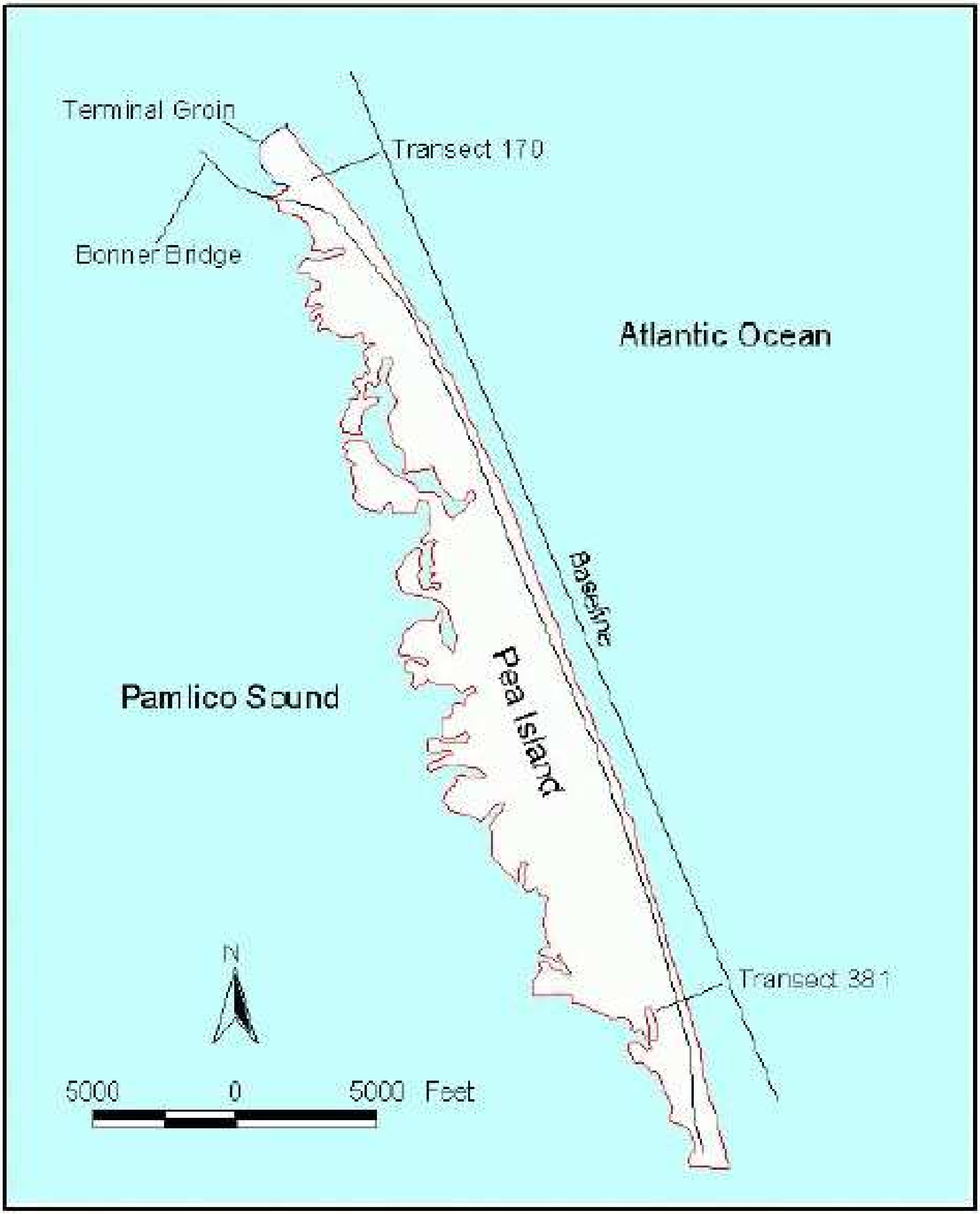
The North Carolina Department of Transportation has collected data using aerial photography and field surveys to determine the changes along the Northern end of

Pea Island, after the construction of the groin. This is an effort to model and predict the nature of the shoreline erosion. Our ultimate goal is to forecast the amount of erosion and thus the position of the shoreline.

The data is in the form of distance from the shoreline to an imaginary baseline. The distance is being measured every 150 ft along the monitored length of the coastline (approximately 7 miles, see Figure 1.1). This baseline is an imaginary line going through the ocean and all distances are measured perpendicular to the baseline. The measurements have been made every 2 months since August, 1991. The shoreline is defined as the boundary between dry sand and wet sand along the beach visible on the aerial photographs. Previous studies have indicated that erosion is mostly caused by storms such as tropical storms or hurricanes. A separate data set containing storm parameters such as wave height, wave time period, tide height etc is also available from the Field Research Facility. This data has been collected at an offshore buoy a few miles into the sea.

The erosion of the shoreline is determined by an increase in the distance of the shoreline to the baseline and a decrease in this distance implies accretion. Here we refer to the two processes of erosion and accretion as shoreline change. Shoreline change could be thought of as being a function of the location on the shoreline and time. We first need to model the spatial and temporal effects and the underlying covariance structure, to be able to forecast the change in the position of the shoreline. We use the non-parametric bootstrap method to obtain estimates and standard errors of the parameters of the covariance structure of the process.

Bootstrap is a data based simulation method for statistical inference, which can be used to compute measures of accuracy of statistical estimators. The basic idea involves sampling with replacement from the original data to produce random samples. Each of these samples is known as a bootstrap sample. From each bootstrap sample an estimate of the parameter of interest is calculated. By repeating this process several times information on the variability of the estimator can be obtained. In this thesis we study resampling methods and propose a new bootstrapping method for spatial data. This is a challenging problem for spatial data as it is difficult to resample and preserve the correlation structure of the data, at the same time. We are also studying


Terminal Groin
Transect 170
Bonner Bridge
Atlantic Ocean
Baseline
Pea Island
Pamlico Sound
N
Transect 381
5000
0
5000
Feet

the properties of the estimators for the mean and the semi-variogram. We also derive asymptotic distributional properties of these estimators.

The Bootstrapping procedure as proposed by Efron is well established as a non-parametric estimator of the variance of a statistic. This method assumes that the observations are independent and identically distributed. This method has been extended by Efron and Tibshirani (Efron & Tibshirani 1993), to observations that are no longer independent but, they require the fitting of a parametric model to the data prior to the application of the bootstrap method. Since, fitting a model may not always be possible this approach would break down for complicated data structures.

To demonstrate the inconsistency of the bootstrap Singh(1981) considered the sample mean $\bar{X_n}$ of $X_1, \ldots, X_n,$. Let $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$ exist and

$$Var(\bar{X_n}) = \frac{\sigma^2}{n} + \frac{2}{n}\sum_{t=1}^{m} Cov(X_1, X_{1+t}) \tag{1.1}$$

then

$$(\bar{X_n} - \mu)/[Var(\bar{X_n})]^{1/2} \to N(0,1) \tag{1.2}$$

If $\bar{X}_n^*$ is the sample mean of the bootstrap sample $X_1^*, \ldots, X_n^*$ that are independent and identically distributed under the empirical distribution function of $X_1, \ldots, X_n$ then

$$Var^*(\bar{X}_n^*) = \frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X_n})^2 \tag{1.3}$$

and

$$(\bar{X}_n^* - \bar{X_n})/[Var^*(\bar{X}_n^*)]^{1/2} \to N(0,1) \tag{1.4}$$

By the Strong Law of Large numbers for m-dependent data it can be shown that

$$\frac{1}{n}\sum_{t=1}^{n}(X_i - \bar{X_n})^2 \to \sigma^2 \tag{1.5}$$

Therefore, the bootstrap variance estimator, $Var^*(\bar{X}_n^*)$, of $\bar{X_n}$ is inconsistent unless

$$\sum_{t=1}^{m} cov(X_1, X_{1+t}) = 0 \tag{1.6}$$

An estimator is consistent when the bias and variance of the estimator tend to 0 as the sample size, n, tends to infinity.

To solve this problem of inconsistency several methods have been proposed and some of them are discussed here in brief.

Carlstein (Carlstein 1986) proposed the use of sub-series of the data and calculating the statistic on each sub-series. Let, there be n data points and the sub- series length be m then it is possible to have $k = \left[\frac{n}{m}\right] - 1$ non-overlapping sub-series or blocks, that preserve the correlation structure of the data. Here, $\left[\frac{n}{m}\right]$ is the greatest integer not greater than $\frac{n}{m}$. Using these replicates of the estimate of the statistic the bootstrap distribution function can be computed. The size of the sub-series and the number of replicates play an important role in the bias and the variance of the estimator. He showed that the variance increases with the number of replicates and the size of the sub-series whereas, the bias decreases with the increasing sub-series size. He concluded that this interaction of the variance and bias should yield an optimal sub-series length for a given n. For the case of a stationary AR(1) sequence he showed that the optimal sub-series length is of the order of $n^{\frac{1}{3}}$.

Kunsch (Kunsch 1989) and Liu and Singh (Liu & Singh 1992) proposed the moving block bootstrap. Kunsch defined the moving block bootstrap as follows. Let there be a sequence $X_1, \ldots, X_N$ of observations from a stationary process. Define blocks of observations taken m at a time as $Y_i = (X_i, \ldots, X_{i+m-1})$, $i = 1 \ldots N - m + 1$. Construct the bootstrap sample by randomly selecting k blocks from $Y_i, i = 1 \ldots N - m + 1$. Now, the bootstrap statistic is constructed based on the empirical distribution function obtained from the bootstrap sample. He also applied a similar procedure to the Jackknife where blocks of m-tuples are down weighted or deleted, such that the down weighted block is of length l. For the Jackknife he showed that the variance of the jackknife estimator of the variance of the sample mean is significantly reduced by using the moving block approach in comparison to using disjoint blocks.

The first few observations and the last few observations have a lesser chance of being selected in the moving block bootstrap procedure. To counter that Politis and Romano (Politis & Romano 1992) suggested the circular block bootstrap. The observations $X_1, \ldots, X_N$ are wrapped on a circle and then blocks are constructed by choosing consecutive observations. Let, $Y_i = (X_i, \ldots, X_{i+m-1})$, $i = 1 \ldots N$ and define $X_{i+m-1} \equiv X_{(i+m-1)mod\ N}$.

Politis and Romano (Politis & Romano 1993) extended Kunsch's approach by suggesting the blocks of blocks resampling approach and it is briefly described as follows. Consider, the sequence $X_1, \ldots, X_N$. Let $B_{i,M,L}$ be the sequence of M consecutive observations starting from $(i-1)L+1$: $X_{(i-1)L+1}, \ldots, X_{(i-1)L+M}$ where, M, L are integer functions of N. Define, $T_{i,M,L} = \phi(B_{i,M,L})$ where $\phi(.)$ is a function from $R^M \to R$. For example, $\phi(B_{i,M,L}) = \frac{(X_{(i-1)L+1}+\ldots+X_{(i-1)L+M})}{M}$.We now have the sequence $T_{i,M,L}$, $i = 1 \ldots Q$, where, $Q = \left[\frac{N-M}{L}\right] + 1$. Now define B_j to be the block of b consecutive $T_{i,M,L}$'s starting from $T_{(j-1)h+1,M,L}$ i.e. $B_j = T_{(j-1)h+1,M,L}, \ldots, T_{(j-1)h+b,M,L}$, $j = 1 \ldots q$, where, $q = \left[\frac{Q-b}{h}\right] + 1$. This is effectively resampling whole blocks (of size b) of blocks (of size M) of the original observations. Equivalently, this can be thought of as resampling bigger blocks of size $(b-1)L + M$ of the observations X_i. They further generalized this procedure from one dimension to n dimensions.

Hall (Hall 1985) suggested resampling methods for spatial patterns that are similar to those described for the single dimension. Here the sampling elements are going to be tiles that contain the pattern observed on a certain area. To construct a bootstrap sample the domain is divided into regions, which could ideally be of any shape but here we will take them to be rectangles of equal area. The pattern to be placed on these sub-regions is selected randomly from a set of possible patterns. There are three ways of doing this- Fixed tiles, Moving tiles and an extension to the moving tiles. For the fixed tile approach the domain is divided into non-overlapping rectangular tiles of equal area such that the area of a tile is equal to the area of a sub-region. The patterns to be placed on the sub-regions are chosen randomly with replacement from this set of tiles. In the moving tile approach instead of partitioning the domain into non-overlapping regions the set of tiles is the set of all possible tiles of a specified area that can be constructed from the domain. Thus, this set of possible patterns is much larger. If the domain were regarded as being a torus and the tiles chosen using the moving tiles approach keeping this in mind then this would be similar to the circular block bootstrap approach.

Another recent approach was given by Brieman (Brieman 1996) as a means for improving the accuracy of estimators of functions $\theta(x)$ of data $x = x_1, x_2, \ldots, x_N$, $\widehat{\theta}(x) = \arg \min\limits_{\theta(x) \in \Theta} L(\theta(x))$.
The procedure is called "**b**ootstrap **agg**regat**ing**" and is referred to using the acronym "bagging". Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The multiple versions are formed by obtaining bootstrap replicates of the data and using these to calculate the bagged estimator. The bagged estimator is calculated by either optimizing the value of $L(\theta(x))$ averaged over the resamples or averaging the resampled values of $\hat{\theta}$. The bagged estimator would be either $\hat{\theta}_{bagg}(x) = \arg \min\limits_{\theta(x) \in \Theta} \dfrac{1}{B} \sum\limits_{b=1}^{B} L(\theta(x_b))$ or $\hat{\theta}_{bagg}(x) = \dfrac{1}{B} \sum\limits_{b=1}^{B} \hat{\theta}(x_b)$, where $\hat{\theta}(x_b)$ is the version of $\hat{\theta}(x)$ computed from the bth bootstrap sample. The objective function $L(\theta(x))$ is a data based estimate of the expected value of some functional such as the negative log-likelihood or some other loss function. The stability of the procedure that is used to construct $L(\theta(x))$ is a critical factor in deciding whether bagging will improve accuracy. Improvements will occur for unstable procedures when a small change in the bootstrap replicate results in a large change in $L(\theta(x))$.

In the bootstrapping procedures outlined earlier it is conceivable that additional bias and variance were introduced into the samples because of placing dissimilar blocks beside each other. To reduce the effect of the boundary between blocks we are smoothing the observations on the boundary of each sub-region. The smoothing is done by recomputing the value of the observations on the boundary as a weighted average of neighbouring observations.

In chapter 2 we present a method to get a better estimate of the semi-variogram using a bootstrap method with smoothing and we study the properties of the proposed estimators for the mean and semi-variogram. We also study asymptotic properties of these estimators.

In chapter 3 results from Monte Carlo simulation studies to study the behaviour of the estimator of the variogram in small samples are presented. Also results from a numerical study conducted to study the behaviour of the mean squared error of the proposed variogram estimator are presented.

In chapter 4 we present an analysis of the erosion data. Here we present a regression analysis, where we model the shoreline change by covariates that are functions of storm parameters and other covariates like the sine and cosine functions to model the periodicity (if present). We also include appropriate terms to model the AR component if present. We analyze the residuals and model them using a parametric model. Analysis of the residuals is important as this gives us valuable information about the underlying covariance structure of the process. For instance if the errors were assumed to be independent and identically distributed when they were actually correlated then this would affect computations of the variance and confidence intervals. The method developed in chapter 2 is used to obtain empirical confidence intervals for the empirical semi-variogram. We then fit a parametric model to the empirical semi-variograms and obtain estimates for the parameters of the fitted parametric model. We compare predictions of the distance of shoreline to baseline from our model, to observed data, and we also simulate the value of the process for time points in the future and compare with results from other techniques in the literature.

To clearly explain the need for studying the residuals and understanding the underlying covariance structure let us consider a simple example that is given by Cressie (Cressie 1993). Let $Z(1), \ldots, Z(n)$ be independent and identically distributed (iid) random variables from the Gaussian distribution with unknown mean μ and known variance σ^2. Then a two sided 95% confidence interval for μ is given by: $\left(\bar{Z} - 1.96\sigma/\sqrt{n}, \bar{Z} + 1.96\sigma/\sqrt{n}\right)$. Now, instead of independent data suppose the data are positively correlated with a correlation that decreases as the separation between the data increases i.e. $Cov(Z(i), Z(j)) = \sigma^2\rho^{|i-j|}$, $i, j = 1, \ldots, n, 0 < \rho < 1$. Based on this the variance of $\bar{Z}$ can be calculated and is given by the following expression.

$$\begin{aligned} Var(\bar{Z}) &= n^{-2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(Z(i), Z(j))\right\} \\ &= \frac{\sigma^2}{n}\left[1 + 2\left\{\rho/(1-\rho)\right\}\left\{1-(1/n)\right\} - 2\left\{\rho/(1-\rho)\right\}^2 (1-\rho(n-1))/n\right] \end{aligned}$$

Based on this variance a two sided 95% confidence interval for μ can be calculated. If for examples' sake, we consider the following values for the parameters, $n = 10$, $\rho = 0.26$ then we can easily see that this confidence interval for μ would be $\left(\bar{Z} - 2.485\sigma/\sqrt{10}, \bar{Z} + 2.485\sigma/\sqrt{10}\right)$ and the confidence interval based on the *iid* assumption would be $\left(\bar{Z} - 1.96\sigma/\sqrt{10}, \bar{Z} + 1.96\sigma/\sqrt{10}\right)$. Thus the *iid* assumption leads to a confidence interval that is too narrow.

Figure 1.2 gives a overview of the work.

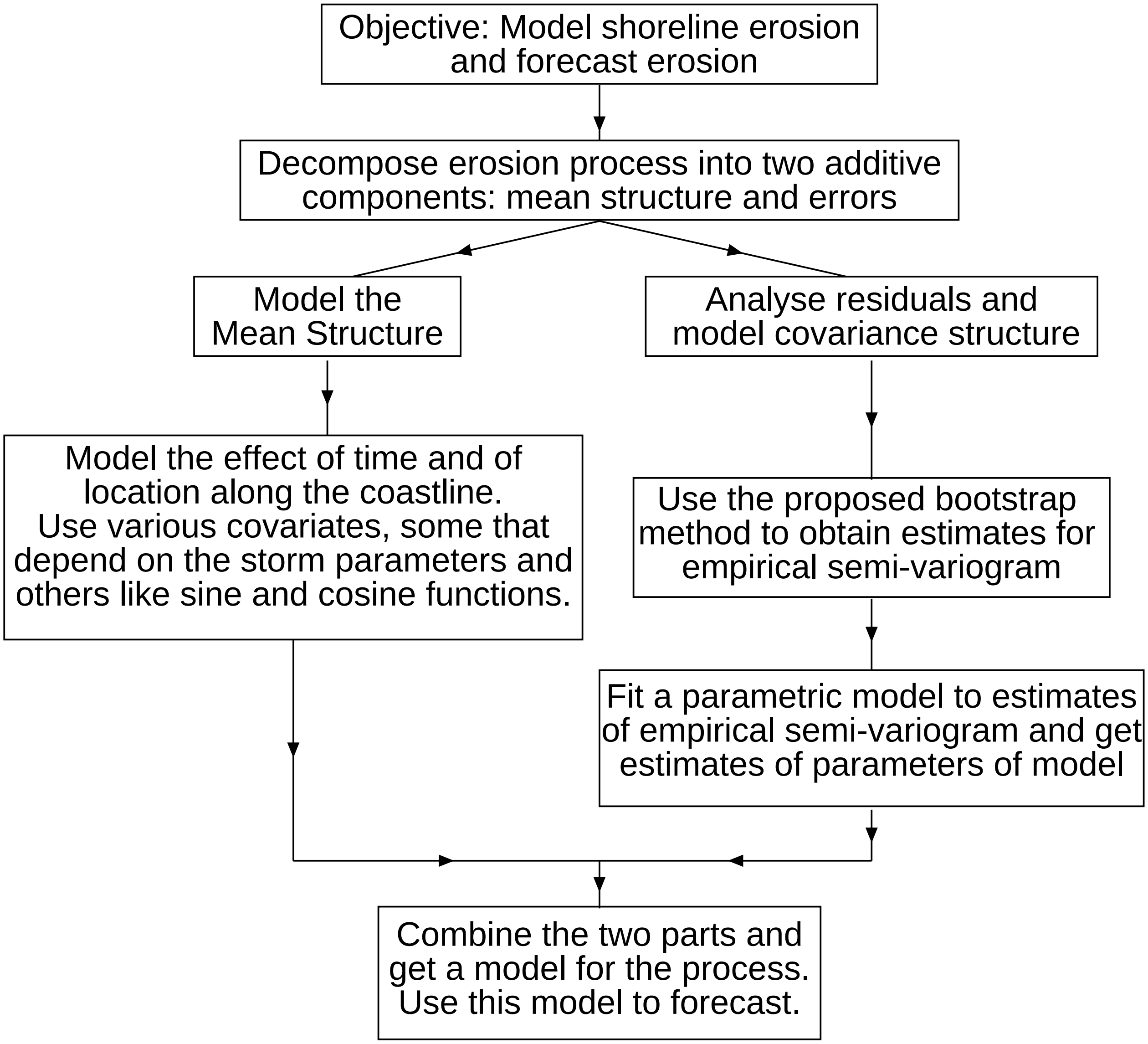

Objective: Model shoreline erosion and forecast erosion
Decompose erosion process into two additive components: mean structure and errors
Model the Mean Structure
Analyse residuals and model covariance structure
Model the effect of time and of location along the coastline. Use various covariates, some that depend on the storm parameters and others like sine and cosine functions.
Use the proposed bootstrap method to obtain estimates for empirical semi-variogram
Fit a parametric model to estimates of empirical semi-variogram and get estimates of parameters of model
Combine the two parts and get a model for the process. Use this model to forecast.

Figure 1.2: Overview

Chapter 2

Proposed Method

2.1 Proposed Method

The method we are proposing is outlined below. Let the area of the region be A and let the number of sub-regions be K, each of area $\frac{A}{K}$. Construct blocks of area $\frac{A}{K}$ from the region and sample K blocks with replacement from the set of all possible blocks that can be constructed. Placing these K blocks on the K sub-regions would give us a manifestation of the bootstrap surface. Now, we smooth the observations by considering a neighbourhood around each observation and computing a weighted average of the observations in this neighbourhood. Figures 2.1 through 2.4 explain the method pictorially. Figure 2.2 shows the construction of blocks. The points referenced by the two arrows indicate the two centers around which the blocks 1 and 2 are constructed. Figure 2.3 shows the construction of the surface using the resampled blocks. Finally, Figure 2.4 shows the smoothing of observations by considering a neighbourhood. The location being smoothed is labeled 's' and its neighbourhood is given by the square labeled $nbd(s)$.

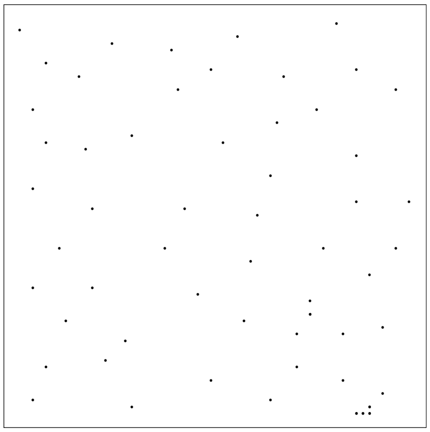


Figure 2.1: Original surface

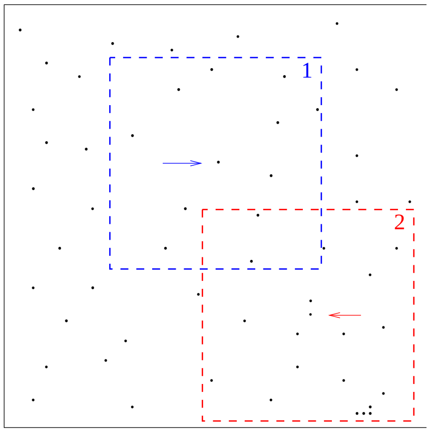


Figure 2.2: Construction of blocks

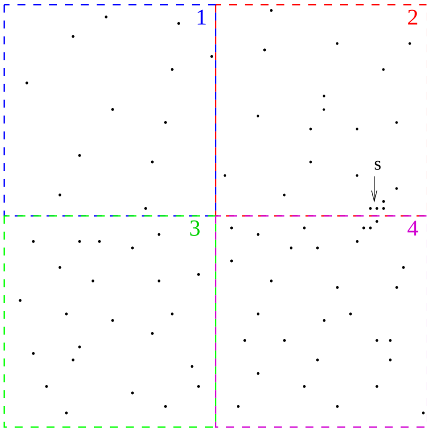


Figure 2.3: Reconstructed surface

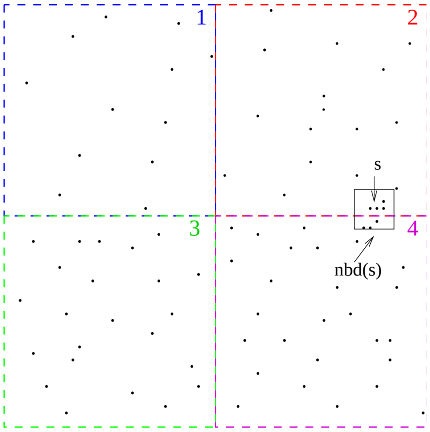


Figure 2.4: Smoothing

The amount of smoothing controls the behaviour of the estimators of statistics along with the size of the sub-regions and the number of sub-regions. We are interested in obtaining the optimal amount of smoothing such that the mean squared error of the estimator is minimized. In the limiting case each neighbourhood contains only one observation and this corresponds to the traditional moving block bootstrap.

Let, $Z(s)$ be the process defined over the domain D with area A where, $s \in D$. Divide the domain into K sub-regions $D_i, i = 1 \dots K$ with area $\frac{A}{K}$ such that $D = \bigcup_{i=1}^{K} D_i$. Select points in D by random sampling with replacement that will serve as the generating centers for the tiles. That is, using these points as the center we construct the tiles of the specified size. In other words we are randomly sampling from the set of all possible tiles that can be constructed from the domain, to maintain the correlation structure of the data. Let there be Q possible tiles of area $\frac{A}{K}$ available from the data.

Let $\{B_i : i = 1 \dots Q\}$ be the set of all possible tiles.

Denote the value of the process $Z(s)$ in tile B_j by $Z_j(s)$ when $s \in B_j$ i.e.,

$$Z(s) = Z_j(\phi_j(s)), s \in B_j, j = 1 \dots Q \tag{2.1}$$

where, $\phi_j(s)$ is a one-one transformation of the position coordinates of s to a frame of reference local to the tile B_j.

Let $Z^*(s), s \in D$ denote the process in the bootstrap sample.

Therefore,

$$\begin{aligned} Z^*(s) &= Z_j^*(\phi_j(s)) \quad \text{in the bootstrap sample in sub-region j} \\ &= Z_m(\phi_m(F_m(s))) \quad \text{for some m in the original sample} \\ &= Z(F_m(s)) \end{aligned}$$

Here, $F_m(.)$ is a function that maps the location s in the bootstrap sample to the location $F_m(s)$ in the original sample, i.e., the value of the process that is observed at the location s in the bootstrap sample is in reality observed at location $F_m(s)$ in the original sample.

Define,

$$Q_j^\star(s') = \begin{cases} Z_j^*(\phi_j(s')) & s' \in D_j \\ 0 & otherwise \end{cases}$$

$$Q_j^\star(s') = \begin{cases} Z_m(\phi_m(F_m(s'))) & s' \in D_j \\ 0 & otherwise \end{cases} \tag{2.2}$$

$$W_j^\star(s', s) = \begin{cases} w_j^\star(\phi_j(s'), \phi_j(s)) & s' \in \{D_j \cap nbd(s)\} \\ 0 & otherwise \end{cases} \tag{2.3}$$

where, $w_j^\star(\phi_j(s'), \phi_j(s))$ is a weighting function that depends upon the relative distance of s' from s. For example, the weighting function could be of the simple form: $w_j^\star(\phi_j(s'), \phi_j(s)) = \frac{1}{\phi_j(s') - \phi_j(s)}$

Further, we also also impose the condition

$$\sum_{s' \in D} \sum_{j=1}^{K} W_j^\star(s', s) = 1 \tag{2.4}$$

where, $nbd(s)$ is defined as a square with s as the centroid and side of length r. Suppose s is a location in sub-region m then, $r = r(\phi_m(s))$ for some m, $1 \leq m \leq K$. Here, r could be an increasing function of the distance of s from the center of the sub-region m, i.e., as the location s becomes closer to the boundary of the sub-region m, the size of the neighbourhood of s becomes larger.

Thus, the smoothed process given by $Z^\star(s)$, at a particular location s can be expressed as a convolution of $Q_j^*(s)$ and $W_j^*(s', s)$

$$Z^\star(s) = \sum_{s' \in D} \sum_{j=1}^{K} W_j^\star(s', s) \times Q_j^\star(s') \tag{2.5}$$

2.2 Results on Mean

In this section we present some properties of the proposed estimator of the sample mean. Propositions 2.2.1 and 2.2.2 deal with the first two moments of the proposed estimator of the sample mean. In proposition 2.2.3 we provide the bootstrap approximation to the sampling distribution of the sample mean.

2.2.1 Definitions and Notation

Here we provide some definitions and the notation that we will be using in the rest of this section.

Let Z(s) be a process defined on a two dimensional regular lattice of size $n \times n$. Let $E_{i,M,L}$ be the square consisting of the points $s = (t_1, t_2)$ such that, $(i_k - 1)l_k + 1 \leq t_k \leq (i_k - 1)l_k + m_k$, k=1,2. Here i, M and L are vectors where, $i = (i_1, i_2)$, $M = (m_1, m_2)$ and $L = (l_1, l_2)$. Here, $i_1, i_2, m_1, m_2, a, l_1, l_2, t_1, t_2$ and n can take only integer values. Let $l_k = 1$ and $m_k = a$ for k=1,2 then $i_k \leq t_k \leq i_k + a - 1$, k=1,2 and, if $t_k \geq n$ then $t_k = t_k(mod)n$. This means that each block has $a \times a$ points, and $i_k = 1 \ldots Q_k$ where, $Q_k = n$, k=1,2. Now, let $B_{i,M,L} = \{Z(s) : s \in E_{i,M,L}\}$=$\{Z_{i,M,L}(s), s \in E_{i,M,L}\}$. Once we have the $B_{i,M,L}$ which, is a square shaped tile, then, we can denote by $s_{uv}^{(i,M,L)}$, $u = 1 \ldots a$, $v = 1 \ldots a$ the location characterized by (u,v) in the tile $E_{i,M,L}$. The total number of the $B_{i,m,L}$ blocks available from the data is $Q_1 \times Q_2 = n^2 = N$. The weights, $w_j^*(.,.)$, are a function only of the locations and the distance between them and not of the observed data values at those locations. The weights are being calculated as if the surface were a torus, as this allows us to maintain the symmetry in the weights. K blocks (or tiles) are selected randomly with replacement from the set of all possible blocks $\{B_{i,M,L}, i = (i_1, i_2), i_1 = 1 \ldots n, , i_2 = 1 \ldots n\}$ and then placed on the K sub-regions, $D_j, j = 1 \ldots K$.

Define,

$$T_{i,M,L} = b_{11} Z_{i,M,L}(s_{11}^{(i,M,L)}) + b_{12} Z_{i,M,L}(s_{12}^{(i,M,L)}) + \ldots + b_{aa} Z_{i,M,L}(s_{aa}^{(i,M,L)}) \quad (2.6)$$

$T_{i,M,L}$ is a function of observations just in the block $B_{i,M,L}$.

Now,

$$b_{ij} = \frac{1}{a^2} \sum_{s \in D} w_p^*(s_{ij}^p, s) I(\text{nbd(s) contains } s_{ij}^p \text{ and } s_{ij}^p \in D_p) \quad (2.7)$$

and

$$\sum_{s \in D} \sum_{s_{ij}^p} \sum_{p=1}^{K} w_p^*(s_{ij}^p, s) I(s_{ij}^p \in nbd(s) \cap D_p) = N \quad (2.8)$$

The b_{ij}'s are the weights at a location s_{ij}^p in sub-region p. The summation is taken over all the locations s whose neighbourhoods contain s_{ij}^p. Due to the regularity of the lattice and the symmetry and regularity of the tiles the b_{ij}'s do not depend on which sub-region is considered, and, so if $T_{i,M,L}$ is viewed as a transformation of the block $B_{i,M,L}$ then this transformation is independent of i and the sub-region. Any sub-region p has a^2 locations characterized by $s_{ij}^p = (i,j)$ and the superscript is used to denote the sub-region.

Define, blocks of $T_{i,M,L}$ as $\mathcal{B}_j$=$\{T_{i,M,L}, i \in E_{j,b,h}\}$, where, $E_{j,b,h}$ is the rectangle consisting of the points $(j_k - 1)h_k + 1 \leq i_k \leq (j_k - 1)h_k + b_k$, $j_k = 1 \dots q_k$, k=1,2 and $j = (j_1, j_2)$, $b = (b_1, b_2)$ and $h = (h_1, h_2)$. Let $h_1 = h_2 = 1$ then $j_k \leq i_k \leq j_k + (b_k - 1)$, $j_k = 1 \dots q_k$, k=1,2. Let $\bar{T^*}$ be the average of all the $T_{i,M,L}$'s found in the resampled blocks. Therefore, $\bar{T^*}$ is the average of $K'(= lb)$, $T_{i,M,L}$'s, where, l=number of blocks of $T_{i,M,L}$ and $b = b_1 b_2$ is the number of $T_{i,M,L}$ in each block.

Define,

$$T_i^* = b_{11} Z_i^*(s_{11}^{(i)}) + b_{12} Z_i^*(s_{12}^{(i)}) + \ldots + b_{aa} Z_i^*(s_{aa}^{(i)}) \tag{2.9}$$

,$i = 1 \dots K'$.

then,

$$\bar{T^*} = \frac{1}{K'} \sum_{i=1}^{K'} T_i^* \tag{2.10}$$

Definition 2.2.1.1 (m-dependence) *The random variables in the family $\{Z(s), s \in D\}$ are said to be m-dependent if for all subsets C_1 and C_2 of D for which $d(C_1, C_2) > m$, $\{Z(a), a \in C_1\}$ and $\{Z(b), b \in C_2\}$ are independent families. Here, $d(C_1, C_2)$=$inf\{||a - b||,\ a \in C_1,\ b \in C_2\}$ and $||a|| = max\{|a_1|, |a_2|\}$*

Definition 2.2.1.2 (Second order stationarity) *A random process Z(s),$s \in D$ is said to be second order stationary(weakly stationary or stationary in the wide sense) if*

a) $E(Z(s)) = \mu$, for all $s \in D$

b) $Cov(Z(s), Z(t)) = C(s - t)$, for all $s, t \in D$.

The function C is known as the stationary covariance function.

2.2.2 Assumptions

In the following section we make frequent use of these assumptions. We state the assumptions here for quick reference and we justify the need of these assumptions as needed in the following results.

A0 : Z(s) is a Gaussian, second order stationary and m-dependent process. $E\left[Z(s)\right] = \mu$ and $Var\left[Z(s)\right] = \sigma^2$.

A1 : $b \rightarrow \infty$

The assumption [A0] simplifies all calculations involved. Assumption [A1] simply stated implies that the number of $T_{i,M,L}$'s in a resampled block goes to infinity.

2.2.3 Main Results

In this section the main results are outlined and the proofs are presented in the appendix. The main results presented in this section are the bootstrap approximation of the sampling distribution of the sample mean and also results regarding the variance of the bootstrap estimator.

Lemma 2.2.1 *Under the assumption A0* $E\left[T_{i,M,L}\right] = \mu$ *and* $E\left[\frac{1}{N}\sum_i T_{i,M,L}\right] = \mu$.

The next two results are useful when showing the result on the bootstrap approximation of the sampling distribution of the sample mean that is given later in this section.

Lemma 2.2.2 *Assuming A0 and if p,* δ*, C are constants such that p is an integer greater than 2,* $0 < \delta \leq 2$ *and* $C > 0$ *then for any M*
$E\left[|T_{i,M,L}|^{2p+\delta}\right] < C$.

Lemma 2.2.3 *Under the assumption A0, and using lemma 2.2.1 and 2.2.2 we have that*

$$\sqrt{N}\left(\bar{T} - E\left[\bar{T}\right]\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_\infty) \tag{2.11}$$

where, $\lim_{N\rightarrow\infty} Var\left(N^{\frac{1}{2}}\bar{T}\right) = \sigma^2_\infty$.

The following is a result regarding the variance of the sample mean for m-dependent data that was established by Rosén(Rosén 1968).

Lemma 2.2.4 *Under the assumption A0, Rosén(Rosén 1968) showed that*

$$Var\left[\bar{Z}\right] = \frac{1}{N} \sum_{\substack{\boldsymbol{t}=(t_1,t_2) \\ |t_1| \le m, |t_2| \le m}} \sigma(\boldsymbol{t}) \left(1 - \frac{|t_1|}{n}\right)^{+} \left(1 - \frac{|t_2|}{n}\right)^{+} \tag{2.12}$$

where, $\sigma(\boldsymbol{t}) = Cov(Z(\boldsymbol{s}), Z(\boldsymbol{s}+\boldsymbol{t}))$ *and for i=1,2*

$$\left(1 - \frac{t_i}{n}\right)^{+} = \begin{cases} 0, & \left(1 - \frac{t_i}{n}\right) < 0, \\ \left(1 - \frac{t_i}{n}\right), & otherwise \end{cases}$$

Proposition 2.2.1 *Under the assumption A0 it can be shown that* $E\left[Z^{\star}(s)\right] = \mu$ *and* $E\left[\bar{Z}^{\star}\right] = \mu$

Proposition 2.2.2 *It can be shown that*

$$Var\left[\bar{Z}^{\star}\right] = O\left(\left(1 - \frac{1}{K'}\right) Var\left[\bar{Z}\right]\right) \tag{2.13}$$

Using the previously proven results and using the results proven by Politis and Romano we give here a result regarding the bootstrap approximation of the sampling distribution of the sample mean. To be able to use Politis and Romanos result we need to make two additional assumptions: [A2]: $b = o(N^{\frac{2}{3}})$ and [A3]: $b = o(K')$

Proposition 2.2.3 *Under the assumptions A0-A3 and using lemma2.2.1 - lemma2.2.3 the following result holds. Here,* E^{*} *and* Var^{*} *are the expectation and variance under the resampling probability* P^{*}.

$$\sup_{x} \left| P^{*}\left(\sqrt{K'} \frac{\bar{T}^{*} - E^{*}\bar{T}^{*}}{\sqrt{Var^{*}\left(\sqrt{K'}\bar{T}^{*}\right)}} \le x\right) - P\left(\sqrt{N}\frac{\bar{Z} - \mu}{\sigma_{\infty}} \le x\right) \right| \xrightarrow{p} 0 \tag{2.14}$$

2.3 Variogram

In this section we present some properties of the proposed estimator of the semi-variogram. We begin with some basic results that we will use later to prove other results. The main results that we establish in this section are the order of the first two moments of the proposed estimator of the variogram. We use these to calculate the mean squared error of the proposed variogram estimator. We want to be able to identify the optimal amount of smoothing so as to minimize the mean squared error of the estimator, and in this connection we present results of a numerical study to numerically find the optimal values for the smoothing parameter that minimize the mean squared error in chapter 4. Here we also establish some results for the traditional bootstrap estimator for the variogram. We make use of the results already proved by Politis and Romano in their 1993 paper.

2.3.1 Definitions and Notation

Let $E_{i,M,L}$ be the rectangular lattice consisting of the points $s = (t_1, t_2)$ such that, $(i_k - 1)l_k + 1 \leq t_k \leq (i_k - 1)l_k + m_k$, k=1,2. Let $l_k = 1$ and $m_k = h_k + 1$ for k=1,2 then $i_k \leq t_k \leq i_k + h_k$, k=1,2. This means that each block has $(h_1 + 1) \times (h_2 + 1)$ points, and $i_k = 1 \ldots Q_k$ where, $Q_k = n - h_k$, k=1,2. Note here that $\boldsymbol{i}$, $\boldsymbol{M}$,$\boldsymbol{L}$ and $\boldsymbol{h}$ are vectors where, $M = (m_1, m_2)$, $\boldsymbol{h}$=(h_1, h_2), $L = (l_1, l_2) = (1, 1)$ and $i = (i_1, i_2)$. So, $B_{i,M,L} = \{Z(s) : s \in E_{i,M,L}\}$=$\{Z_{i,M,L}(s), s \in E_{i,M,L}\}$. The total number of the $B_{i,m,L}$ blocks available from the data is $Q_1 \times Q_2 = |N(h)|$, where $|N(h)|$ is the number of pairs in the set N(h)=$\{(s,t) : s - t = \boldsymbol{h}\}$. Define, blocks of $T_{i,M,L}$ as $\mathcal{B}_j$=$\{T_{i,M,L}, i \in E_{j,b,g}\}$, where, $E_{j,b,g}$ is the rectangle consisting of the points $(j_k - 1)g_k + 1 \leq i_k \leq (j_k - 1)g_k + b_k$, $i_k = i_k(mod)Q_k$, k=1,2 and $j = (j_1, j_2)$, $b = (b_1, b_2)$ and $g = (g_1, g_2)$. Let $g_1 = g_2 = 1$ then $j_k \leq i_k \leq j_k + (b_k - 1)$, $i_k = i_k(mod)Q_k$, k=1,2. The blocks of $T_{i,M,L}$, $\mathcal{B}_j$, are constructed by wrapping the $T_{i,M,L}$'s on a torus and then constructing rectangles of size b_1 by b_2. Each block $\mathcal{B}_j$ has b=$b_1 b_2$, $T_{i,M,L}$'s.

Let, $T_{i,M,L} = (Z_{i,M,L}(s) - Z_{i,M,L}(s + \boldsymbol{h}))^2$, where, s and s+$\boldsymbol{h}$ both are in $E_{i,M,L}$.Then $\bar{T}$ =sample variogram for lag $\boldsymbol{h}$=(h_1, h_2).

2.3.2 Assumptions

In the following section we make frequent use of these assumptions. The assumptions are stated here for easy reference however, we justify the need for these assumptions as we go along.

B0 : Z(s) is a Gaussian, second order stationary and m-dependent process. $E\left[Z(s)\right] = \mu$ and $Var\left[Z(s)\right] = \sigma^2$.

B1 : $b \to \infty$

B2 : bK$\approx |N(h)|$

The assumption [B0] simplifies calculations. Further if a process is from a normal distribution and is weakly stationary then together these imply that it is strongly stationary. Strong stationarity is needed for some of the results that are presented in the following section. The assumption [B1] implies that the number of $T_{i,M,L}$'s in the resampled blocks goes to infinity, i.e. the block size goes to infinity.

2.3.3 Main Results

The first two results are properties of jointly normally distributed random variables that come in handy when calculating higher order moments. The first result follows easily from the second as can be seen in the proof of the results.

Lemma 2.3.1 *Let X and Y be jointly normally distributed random variables with zero means then,* $Cov\left[X^2, Y^2\right] = 2\left\{Cov\left(X, Y\right)\right\}^2$

Lemma 2.3.2 *Let* X_1, X_2, X_3 *and* X_4 *be jointly normally distributed random variables with zero means then,*
$E\left[X_1X_2X_3X_4\right] = E\left[X_1X_2\right]E\left[X_3X_4\right] + E\left[X_1X_3\right]E\left[X_2X_4\right] + E\left[X_1X_4\right]E\left[X_2X_3\right]$

Lemma 2.3.3 *Assuming B0 and if p, δ, C are constants such that p is an integer greater than 2,* $0 < \delta \leq 2$ *and* $C > 0$ *then for any M*
$E\left[|T_{i,M,L}|^{2p+\delta}\right] < C$.

The following result is similar to a result that was shown for the sample mean in the previous section. This result is useful in establishing the bootstrap approximation to the sampling distribution of the traditional block bootstrap variogram estimator, that is given later in this section.

Lemma 2.3.4 *Under the assumption B0, and using lemma 2.3.3 we have that*

$$\sqrt{|N(h)|}\left(\bar{T} - E\left[\bar{T}\right]\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_\infty) \tag{2.15}$$

where, $\lim\limits_{|N(h)|\to\infty} Var\left(|N(h)|^{\frac{1}{2}}\bar{T}\right) = \sigma^2_\infty.$

Sample with replacement from the set $\{\mathcal{B}_j\}$ and get K blocks $\mathcal{B}_j$. Then let $\bar{T}^*$ be the average of all the $l = Kb$ observations $T_{i,M,L}$'s found in the resampled blocks. Let E^* and Var^* denote the expectation and variance under the resampling probability P^*. We know that

$$\begin{aligned} 2\widehat{\gamma^*_{ns}(\boldsymbol{h})} &= \frac{1}{|N(h)|}\sum_{N(h)}\left(Z^*(s) - Z^*(t)\right) \\ &= \frac{1}{|N(h)|}\left[\sum_{N(h)}\sum_{i=1}^{K}\left(Z^*_i(s) - Z^*_i(t)\right)^2 I_{(s,t)\in D_i}\right. \\ &\quad \left. + \sum_{N(h)}\sum_{i=1,i}^{K}\sum_{\neq j=1}^{K}\left(Z^*_i(s) - Z^*_j(t)\right)^2 I_{(s\in D_i)\text{ and } t\in D_j)}\right] \end{aligned} \tag{2.16}$$

where, N(h)≡ $\{(s,t) : s - t = \ \boldsymbol{h}\ \}$. The data are on a regular lattice and all the tiles and the sub-regions have the same shape and size and b=number of pairs (s,t) in a sub-region D_i such that $s - t = \ \boldsymbol{h}$ (for a fixed lag $\boldsymbol{h}$)=number of $T_{i,M,L}$'s in the resampled block $\mathcal{B}_j$. Then,

$$2\widehat{\gamma^*_{ns}(\boldsymbol{h})} = \frac{1}{|N(h)|}\left[bK\bar{T}^* + \sum_{N(h)}\sum_{i=1,i}^{K}\sum_{\neq j=1}^{K}\left(Z^*_i(s) - Z^*_j(t)\right)^2 I_{(s\in D_i)\text{ and } (t\in D_j)}\right] \tag{2.17}$$

Now, if we let h be small in comparison to the size of the block such that the number of terms where $s \in D_i$ and $t \in D_j$ for some i,j and $i \neq j$ is negligible compared to

$|N(\mathrm{h})|$, that is, in other words, bK$\approx |N(\mathrm{h})|$ then we can say that

$$2\widehat{\gamma^*_{ns}(\boldsymbol{h})} \approx \bar{T}^* \tag{2.18}$$

For the result of the following proposition to hold we need to make two additional assumptions: [**B3**]: $b = o(|N(h)|^{\frac{2}{3}})$ and [**B4**]: $b = o(l)$. These two assumptions are needed to satisfy the requirements of a result proved by Politis and Romano (Politis & Romano 1993), that we use in showing the following result.

Proposition 2.3.1 *Under the assumptions B0-B4 and using lemmas 2.3.3 and 2.3.4*

$$\sup_x \left| P^*\left\{\sqrt{l}\frac{\bar{T}^* - E^*\bar{T}^*}{Var^*(\sqrt{l}\bar{T}^*)} \le x\right\} - P\left\{\sqrt{|N(h)|}\frac{\bar{T} - 2\gamma(h)}{\sigma_\infty} \le x\right\}\right| \xrightarrow{p} 0 \tag{2.19}$$

Proposition 2.3.2 *Under the assumption B0*

$$O\left(\frac{Var(\widehat{\gamma^*_s(\boldsymbol{h})})}{Var(\widehat{\gamma^*_{ns}(\boldsymbol{h})})}\right) = 1 \tag{2.20}$$

With the intent of finding an optimal value for the smoothing parameter we now study the mean squared error of the proposed variogram estimator. We are interested in finding an expression for the mean squared error for our proposed variogram estimator. To do this we calculate the bias and the variance.

Construct the rectangle $E_{\boldsymbol{i,M,L}}$ consisting of the points $s = (s_1, s_2)$ such that $(i_k - 1)l_k + 1 \le s_k \le (i_k - 1)l_k + m_k$, and if $s_k > n$ then $s_k = s_k(mod)n$, k=1,2. This means that each rectangular tile has $p = m_1 \times m_2$ points and $i_k = 1 \ldots Q_k$, $Q_k = n$. Here, $\boldsymbol{M}$=(m_1, m_2), and $\boldsymbol{L}$=$(l_1, l_2) = (1, 1)$. The total number of rectangular tiles that can be constructed are $N = n \times n$. We sample with replacement from these N tiles and select K tiles to place on the K sub-regions and reconstruct the surface. We make some assumptions to simplify the calculations and they are stated below.

[**B5**]: We are considering pairs of locations $(s, t) \in N(h)$ such that the same sub-regions contain the neighbourhoods of s and t and further the neighbourhoods are contained in two sub-regions. That is we are ignoring those pairs of location (s, t) where s and t belong to two different sub-regions or if the total number of sub-regions spanned by the neighbourhoods of s and of t is greater than 2. If we let b be the number of pairs (s, t) that belong to $N(h)$ and are in some sub-region i then, $bK \approx |N(h)|$. Since, all the sub-regions are of the same size they contain the same number of pairs (s, t).

[B6]: $Z(s)$ is a gaussian process with mean 0 and variance σ^2, $Z(s)$ is m-dependent and second order stationary.

[B7]: The number of tiles, $N = n \times n$, that can be constructed from the domain, goes to infinity.

[B8]: The sub-regions are of regular shape and size. The data is on a regular lattice.

Assumption [B5] is a tenable assumption since we are interested in short lags. If the lag is assumed to be small compared to the dimension of the sub-regions then this assumption would be ok. For short lags the number of pairs that we are ignoring is going to be negligible compared to the total number of pairs $|N(h)|$.

Assumptions [B6] and [B8] help in reducing the complexity of the calculations. The total number of points in the domain is $N = n \times n$ and this is equal to Kp, where K is the number of sub-regions that the domain can be sub-divided into and p is the number of points in a sub-region.

Proposition 2.3.3 *Using the set up described above and the assumptions [B5]-[B8] we have the following results for the bias and the variance of the proposed estimator of the variogram.*

$$Bias \;=\; \frac{1}{|N(h)|} \sum_{(s,t)\in N(h)} \big(\sigma_1^2(s,t) + \sigma_2^2(s,t)\big) - 2\gamma(h) + o(\frac{1}{Kp}) \tag{2.21}$$

$$\begin{aligned} Var(2\widehat{\gamma_s(h)}) &= \frac{2}{|N(h)|^2}\sum_{(s,t)\in N(h)}\left(\sigma_1^2(s,t)+\sigma_2^2(s,t)\right)^2 + \\ &\quad \frac{1}{|N(h)|^2}\sum_{\substack{(s,t)\neq(u,v)\\(s,t)\in N(h)\\(u,v)\in N(h)}} H(s,t,u,v)+o(\frac{1}{K^2p^2})+o(\frac{1}{K^3p^3}) \end{aligned} \quad (2.22)$$

where, $\sigma_1^2(s,t)$, $\sigma_2^2(s,t)$ are given in equations (2.84) and (2.88) respectively, and $H(s,t,u,v)$ is given by equations (2.110) or (2.111).

The bias expression in (2.21) can be written as :

$$Bias = \frac{1}{|N(h)|}\sum_{(s,t)\in N(h)}\left(\sigma_1^2(s,t)-2\gamma(h)\right)+\frac{1}{|N(h)|}\sum_{(s,t)\in N(h)}\sigma_2^2(s,t)+o(\frac{1}{Kp})$$

$\sigma_2^2(s,t)$ is non-zero only for those pairs $(s,t)\in N(h)$ for which either s or t or both are on the boundary and are smoothed. If we let ρ be the percentage of smoothing and let $a(\rho)(\subset N(h))$ denote this set of pairs then,

$$Bias = \frac{1}{|N(h)|}\sum_{(s,t)\in N(h)}\left(\sigma_1^2(s,t)-2\gamma(h)\right)+\frac{1}{|N(h)|}\sum_{(s,t)\in a(\rho)}\sigma_2^2(s,t)+o(\frac{1}{Kp})$$

The above bias is of the order of $O(1)+O(\rho)+o\left(\frac{1}{Kp}\right)$.

Consider, the j^{th} sub-region in the bootstrap sample. For a location s in this sub-region the value of the process at s is given by $Z_j^\star(s)=\sum_{s'\in D}\sum_{i=1}^{K}Z_i^*(s')w_i^\star(s',s)I_i(s',s)$, where,

$$I_j(s',s)=\begin{cases}1, & s'\in nbd(s)\cap D_j,\\ 0, & otherwise\end{cases}$$

$Z_j^*(s)$ can be expressed as being equal to

$$\sum_{s'\in D}Z_j^*(s')w_j^\star(s',s)I_j(s',s)+\sum_{s'\in D}\sum_{\substack{i=1\\i\neq j}}^{K}Z_i^*(s')w_i^\star(s',s)I_i(s',s).$$

Let, $|n(h)|$ be the number of pairs $(s,t)\in N(h)$ that lie in the j^{th} sub-region. Considering, only those locations that are in the j^{th} sub-region we suggest a correction term for the bias as follows.

Denote, the sum $\sum_{s'\in D} Z_j^*(s')w_j^\star(s',s)I_j(s',s)$ by $Z_j^\dagger(s)$. We propose the correction term $E\left\{\frac{1}{|n(h)|}\sum_{(s,t)\in n(h)}\left[Z_j^\dagger(s)-Z_j^\dagger(t)\right]^2\right\}-2\gamma(h)$ and this term depends only on the value of the process in sub-region j. Since, the sub-regions are symmetrical and identical and $K|n(h)|=|N(h)|$ the above correction term is equivalent to $E\left\{\frac{1}{|N(h)|}\sum_{(s,t)\in N(h)}\left[Z_j^\dagger(s)-Z_j^\dagger(t)\right]^2\right\}-2\gamma(h)$. This correction term is subtracted from the bias to obtain the adjusted bias.

The need for a correction term to adjust the bias arises as the bias of the proposed variogram estimator dominates the variance and thus controls the behaviour of the mean squared error. We shall study the behaviour of the mean squared error as a function of the amount of smoothing using a numerical study, to further illustrate this, and we present the results of the numerical study in the following chapter.

Using this correction term the adjusted bias is easily calculated as follows.

$$\begin{aligned} Bias &= \frac{1}{|N(h)|}E\left\{\sum_{(s,t)\in N(h)}\left[(Z^\star(s)-Z^\star(t))^2\right]\right\}- \\ &\quad \frac{1}{|N(h)|}E\left\{\sum_{(s,t)\in N(h)}\left[Z_j^\dagger(s)-Z_j^\dagger(t)\right]^2\right\} \end{aligned} \tag{2.23}$$

The second term $E\left\{\left[Z_j^\dagger(s)-Z_j^\dagger(t)\right]^2\right\}$ is equal to $\sigma_1^2(s,t)$ or $\sigma_2^2(s,t)$ depending upon whether the pair (s,t) is in the sub-region j_1 or j_2. If we assume as before that the sub-region j_1 contains the locations s and t then the above expectation is equal to $\sigma_1^2(s,t)$ where, $\sigma_1^2(s,t)$ is given in equation (2.84).

Therefore, the adjusted bias becomes equal to

$$\frac{1}{|N(h)|}\sum_{(s,t)\in N(h)}\sigma_2^2(s,t)+o\left(\frac{1}{Kp}\right) \tag{2.24}$$

The adjusted bias given in (2.24) is of the order of $O(\rho)+o\left(\frac{1}{N}\right)$. So as the smoothing increases the bias increases, and for no smoothing the bias is zero. This is in agreement with our understanding of the process as the existing estimator of the variogram is an unbiased estimator. The variance given by the expression in (2.22) can be written as:

$$\frac{2}{|N(h)|^2} \sum_{(s,t)\in N(h)-a(\rho)} \left(\sigma_1^2(s,t)+\sigma_2^2(s,t)\right)^2 + \frac{2}{|N(h)|^2} \sum_{(s,t)\in a(\rho)} \left(\sigma_1^2(s,t)+\sigma_2^2(s,t)\right)^2$$
$$+\frac{1}{|N(h)|^2} \sum_{\substack{(s,t)\neq(u,v)\\(s,t)\in N(h)\\(u,v) in N(h)}} H(s,t,u,v) + o\left(\frac{1}{K^2p^2}\right) + o\left(\frac{1}{K^3p^3}\right) \tag{2.25}$$

The variance decreases as the smoothing increases. If we look at the expression in (2.25) we see that as smoothing increases the number of terms in the set $N(h)-a(\rho)$ decrease and the number of terms in the set $a(\rho)$ increase. When $(s,t)\in N(h)-a(\rho)$ then $\sigma_1^2(s,t)+\sigma_2^2(s,t) = 2C(0)-2C(h)$ and when $(s,t)\in a(\rho)$ then $\sigma_1^2(s,t)+\sigma_2^2(s,t) < 2C(0)-2C(h)$. This implies that the variance decreases as the smoothing increases. Using the expressions for the adjusted bias and the variance we can compute the mean squared error.

$$\begin{aligned} \text{MSE} &= \text{Bias}^2 + \text{Variance} \\ &= \left\{\frac{1}{|N(h)|} \sum_{(s,t)\in N(h)} \sigma_2^2(s,t)\right\}^2 + Var(2\widehat{\gamma_s(h)}) \end{aligned} \tag{2.26}$$

where, the expression for $Var(2\widehat{\gamma_s(h)})$ is given in equation (2.22).

2.4 Appendix

Proof of lemma 2.2.1:

$$b_{ij} = \frac{1}{a^2}\sum_{s\in D} w_p^*(s_{ij}^p, s)I(\text{nbd(s) contains } s_{ij}^p \text{ and } s_{ij}^p \in D_p) \qquad (2.27)$$

We know that the sum of the weights over all locations in a neighbourhood equals 1. Thus the sum of the weights over all the locations in the domain equals the number of locations. We can write this as follows:

$$\sum_{s\in D}\sum_{s_{ij}^p}\sum_{p=1}^{K} w_p^*(s_{ij}^p, s)I(s_{ij}^p \in nbd(s)\cap D_p) = N \qquad (2.28)$$

Now,

$$\begin{aligned} I(s_{ij}^p \in nbd(s)\cap D_p) &= I(s_{ij}^p \in nbd(s))I(s_{ij}^p \in D_p) \\ &= I(\text{nbd(s) contains } s_{ij}^p)I(s_{ij}^p \in D_p) \qquad (2.29) \end{aligned}$$

Using equations 2.28 and 2.29 we can work through the algebra as shown in the following steps Then,

$$\begin{aligned} N &= \sum_{s\in D}\sum_{s_{ij}^p}\sum_{p=1}^{K} w_p^*(s_{ij}^p, s)I(s_{ij}^p \in nbd(s)\cap D_p) \\ &= \sum_{s\in D}\sum_{s_{ij}^p}\sum_{p=1}^{K} w_p^*(s_{ij}^p, s)I(\text{nbd(s) contains } s_{ij}^p)I(s_{ij}^p \in D_p) \\ &= \sum_{s_{ij}^p}\sum_{p=1}^{K} I(s_{ij}^p \in D_p)\sum_{s\in D} w_p^*(s_{ij}^p, s)I(\text{nbd(s) contains } s_{ij}^p) \\ &= \sum_{s_{ij}^p}\sum_{p=1}^{K} I(s_{ij}^p \in D_p)b_{ij}a^2 \\ &= Ka^2[b_{11} + b_{12} + b_{13} + \cdots + b_{aa}] \qquad (2.30) \end{aligned}$$

to obtain

$$[b_{11} + b_{12} + b_{13} + \cdots + b_{aa}] = N/(Ka^2) = 1 \qquad (2.31)$$

Using the above result we can show that $\frac{1}{N}\sum_i T_{i,M,L} = \bar{Z}$ as shown in the following steps.

$$\begin{aligned}
\frac{1}{N}\sum_i T_{i,M,L} = \quad & \frac{1}{N}\sum_i \Big[b_{11}Z_{i,M,L}(s_{11}^{(i,M,L)}) + b_{12}Z_{i,M,L}(s_{12}^{(i,M,L)}) + \cdots \\
& +b_{aa}Z_{i,M,L}(s_{aa}^{(i,M,L)})\Big] \\
= \quad & \frac{1}{N}\Bigg[b_{11}\sum_i Z_{i,M,L}(s_{11}^{(i,M,L)}) + b_{12}\sum_i Z_{i,M,L}(s_{12}^{(i,M,L)}) + \cdots \\
& +b_{aa}\sum_i Z_{i,M,L}(s_{aa}^{(i,M,L)})\Bigg] \\
= \quad & \big[b_{11}\bar{Z} + b_{12}\bar{Z} + \cdots + b_{aa}\bar{Z}\big] \\
= \quad & \bar{Z} \qquad (2.32)
\end{aligned}$$

Using the above results it is easy to show that $E[T_{i,M,L}] = \mu$

$$\begin{aligned}
E[T_{i,M,L}] = \quad & E\Big[b_{11}Z_{i,M,L}(s_{11}^{i,M,L}) + b_{12}Z_{i,M,L}(s_{12}^{i,M,L}) + \cdots \\
& +b_{aa}Z_{i,M,L}(s_{aa}^{i,M,L})\Big] \\
= \quad & b_{11}\mu + b_{12}\mu + \cdots + b_{aa}\mu \\
= \quad & \mu \qquad (2.33)
\end{aligned}$$

Proof of lemma 2.2.2:

$$\begin{aligned}
Var\left[T_{i,M,L}\right] &= \sum_{i=1}^{a} b_{ii}^2\sigma^2 + \sum_{i,j\neq c,d} b_{ij}b_{cd}Cov\left[Z_{i,M,L}(s_{ij}^{i,M,L}), Z_{i,M,L}(s_{cd}^{i,M,L})\right] \\
&= \sigma^2\sum_{i=1}^{a} b_{ii}^2 + \sum_{i,j\neq c,d} b_{ij}b_{cd}C\left(s_{ij}^{i,M,L} - s_{cd}^{i,M,L}\right)
\end{aligned} \qquad (2.34)$$

and using the fact that $\sum_{i,j=1}^{a} b_{ij} = 1$ and $b_{ij} > 0$ we can show that

$$Var\left[T_{i,M,L}\right] = V \leq \sigma^2 \qquad (2.35)$$

Therefore, $T_{i,M,L} \sim N(\mu, V)$, where, V is given by the expression in 2.34.

Now consider $E\left[|T_{i,M,L}|^{2p+\delta}\right]$. For simplicity of notations sake denote $T_{i,M,L}$ by X.

Then,

$$E\left[|X|^{2p+\delta}\right] = \int_{-\infty}^{\infty} |x|^{2p+\delta} f_X(x)\ \mathrm{dx} \tag{2.36}$$

where, $f_X(x)$ is the density function of X.

Now,

$$|x|^{2p+\delta} \leq \begin{cases} |x|^{2p+2} & |x| \geq 1 \\ 1 & |x| \leq 1 \end{cases} \tag{2.37}$$

Then,

$$\begin{aligned} E\left[|X|^{2p+\delta}\right] &= \int_{-\infty}^{\infty} |x|^{2p+\delta} f_X(x)\ \mathrm{dx} \\ &= \int_{-\infty}^{-1} |x|^{2p+\delta} f_X(x)\ \mathrm{dx} + \int_{-1}^{1} |x|^{2p+\delta} f_X(x)\ \mathrm{dx} + \int_{1}^{\infty} |x|^{2p+\delta} f_X(x)\ \mathrm{dx} \\ &< \int_{-\infty}^{\infty} |x|^{2p+2} f_X(x)\ \mathrm{dx} + 1 \\ &= 1 + E\left[|X|^{2p+2}\right] \end{aligned} \tag{2.38}$$

Now, $E\left[|X|^{2p+2}\right]$ can be expressed as a function of μ and V and since, $V < \sigma^2$ it is easy to show that $E\left[|X|^{2p+2}\right]$ is less than some constant, $C > 0$.

Proof of lemma 2.2.3:

From lemma 2.2.4 we have the result for the variance of $\bar{Z}$.

$$\begin{aligned} Var\left(\bar{T}\sqrt{N}\right) &= Var\left(\bar{Z}\sqrt{N}\right) \\ &= \sum_{\substack{\boldsymbol{t}=(t_1,t_2) \\ |t_1|\leq m, |t_2|\leq m}} \sigma(\boldsymbol{t}) \left(1 - \frac{|t_1|}{n}\right)^+ \left(1 - \frac{|t_2|}{n}\right)^+ \end{aligned} \tag{2.39}$$

Let $\lim_{N\to\infty} Var\left(\bar{T}\sqrt{N}\right) = \sigma^2$. Sufficient conditions for the result of this lemma to hold are the result of lemma 2.2.2, m-dependence and the variance condition that we discussed above. Using a result proved by Tikhomirov(Tikhomirov 1980) about the central limit theorem for dependent data we can easily show that the conditions of the result are satisfied and that the desired result follows.

Proof of lemma 2.2.4:

This result has been proved by Rosén(Rosén 1968).

Proof of proposition 2.2.1: Using conditional probabilities we first show that $E\left[Z^*(s)\right] = \mu$

$$\begin{aligned}
E\left[Z^*(s)\right] &= E\left[Z_j^*(\phi_j(s))\right] \\
&= E\left[E\left[Z_j^*(\phi_j(s)|Z(t), t \in D\right]\right] \\
&= E\left[E\left[\sum_{i=1}^{N} Z_i(\phi_j(s)) I(i^{th} \text{ tile is chosen to be in the } j^{th} \text{ sub-region})\right]\right] \\
&= E\left[\frac{1}{N}\sum_{s \in D} Z(s)\right] \\
&= \frac{1}{N} N\mu \\
&= \mu
\end{aligned} \tag{2.40}$$

Using the above result we show that $E\left[\bar{Z}^\star\right] = \mu$

$$\begin{aligned}
E\left[\bar{Z}^\star\right] = &\quad \frac{1}{N} E\left[\sum_{s \in D}\sum_{s' \in D}\sum_{j=1}^{K} Z_j^*(s') w_j^*(s,s') I(s' \in nbd(s) \cap D_j)\right] \\
= &\quad \frac{\mu}{N}\sum_{s \in D}\sum_{s' \in D}\sum_{j=1}^{K} w_j^*(s,s') I(s' \in nbd(s) \cap D_j) \\
= &\quad \mu
\end{aligned} \tag{2.41}$$

Proof of Proposition 2.2.2:

$$Var\left[\bar{Z}^\star\right] = E\left[Var\left[\bar{Z}^\star | Z(s), s \in D\right]\right] + Var\left[E\left[\bar{Z}^\star | Z(s), s \in D\right]\right] \tag{2.42}$$

$$\begin{aligned}
Var\left[\bar{Z}^\star | Z(s), s \in D\right] = &\quad Var\left[\bar{T}^* | Z(s), s \in D\right] \\
= &\quad \frac{1}{K'^2} Var\left[\sum_{i=1}^{K'} T_i^* | Z(s), s \in D\right]
\end{aligned} \tag{2.43}$$

Since, the T_i^* are independent and identical we can write

$$\begin{aligned} Var\left[\bar{Z}^\star|Z(s),s\in D\right] = & \quad \frac{1}{K'}Var\left[T_1^*|Z(s),s\in D\right] \\ = & \quad \frac{1}{NK'}\sum_{\substack{\boldsymbol{i}=(i_1,i_2)\\ i_1=1\cdots n\\ i_2=1\cdots n}}\left[T_{\boldsymbol{i},M,L}-\bar{T}\right]^2 \\ = & \quad \frac{1}{NK'}\sum_{\substack{\boldsymbol{i}=(i_1,i_2)\\ i_1=1\cdots n\\ i_2=1\cdots n}}\left[T_{\boldsymbol{i},M,L}-\bar{Z}\right]^2 \end{aligned} \tag{2.44}$$

So,

$$Var\left[\bar{Z}^\star\right] = \quad Var\left[\bar{Z}\right]+E\left[\frac{1}{NK'}\sum_{\substack{\boldsymbol{i}=(i_1,i_2)\\ i_1=1\cdots n\\ i_2=1\cdots n}}\left[T_{\boldsymbol{i},M,L}-\bar{Z}\right]^2\right] \tag{2.45}$$

$$\begin{aligned} E\left[\frac{1}{NK'}\sum_{\substack{\boldsymbol{i}=(i_1,i_2)\\ i_1=1\cdots n\\ i_2=1\cdots n}}\left[T_{\boldsymbol{i},M,L}-\bar{Z}\right]^2\right] &= E\frac{1}{NK'}\left[\sum_{\substack{\boldsymbol{i}=(i_1,i_2)\\ i_1=1\cdots n\\ i_2=1\cdots n}}\left(T^2_{\boldsymbol{i},M,L}\right)\right]-E\frac{1}{NK'}\left[N(\bar{Z})^2\right] \\ &= \frac{1}{K'}E\left[T^2_{1,M,L}\right]-\frac{1}{K'a^4}E\left[(\bar{Z})^2\right] \end{aligned} \tag{2.46}$$

$$\begin{aligned} E\left[T^2_{1,M,L}\right] = & \quad Var\left[T_{1,M,L}\right]+\mu^2 \\ = & \quad \mu^2+\sum_{(p,q)}b^2_{pq}\sigma^2+\sum_{(p,q)\neq(r,s)}b_{pq}b_{rs}C(s^{1,M,L}_{pq}-s^{1,M,L}_{rs}) \end{aligned} \tag{2.47}$$

where, 1 ≤ p,q,r,s ≤ a.

$$\begin{aligned} E\left[\bar{Z}^2\right] &= Var\left[\bar{Z}\right]+\left(E\bar{Z}\right)^2 \\ &= Var\left[\bar{Z}\right]+\mu^2 \end{aligned} \tag{2.48}$$

Therefore,

$$\begin{aligned} Var\left[\bar{Z}^{\star}\right] &= Var\left[\bar{Z}\right] + \frac{1}{K'}\left[\mu^2 + \sum_{(p,q)} b_{pq}^2\sigma^2 + \sum_{(p,q)\neq(r,s)} b_{pq}b_{rs}C(s_{pq}^{1,M,L} - s_{rs}^{1,M,L})\right] \\ &\quad -\frac{1}{K'}\left[Var\left[\bar{Z}\right] + \mu^2\right] \\ &= (1-\frac{1}{K'})Var\left[\bar{Z}\right] + \frac{1}{K'}\left[\sum_{(p,q)} b_{pq}^2\sigma^2 + \sum_{(p,q)\neq(r,s)} b_{pq}b_{rs}C(s_{pq}^{1,M,L} - s_{rs}^{1,M,L})\right] \end{aligned} \tag{2.49}$$

The second term of the above sum goes to 0 as $K' \to \infty$. Thus, from here we can say that $Var\left[\bar{Z}^{\star}\right] = O\left((1-\frac{1}{K'})Var\left[\bar{Z}\right]\right)$.

<u>Proof of Proposition 2.2.3</u>:

This result can be proved by using assumptions [**A0**] through [**A3**], lemmas 2.2.1, 2.2.2 and 2.2.3 and using Theorem 4 of Politis and Romano's (Politis & Romano 1993) paper.

<u>Proof of lemma 2.3.1 and 2.3.2</u>:

To prove lemma 2.3.1 we first need to prove lemma 2.3.2.

$$E\left[X_1X_2X_3X_4\right] = \frac{\partial^4}{\partial u_1\partial u_2\partial u_3\partial u_4}\varphi_{X_1,X_2,X_3,X_4}(0,0,0,0) \tag{2.50}$$

where, $\varphi_{X_1,X_2,X_3,X_4}(0,0,0,0)$ is the joint characteristic function of X_1, X_2, X_3 and X_4.

$$\varphi_{X_1,X_2,X_3,X_4}(u_1,u_2,u_3,u_4) = exp\{\frac{-1}{2}\sum_{i,j=1}^{4} u_1u_2E\left[X_iX_j\right]\} \tag{2.51}$$

Let φ denote the right hand side of equation 2.50 and define $L_i = \sum_{j=1}^{4} u_jE\left[X_iX_j\right]$

then

$$
\begin{aligned}
\frac{\partial \varphi}{\partial u_1} &= -\varphi L_1 \\
\frac{\partial^2 \varphi}{\partial u_1 \partial u_2} &= \{L_1 L_2 - E\left[X_1 X_2\right]\}\varphi \\
\frac{\partial^3 \varphi}{\partial u_1 \partial u_2 \partial u_3} &= \{-L_1 L_2 L_3 + L_3 E\left[X_1 X_2\right] + L_2 E\left[X_1 X_3\right] + L_1 E\left[X_2 X_3\right]\}\varphi \\
\frac{\partial^4 \varphi}{\partial u_1 \partial u_2 \partial u_3 \partial u_4} &= \{L_1 L_2 L_3 L_4 - L_1 L_2 E\left[X_3 X_4\right] - L_1 L_3 E\left[X_2 X_4\right] \\
&\quad - L_1 L_4 E\left[X_2 X_3\right] - L_2 L_3 E\left[X_1 X_4\right] - L_2 L_4 E\left[X_1 X_3\right] \\
&\quad - L_3 L_4 E\left[X_1 X_2\right] + E\left[X_1 X_2\right] E\left[X_3 X_4\right] + E\left[X_1 X_3\right] E\left[X_2 X_4\right] \\
&\quad + E\left[X_1 X_4\right] E\left[X_2 X_3\right]\}\varphi
\end{aligned}
\tag{2.52}
$$

From equations 2.50 and 2.52 we get the desired result given in lemma 2.3.2.

Now,

$$Cov\left[X_1 X_2, X_3 X_4\right] = E\left[X_1 X_2 X_3 X_4\right] - E\left[X_1 X_2\right] E\left[X_3 X_4\right] \tag{2.53}$$

Using the result of lemma 2.3.2 in the above equation we get that

$$
\begin{aligned}
Cov\left[X_1 X_2, X_3 X_4\right] &= E\left[X_1 X_3\right] E\left[X_2 X_4\right] + E\left[X_1 X_4\right] E\left[X_2 X_3\right] \\
&= Cov\left[X_1, X_3\right] Cov\left[X_2, X_4\right] + Cov\left[X_1, X_4\right] Cov\left[X_2, X_3\right]
\end{aligned}
\tag{2.54}
$$

Now using the above equation if we let $X_1 = X_2$ and $X_3 = X_4$, we can easily see that

$$Cov\left[X_1^2, X_3^2\right] = 2\{Cov\left[X_1, X_3\right]\}^2 \tag{2.55}$$

thus proving lemma 2.3.1.

<u>Proof of lemma 2.3.3</u>:

Let X denote the quantity $Z_{i,M,L}(s) - Z_{i,M,L}(s + \mathbf{h})$. Then it is easy to show that $X \sim N(0, 2\sigma^2 - 2C(h))$.

Now consider $E\left[|T_{i,M,L}|^{2p+\delta}\right]$.

$$\begin{aligned} E\left[|T_{i,M,L}|^{2p+\delta}\right] &= E\left[|X|^{2(2p+\delta)}\right] \\ &= E\left[X^{2(2p+\delta)}\right] \end{aligned} \tag{2.56}$$

and as in the proof of lemma 2.2.2 we can show that this expectation is less than a constant quantity, since, $\sigma^2 - C(h) < \sigma^2$.

<u>**Proof of lemma 2.3.4**</u>: Davis and Borgman(Davis & Borgman 1982) have shown that the $T_{i,M,L}$ are $m + |\boldsymbol{h}|$-dependent, where $|\boldsymbol{h}|$ denotes the length of $\boldsymbol{h}$. Also $T_{i,M,L}$ are stationary since, Z(s) are stationary(since, Z(s) are weakly stationary and are normally distributed).

Further, they also showed that under the assumptions **[B0]** and **[B5]**,

$$Var(\frac{\bar{T}}{2}) = Var(\widehat{\gamma(\boldsymbol{h})}) = \frac{1}{|N(h)|} \sum_{\substack{\boldsymbol{i}=(i_1,i_2)\\ |i_1|,|i_2| \leq (m+|\boldsymbol{h}|)}} \sigma(\boldsymbol{i}) \left(1 - \frac{|i_1|}{Q_1}\right)^+ \left(1 - \frac{|i_2|}{Q_2}\right)^+ \tag{2.57}$$

where, $\sigma(\boldsymbol{i}) = Cov(T_{1,M,L}, T_{1+i,M,L})$

Then

$$Var(\frac{\sqrt{|N(h)|}}{2}\bar{T}) = \sum_{\substack{\boldsymbol{i}=(i_1,i_2)\\ |i_1|,|i_2| \leq (m+|\boldsymbol{h}|)}} \sigma(\boldsymbol{i}) \left(1 - \frac{|i_1|}{Q_1}\right)^+ \left(1 - \frac{|i_2|}{Q_2}\right)^+ \tag{2.58}$$

$E\left[\overline{T}\right] = 2\gamma(h)$ by the definition of $\gamma(h)$. Let $\lim_{|N(h)| \to \infty} Var(\sqrt{N(h)}\bar{T}) = \sigma^2_\infty$ then using the same method as outlined in the proof of lemma 2.2.3 we can show that

$$\sqrt{|N(h)|}\left(\overline{T} - E\left[\overline{T}\right]\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_\infty) \tag{2.59}$$

<u>**Proof of Proposition 2.3.1**</u>:

Using the result of lemma 2.3.3 and lemma 2.3.4 and assumptions **[B0]** - **[B4]** the result of this proposition follows directly from an application of theorem 4 of Politis and Romano (Politis & Romano 1993).

<u>**Proof of Proposition 2.3.2**</u>:

Consider the difference of the variograms, $2\widehat{\gamma^*_{ns}(\boldsymbol{h})} - 2\widehat{\gamma^*_{s}(\boldsymbol{h})}$

$$\begin{aligned} & 2\widehat{\gamma^*_{ns}(\boldsymbol{h})} - 2\widehat{\gamma^*_{s}(\boldsymbol{h})} \\ = & \frac{1}{|N(h)|}\sum_{N(h)} \left(Z^*(s) - Z^*(t)\right)^2 \\ - & \frac{1}{|N(h)|}\sum_{N(h)} \left[\sum_{s'\in D}\sum_{j=1}^{K} Z^*_j(s)w^*_j(s',s)I_{(s',nbd(s)\cap D_j)} - \sum_{t'\in D}\sum_{j=1}^{K} Z^*_j(t)w^*_j(t',t)I_{(t',nbd(t)\cap D_j)}\right]^2 \end{aligned} \tag{2.60}$$

For ease of notation let us define

$$I_1 = I_{(s',nbd(s)\cap D_j)} = \begin{cases} 1, & s' \in nbd(s) \cap D_j, \\ 0, & otherwise \end{cases}$$

and

$$I_2 = I_{(t',nbd(t)\cap D_j)} = \begin{cases} 1, & s' \in nbd(t) \cap D_j, \\ 0, & otherwise \end{cases}$$

Then we can write the above equation as

$$\begin{aligned} & 2\widehat{\gamma^*_{ns}(\boldsymbol{h})} - 2\widehat{\gamma^*_{s}(\boldsymbol{h})} \\ = & \frac{1}{|N(h)|}\sum_{N(h)} \\ & \left[Z^*(s) - Z^*(t) - \sum_{s'\in D}\sum_{j=1}^{K} Z^*_j(s')w^*_j(s',s)I_1 + \sum_{t'\in D}\sum_{j=1}^{K} Z^*_j(t')w^*_j(t',t)I_2\right] \\ \times & \left[Z^*(s) - Z^*(t) + \sum_{s'\in D}\sum_{j=1}^{K} Z^*_j(s)w^*_j(s',s)I_1 - \sum_{t'\in D}\sum_{j=1}^{K} Z^*_j(t')w^*_j(t',t)I_2\right] \end{aligned} \tag{2.61}$$

Now, consider

$$\left|Z^*(s) - Z^*(t) - \sum_{s'\in D}\sum_{j=1}^{K} Z^*_j(s')w^*_j(s',s)I_1 + \sum_{t'\in D}\sum_{j=1}^{K} Z^*_j(t')w^*_j(t',t)I_2\right| \tag{2.62}$$

Here,
$Z^*(s), Z^*(t), Z_j^*(t'), Z_j^*(s')$ and $Z_j^*(t') \leq \max_{s\in D}|Z(s)|$,
$0 \leq w_j^*(s', s), w_j^*(t', t) \leq 1$
Therefore,

$$\begin{aligned}
&\left|Z^*(s) - Z^*(t) - \sum_{s'\in D}\sum_{j=1}^{K} Z_j^*(s')w_j^*(s',s)I_1 + \sum_{t'\in D}\sum_{j=1}^{K} Z_j^*(t')w_j^*(t',t)I_2\right| \\
\leq\ & |Z^*(s)| + |Z^*(t)| + \left|\sum_{s'\in D}\sum_{j=1}^{K} Z_j^*(s')w_j^*(s',s)I_1\right| + \left|\sum_{t'\in D}\sum_{j=1}^{K} Z_j^*(t')w_j^*(t',t)I_2\right| \\
\leq\ & 2(r+1)\max_{s\in D}|Z(s)| \qquad (2.63)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left|Z^*(s) - Z^*(t) + \sum_{s'\in D}\sum_{j=1}^{K} Z_j^*(s')w_j^*(s',s)I_1 - \sum_{t'\in D}\sum_{j=1}^{K} Z_j^*(t')w_j^*(t',t)I_2\right| \\
\leq\ & |Z^*(s)| + |Z^*(t)| + \left|\sum_{s'\in D}\sum_{j=1}^{K} Z_j^*(s')w_j^*(s',s)I_1\right| + \left|\sum_{t'\in D}\sum_{j=1}^{K} Z_j^*(t')w_j^*(t',t)I_2\right| \\
\leq\ & 2(r+1)\max_{s\in D}|Z(s)| \qquad (2.64)
\end{aligned}$$

Here, r is the number of points in the neighbourhood of s.

Therefore,

$$\left|2\widehat{\gamma_{ns}^*(\boldsymbol{h})} - 2\widehat{\gamma_s^*(\boldsymbol{h})}\right| \leq 4(r+1)^2\left(\max_{s\in D}|Z(s)|\right)^2 \qquad (2.65)$$

From this we get that

$$\widehat{\gamma_{ns}^*(\boldsymbol{h})} - 2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)^2 \leq \widehat{\gamma_s^*(\boldsymbol{h})} \leq \widehat{\gamma_{ns}^*(\boldsymbol{h})} + 2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)^2 \qquad (2.66)$$

$$\begin{aligned}
Var\left[\widehat{\gamma_{ns}^*(\boldsymbol{h})} - 2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)^2\right] &\leq Var\left[\widehat{\gamma_s^*(\boldsymbol{h})}\right] \\
&\leq Var\left[\widehat{\gamma_{ns}^*(\boldsymbol{h})} + 2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)^2\right]
\end{aligned} \qquad (2.67)$$

$\Rightarrow$

$$\begin{aligned} & Var\left[\widehat{\gamma^*_{ns}(\boldsymbol{h})}\right] + Var\left[2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)^2\right] \\ & - 2Cov\left(\widehat{\gamma^*_{ns}(\boldsymbol{h})}, 2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)\right) \leq Var\left[\widehat{\gamma^*_{s}(\boldsymbol{h})}\right] \leq Var\left[\widehat{\gamma^*_{ns}(\boldsymbol{h})}\right] + \\ & \quad Var\left[2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)^2\right] + 2Cov\left(\widehat{\gamma^*_{ns}(\boldsymbol{h})}, 2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)\right) \end{aligned} \tag{2.68}$$

$$\begin{aligned} & \left|Cov\left(\widehat{\gamma^*_{ns}(\boldsymbol{h})}, 2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)\right)\right| \leq \\ & \qquad \sqrt{Var\left[\widehat{\gamma^*_{ns}(\boldsymbol{h})}\right] Var\left[2(r+1)^2\left(\max_{s\in D}|Z(s)|\right)^2\right]} \end{aligned} \tag{2.69}$$

It is not difficult to show that if $M_N = \max_{s\in D}|Z(s)|$ then under the assumption **[B0]**

$$P(a_N(M_N - b_N) \leq x) \to e^{-e^{-x}} \text{ for some constants } a_N \text{ and } b_N \tag{2.70}$$

Therefore, $O\left(\dfrac{Var\left[\widehat{\gamma^*_{s}(\boldsymbol{h})}\right]}{Var\left[\widehat{\gamma^*_{ns}(\boldsymbol{h})}\right]}\right) = 1.$

<u>**Proof of proposition 2.3.3**</u>:

We first calculate the variance of the estimator.

$$\begin{aligned} Var(2\widehat{\gamma^\star_s(h)}) &= Var\left(\frac{1}{|N(h)|}\sum_{(s,t)\in N(h)}(Z^\star(s) - Z^\star(t))^2\right) \\ &= \frac{1}{N(h)^2}\sum_{(s,t)\in N(h)} Var\left((Z^\star(s) - Z^\star(t))^2\right) \\ &\quad + \frac{1}{N(h)^2}\sum_{\substack{(s,t)\neq(u,v)\\(s,t)\in N(h)\\(u,v)\in N(h)}} Cov\left[(Z^\star(s) - Z^\star(t))^2, (Z^\star(u) - Z^\star(v))^2\right] \end{aligned} \tag{2.71}$$

Here, $N(h) = \{(s,t) : s \in D, t \in D \text{ and } |s-t| = h\}$ and $|N(h)|$ is the number of pairs in the set $N(h)$. Consider the variance term $Var\left[(Z^\star(s) - Z^\star(t))^2\right]$.

$$Var\left((Z^\star(s) - Z^\star(t))^2\right) = E\left[(Z^\star(s) - Z^\star(t))^4\right] - \left\{E\left[(Z^\star(s) - Z^\star(t))^2\right]\right\}^2 \tag{2.72}$$

Conditioning on the data the first expectation can be written as follows:

$$\begin{aligned} & E\left[(Z^\star(s) - Z^\star(t))^4\right] \\ = \;& E\left\{E\left[(Z^\star(s) - Z^\star(t))^4 \middle| Z(s)\right]\right\} \\ = \;& E\left\{E\left[\left(\sum_{s'\in D}\sum_{j=1}^{K} Z_j^*(s')w_j^*(s',s)I_j(s',s) - \sum_{t'\in D}\sum_{j=1}^{K} Z_j^*(t')w_j^*(t',t)I_j(t',t)\right)^4 \middle| Z(s)\right]\right\} \end{aligned} \tag{2.73}$$

where,

$$I_j(s',s) = \begin{cases} 1, & s' \in nbd(s) \cap D_j, \\ 0, & otherwise \end{cases}$$

and

$$I_j(t',t) = \begin{cases} 1, & t' \in nbd(t) \cap D_j, \\ 0, & otherwise \end{cases}$$

Using [B5] we can rewrite the quantity whose expected value we seek in the following manner. Let j_1 and j_2 be the two sub-regions that contain the neighbourhoods of s and t. Without any loss of generality we assume that j_1 is the sub-region that contains the two locations s and t as well.

$$
\begin{aligned}
&\left(\sum_{s'\in D}\sum_{j=1}^{K} Z_j^*(s')w_j^*(s',s)I_j(s',s) - \sum_{t'\in D}\sum_{j=1}^{K} Z_j^*(t')w_j^*(t',t)I_j(t',t)\right)^4 \\
= &\left[\left(\sum_{s'\in D} Z_{j_1}^*(s')w_{j_1}^*(s',s)I_{j_1}(s',s) - \sum_{t'\in D} Z_{j_1}^*(t')w_{j_1}^*(t',t)I_{j_1}(t',t)\right)\right. \\
&\left. + \left(\sum_{s'''\in D} Z_{j_2}^*(s''')w_{j_2}^*(s''',s)I_{j_2}(s''',s) - \sum_{t'''\in D} Z_{j_2}^*(t''')w_{j_2}^*(t''',t)I_{j_2}(t''',t)\right)\right]^4 \\
= \quad & \left(A^*(s',s,j_1) - A^*(t',t,j_1)\right)^4 + \left(A^*(s''',s,j_2) - A^*(t''',t,j_2)\right)^4 \\
+ \quad & 6\left(A^*(s',s,j_1) - A^*(t',t,j_1)\right)^2 \left(A^*(s''',s,j_2)^* - A^*(t''',t,j_2)\right)^2 \\
+ \quad & 4\left(A^*(s',s,j_1) - A^*(t',t,j_1)\right)^3 \left(A^*(s''',s,j_2) - A^*(t''',t,j_2)\right) \\
+ \quad & 4\left(A^*(s',s,j_1) - A^*(t',t,j_1)\right) \left(A^*(s''',s,j_2) - A^*(t''',t,j_2)\right)^3
\end{aligned}
\tag{2.74}
$$

where,

$$
A^*(s',s,j_1) = \left(\sum_{s'\in D} Z_{j_1}^*(s')w_{j_1}^*(s',s)I_{j_1}(s',s)\right) \tag{2.75}
$$

Using the expanded from given in equation (2.74) the conditional expectation can be written as

$$
E\left\{E\left[\left(\sum_{s'\in D}\sum_{j=1}^{K} Z_j^*(s')w_j^*(s',s)I_j(s',s) - \sum_{t'\in D}\sum_{j=1}^{K} Z_j^*(t')w_j^*(t',t)I_j(t',t)\right)^4 \middle| Z(s)\right]\right\}
$$

$$
\begin{align}
= \quad & E\Big\{E\Big[\Big[\left(A^*(s',s,j_1) - A^*(t',t,j_1)\right)^4 + \left(A^*(s''',s,j_2) - A^*(t''',t,j_2)\right)^4 \tag{2.76} \\
+ \quad & 6\left(A^*(s',s,j_1) - A^*(t',t,j_1)\right)^2 \left(A^*(s''',s,j_2) - A^*(t''',t,j_2)\right)^2 \notag \\
+ \quad & 4\left(A^*(s',s,j_1) - A^*(t',t,j_1)\right)^3 \left(A^*(s''',s,j_2) - A^*(t''',t,j_2)\right) \tag{2.77} \\
+ \quad & 4\ \left(A^*(s',s,j_1) - A^*(t',t,j_1)\right) \left(A^*(s''',s,j_2) - A^*(t''',t,j_2)\right)^3\Big] \Big| Z(s)\Big]\Big\} \tag{2.78}
\end{align}
$$

Utilizing the fact that there are N tiles to choose from the above expectation can be evaluated as shown below. We will evaluate the expectation term by term.

$$E\left\{E\left[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)^4 \middle| Z(s)\right]\right\} = E\left\{\frac{1}{N}\sum_{i_1=1}^{N}\left[A(s',s,i_1)-A(t',t,i_1)\right]^4\right\} \tag{2.79}$$

where,

$$A(s',s,i_1) \;=\; \sum_{s'\in D} Z_{i_1}(s')w^*_{j_1}(s',s)I_{j_1}(s',s) \tag{2.80}$$

Using **[B6]** we can easily see that $(A(s',s,i_1)-A(t',t,i_1)) \sim \textit{Normal}$ with 0 mean and variance given by the following expression.

$$\begin{aligned} Var\left[A(s',s,i_1)-A(t',t,i_1)\right] = \quad & Var\left[A(s',s,i_1)\right] + Var\left[A(t',t,i_1)\right] \\ & -2Cov\left(A(s',s,i_1),A(t',t,i_1)\right) \end{aligned} \tag{2.81}$$

$$\begin{aligned} Var\left(A(s',s,i_1)\right) \;=\; & \sum_{s'\in D} Var\left(Z_{i_1}(s')w^*_{j_1}(s',s)I_{j_1}(s',s)\right) + \\ & \sum_{\substack{s'\neq s''\\ s'\in D\\ s''\in D}} Cov\left(Z_{i_1}(s')w^*_{j_1}(s',s)I_{j_1}(s',s), Z_{i_1}(s'')w^*_{j_1}(s'',s)I_{j_1}(s'',s)\right) \end{aligned} \tag{2.82}$$

Similarly,

$$\begin{aligned} Var\left(A(t',t,i_1)\right) \;=\; & \sum_{t'\in D} Var\left(Z_{i_1}(t')w^*_{j_1}(t',t)I_{j_1}(t',t)\right) + \\ & \sum_{\substack{t'\neq t''\\ t'\in D\\ t''\in D}} Cov\left(Z_{i_1}(t')w^*_{j_1}(t',t)I_{j_1}(t',t), Z_{i_1}(t'')w^*_{j_1}(t'',t)I_{j_1}(t'',t)\right) \end{aligned} \tag{2.83}$$

and now combining these we get the variance of $(A(s',s,i_1)-A(t',t,i_1))$ which we denote by $\sigma_1^2(s,t)$

$$\begin{aligned}
\sigma_1^2(s,t) \quad = \quad & \sigma^2 \left[\sum_{s' \in D} \left(w^*_{j_1}(s',s) \right)^2 I_{j_1}(s',s) + \sum_{t' \in D} \left(w^*_{j_1}(t',t) \right)^2 I_{j_1}(t',t) \right] + \\
& \sum_{\substack{s' \neq s'' \\ s' \in D \\ s'' \in D}} C\left(s' - s''\right) w^*_{j_1}(s',s) w^*_{j_1}(s'',s) I_{j_1}(s',s) I_{j_1}(s'',s) + \\
& \sum_{\substack{t' \neq t'' \\ t' \in D \\ t'' \in D}} C\left(t' - t''\right) w^*_{j_1}(t',t) w^*_{j_1}(t'',t) I_{j_1}(t',t) I_{j_1}(t'',t) - \\
& \sum_{\substack{s' \in D \\ t' \in D}} 2C\left(s' - t'\right) w^*_{j_1}(s',s) w^*_{j_1}(t',t) I_{j_1}(s',s) I_{j_1}(t',t) \qquad (2.84)
\end{aligned}$$

Using Stein's lemma it is easy to show that

$$E\left[\left(A(s',s,i_1) - A(t',t,i_1) \right)^4 \right] = 3\sigma_1^4(s,t) \qquad (2.85)$$

Note that this variance doesn't depend upon the tile index i_1 . Using this expression and equation (2.79) we can now say that

$$E\left\{ E\left[\left(A^*(s',s,j_1) - A^*(t',t,j_1) \right)^4 \middle| Z(s) \right] \right\} = 3\sigma_1^4(s,t) \qquad (2.86)$$

Similarly we can show that

$$E\left\{ E\left[\left(A^*(s',s,j_2) - A^*(t',t,j_2) \right)^4 \middle| Z(s) \right] \right\} = 3\sigma_2^4(s,t) \qquad (2.87)$$

where,

$$\begin{aligned}
\sigma_2^2(s,t) \quad = \quad & \sigma^2 \left[\sum_{s''' \in D} \left(w^*_{j_2}(s''',s) \right)^2 I_{j_2}(s''',s) + \sum_{t''' \in D} \left(w^*_{j_2}(t''',t) \right)^2 I_{j_2}(t''',t) \right] + \\
& \sum_{\substack{s''' \neq s'''' \\ s''' \in D \\ s'''' \in D}} C\left(s''' - s''''\right) w^*_{j_2}(s''',s) w^*_{j_2}(s'''',s) I_{j_2}(s''',s) I_{j_2}(s'''',s) + \\
& \sum_{\substack{t''' \neq t'''' \\ t''' \in D \\ t'''' \in D}} C\left(t''' - t''''\right) w^*_{j_2}(t''',t) w^*_{j_2}(t'''',t) I_{j_2}(t''',t) I_{j_2}(t'''',t) - \\
& \sum_{\substack{s''' \in D \\ t''' \in D}} 2C\left(s''' - t'''\right) w^*_{j_2}(s''',s) w^*_{j_2}(t''',t) I_{j_2}(s''',s) I_{j_2}(t''',t) \qquad (2.88)
\end{aligned}$$

Now let us find an expression for
$E\left[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)^2\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)^2\right]$

$$\begin{aligned}
& E\left\{E\left[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)^2\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)^2 \Big| Z(s)\right]\right\} \\
= \;& E\left\{\frac{1}{N^2}\left[\sum_{i_1=1}^{N}\left(A(s',s,i_1)-A(t',t,i_1)\right)^2\right]\left[\sum_{i_2=1}^{N}\left(A(s''',s,i_2)-A(t''',t,i_2)\right)^2\right]\right\} \\
= \;& \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}E\left[\left(A(s',s,i_1)-A(t',t,i_1)\right)^2\left(A(s''',s,i_2)-A(t''',t,i_2)\right)^2\right] \\
= \;& \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\left\{Cov\left(\left(A(s',s,i_1)-A(t',t,i_1)\right)^2,\left(A(s''',s,i_2)-A(t''',t,i_2)\right)^2\right)\right. \\
& \left.+E\left(A(s',s,i_1)-A(t',t,i_1)\right)^2E\left(A(s''',s,i_2)-A(t''',t,i_2)\right)^2\right\} \\
= \;& \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\left\{2\left[Cov\left(\left(A(s',s,i_1)-A(t',t,i_1)\right),\left(A(s''',s,i_2)-A(t''',t,i_2)\right)\right)\right]^2\right. \\
& \left.+E\left(A(s',s,i_1)-A(t',t,i_1)\right)^2E\left(A(s''',s,i_2)-A(t''',t,i_2)\right)^2\right\}
\end{aligned} \tag{2.89}$$

It is easy to show that the term
$\frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\left\{2\left[Cov\left(\left(A(s',s,i_1)-A(t',t,i_1)\right),\left(A(s''',s,i_2)-A(t''',t,i_2)\right)\right)\right]^2\right\}$ is $o(\frac{1}{N})$.
If we fix i_1 then let there be N_1 terms $(A(s''',s,i_2)-A(t''',t,i_2))$ such that the above covariance is non-zero. Then we can bound this sum of covariances by
$\frac{2}{N^2}NN_1Var\left(\left(A(s',s,i_1)-A(t',t,i_1)\right)\right)Var\left(\left(A(s''',s,i_2)-A(t''',t,i_2)\right)\right)$. Since, the variances are finite and N_1 is finite this bound goes to 0 as $N\to\infty$.

The term $\frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\left\{E\left(A(s',s,i_1)-A(t',t,i_1)\right)^2E\left(A(s''',s,i_2)-A(t''',t,i_2)\right)^2\right\}$ is equal to the quantity $\sigma_1^2(s,t)\sigma_2^2(s,t)$.

Therefore,

$$E\left[6\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)^2\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)^2\right]=6\sigma_1^2(s,t)\sigma_2^2(s,t) \tag{2.90}$$

Let us now consider the term
$E\left\{E\left[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)^3\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)\Big|Z(s)\right]\right\}$.

$$
\begin{aligned}
& E\left\{E\left[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)^3\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)\Big|Z(s)\right]\right\} \\
= & \frac{1}{N^2}E\left\{\left[\sum_{i_1=1}^{N}\left(A(s',s,i_1)-A(t',t,i_1)\right)^3\right]\left[\sum_{i_2=1}^{N}\left(A(s''',s,i_2)-A(t''',t,i_2)\right)\right]\right\} \\
= & \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}E\left[\left(A(s',s,i_1)-A(t',t,i_1)\right)^3\left(A(s''',s,i_2)-A(t''',t,i_2)\right)\right]
\end{aligned}
\tag{2.91}
$$

Using lemma 2.3.2 here since, $A(s',s,i_1)-A(t',t,i_1)$ and $A(s''',s,i_2)-A(t''',t,i_2)$ are jointly normally distributed we can write the last line of the above equation as

$$
\begin{aligned}
= & \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\left\{3E\left[\Big(A(s',s,i_1)-A(t',t,i_1)\Big)^2\right]\times\right. \\
& \left. E\Big[\left(A(s',s,i_1)-A(t',t,i_1)\right)\left(A(s''',s,i_2)-A(t''',t,i_2)\right)\Big]\right\} \\
= & \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\left\{3E\left[\Big(A(s',s,i_1)-A(t',t,i_1)\Big)^2\right]\times\right. \\
& \left. Cov\bigg(\Big(A^*(s',s,i_1)-A^*(t',t,i_1)\Big),\Big(A^*(s''',s,i_2)-A^*(t''',t,i_2)\Big)\bigg)\right\}
\end{aligned}
\tag{2.92}
$$

By a similar argument as that used for the previous term, we can show that this quantity is $o\left(\frac{1}{N}\right)$.

So,

$$
\begin{aligned}
& E\left\{E\left[4\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)^3\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big)\Big|Z(s)\right]\right\} \\
& \text{is } o\left(\frac{1}{N}\right)
\end{aligned}
\tag{2.93}
$$

And similarly for the remaining term, we can show that

$$E\left\{E\left[4\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big)^3\middle|Z(s)\right]\right\}$$
$$\text{is } o\left(\frac{1}{N}\right) \tag{2.94}$$

Combining everything from equations (2.86), (2.87), (2.90) ,(2.93) and (2.94) we get that

$$\begin{aligned} & E\left[\left(\sum_{s'\in D}\sum_{j=1}^{K}Z_j^*(s')w_j^*(s',s)I_j(s',s)-\sum_{t'\in D}\sum_{j=1}^{K}Z_j^*(t')w_j^*(t',t)I_j(t',t)\right)^4\right] \\ = & \; 3\sigma_1^4(s,t)+3\sigma_2^4(s,t)+6\sigma_1^2(s,t)\sigma_2^2(s,t)+o\left(\frac{1}{N}\right) \end{aligned} \tag{2.95}$$

Next let us consider the second term in the expression in (2.72).

$$\begin{aligned} & E\left[\left(\sum_{s'\in D}\sum_{j=1}^{K}Z_j^*(s')w_j^*(s',s)I_j(s',s)-\sum_{t'\in D}\sum_{j=1}^{K}Z_j^*(t')w_j^*(t',t)I_j(t',t)\right)^2\right] \\ = & \; E\left\{E\left[\left(\sum_{s'\in D}\sum_{j=1}^{K}Z_j^*(s')w_j^*(s',s)I_j(s',s)-\sum_{t'\in D}\sum_{j=1}^{K}Z_j^*(t')w_j^*(t',t)I_j(t',t)\right)^2\middle|Z(s)\right]\right\} \\ = & \; E\Big\{E\Big[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)^2+\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)^2 \\ & +2\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)\Big|Z(s)\Big]\Big\} \end{aligned} \tag{2.96}$$

$$\begin{aligned} E\left\{E\left[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)^2\middle|Z(s)\right]\right\} &= E\left\{\frac{1}{N}\sum_{i_1=1}^{N}\left(A(s',s,i_1)-A(t',t,i_1)\right)^2\right\} \\ &= \frac{1}{N}\sum_{i_1=1}^{N}Var\left[\left(A(s',s,i_1)-A(t',t,i_1)\right)\right] \\ &= \sigma_1^2(s,t) \end{aligned} \tag{2.97}$$

and

$$\begin{aligned} E\left\{E\left[\left(A^*(s''', s, j_2) - A^*(t''', t, j_2)\right)^2 \Big| Z(s)\right]\right\} &= E\left\{\frac{1}{N}\sum_{i_2=1}^{N}\left(A(s''', s, i_2) - A(t''', t, i_2)\right)^2\right\} \\ &= \frac{1}{N}\sum_{i_2=1}^{N} Var\left[\left(A(s''', s, i_2) - A(t''', t, i_2)\right)\right] \\ &= \sigma_2^2(s,t) \end{aligned} \qquad (2.98)$$

and

$$\begin{aligned} & E\left\{\left[\left(A^*(s', s, j_1) - A^*(t', t, j_1)\right)\left(A^*(s''', s, j_2) - A^*(t''', t, j_2)\right) \Big| Z(s)\right]\right\} \\ = & \frac{1}{N^2}E\left\{\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\left(A(s', s, i_1) - A(t', t, i_1)\right)\left(A(s''', s, i_2) - A(t''', t, i_2)\right)\right\} \\ = & \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N} Cov\left(\left(A(s', s, i_1) - A(t', t, i_1)\right), \left(A(s''', s, i_2) - A(t''', t, i_2)\right)\right) \end{aligned} \qquad (2.99)$$

and we can show that this is $o\left(\frac{1}{N}\right)$. So,

$$\begin{aligned} & E\left[\left(\sum_{s'\in D}\sum_{j=1}^{K} Z_j^*(s')w_j^*(s', s)I_j(s', s) - \sum_{t'\in D}\sum_{j=1}^{K} Z_j^*(t')w_j^*(t', t)I_j(t', t)\right)^2\right] \\ & \qquad = \left(\sigma_1^2(s,t) + \sigma_2^2(s,t)\right) + o\left(\frac{1}{N}\right) \end{aligned}$$

Using the equations (2.95) and (2.100) and substituting in (2.72)

$$Var\left[(Z^\star(s) - Z^\star(t))^2\right] = 2\left(\sigma_1^2(s,t) + \sigma_2^2(s,t)\right)^2 + o\left(\frac{1}{N}\right) + o\left(\frac{1}{N^2}\right) \qquad (2.100)$$

Now let us consider the covariance terms in the right hand side of the equation (2.71)

Consider the covariance term

$$Cov\left[\left(B^*(s',s,j)-B^*(t',t,j)\right)^2,\left(B^*(u',u,k)-B^*(v',v,k)\right)^2\right]$$

where, $B^*(s',s,j)=\sum_{s'\in D}\sum_{j=1}^{K}Z_j^*(s')w_j^*(s',s)I_j(s',s)$, $(s,t)\neq(u,v)$ and $(s,t)\in N(h)$ and $(u,v)\in N(h)$.

$$\begin{aligned}
& Cov\left[\left(B^*(s',s,j)-B^*(t',t,j)\right)^2,\left(B^*(u',u,k)-B^*(v',v,k)\right)^2\right] \\
= \; & Cov\Bigg[\Big(A^*(s',s,j_1)-A^*(t',t,j_1)+A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big)^2, \\
& \Big(A^*(u',u,k_1)-A^*(v',v,k_1)+A^*(u''',u,k_2)-A^*(v''',v,k_2)\Big)^2\Bigg] \\
= \; & Cov\Bigg\{\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)^2+\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big)^2+ \\
& 2\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big), \\
& \Big(A^*(u',u,j_1)-A^*(v',v,j_1)\Big)^2+\Big(A^*(u''',u,j_2)-A^*(v''',v,j_2)\Big)^2+ \\
& 2\Big(A^*(u',u,j_1)-A^*(v',v,j_1)\Big)\Big(A^*(u''',u,j_2)-A^*(v''',v,j_2)\Big)\Bigg\}
\end{aligned} \tag{2.101}$$

Now let us consider two cases:

Case 1: $j_1=k_1$ and $j_2=k_2$, that is the two pairs (s,t) and (u,v) belong to the same sub-region and have neighbourhoods contained by the same two sub-regions.

Case 2: $j_1=k_1$ but $j_2\neq k_2$, that is the two pairs belong to the same sub-region and one of the two sub-regions that contains the neighbourhoods is the same for both the pairs.

Most of the pairs of locations will fall in one of these two cases.

Let us evaluate the covariance under the scenario outlined in Case 1.

It is easy to show that the only non-negligible terms in the covariance expression given in (2.101) are
$Cov\Big[\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)^2,\Big(A^*(u',u,j_1)-A^*(v',v,j_1)\Big)^2\Big]$,
$Cov\Big[\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big)^2,\Big(A^*(u''',u,j_2)-A^*(v''',v,j_2)\Big)^2\Big]$ and
$4Cov\Big[\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big),$

$$\Big(A^*(u',u,j_1)-A^*(v',v,j_1)\Big)\Big(A^*(u''',u,j_2)-A^*(v''',v,j_2)\Big)\Big].$$

We evaluate these three covariances and the derivations are shown below.

$$\begin{aligned}
& Cov\left[\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)^2,\Big(A^*(u',u,j_1)-A^*(v',v,j_1)\Big)^2\right]\\
= & Cov\left[\Big(A(s',s,i_1)-A(t',t,i_1)\Big)^2,\Big(A(u',u,i_1)-A(v',v,i_1)\Big)^2\right]\\
= & 2\Big\{Cov\Big[\left(A(s',s,i_1)-A(t',t,i_1)\right),\left(A(u',u,i_1)-A(v',v,i_1)\right)\Big]\Big\}^2\\
= & 2\Bigg\{\sum_{s'\in D}\sum_{u'\in D}w^*_{j_1}(s',s)w^*_{j_1}(u',u)C(s'-u')-\sum_{s'\in D}\sum_{v'\in D}w^*_{j_1}(s',s)w^*_{j_1}(v',v)C(s'-v')\\
& -\sum_{t'\in D}\sum_{u'\in D}w^*_{j_1}(t',t)w^*_{j_1}(u',u)C(t'-u')+\sum_{t'\in D}\sum_{v'\in D}w^*_{j_1}(t',t)w^*_{j_1}(v',v)C(t'-v')\Bigg\}^2
\end{aligned} \tag{2.102}$$

and similarly,

$$\begin{aligned}
& Cov\Big[\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big)^2,\Big(A^*(u''',u,j_2)-A^*(v''',v,j_2)\Big)^2\Big]=\\
& 2\Bigg\{\sum_{s'''\in D}\sum_{u'''\in D}w^*_{j_2}(s''',s)w^*_{j_2}(u''',u)C(s'''-u''')-\sum_{s'''\in D}\sum_{v'''\in D}w^*_{j_2}(s''',s)w^*_{j_2}(v''',v)C(s'''-v''')\\
& -\sum_{t'''\in D}\sum_{u'''\in D}w^*_{j_2}(t''',t)w^*_{j_2}(u''',u)C(t'''-u''')+\sum_{t'''\in D}\sum_{v'''\in D}w^*_{j_2}(t''',t)w^*_{j_2}(v''',v)C(t'''-v''')\Bigg\}^2
\end{aligned} \tag{2.103}$$

$$\begin{aligned}
& Cov\Big[\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big),\\
& \qquad\Big(A^*(u',u,j_1)-A^*(v',v,j_1)\Big)\Big(A^*(u''',u,j_2)-A^*(v''',v,j_2)\Big)\Big]\\
= & E\Big[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)\\
& \quad\left(A^*(u',u,j_1)-A^*(v',v,j_1)\right)\left(A^*(u''',u,j_2)-A^*(v''',v,j_2)\right)\Big]-\\
& \Big\{E\Big[\left(A^*(s',s,j_1)-A^*(t',t,j_1)\right)\left(A^*(s''',s,j_2)-A^*(t''',t,j_2)\right)\Big]\\
& \quad E\Big[\left(A^*(u',u,j_1)-A^*(v',v,j_1)\right)\left(A^*(u''',u,j_2)-A^*(v''',v,j_2)\right)\Big]\Big\}
\end{aligned} \tag{2.104}$$

It is easy to show that

$$E\Big[\,(A^*(s',s,j_1)-A^*(t',t,j_1))\,(A^*(s''',s,j_2)-A^*(t''',t,j_2))\,\Big] \\ \text{is } o\left(\frac{1}{N}\right) \tag{2.105}$$

and

$$E\Big[\,(A^*(u',u,j_1)-A^*(v',v,j_1))\,(A^*(u''',u,j_2)-A^*(v''',v,j_2))\,\Big] \\ \text{is } o\left(\frac{1}{N}\right) \tag{2.106}$$

$$\begin{aligned} & E\Bigg[\Big(A^*(s',s,j_1)-A^*(t',t,j_1)\Big)\Big(A^*(s''',s,j_2)-A^*(t''',t,j_2)\Big) \\ & \Big(A^*(u',u,j_1)-A^*(v',v,j_1)\Big)\Big(A^*(u''',u,j_2)-A^*(v''',v,j_2)\Big)\Bigg] \\ = \;& E\Bigg\{E\Big[\,(A^*(s',s,j_1)-A^*(t',t,j_1))\,(A^*(s''',s,j_2)-A^*(t''',t,j_2)) \\ & (A^*(u',u,j_1)-A^*(v',v,j_1))\,(A^*(u''',u,j_2)-A^*(v''',v,j_2))\,\Big|\,Z(s)\Big]\Bigg\} \\ = \;& \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}E\Big[\,(A(s',s,i_1)-A(t',t,i_1))\,(A(s''',s,i_2)-A(t''',t,i_2)) \\ & \qquad (A(u',u,i_1)-A(v',v,i_1))\,(A(u''',u,i_2)-A(v''',v,i_2))\Big] \end{aligned} \tag{2.107}$$

and this can be written as the sum of the product of second moments as shown below

$$\begin{aligned}\frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N} \quad & \Bigg\{E\Big[\Big(A(s',s,i_1)-A(t',t,i_1)\Big)\Big(A(s''',s,i_2)-A(t''',t,i_2)\Big)\Big] \\ & E\Big[\Big(A(u',u,i_1)-A(v',v,i_1)\Big)\Big(A(u''',u,i_2)-A(v''',v,i_2)\Big)\Big] \\ + \quad & E\Big[(A(s',s,i_1)-A(t',t,i_1))\,(A(u',u,i_1)-A(v',v,i_1))\Big] \\ & E\Big[\Big(A(s''',s,i_2)-A(t''',t,i_2)\Big)\Big(A(u''',u,i_2)-A(v''',v,i_2)\Big)\Big] \\ + \quad & E\Big[\Big(A(s',s,i_1)-A(t',t,i_1)\Big)\Big(A(u''',u,i_2)-A(v''',v,i_2)\Big)\Big] \\ & E\Big[\Big(A(u',u,i_1)-A(v',v,i_1)\Big)\Big(A(s''',s,i_2)-A(t''',t,i_2)\Big)\Big]\Bigg\}\end{aligned} \tag{2.108}$$

and this can be easily shown to be equal to (as the other two terms tend to 0 as $N \to \infty$)

$$\begin{aligned}&\frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\Bigg\{E\Big[\Big(A(s',s,i_1)-A(t',t,i_1)\Big)\Big(A(u',u,i_1)-A(v',v,i_1)\Big)\Big] \\ &\qquad E\Big[\Big(A(s''',s,i_2)-A(t''',t,i_2)\Big)\Big(A(u''',u,i_2)-A(v''',v,i_2)\Big)\Big]\Bigg\} \\ &= \Bigg\{\sum_{s'\in D}\sum_{u'\in D} w^*_{j_1}(s',s)w^*_{j_1}(u',u))C(s'-u') - \sum_{s'\in D}\sum_{v'\in D} w^*_{j_1}(s',s)w^*_{j_1}(v',v)C(s'-v') \\ &- \sum_{t'\in D}\sum_{u'\in D} w^*_{j_1}(t',t)w^*_{j_1}(u',u)C(t'-u') + \sum_{t'\in D}\sum_{v'\in D} w^*_{j_1}(t',t)w^*_{j_1}(v',v)C(t'-v')\Bigg\} \times \\ &\Bigg\{\sum_{s'''\in D}\sum_{u'''\in D} w^*_{j_2}(s''',s)w^*_{j_2}(u''',u)C(s'''-u''') - \sum_{s'''\in D}\sum_{v'''\in D} w^*_{j_2}(s''',s)w^*_{j_2}(v''',v))C(s'''-v''') - \\ &\sum_{t'''\in D}\sum_{u'''\in D} w^*_{j_2}(t''',t)w^*_{j_2}(u''',u)C(t'''-u''') + \sum_{t'''\in D}\sum_{v'''\in D} w^*_{j_2}(t''',t)w^*_{j_2}(v''',v))C(t'''-v''')\Bigg\}\end{aligned} \tag{2.109}$$

Combining the expressions in (2.102), (2.103) and (2.109) we get the complete form of the covariance under case 1 as :

$$
\begin{aligned}
&= 2\Bigg\{\sum_{s'\in D}\sum_{u'\in D} w^*_{j_1}(s',s)w^*_{j_1}(u',u)C(s'-u') - \sum_{s'\in D}\sum_{v'\in D} w^*_{j_1}(s',s)w^*_{j_1}(v',v)C(s'-v') \\
&- \sum_{t'\in D}\sum_{u'\in D} w^*_{j_1}(t',t)w^*_{j_1}(u',u)C(t'-u') + \sum_{t'\in D}\sum_{v'\in D} w^*_{j_1}(t',t)w^*_{j_1}(v',v)C(t'-v')\Bigg\}^2 + \\
&2\Bigg\{\sum_{s'''\in D}\sum_{u'''\in D} w^*_{j_2}(s''',s)w^*_{j_2}(u''',u)C(s'''-u''') - \sum_{s'''\in D}\sum_{v'''\in D} w^*_{j_2}(s''',s)w^*_{j_2}(v''',v)C(s'''-v''') \\
&- \sum_{t'''\in D}\sum_{u'''\in D} w^*_{j_2}(t''',t)w^*_{j_2}(u''',u)C(t'''-u''') + \sum_{t'''\in D}\sum_{v'''\in D} w^*_{j_2}(t''',t)w^*_{j_2}(v''',v)C(t'''-v''')\Bigg\}^2 \\
&+ \Bigg\{\sum_{s'\in D}\sum_{u'\in D} w^*_{j_1}(s',s)w^*_{j_1}(u',u)C(s'-u') - \sum_{s'\in D}\sum_{v'\in D} w^*_{j_1}(s',s)w^*_{j_1}(v',v)C(s'-v') \\
&- \sum_{t'\in D}\sum_{u'\in D} w^*_{j_1}(t',t)w^*_{j_1}(u',u)C(t'-u') + \sum_{t'\in D}\sum_{v'\in D} w^*_{j_1}(t',t)w^*_{j_1}(v',v)C(t'-v')\Bigg\} \times \\
&\Bigg\{\sum_{s'''\in D}\sum_{u'''\in D} w^*_{j_2}(s''',s)w^*_{j_2}(u''',u))C(s'''-u''') - \sum_{s'''\in D}\sum_{v'''\in D} w^*_{j_2}(s''',s)w^*_{j_2}(v''',v))C(s'''-v''') \\
&- \sum_{t'''\in D}\sum_{u'''\in D} w^*_{j_2}(t''',t)w^*_{j_2}(u''',u)C(t'''-u''') + \sum_{t'''\in D}\sum_{v'''\in D} w^*_{j_2}(t''',t)w^*_{j_2}(v''',v)C(t'''-v''')\Bigg\}
\end{aligned}
\tag{2.110}
$$

The covariance under case 2 can be similarly calculated using the expression given in (2.101) and is given by the expression below:

$$
\begin{aligned}
&2\Bigg\{\sum_{s'\in D}\sum_{u'\in D} w^*_{j_1}(s',s)w^*_{j_1}(u',u)C(s'-u') - \sum_{s'\in D}\sum_{v'\in D} w^*_{j_1}(s',s)w^*_{j_1}(v',v)C(s'-v') \\
&- \sum_{t'\in D}\sum_{u'\in D} w^*_{j_1}(t',t)w^*_{j_1}(u',u)C(t'-u') + \sum_{t'\in D}\sum_{v'\in D} w^*_{j_1}(t',t)w^*_{j_1}(v',v)C(t'-v')\Bigg\}^2
\end{aligned}
\tag{2.111}
$$

Combining the variances (2.100) and the covariances (2.110) or (2.111) we get the variance of our proposed estimator.

$$\begin{aligned} Var\left(2\widehat{\gamma_s^*(h)}\right) = & \quad \frac{1}{|N(h)|^2} \sum_{(s,t)\in N(h)} 2\left(\sigma_1^2(s,t)+\sigma_2^2(s,t)\right)^2 + \\ & \quad \frac{1}{|N(h)|^2} \sum_{\substack{(s,t)\neq(u,v)\\(s,t)\in N(h)\\(u,v)\in N(h)}} H(s,t,u,v) + o\left(\frac{1}{N|N(h)|}\right) + o\left(\frac{1}{N^2|N(h)|}\right) \\ = & \quad \frac{1}{|N(h)|^2} \sum_{(s,t)\in N(h)} 2\left(\sigma_1^2(s,t)+\sigma_2^2(s,t)\right)^2 \\ + & \quad \frac{1}{|N(h)|^2} \sum_{\substack{(s,t)\neq(u,v)\\(s,t)\in N(h)\\(u,v)\in N(h)}} H(s,t,u,v) + o\left(\frac{1}{K^2p^2}\right) + o\left(\frac{1}{K^3p^3}\right) \quad (2.112) \end{aligned}$$

where, $H(s,t,u,v)$ takes the value given in expression (2.110) or in expression (2.111) and the expressions for $\sigma_1^2(s,t)$ and $\sigma_2^2(s,t)$ are given in (2.84) and (2.88) respectively.

$$\begin{aligned} Bias &= \frac{1}{|N(h)|} E\left\{\sum_{(s,t)\in N(h)} \left[\left(Z^\star(s)-Z^\star(t)\right)^2\right]\right\} - 2\gamma(h) \\ &= \frac{1}{|N(h)|} \sum_{(s,t)\in N(h)} E\left[\left(Z^\star(s)-Z^\star(t)\right)^2\right] - 2\gamma(h) \end{aligned} \quad (2.113)$$

We calculated the above expected value in equation (2.100) and using that expression here we get

$$\begin{aligned} Bias &= \frac{1}{|N(h)|} \sum_{(s,t)\in N(h)} \left(\sigma_1^2(s,t)+\sigma_2^2(s,t)\right) - 2\gamma(h) + o\left(\frac{1}{N}\right) \\ &= \frac{1}{|N(h)|} \sum_{(s,t)\in N(h)} \left(\sigma_1^2(s,t)+\sigma_2^2(s,t)\right) - 2\gamma(h) + o\left(\frac{1}{Kp}\right) \quad (2.114) \end{aligned}$$

The expressions for $\sigma_1^2(s,t)$ and $\sigma_2^2(s,t)$ are given in equations (2.84) and (2.88) respectively.

Chapter 3

Simulation Study

3.1 The objectives of the Simulation Study

The main objective of the simulation study is to compare bootstrap percentile confidence intervals for the traditional and proposed bootstrap methods. To conduct this study we calculate the coverage for the true semi-variogram using empirical confidence intervals of the sample semi-variograms generated using the two methods.

We are interested in investigating how the coverage changes as the settings of the range, sill and nugget of the covariance model are changed. We are also interested in the performance of the proposed method based upon different amounts of smoothing and different sizes of the neighbourhood for locations on the boundary.

3.2 Simulation Procedure

The following steps outline the simulation procedure :

Step 1 : A sample of 400 values is generated on an irregular grid of dimensions 35 units by 35 units. The data are generated from the normal distribution with an underlying covariance structure given by the matern covariance model (3.1).

Step 2 : The proposed method and the existing method are applied to this sample to give us a bootstrap surface from each of the two methods.

Step 3 : For each of the two surfaces the empirical semi-variogram is calculated.

Step 4 : Steps 2 and 3 are repeated 100 times to give us 100 estimates of the empirical semi-variogram from each method.

Step 5 : Using the 100 estimates from Step 4 empirical 95% confidence intervals are calculated for each method.

Step 6 : Steps 1 through 5 are repeated a hundred times to give us 100 95% confidence intervals for each method.

Step 7 : Coverage is calculated using the 100 confidence intervals generated in Step 6 for both the methods.

For the proposed method the two factors - the amount of smoothing and the size of the neighbourhood of the locations on the boundary are held at the following levels:

a) Smoothing - 5%, 15%, 80%

b) Size of the neighbourhood - 4 sq. units, 16 sq. units.

Steps 1 though 7 are repeated for the following settings of the parameters of the matern covariance model and the 6 combinations of the amount of smoothing and the size of the neighbourhood specified above.

Simulation set 1 : Range=3, partial sill=4, nugget=0.2, shape=0.5
Simulation set 2 : Range=9, partial sill=4, nugget=0.2, shape=0.5
Simulation set 3 : Range=3, partial sill=8, nugget=0.2, shape=0.5
Simulation set 4 : Range=3, partial sill=4, nugget=2, shape=0.5

The block size is kept fixed at 7 units by 7 units.

The matern class was originally given by Matern and is best defined in terms of its isotropic covariance: $C(h) = C_0(\|h\|)$ where, $C_0(0) = 1$ and

$$C_0(t) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{2\sqrt{\theta_2} t}{\theta_1} \right)^{\theta_2} \mathcal{K}_{\theta_2} \left(\frac{2\sqrt{\theta_2} t}{\theta_1} \right) \tag{3.1}$$

The function $\Gamma(.)$ is the usual gamma function and $\mathcal{K}_{\theta_2}$ is the modified bessel function of the third kind of order θ_2. As special cases of this class, $\theta_2 = \frac{1}{2}$ corresponds to the exponential form of the semi-variogram and the limit $\theta_2 \to \infty$ corresponds to the Gaussian form.

3.3 Conclusions and discussion of results

The plots of coverage versus the lag number are shown in figures 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7 and 3.8.

- At shorter lags the coverage obtained using the proposed method is better than the coverage obtained using the traditional method. The behaviour of the spatial processes at shorter lag distances is the most important factor for prediction purposes (Stein 1995).

- If we over smooth we lose too much information and we get worse coverage using the proposed method. This is because, by increasing the amount of smoothing the variability in the data is reduced. Also by over smoothing we lose important information about the spatial structure of the process.

- When the range of the process is larger than the block size, the coverage obtained from the proposed method is better than that obtained from the traditional method. This is because we force some discontinuity that is not real with the traditional method and this is more apparent for larger values of the range. This suggests that the amount of smoothing and the size of the block together affect the coverage. However, both methods gave coverage less than 95%.

- The coverage does not seem to be affected by changing the value of the partial sill.

- When the nugget of the process is large compared to the sill of the process then the coverage from the proposed method is very sensitive to the amount of smoothing. The nugget measures the measurement error and when we increase the amount of smoothing we lose this information.

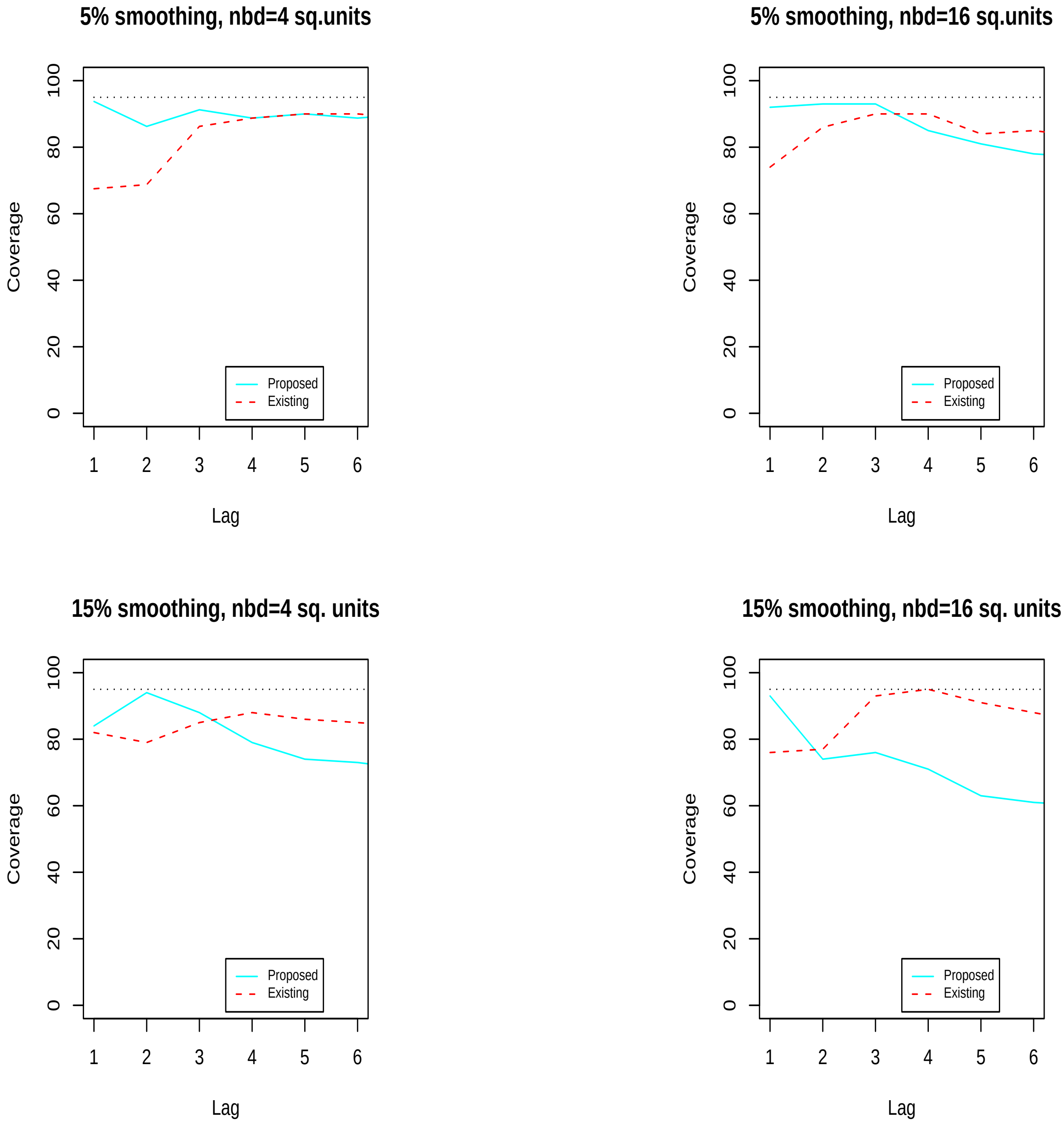


Figure 3.1: Simulation set 1

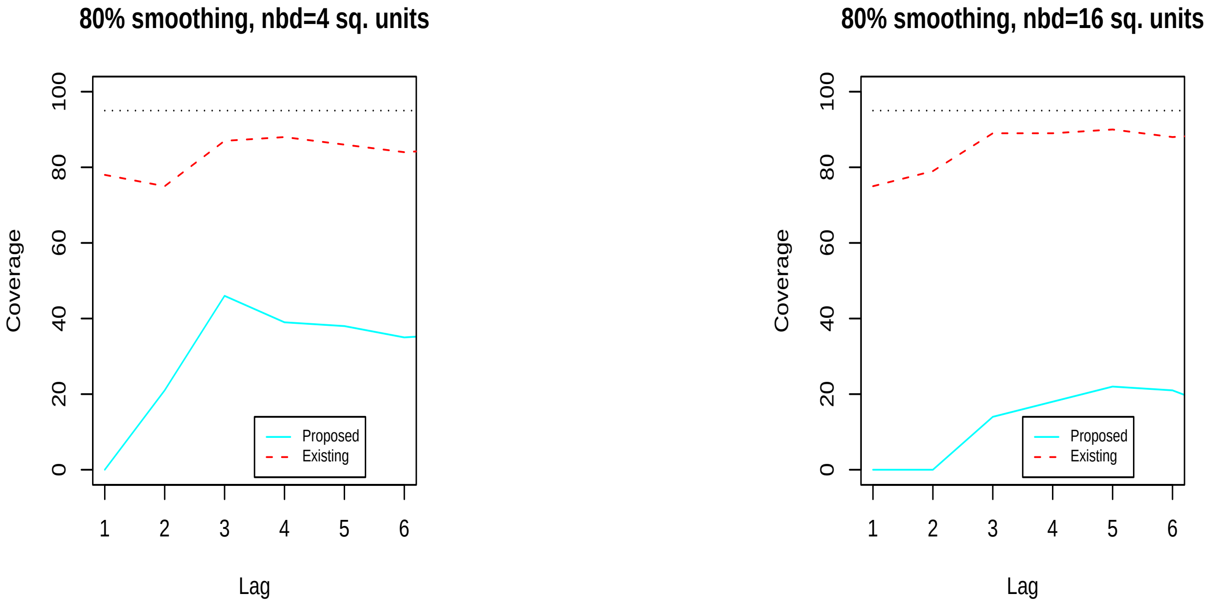
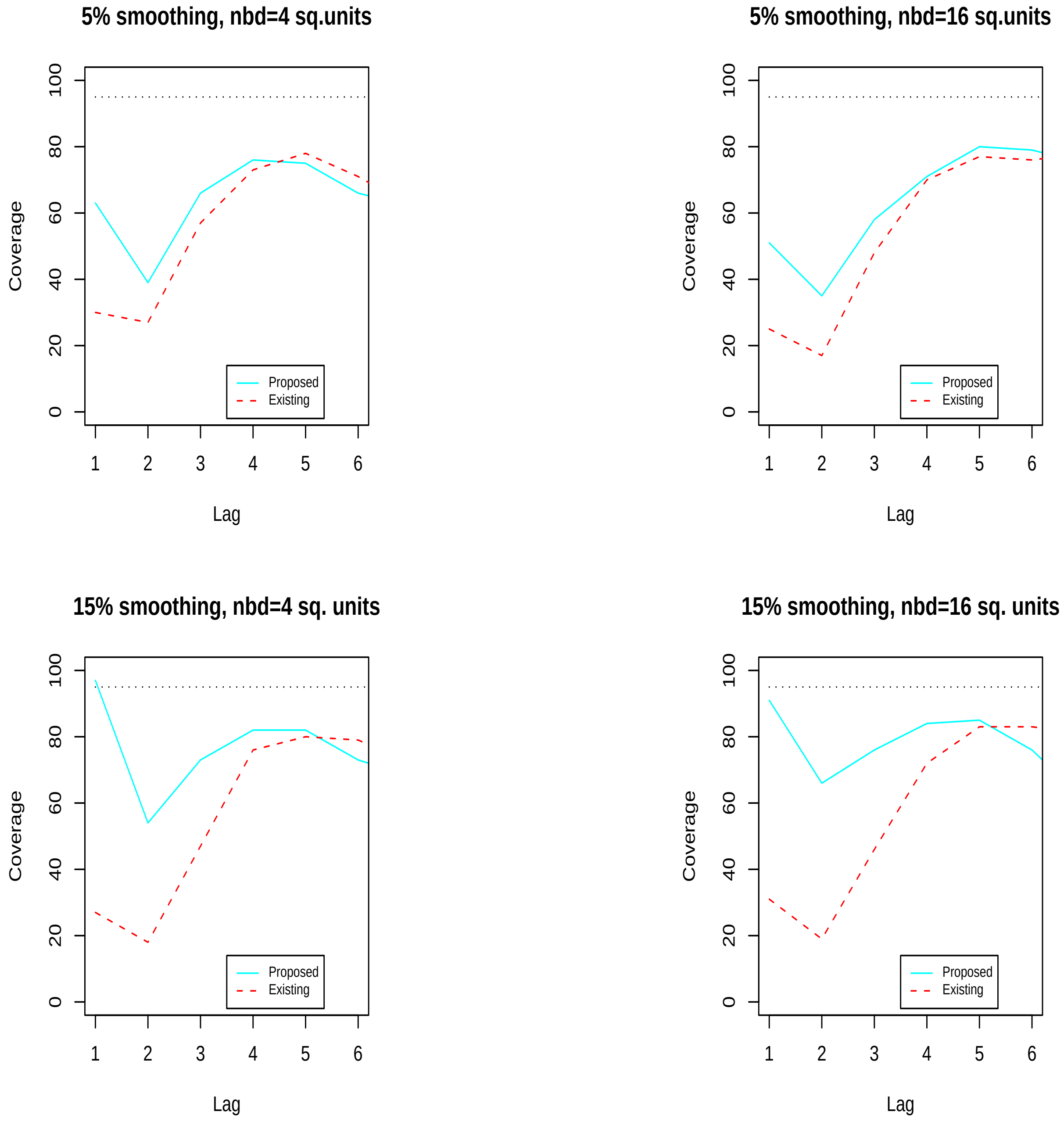


Figure 3.2: Continuation of results from Simulation set 1.



5% smoothing, nbd=4 sq.units

5% smoothing, nbd=16 sq.units

15% smoothing, nbd=4 sq. units

15% smoothing, nbd=16 sq. units

Figure 3.3: Simulation set 2.

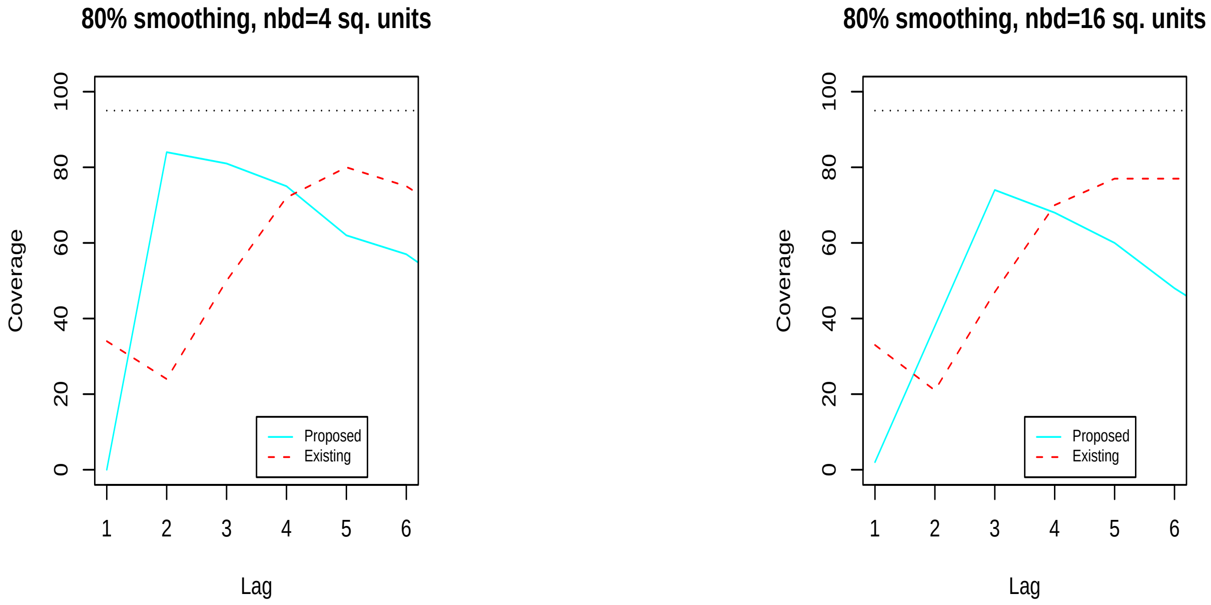


Figure 3.4: Continuation of results from Simulation set 2

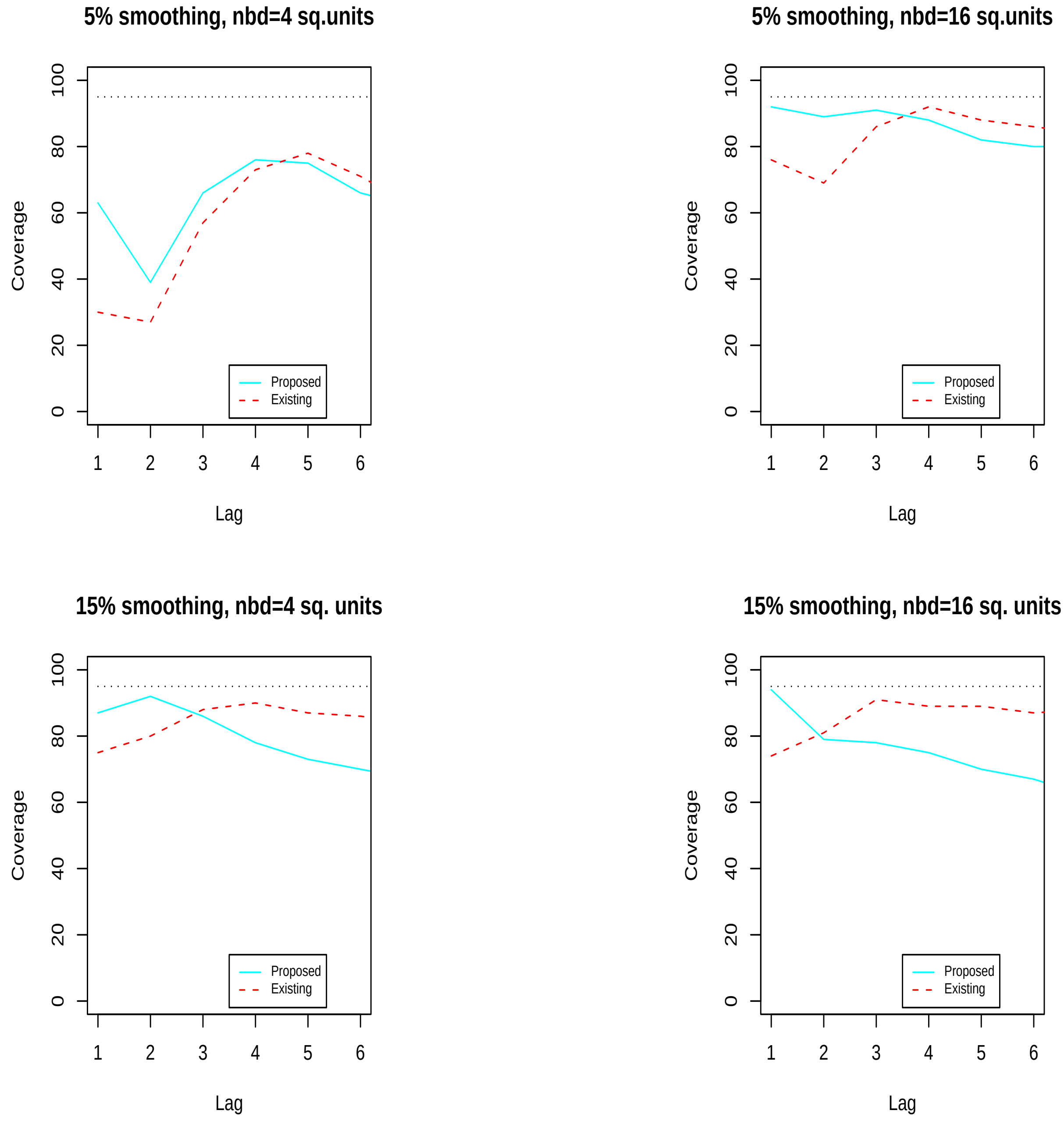
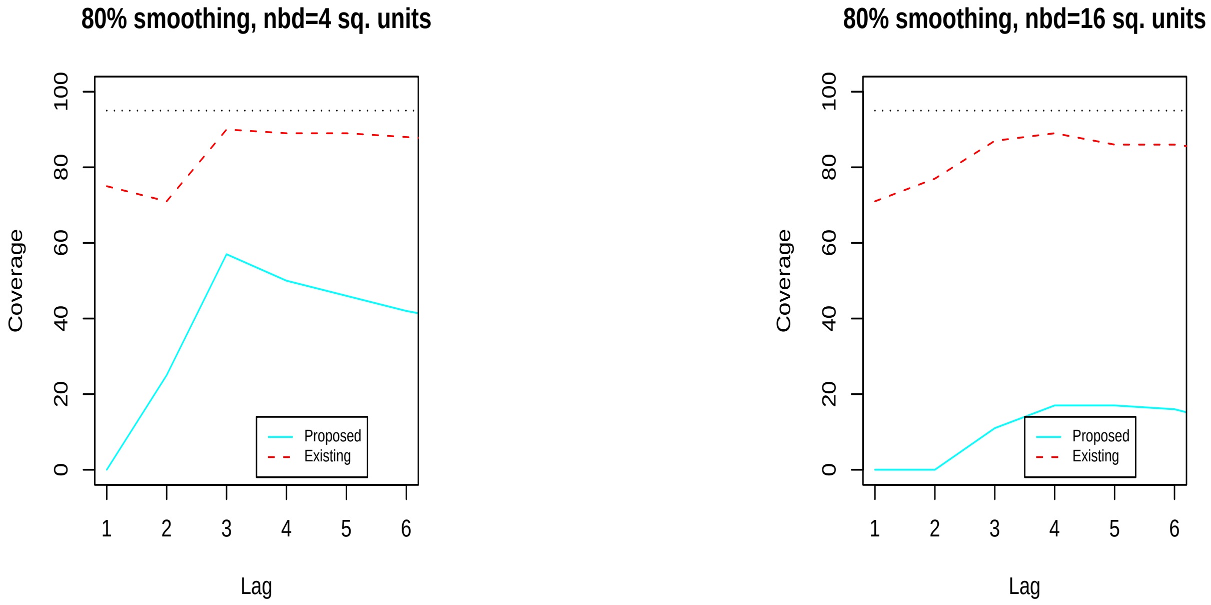


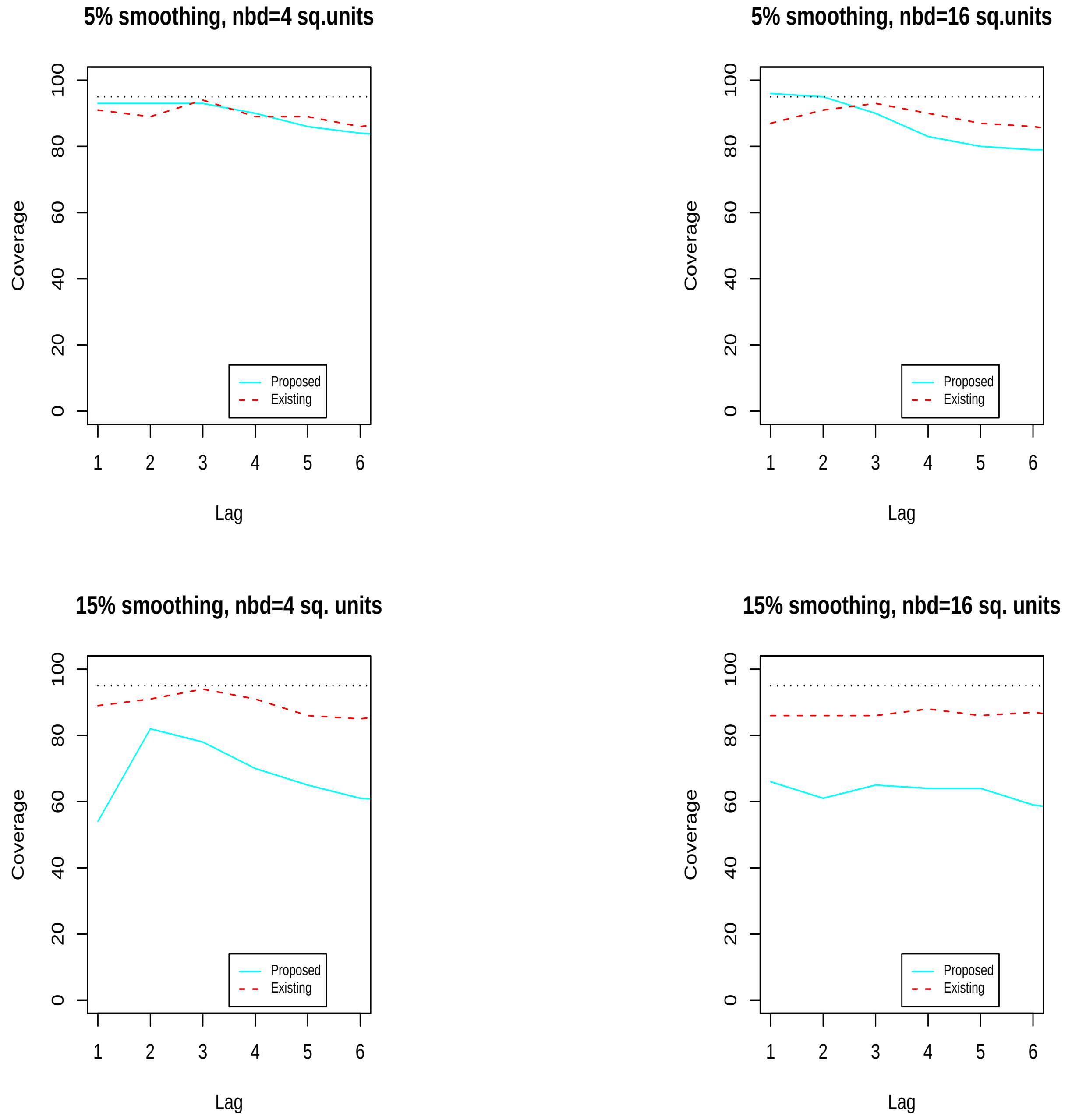
Figure 3.5: Simulation set 3



80% smoothing, nbd=4 sq. units

80% smoothing, nbd=16 sq. units

Figure 3.6: Continuation of results from Simulation set 3.



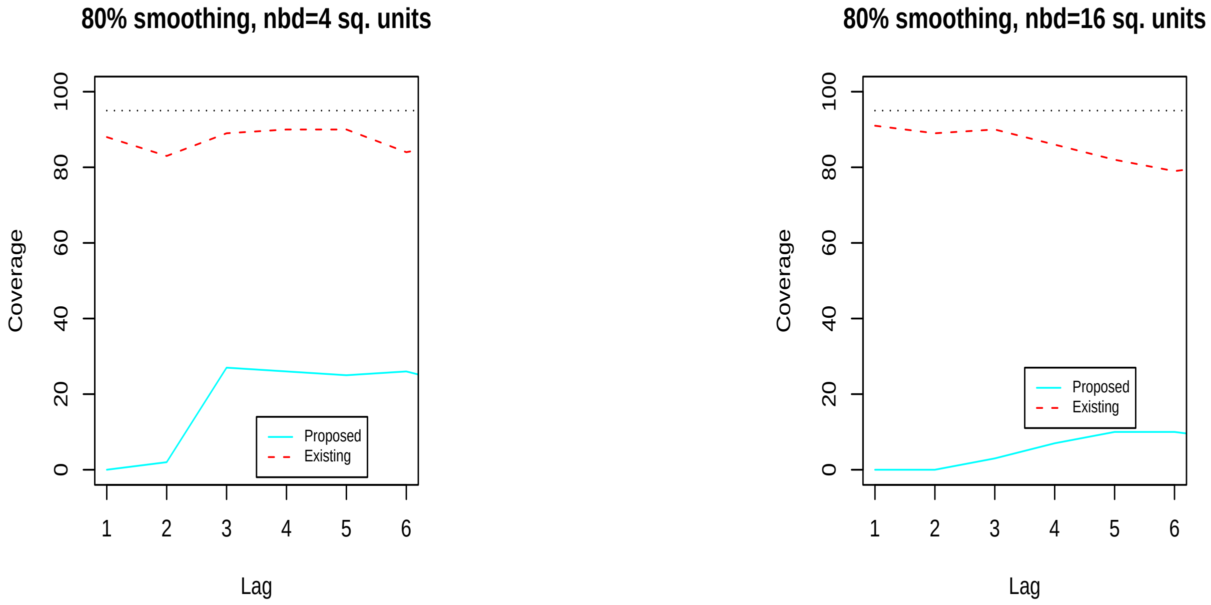
5% smoothing, nbd=4 sq.units

5% smoothing, nbd=16 sq.units

15% smoothing, nbd=4 sq. units

15% smoothing, nbd=16 sq. units

Figure 3.7: Simulation set 4


80% smoothing, nbd=4 sq. units

80% smoothing, nbd=16 sq. units

Figure 3.8: Continuation of results from Simulation set 4.

3.4 The objectives of the numerical study

The objective of this numerical study is to obtain the values for the smoothing parameters that minimize the mean squared error of the proposed variogram estimator. As mentioned before, the smoothing parameters are implicitly involved in the expression for the mean squared error and it is difficult to solve and obtain explicit expressions for the smoothing parameters. Thus we resort to this numerical technique to obtain a solution.

We are interested in studying the behaviour of the mean squared error as the smoothing parameter is varied for different settings of the range, sill and nugget parameters. The underlying covariance structure for the data was assumed to be given by the matern covariance model. We set the shape parameter of the matern model (3.1) at 0.5 for all the results.

Working under the assumptions [B5]-[B8] used in the derivation of the expression for the mean squared error we calculate the value of the mean squared error for different settings of the various parameters and variables involved. The parameters and variables that we vary are the : sample size, range, partial sill, nugget and the two smoothing parameters. Results are in the form of plots of the value of the mse vs the smoothing , for the various settings of the sample size, range, partial sill and nugget. Along with the mean squared error we also give plots of the variance and the adjusted bias vs the smoothing .

The following settings were considered for the various parameters.

- grid size: 24 by 24 and 30 by 30.
- range: 1.5, 2, 2.5,3.5
- partial sill: 4, 8
- nugget: 0, 0.1, 0.5

We present the results for the mse for the first two lags in the following figures. It is important to remember that it is most important to understand the behaviour of a spatial process at short lags and thus in our study we have been focusing on the performance of the proposed variogram estimator at short lags.

3.5 Conclusions and discussion of results

The results are presented in figures 3.9 through 3.24.

- The traditional block bootstrap estimator for the variogram is an unbiased estimator and thus when we introduce smoothing we observe an increase in the bias. The bias increases as the smoothing increases. However, smoothing also affects the variance of the estimator, which decreases as smoothing increases. This suggests that it must be possible to find the optimal amount of smoothing that would minimize the mean squared error. Our estimator reduces to the existing estimator when there is no smoothing and we see from the figures that for no smoothing the bias is almost zero.

- We observe that for higher values of the partial sill and thus, the sill, that larger amount of smoothing is needed to minimize the mse.

- The mse is very sensitive to the nugget. Smoothing reduces the mean squared error and the variance dominates the bias to the extent that we are unable to observe the minima for the mse for the case when lag distance is equal to 0. However, for lag distance is equal to 1.4 we do observe that higher value of the nugget results in more smoothing being needed to minimize the mse.

- We observe a clear minima of the mse for about 10% smoothing in most cases when the sill of the process is not very large.

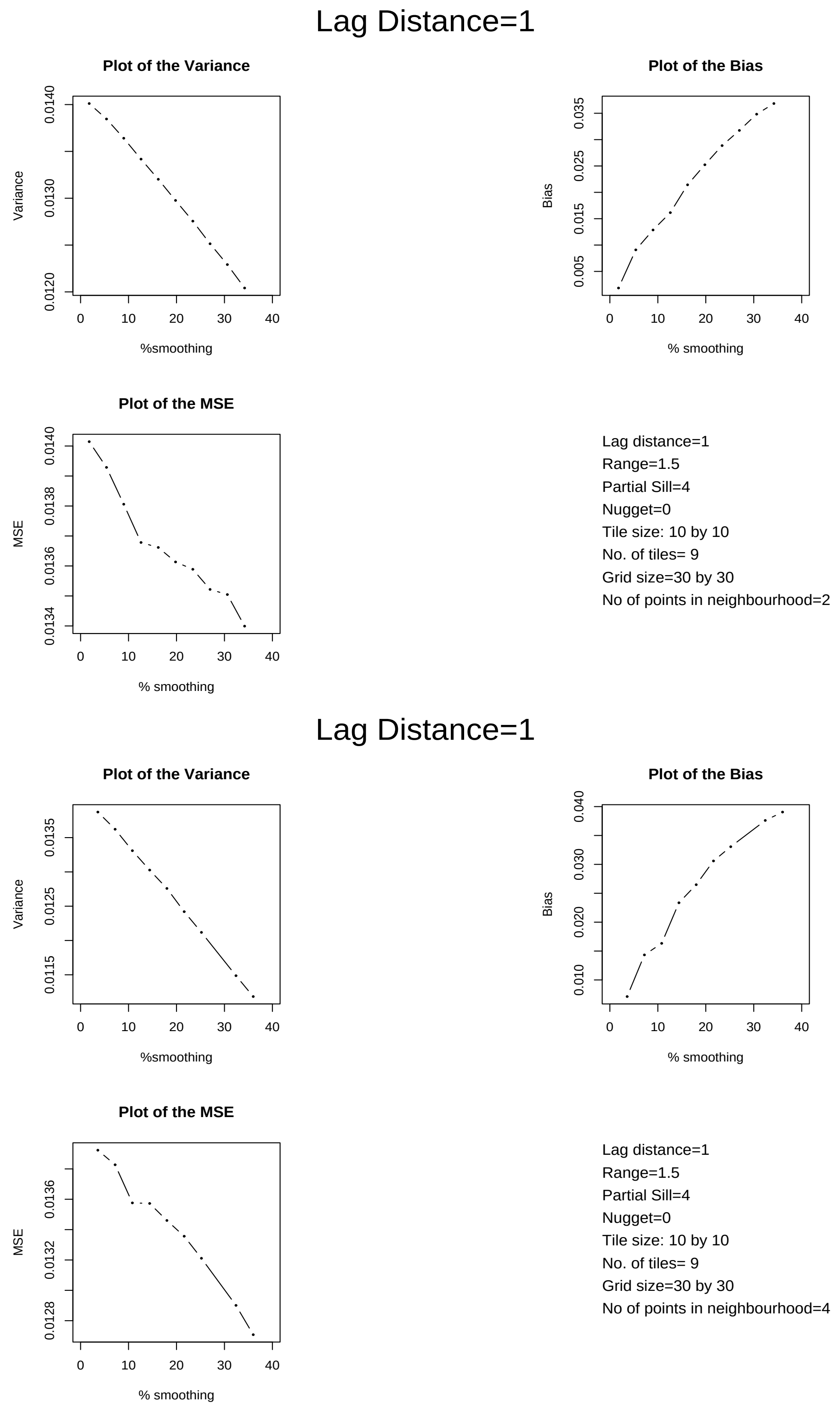
- The variance of the process is of the same order of magnitude as the bias and thus, the mse is affected more by the value of the variance. This can be seen in cases when the sill is large as the decreasing variance controls the behaviour of the mse and we see that the mse decreases with increasing smoothing too.

- The mse is sensitive to changes in the range of the process. When the range is too small then we do not observe the minimum of the mse as the mse keeps decreasing with smoothing. This could perhaps be due to the fact that we are over smoothing the process. When the range is large then this means that the process is smoother and thus the variance of the estimator decreases and is dominated by the bias.

- The minimization of the mse is more obvious for the lag distance of 1.4 as we observe that the mse is minimized for about 10% of smoothing in most cases.

- The size of the neighborhood does not appear to play a big role. Although from the graphs for lag distance equal to 1.4 we do see that when the neighbourhood is larger the amount of smoothing needed to minimize the mse is lesser.

Based on these results we would suggest a smoothing of about 10% of the points on the boundary with at least 2 points in the neighbourhood of each point on the boundary.



Lag distance=1
Range=1.5
Partial Sill=4
Nugget=0
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=2

Lag distance=1
Range=1.5
Partial Sill=4
Nugget=0
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=4

Figure 3.9: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size

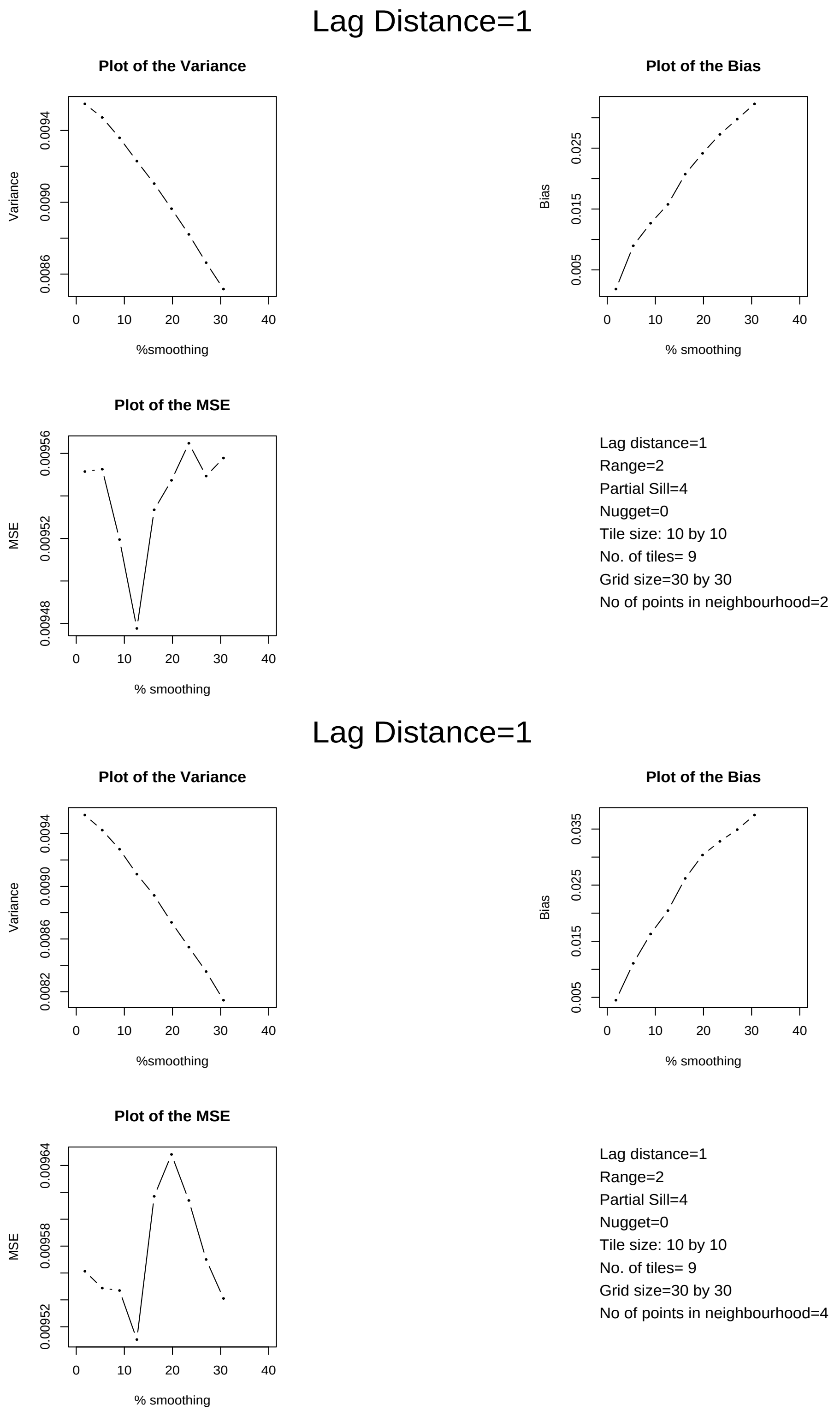


Figure 3.10: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size

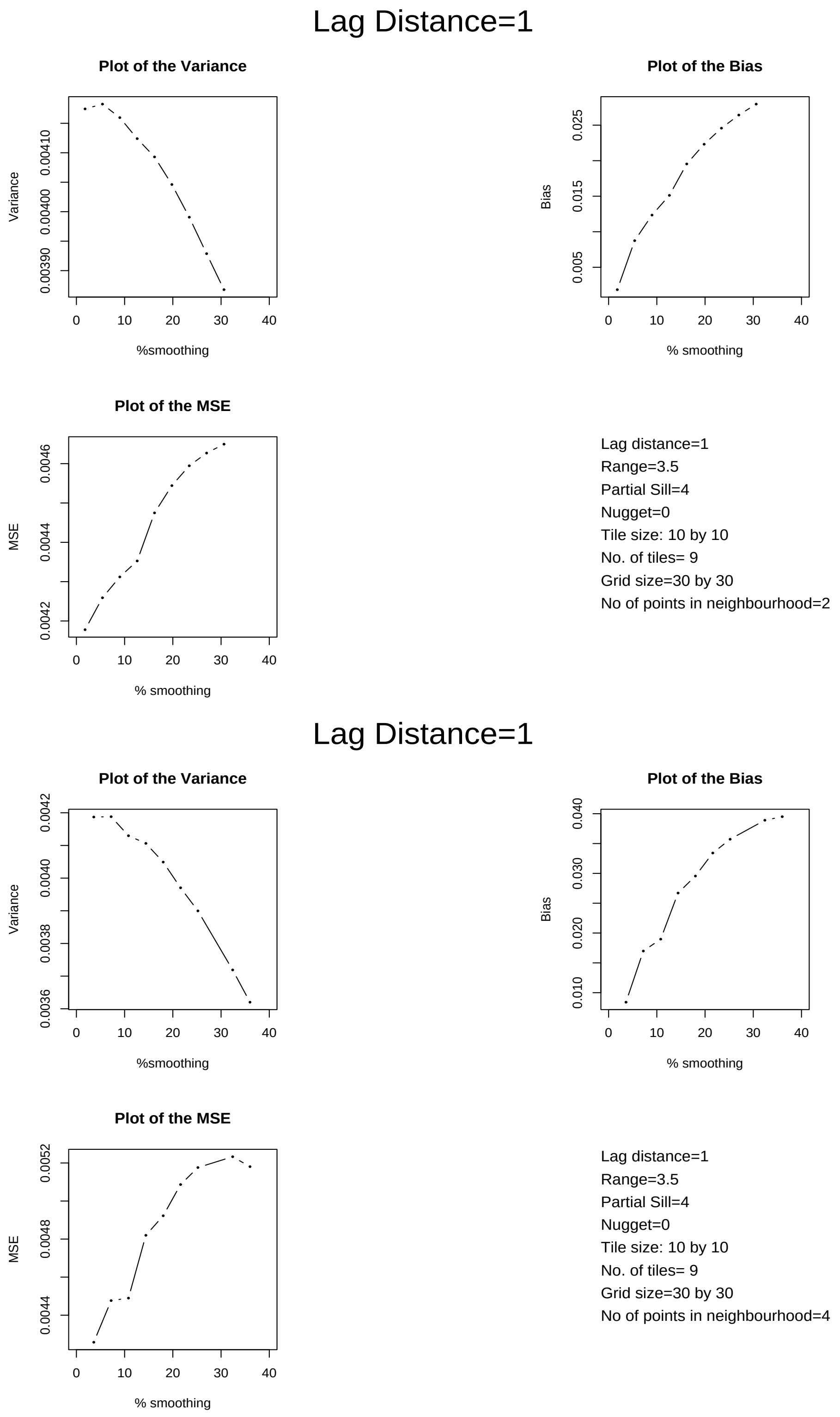
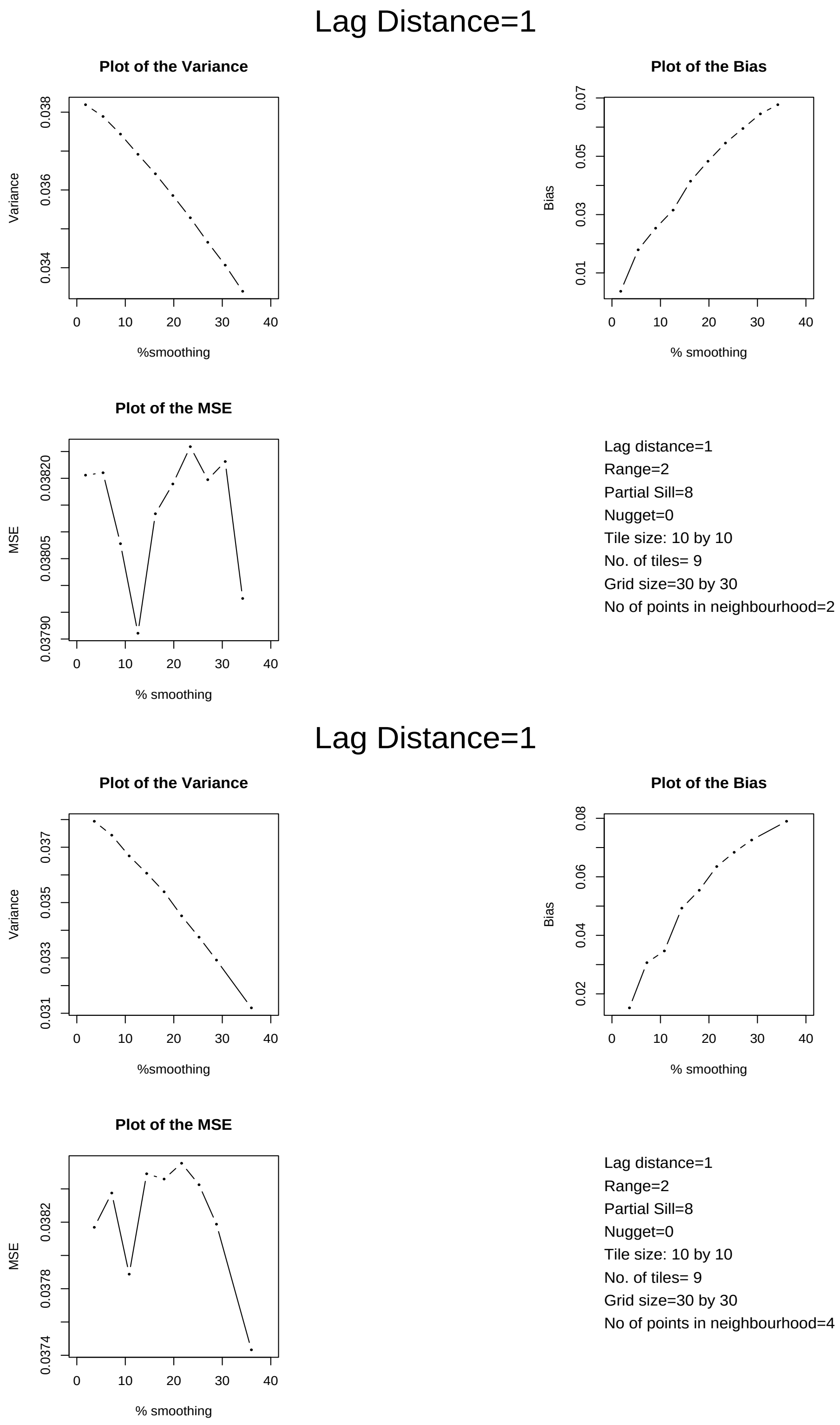


Figure 3.11: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size


Lag Distance=1

Lag distance=1
Range=2
Partial Sill=8
Nugget=0
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=2

Lag Distance=1

Lag distance=1
Range=2
Partial Sill=8
Nugget=0
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=4

Figure 3.12: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size

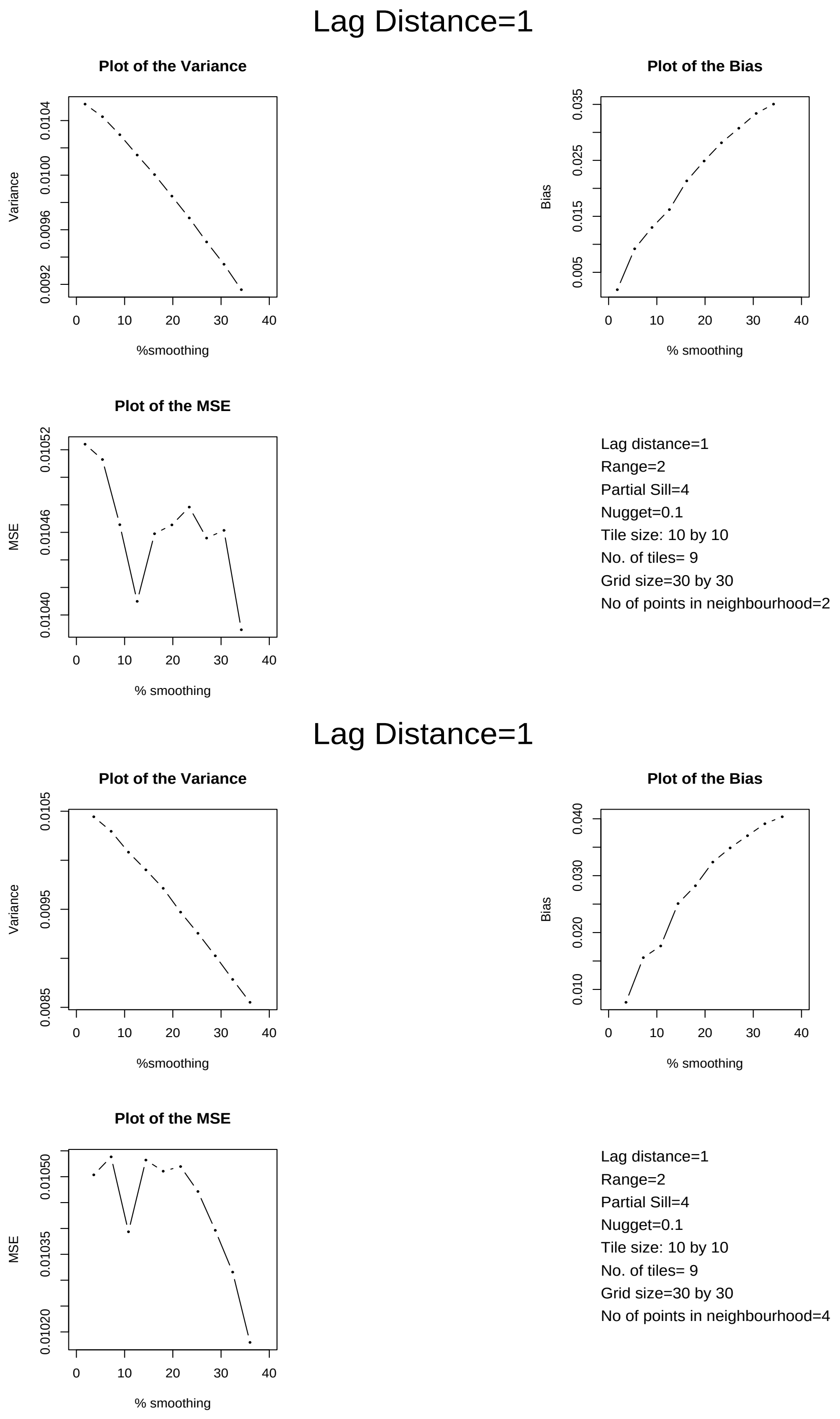


Figure 3.13: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size

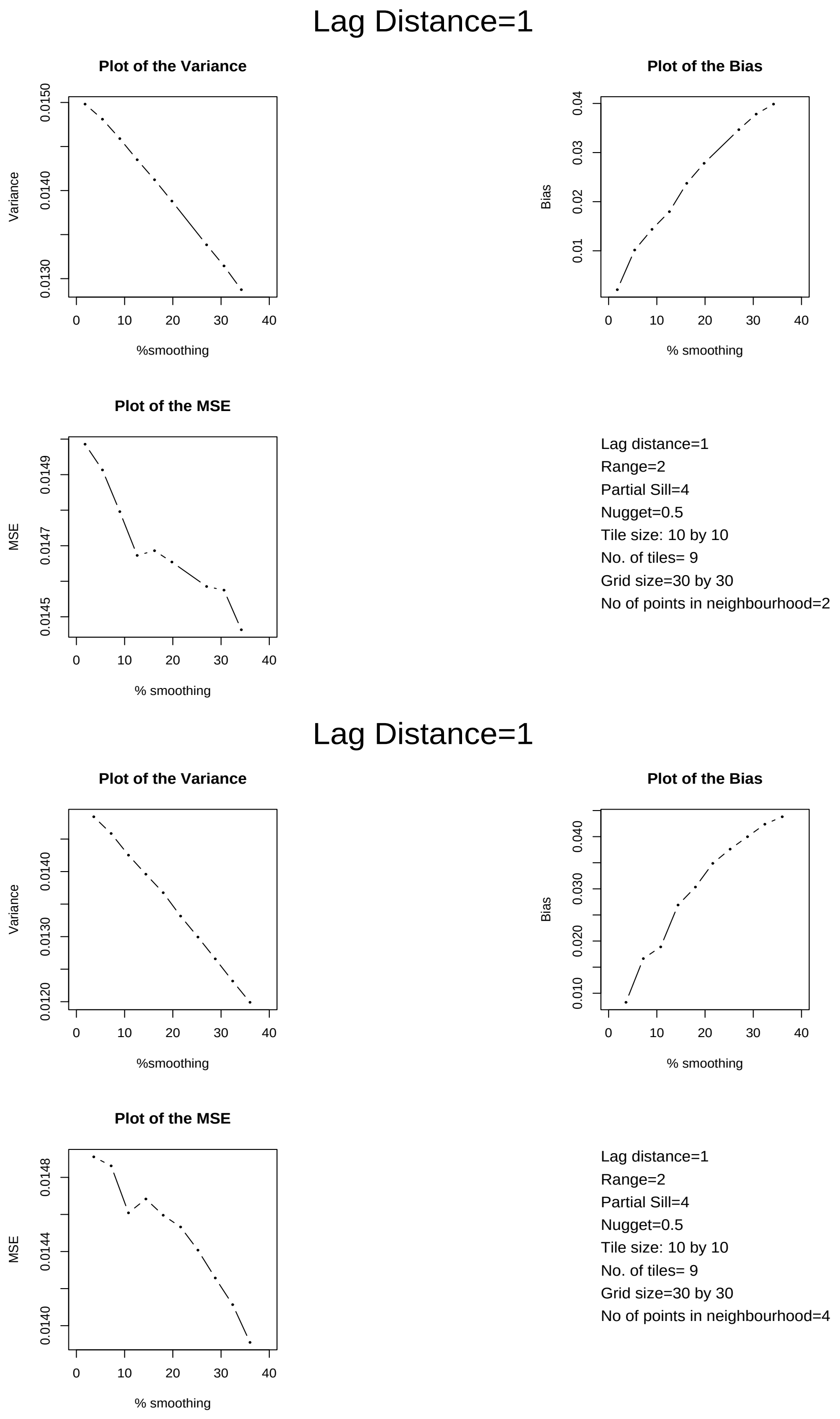
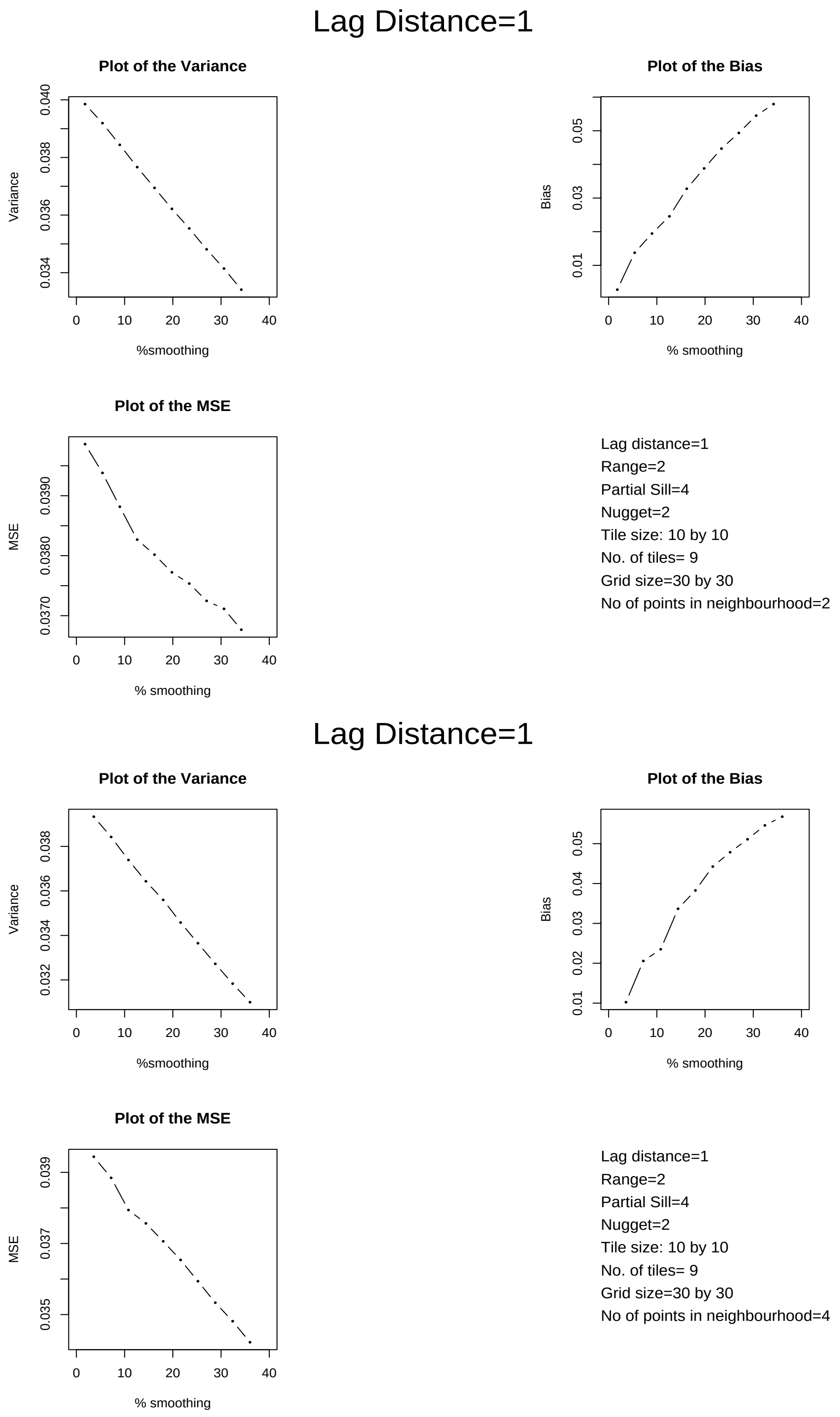


Figure 3.14: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size



Lag distance=1
Range=2
Partial Sill=4
Nugget=2
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=2

Lag distance=1
Range=2
Partial Sill=4
Nugget=2
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=4

Figure 3.15: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size

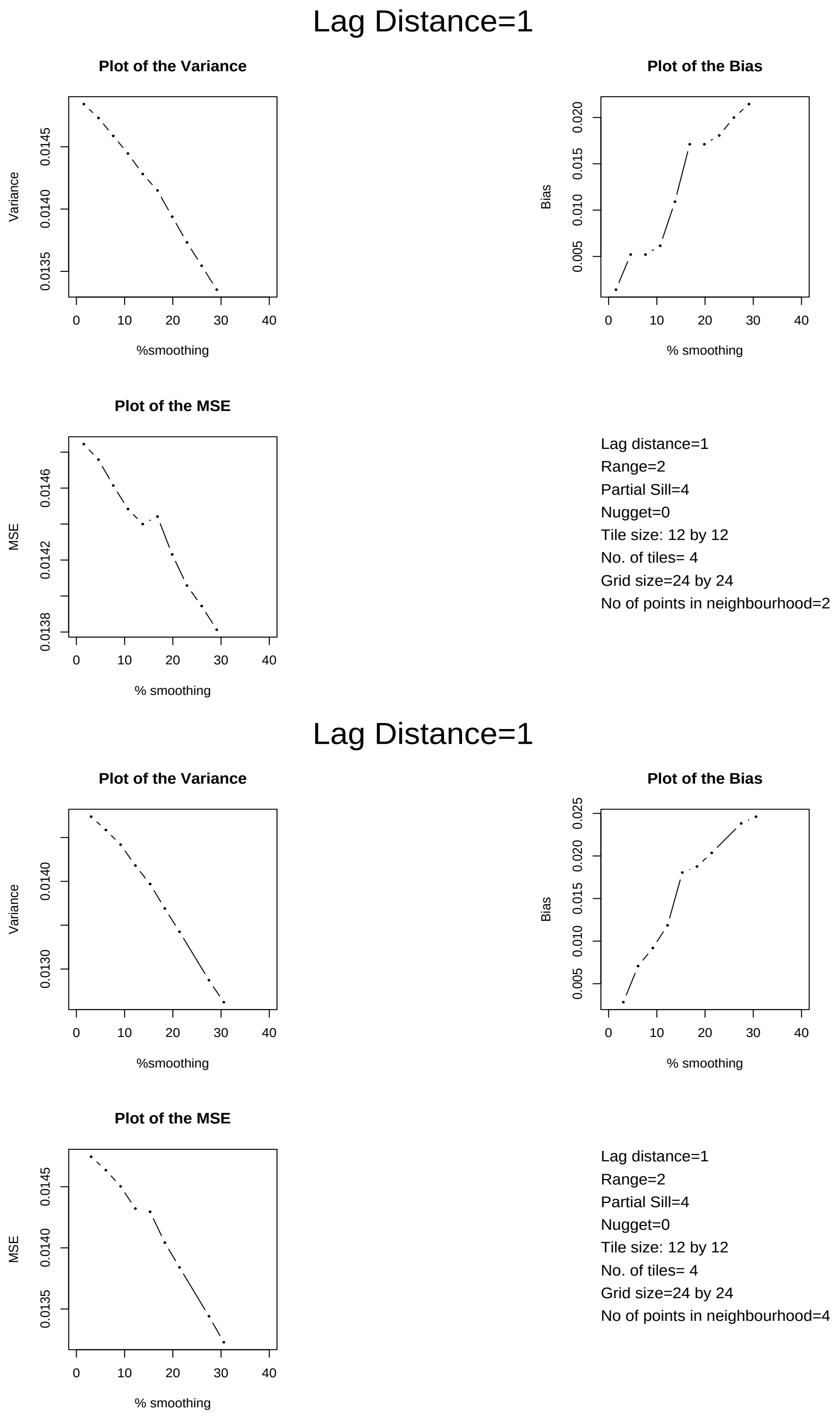
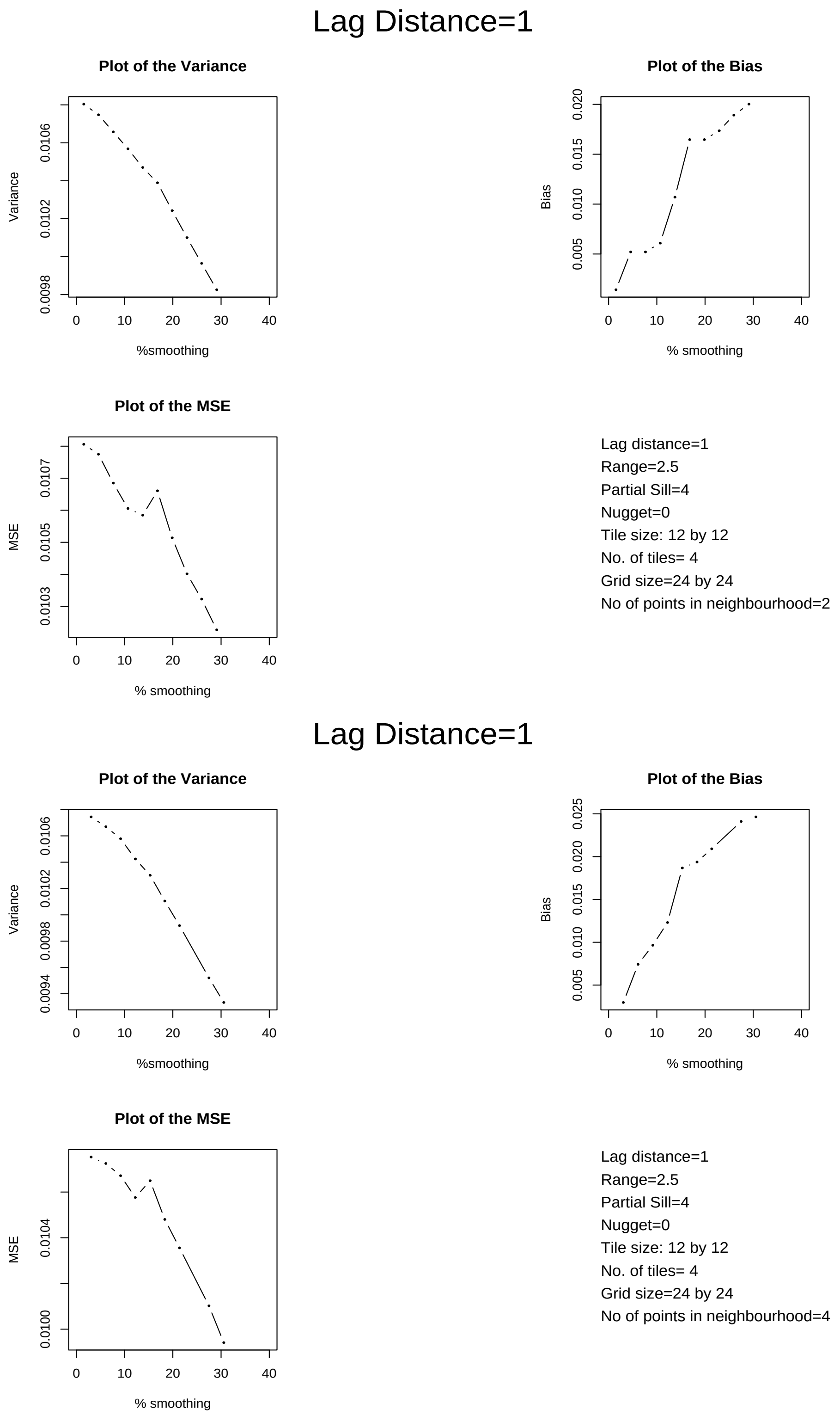


Figure 3.16: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size



Lag Distance=1

Plot of the Variance

Plot of the Bias

Plot of the MSE

Lag distance=1
Range=2.5
Partial Sill=4
Nugget=0
Tile size: 12 by 12
No. of tiles= 4
Grid size=24 by 24
No of points in neighbourhood=2

Lag Distance=1

Plot of the Variance

Plot of the Bias

Plot of the MSE

Lag distance=1
Range=2.5
Partial Sill=4
Nugget=0
Tile size: 12 by 12
No. of tiles= 4
Grid size=24 by 24
No of points in neighbourhood=4

Figure 3.17: Results for MSE for lag distance=1 for specified settings of the range, partial sill, nugget and sample size

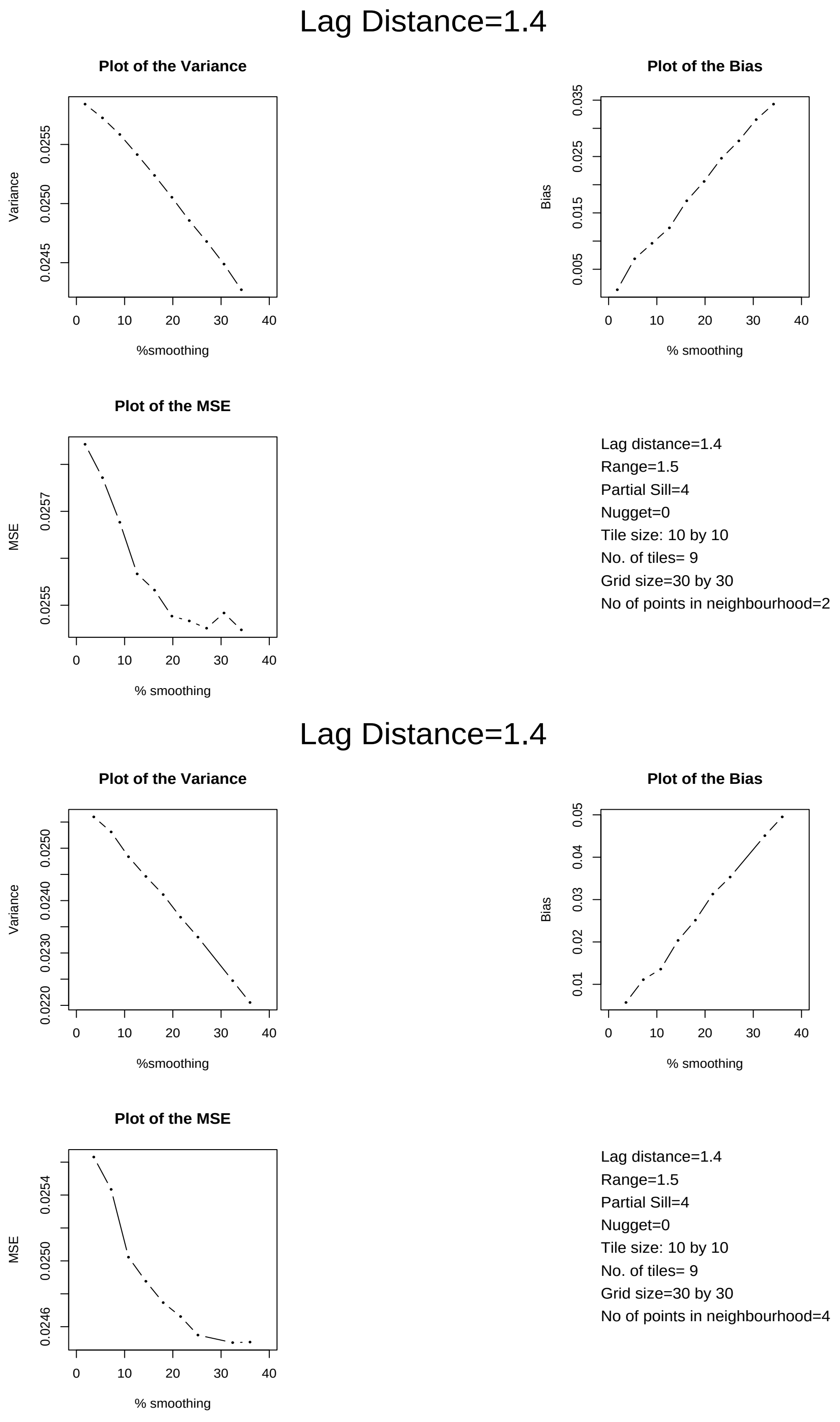
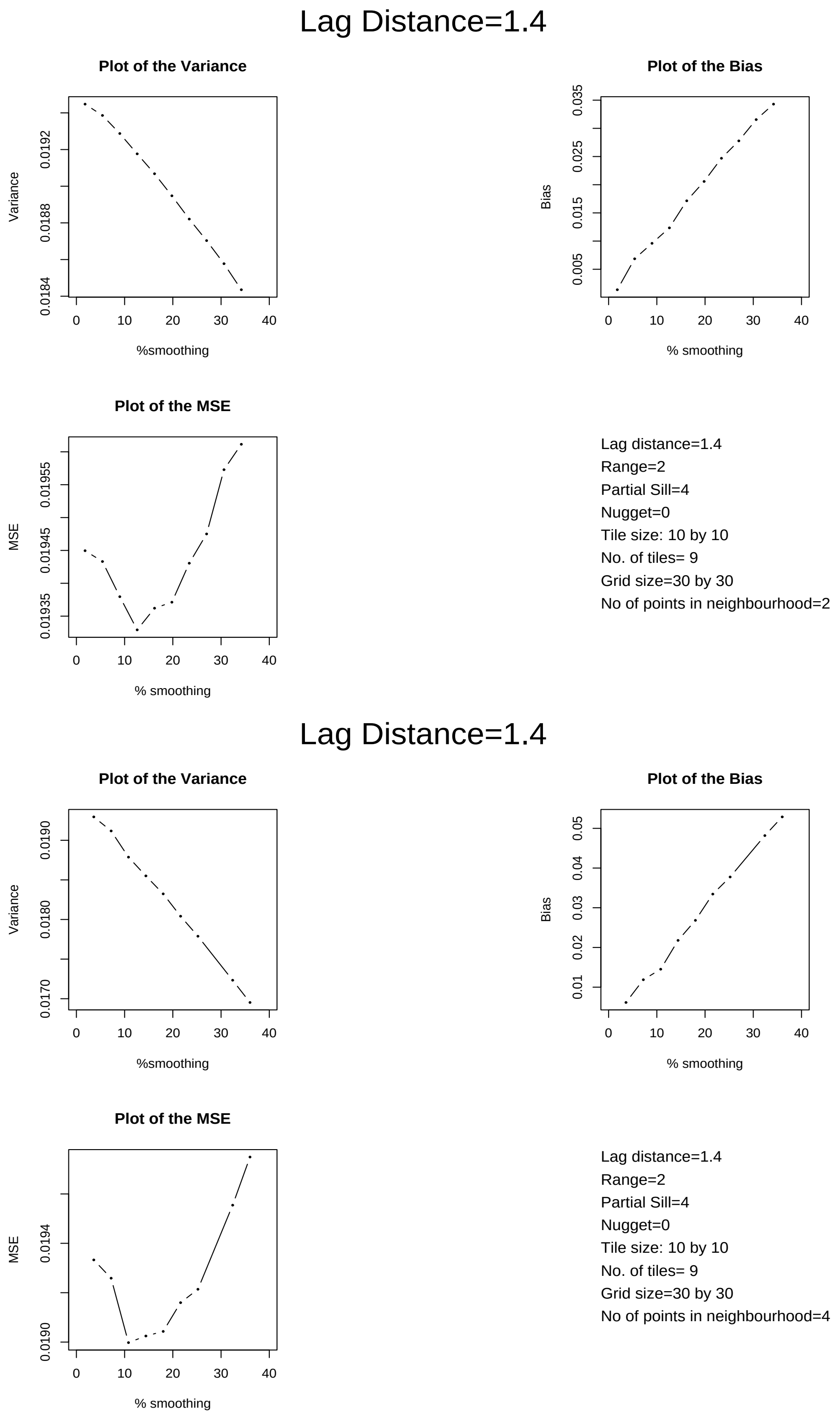


Figure 3.18: Results for MSE for lag distance=1.4 for specified settings of the range, partial sill, nugget and sample size



Lag distance=1.4
Range=2
Partial Sill=4
Nugget=0
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=2

Lag distance=1.4
Range=2
Partial Sill=4
Nugget=0
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=4

Figure 3.19: Results for MSE for lag distance=1.4 for specified settings of the range, partial sill, nugget and sample size

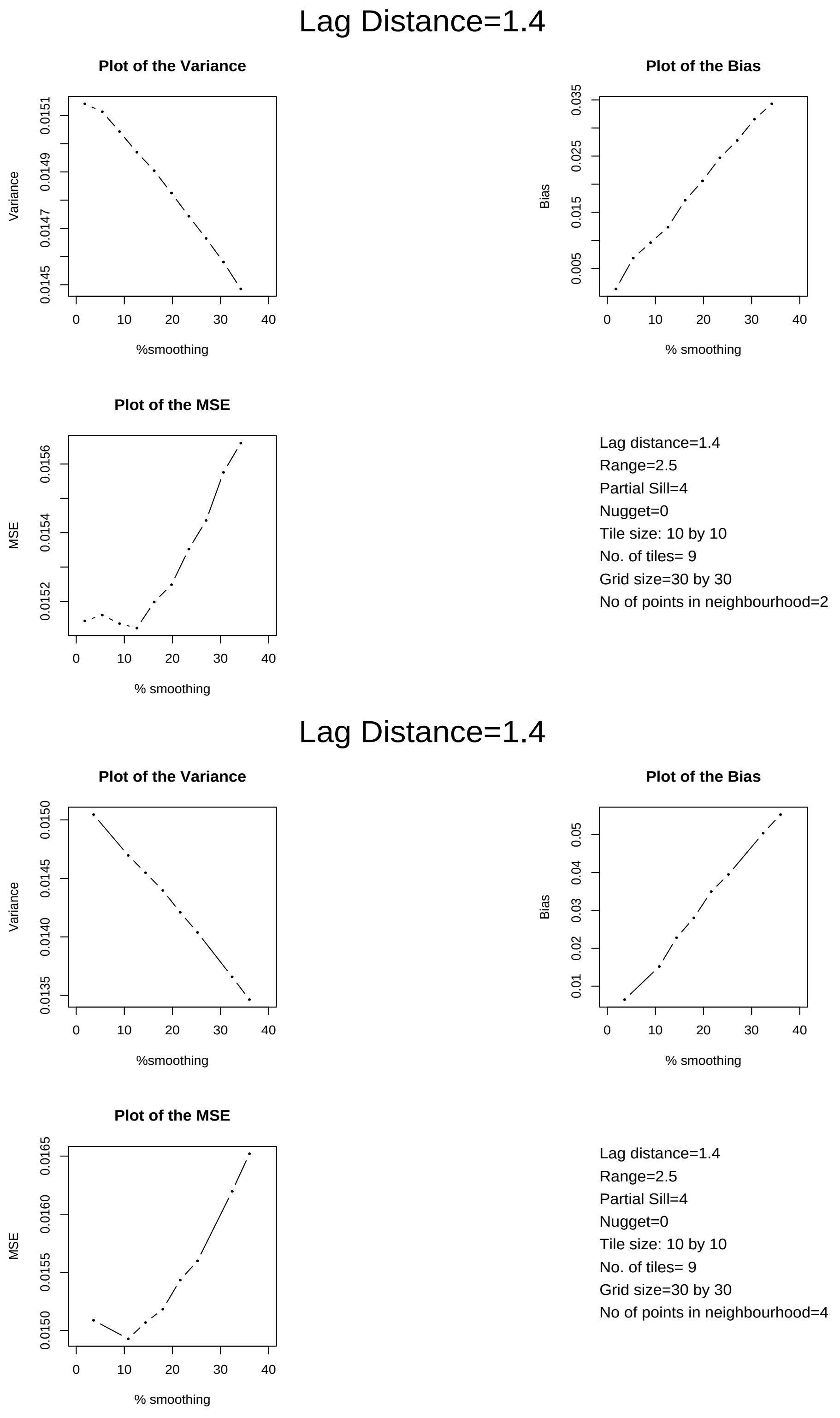


Figure 3.20: Results for MSE for lag distance=1.4 for specified settings of the range, partial sill, nugget and sample size

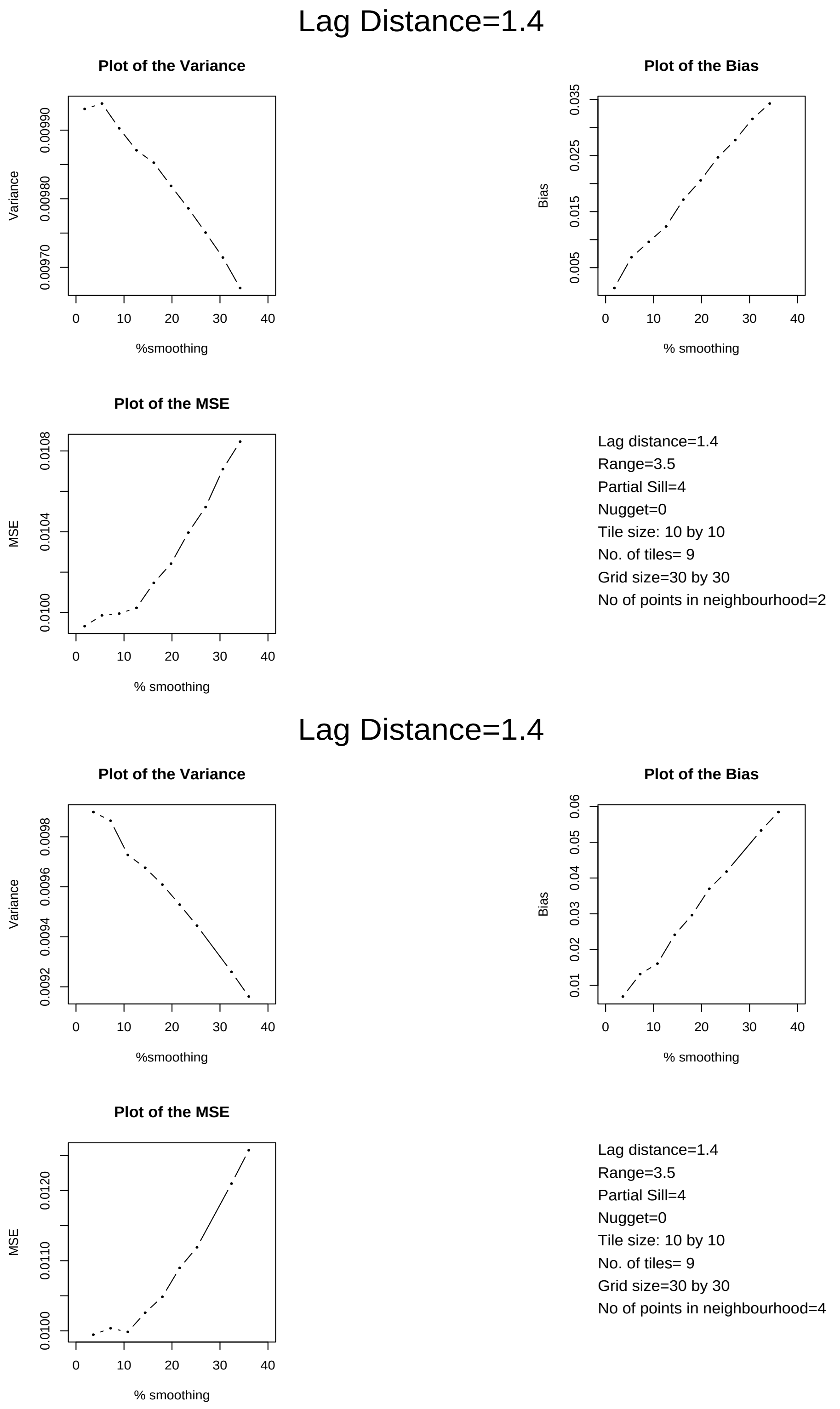


Figure 3.21: Results for MSE for lag distance=1.4 for specified settings of the range, partial sill, nugget and sample size

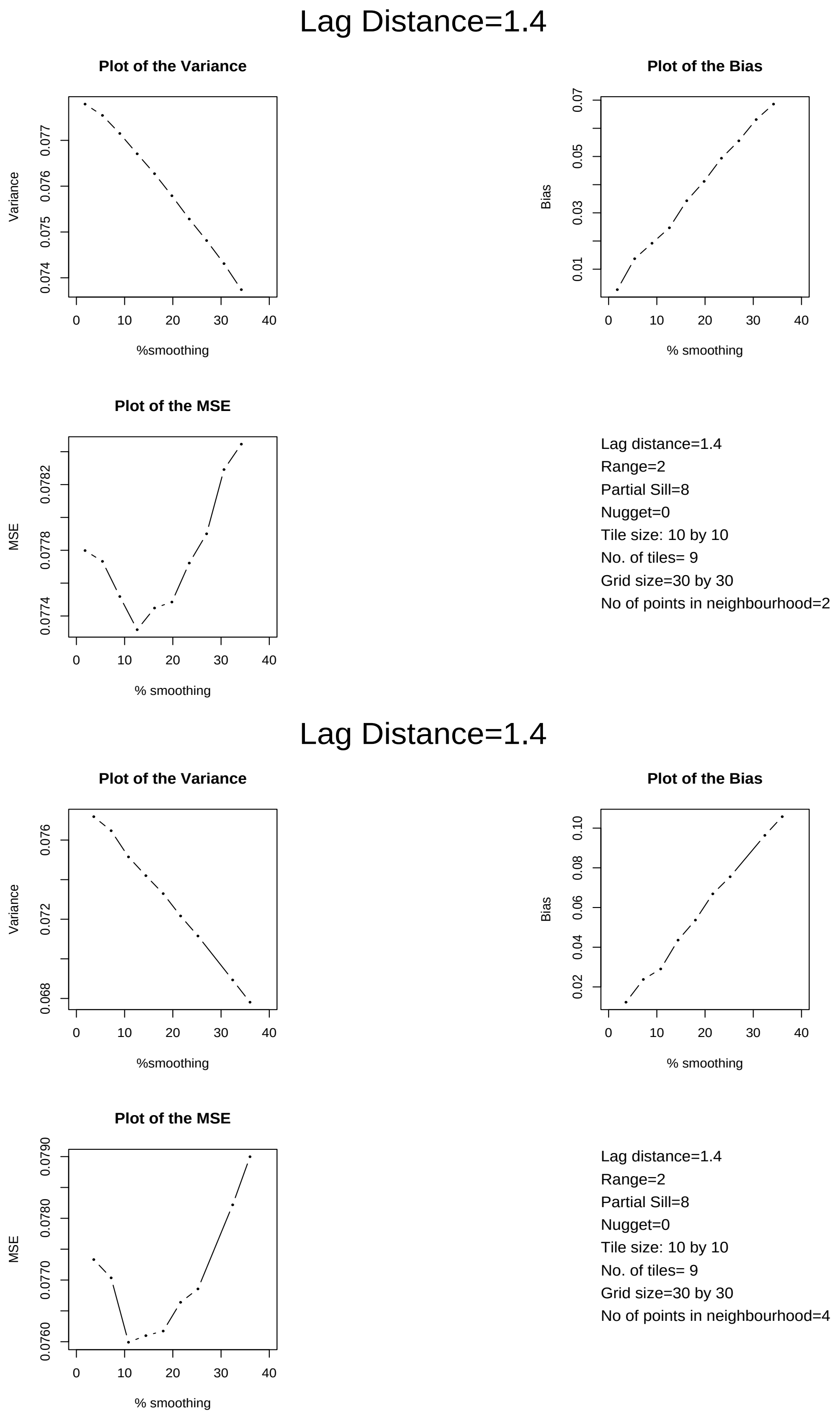


Figure 3.22: Results for MSE for lag distance=1.4 for specified settings of the range, partial sill, nugget and sample size

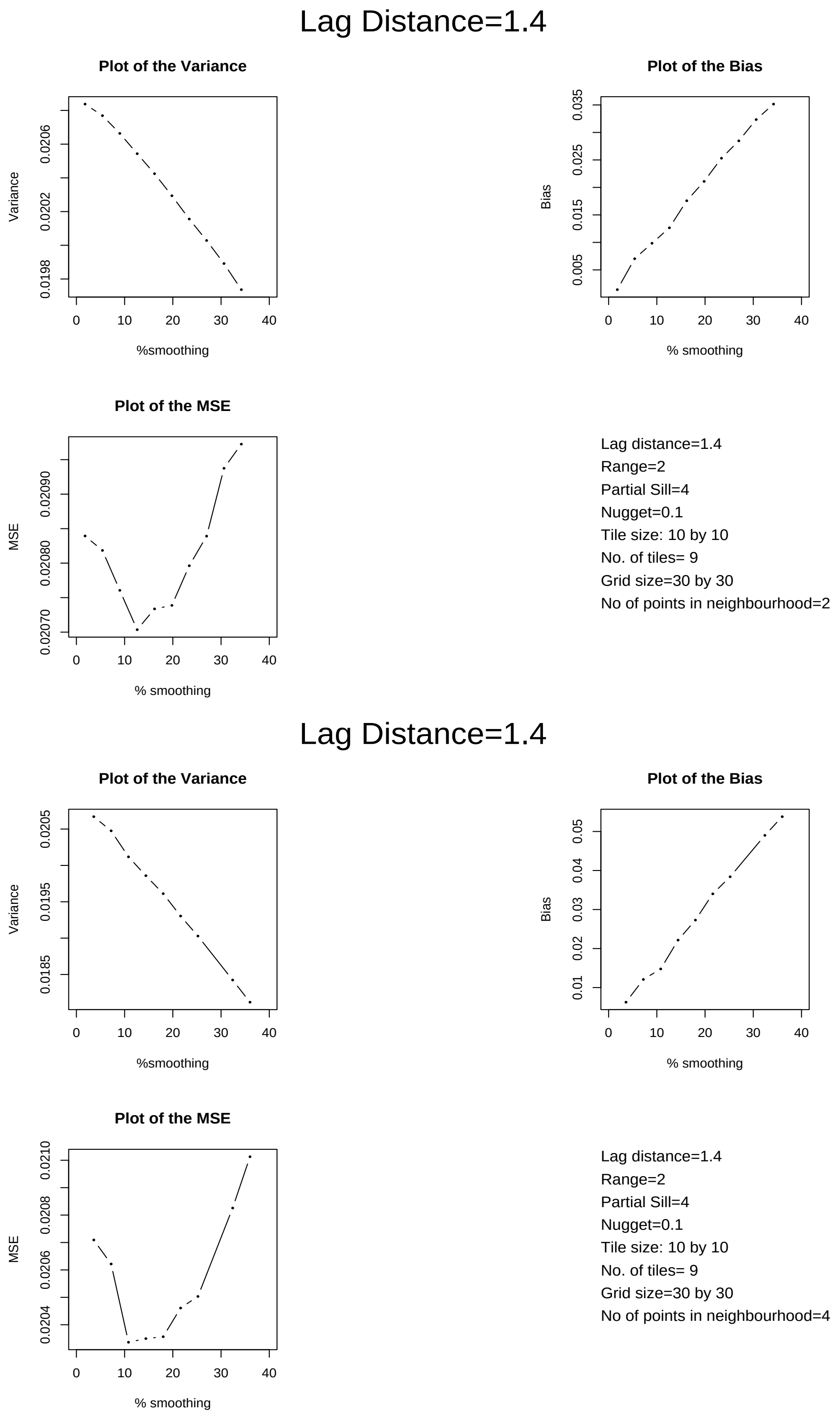
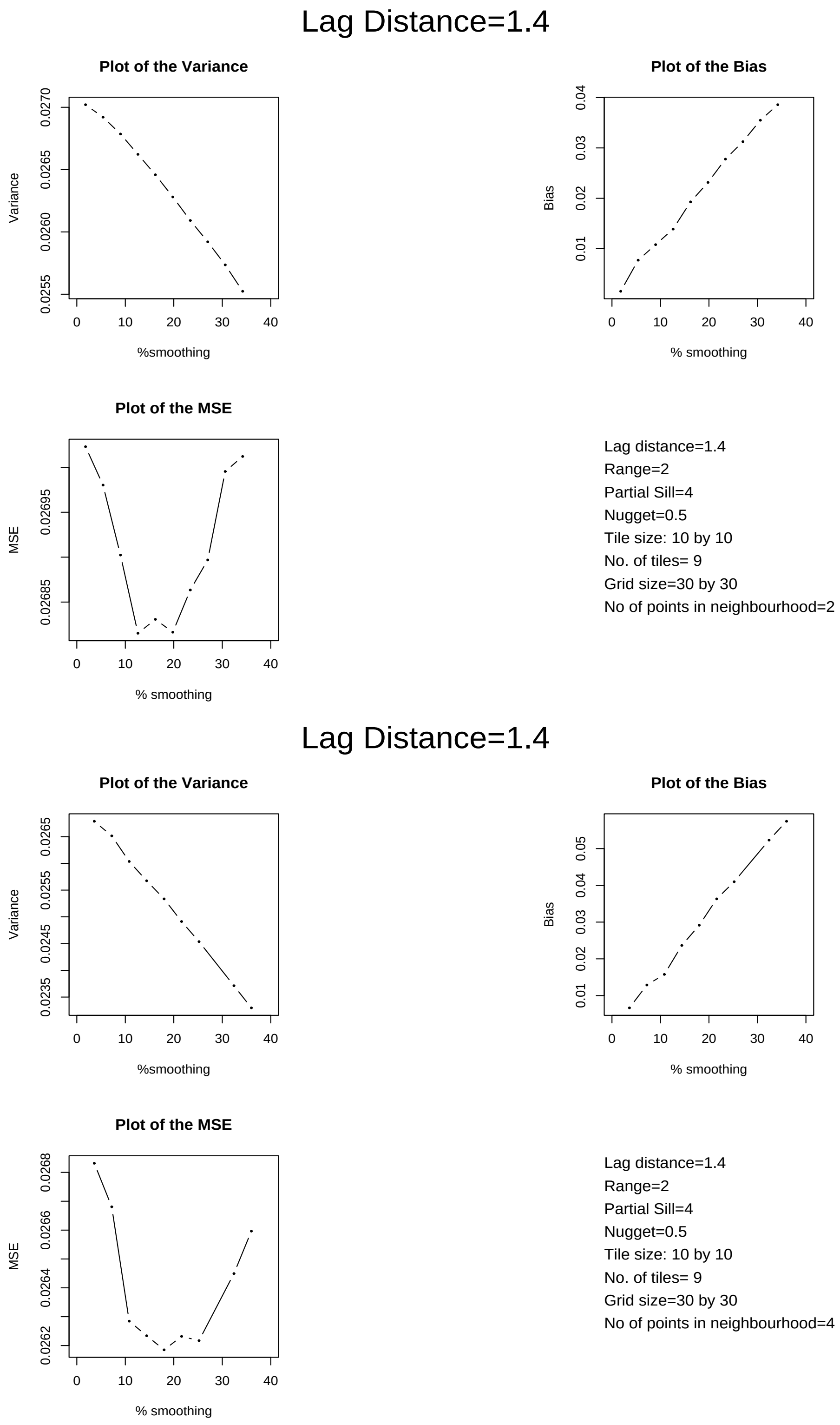


Figure 3.23: Results for MSE for lag distance=1.4 for specified settings of the range, partial sill, nugget and sample size



Lag distance=1.4
Range=2
Partial Sill=4
Nugget=0.5
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=2

Lag distance=1.4
Range=2
Partial Sill=4
Nugget=0.5
Tile size: 10 by 10
No. of tiles= 9
Grid size=30 by 30
No of points in neighbourhood=4

Figure 3.24: Results for MSE for lag distance=1.4 for specified settings of the range, partial sill, nugget and sample size

Chapter 4

Analysis of Coastal Data

4.1 Background

The North Carolina Department of Transportation (NC DOT) finished construction of the terminal groin on the south side of Oregon Inlet, NC in March 1991. To determine the changes along the northern end of Pea Island after the construction of the groin, a monitoring program was initiated. The data has been collected by the NC DOT and is in the form of distance of the shoreline to the baseline at different locations along the 6.5 mile long coastline. The locations at which the measurements are made are regularly spaced, 150 feet apart and are referred to as transects. The distance is measured perpendicular to the baseline. Further, these measurements have been made at two month intervals starting in August of 1991 continuing up till the present time. Thus the data can be presented in the form of a two dimensional table with one axis representing time and the other representing the transect. For a particular transect an increase in the distance implies erosion and a decrease implies accretion. We refer to this erosion/accretion process as shoreline change.

Several basic criteria must be met for a given physical feature to be considered a satisfactory shoreline indicator. The shoreline is a line on a map where the land and water meet. For the purpose of this study the shoreline was defined as the edge of wet sand visible on the aerial photographs. The accuracy of our data depends heavily on the accuracy of the shoreline mapping. The shoreline position is digitized directly from the aerial photographs by the Photogrammetry Unit of NC DOT. The digitized data are then transferred to North Carolina State University and computer processed to determine the distance from the baseline at the 150 foot intervals.

4.1.1 Definitions

Definitions of some terms that are related to coastal engineering and are relevant to this work are given below.

- *Storm surge* is simply water that is pushed toward the shore by the force of the winds swirling around the storm. It can be calculated as the difference between the actual tide height and the predicted tide height.

- *SEPI* is an acronym for the Storm Erosion Potential Index that was proposed by Zhang et al. (Zhang, Douglas & Leatherman 2001) as measure of the intensity of a storm. It is given by $SEPI = \sum_{t=0}^{t_D} S_{2SD}(t)H_{MHHW}(t)\delta(t)$, where, t_D is the integer number of hours of storm duration, S_{2SD} is the hourly value of storm surge that is greater than 2 standard deviations of the annual surge level, $H_{MHHW}(t)$ is the water level greater than MHHW and $\delta(t)$ is taken to be an hour.

- R_{HIGH} (Sallenger 2000) is an empirical index and is defined as follows: $R_{HIGH} = H_0(0.83\zeta_0 + 0.2)$, where, H_0 is the deep water significant wave height, ζ_0 is the Irribarren number. The Irribarren number is given by $\beta/(H_0/L_0)^0.5$, where β is the beach slope and L_0 is the deep water wave length.

- *Kriebel's index* (Kriebel, Dalrymple, Pratt & Sakovich 1996) measures the storm intensity and is defined as follows. If S is the local storm surge height(maximum storm surge+ maximum tidal height) at a given coastal location, H is the offshore significant wave height and t is the duration of the storm event then Kriebel's Index= $SH(t/12)^{0.3}$.

The Field Research Facility (FRF) , which is a part of the Civil Engineering branch of the US Army, has made publicly available, through their web site an enormous amount of data on parameters such as wave height, wave time period, tide height, surge etc.

In our analysis we modeled the shoreline change process using the information available on storms and related parameters. Let us denote the shoreline change process by $Y(s,t)$, where, s represents the spatial location along the coastline and t represents the temporal location of the observation. This process can be expressed as a sum of two components

$$Y(s,t) = \mu(s,t) + \epsilon(s,t) \ , \ \epsilon \sim (\mathbf{0}, \Sigma_\epsilon) \tag{4.1}$$

Here, $\mu(s,t)$ represents the mean structure and $\epsilon(s,t)$ represents the error. Now, the mean structure can be decomposed into additive components c, $f(t)$, $g(s)$. All the three components c, $f(t)$ and $g(t)$ have the same units of distance as the process $Y(s,t)$. Our analysis has three distinct steps:

- De-trend (decompose) the data.
- Model the components $f(t)$ and $g(s)$.
- Model the underlying covariance structure Σ_ϵ.

We investigated for interaction between time and space and found that there was not any significant interaction. Preliminary analysis indicated that the underlying covariance process was non-stationary. However, the covariance process was found to be locally stationary over sections of the island. So we divided the island into three regions that we shall refer to as the Northern ($\approx$ 2.2 km), Central ($\approx$ 1.8 km) and Southern ($\approx$ 4.6 km) regions from here on. The analysis is now done for each of the three regions separately and we present the results over the next few sections. As the first step in our analysis we need to detrend our data. Median polish is a technique that is used to remove trend in gridded data and in the following section we discuss this technique.

4.2 Median Polish

Median Polish is a method for removing trend in gridded data. Here we use it to extract the trend due to the location of the transect and the time of observation. The decomposition of the data is additive: data=all+row+col+residual.
The algorithm successively sweeps medians out of rows , then columns, then rows, then columns, and so on, accumulating them in "row", "column" and "all" registers and leaves behind the table of residuals, where, "all" is the global mean effect, "row" and "col" are the row and column effects respectively.

Below we describe an iterative approach to obtain the median polish for a spatial process Z.

1. Assume data to be on a $p \times q$ rectangular grid $\{(x_l, y_k) : k = 1, \cdots, p; l = 1, \cdots, q\}$. Regard the grid nodes as cells in a 2-way table.

2. Operate iteratively on the data. Alternately subtracting row medians and column medians and accumulating these medians in an extra column and row of cells.

3. Repeat this procedure until another iteration produces virtually no change.

4. Final entries in the extra cells are the median polish estimates or row effects $r_1, \cdots, r_p$, column effects $c_1, \cdots, c_q$ and an overall effect a.

5. The final entries in the body of the table are residuals, $\widehat{e_{kl}}$, such that $Z(x_l, y_k) = \widehat{a} + \widehat{r_k} + \widehat{c_l} + \widehat{e_{kl}}$

The main effects - row effects and the column effects, are the quantities that minimize $\sum_{l=1}^{q}\sum_{k=1}^{p}|Z_{lk} - a - r_k - c_l|$.

We choose to use this technique over ordinary least squares as medians have the property of being resistant to outliers and also it has been shown by Cressie that the residuals are less biased in their estimation of the second order spatial dependence structure as compared to the residuals from ordinary least squares.

Putting this in context of the decomposition that was given in (4.1) we can express the shoreline change process as follows:

$$Y(s,t) = C + f(t) + g(s) + \epsilon(s,t) \ , \ \epsilon \sim (\mathbf{0}, \Sigma_\epsilon) \qquad (4.2)$$

where, $f(t)$ gives the main effect of time, $g(s)$ gives the main effect of the spatial location (location on the coastline) and C is a constant overall mean. In our data if we let time be along the y-axis and the spatial location along the x-axis then the row effect obtained from the median polish would correspond to the main effect of time and the column effect would correspond to the main effect of spatial location. We denote the row effect (main effect of time) by $f(t)$ and the column effect (main effect of spatial location) by $g(s)$. We have 48 values for $f(t)$, $t = 1, \ldots, 48$ and for $g(s)$ we have 48 values for the northern region, 40 values for the central region and 100 values for the southern region. The number of values for $g(s)$ is equal to the number of transects that lie within a region. The next section discusses the modeling of the main effects of time and of the spatial location obtained from the application of median polish to the data. In section 5.4 we discuss the analysis of the residuals.

4.3 Modeling the mean structure

Figures 4.1, 4.2 and 4.3 show plots of the temporal effect for the three regions along with the 95% confidence intervals, and Figures 4.4, 4.5 and 4.6 shows plots of the spatial effect for the three regions. To illustrate the calculation of the variance and the confidence intervals consider the following example:
Let $Y(t)$ be a stationary, AR(1) process given by $Y(t) = \boldsymbol{\alpha X}(\mathrm{t}){+}\beta Y(t-1) + e(t)$, where $\boldsymbol{X}(t)$ is a vector of covariates. The errors $e(t)$ are *iid* and come from a Gaussian distribution with mean 0 and variance σ^2. Given data that corresponds to this process we can obtain estimates for the parameters $\boldsymbol{\alpha}$ and β and for the variance σ^2. Denote the estimates by $\widehat{\boldsymbol{\alpha}}$, $\widehat{\beta}$ and $\widehat{\sigma^2}$. Using these estimates for $\boldsymbol{\alpha}$ and β we can obtain estimates for the process $Y(t)$. $\widehat{Y}(t) = \widehat{\boldsymbol{\alpha}}\boldsymbol{X}(\mathrm{t}){+}\widehat{\beta}Y(t-1)$ for $t = 2, \cdots n$ and $\widehat{Y}(1) = Y(1)$. Then the variance of these estimates $\widehat{Y}(t)$ can be approximated by $Var(Y(t) - \widehat{Y}(t)) = \sigma^2$. This approximation is valid if we make the assumption that the estimates for the parameters are consistent. Since the errors are normally distributed the confidence intervals would be given by $(\widehat{Y}(t) - 1.96\sigma, \widehat{Y}(t) + 1.96\sigma)$. For predicting at one period in the future we have the linear predictor given by $\widehat{Y}_{n+1} = \widehat{\boldsymbol{\alpha}}\boldsymbol{X_{n+1}} + \widehat{\beta}Y_n$. In general we can build predictors for s periods ahead by sub-

stituting the predictors for earlier periods. Therefore, for predicting at some period, s, in the future we have the predictor given by $\widehat{Y}_{n+s} = \widehat{\boldsymbol{\alpha}}\boldsymbol{X_{n+s}} + \widehat{\beta}\widehat{Y}_{n+s-1}$. The variance of the prediction error is approximately given by $\widehat{\sigma^2}\left(1 + \widehat{\beta}^2 + \widehat{\beta}^4 + \cdots + \widehat{\beta}^{2(s-1)}\right)$. Here again we need to make the assumption that the estimates for the parameters are consistent. This variance is an approximation because it ignores the estimation error.

The curvature of the island is reflected in the spatial effect. The distance of the shoreline to the baseline decreases as we approach the central part of the island and then increases as we go away from the central part in either direction. We modeled the temporal effect using some covariates, defined below, based on the wave and tide data amongst others. We investigated several covariates that could possibly help in explaining the process. However, only those that are relevant to the model are given in the following tables. Some covariates that we considered but that are not a part of our model are: the Storm Erosion Potential Index(SEPI), Irribarren number, R_{HIGH} and lag(1) values of some of the covariates.

For the Northern region the linear regression models used to model the temporal effect and the spatial effect are given in table 4.1. The spatial effect was modeled using a polynomial in distance along the coastline

$$f(t) = \beta_0 + \beta_1 X_1(t) + \beta_2 X_2(t) + \beta_3 X_3(t) + \beta_4 X(t) + \delta(t) \tag{4.3}$$

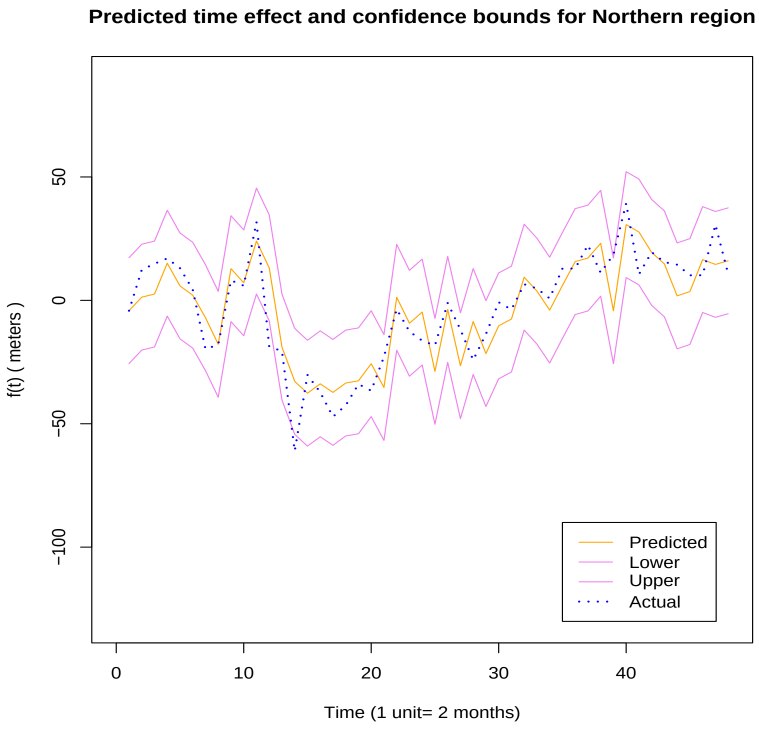


Figure 4.1: Northern Region Temporal Effect

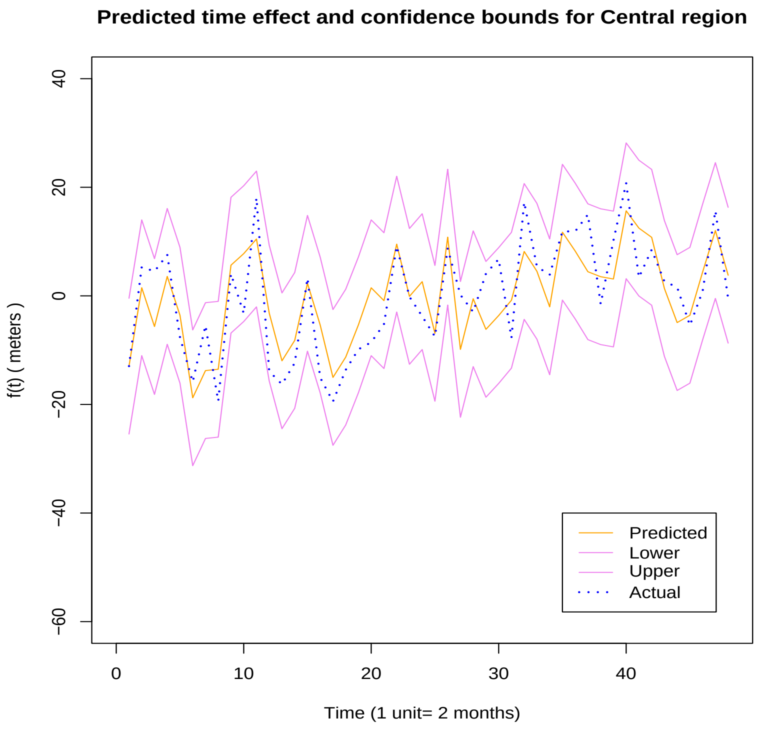


Figure 4.2: Central Region Temporal Effect

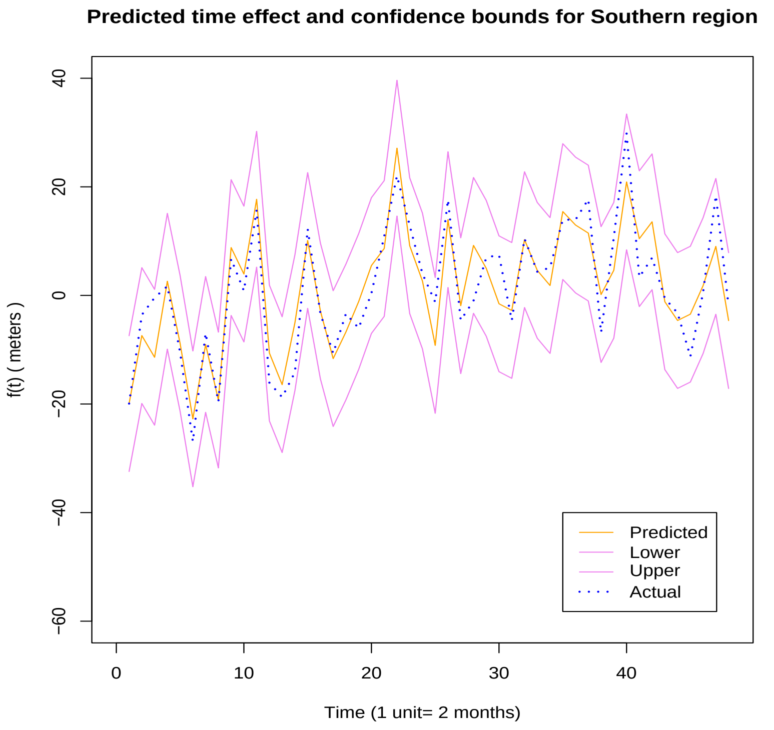


Figure 4.3: Southern Region Temporal Effect

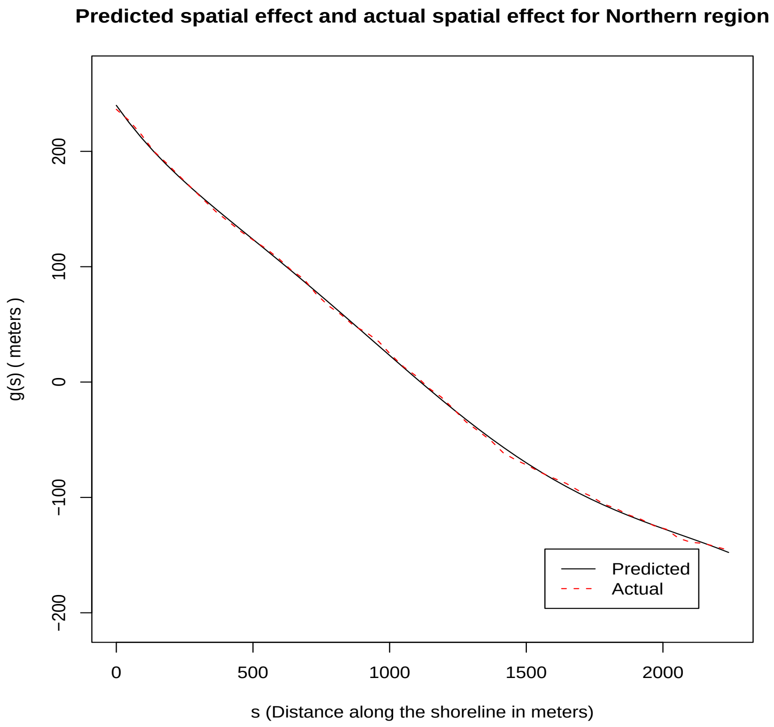


Figure 4.4: Northern Region Spatial Effect

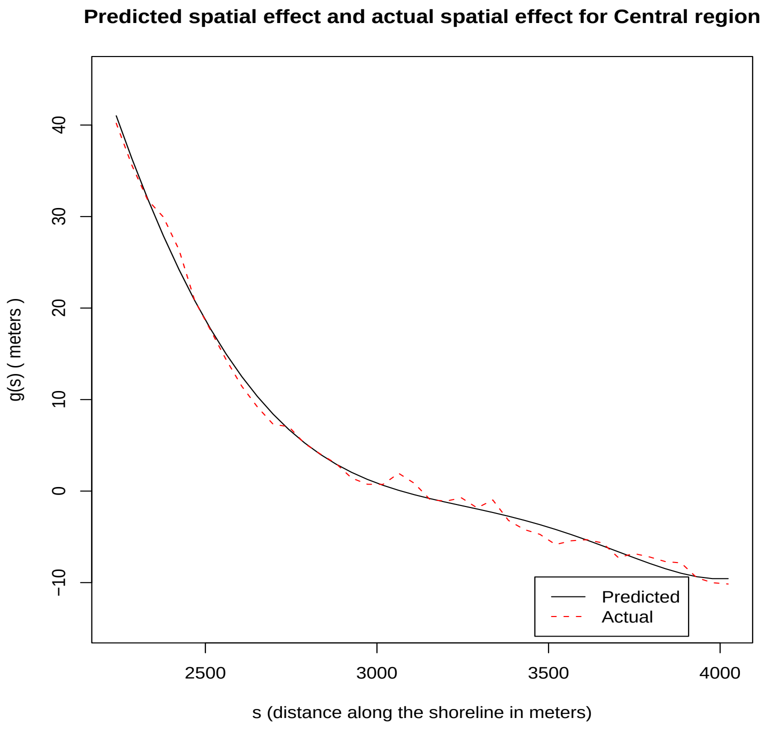


Figure 4.5: Central Region Spatial Effect

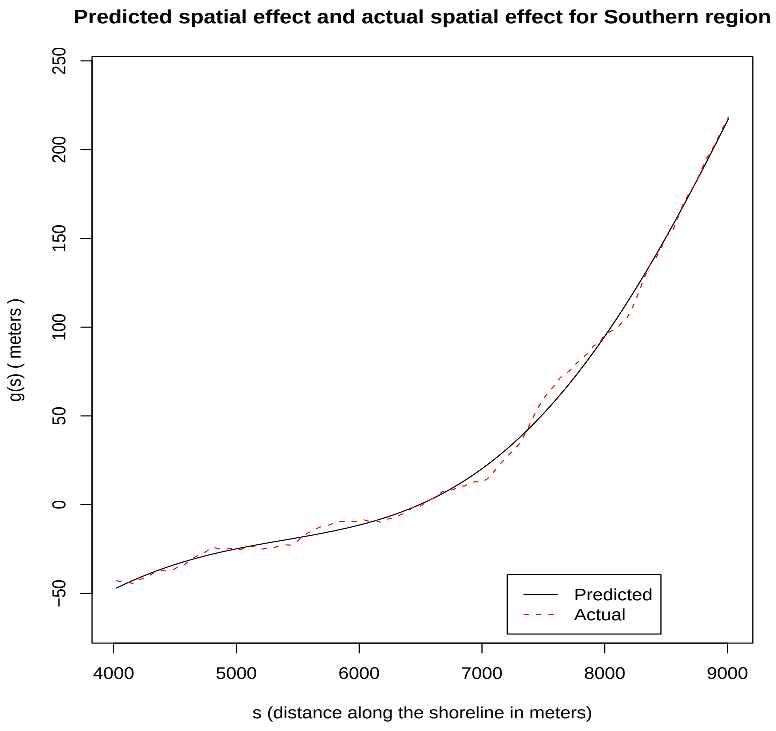


Figure 4.6: Southern Region Spatial Effect

$X_1(t)$:	Maximum surge over one hour immediately prior to the time of photography for the observation at time t.
$X_2(t)$:	Maximum surge over one hour immediately prior to the time of photography for the observation at time $t-1$.
$X_3(t)$:	$f(t-1)$.
$X_4(t)$:	The total number of waves with wave height $>$ 2 meters in the time period between the t^{th} and $t-1^{th}$ observations.

Parameter Estimates for temporal effect

Regression Coeff.	**Estimate**	**Std. Error of Estimate**
β_0	-25.5	8.7
β_1	184.0	39.2
β_2	-100.3	45.7
β_3	0.8	0.09
β_4	0.13	0.04

Parameter Estimates for spatial effect

Regression Coeff.	**Estimate($\times 10^{-4}$)**	**Std. Error of Estimate($\times 10^{-4}$)**
Intercept	840.75	8.97
distance	-12435.97	1010.57
$distance^2$	157712.42	36355.36
$distance^3$	-1905799.81	574116.51
$distance^4$	10903473.27	4434461.77
$distance^5$	-27621746.56	16446784.89
$distance^6$	24600604.35	23418803.24

Table 4.1: Regression Coefficient estimates for Northern region

The model for the central region is given below. The spatial effect was modeled using a 5th degree polynomial in distance along the coastline.

$$\begin{aligned} f(t) = \quad & \beta_0 + \beta_1 X_1(t) + \beta_2 X_2(t) + \beta_3 X_3(t) + \beta_4 X_4(t) + \beta_5 X_5(t) + \\ & \beta_6 X_6(t) + \beta_7 X_7(t) + \delta(t) \end{aligned} \tag{4.4}$$

$X_1(t)$: Mean surge over one hour immediately prior to the the time of photography for the observation at time t.
$X_2(t)$: Maximum water level over 6 hours immediately prior to the time of photography for the observation at time t.
$X_3(t)$: Total mean hourly surge in the time period between the t^{th} and $t - 1^{th}$ observations.
$X_4(t)$: $f(t-1)$.
$X_5(t)$: Kriebel's index at time $t - 1$.
$X_6(t)$: Sine term to model an yearly cycle.
$X_7(t)$: Cosine term to model an yearly cycle.

Parameter Estimates

Regression Coeff.	**Estimate**	**Std. Error of Estimate**
β_0	-29.7	15.3
β_1	72.9	36.7
β_2	48.9	25.1
β_3	0.2	0.07
β_4	0.3	0.1
β_5	-4.5	2.2
β_6	-16.6	5.4
β_7	0.9	5.3

Table 4.2: Regression coefficient estimates for Central region

Parameter Estimates

Regression Coeff.	Estimate($\times 10^{-4}$)	Std. Error of Estimate($\times 10^{-4}$)
Intercept	-4854.95	6681.79
distance	116320.36	107072.23
$distance^2$	-924783.85	680425.15
$distance^3$	3361563.77	2143706.47
$distance^4$	-5791436.86	3348923.94
$distance^5$	3845985.19	2075780.19

Table 4.2: (Continued...)Regression Coefficient Estimates for Central region

Given below is the model for the southern region and the parameter estimates. The spatial effect was modeled using a 5th degree polynomial in distance along the coastline.

$$\begin{aligned} f(t) = \quad & \beta_0 + \beta_1 X_1(t) + \beta_2 X_2(t) + \beta_3 X_3(t) + \beta_4 X_4(t) + \beta_5 X_5(t) + \\ & \beta_6 X_6(t) + \beta_7 X_7(t) + \beta_8 X_8(t) + \beta_9 X_9(t) + \beta_{10} X_{10}(t) + \delta(t) \qquad (4.5) \end{aligned}$$

$X_1(t)$: Mean water level over one hour immediately prior to the time of photography for the observation at time $t-1$.

$X_2(t)$: Maximum water level over one hour immediately prior to the time of photography for the observation at time $t-1$.

$X_3(t)$: Maximum water level over six hours immediately prior to the time of photography for the observation at time t.

$X_4(t)$: The total number of waves with wave height $>$ 2 meters in the time period between the t^{th} and $t-1^{th}$ observations.

$X_5(t)$: Kriebel's index at time $t-1$.

$X_6(t)$: Energy of waves with heights between 3 and 4 meters.

$X_7(t)$: Energy of waves with heights greater than 4.5 meters.

$X_8(t)$: $f(t-1)$.

$X_9(t)$: Sine term to model an yearly cycle.

$X_{10}(t)$: Cosine term to model an yearly cycle.

Parameter Estimates

Regression Coeff.	**Estimate**	**Std. Error of estimate**
β_0	-87.6	19.7
β_1	-509.8	116.8
β_2	603.2	145.8
β_3	105.3	14.1
β_4	0.13	0.03
β_5	-4.1	1.9
β_6	-8.3	2.9
β_7	8.4	3.9
β_8	0.33	0.1
β_9	-29.3	5.4
β_{10}	1.4	6.4

Parameter Estimates

Regression Coeff.	**Estimate($\times 10^{-4}$)**	**Std. Error of Estimate($\times 10^{-4}$)**
$Intercept$	322.27	3413.18
$distance$	-13222.63	27242.09
$distance^2$	69889.69	85580.96
$distance^3$	-147747.59	132331.59
$distance^4$	138439.29	100773.78
$distance^5$	-46586.04	30255.96

Table 4.3: Regression Coefficient Estimates for Southern region

4.4 Modeling the underlying covariance structure

We analyzed the table of residuals left behind by median polish to understand the covariance structure of the process. For the Northern region we had a table in two dimensions with 48 rows and 48 columns. Using a simple test we were able to show that the covariance structure is separable. A spatial temporal covariance model is said to be separable when it can be expressed as the product of functions of space and time. The function of space is a stationary spatial covariance model and the function of time is a stationary temporal covariance model. Applying our proposed method to this data by constructing blocks of size 16 rows and 16 columns we obtained estimates for the empirical semi-variogram in the spatial direction as well as the temporal direction. Similarly, for the central region we had a table with 48 rows and 40 columns and we constructed blocks of size 16 rows and 20 columns, and for the southern region we had a table with 48 rows and 100 columns and we constructed blocks of size 16 rows and 25 columns. Using the estimates for the empirical semi-variograms we obtained the median and the empirical 95% confidence intervals. The confidence intervals were calculated using the percentile method. To these estimates we fit the spherical parametric model and the parameter estimates are given in the following tables. The “nlin” procedure in SAS was used to obtain the estimates for the parameters and using this procedure it is possible to obtain negative values for the estimates for the nugget parameter. Since, the nugget is a measure of the variance of the micro-scale process and the measurement error it is not possible for the true value to be negative. Thus, we set the estimated value for the nugget effect to zero when the estimate was negative and correspondingly there is no estimate for the standard error in this situation.

The spherical variogram model is defined as follows:

$$\gamma(\boldsymbol{h},\boldsymbol{\theta}\ ) = \begin{cases} 0, & \boldsymbol{h}=\boldsymbol{0}, \\ c_0 + c_s\{\frac{3\|\boldsymbol{h}\|}{2a} - \frac{\|\boldsymbol{h}\|^3}{2a^3}\}, & 0 < \|\boldsymbol{h}\| \leq a, \\ c_0 + c_s, & \|\boldsymbol{h}\| \geq a \end{cases}$$

where, $\boldsymbol{\theta} = (c_0, c_s, a)$; c_0 is the nugget , c_s is the partial sill and a is the range of the process and $c_0 \geq 0$, $c_s \geq 0$ and $a \geq 0$. The following tables 4.4, 4.5 and 4.6 give the results of the fit using weighted least squares.

Parameter Estimates for Variogram in Northern region

For Lower bound of 95 % CI

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		40.6	5.9
Partial Sill	168.8	5.5		132.3	7.3
Range	386.4	23.0		18.6	2.0

For Median

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		70.6	6.1
Partial Sill	256.6	3.1		189.1	7.8
Range	469.2	9.2		19.8	1.6

For Upper bound of 95 % CI

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		91.8	7.9
Partial Sill	367.1	4.9		275.5	10.7
Range	529.0	13.8		22.8	1.8

Table 4.4: Northern region: Parameter estimates for spherical covariance model

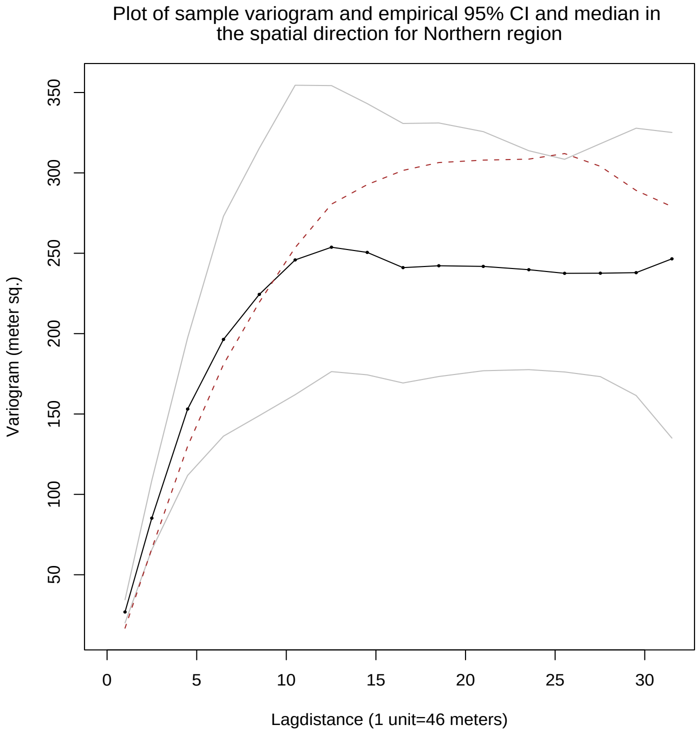


Figure 4.7: Northern Region Variogram in spatial direction

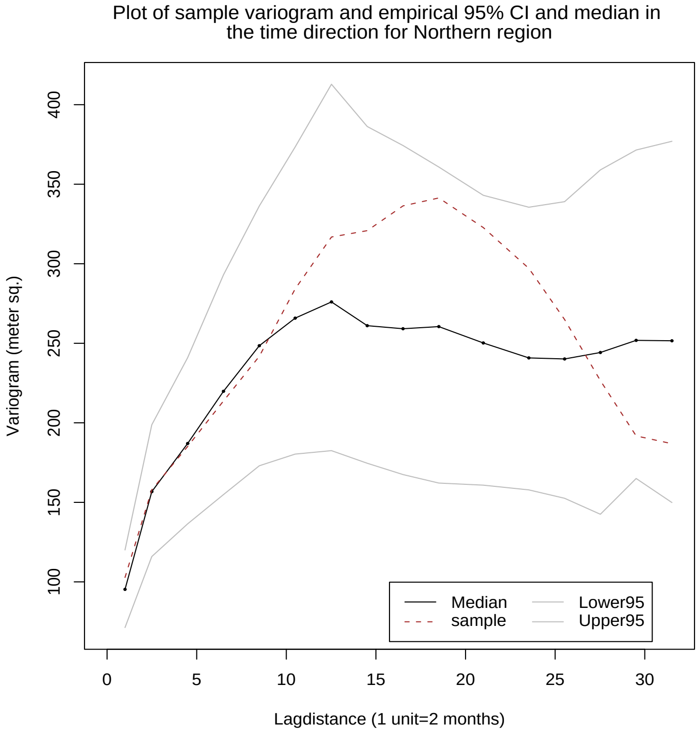


Figure 4.8: Northern Region Variogram in temporal direction

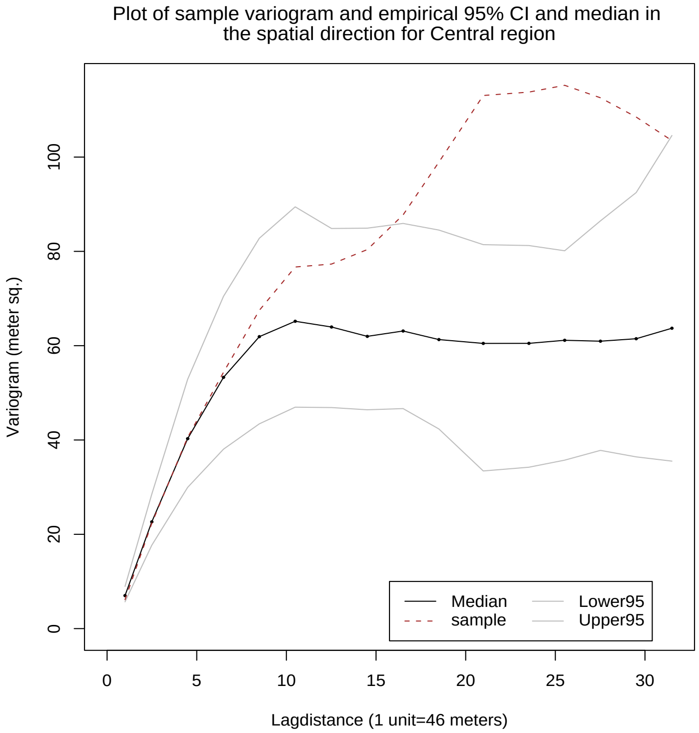


Figure 4.9: Central Region Variogram in spatial direction

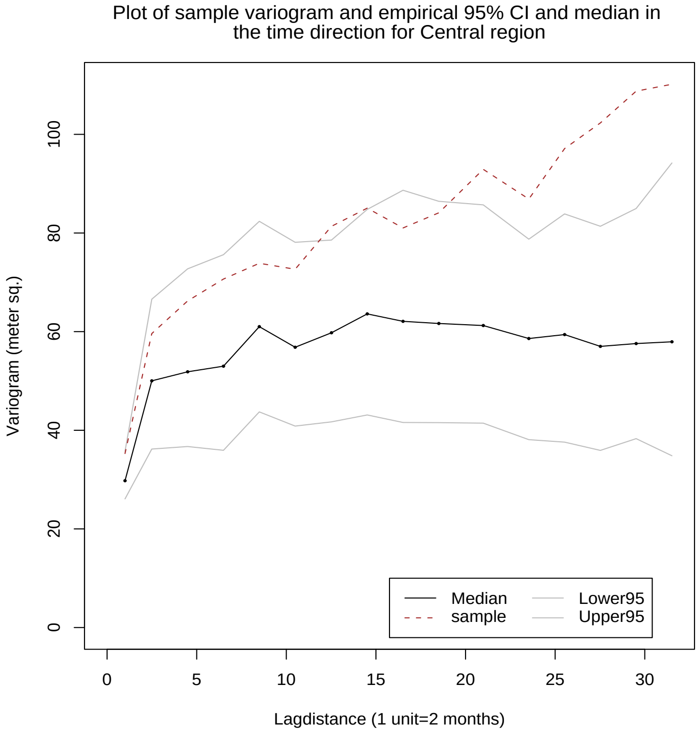


Figure 4.10: Central Region Variogram in temporal direction

Parameter Estimates for Variogram in Central region

For Lower bound of 95 % CI

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		17.4	2.5
Partial Sill	44.1	2.1		21.8	2.5
Range	386.4	36.8		7.4	1.0

For Median

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		12.8	3.3
Partial Sill	65.9	0.6		44.5	3.3
Range	432.4	9.2		7.6	0.8

For Upper bound of 95 % CI

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		10.5	4.3
Partial Sill	90.8	2.0		69.4	4.4
Range	464.6	18.4		8.0	0.6

Table 4.5: Central region: Parameter Estimates for spherical covariance model

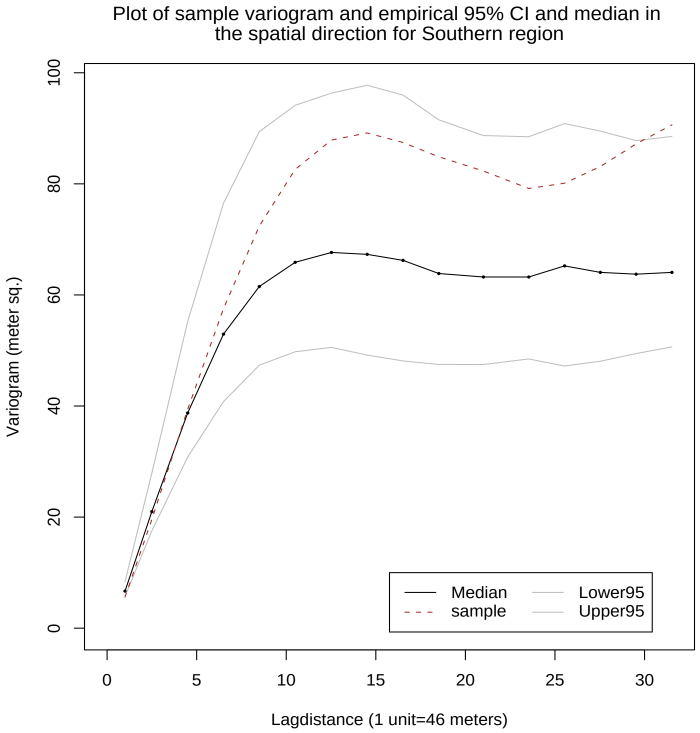


Figure 4.11: Southern Region Variogram in spatial direction

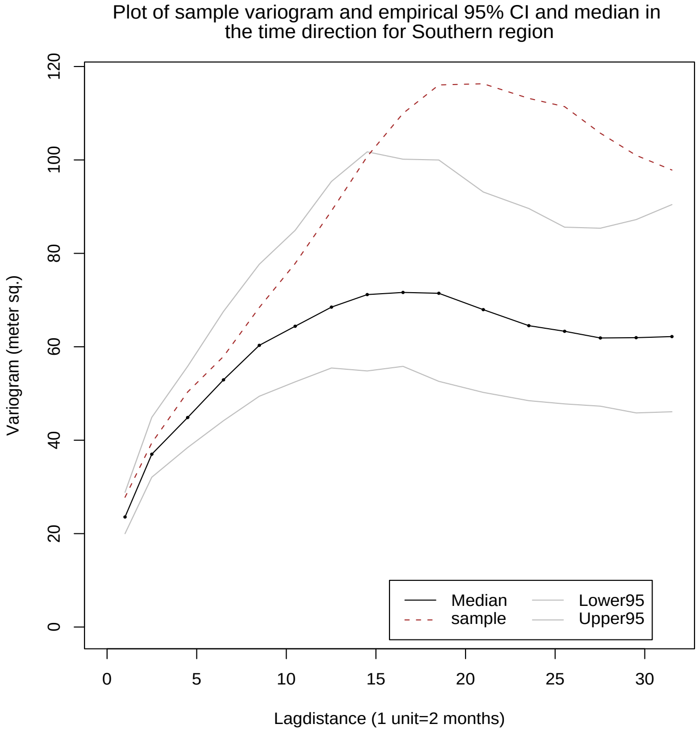


Figure 4.12: Southern Region Variogram in temporal direction

Parameter Estimates for Variogram in Southern region

For Lower bound of 95 % CI

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		14.9	1.4
Partial Sill	50.8	0.5		36.3	1.7
Range	446.2	9.2		18.8	1.8

For Median

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		18.3	1.3
Partial Sill	68.5	0.6		48.6	1.8
Range	478.4	9.2		23.4	1.8

For Upper bound of 95 % CI

	Spatial direction			Temporal direction	
Parameter	Estimate	Std. Error		Estimate	Std. Error
Nugget	0.0	.		22.3	1.6
Partial Sill	99.4	1.4		71.8	2.5
Range	469.2	13.8		28.0	1.8

Table 4.6: Southern region: Parameter estimates for spherical covariance model

4.5 Forecasting the distance of the shoreline to the baseline

Using the model (4.2) we obtained predictions for the distance of the shoreline to the baseline for a period of time for which we had observed values for the covariates as well as observed values for the distance of the shoreline to the baseline. We used the observed values of the covariates to obtain predicted values for the distance of the shoreline to the baseline using our model to compare with the actual distance of the shoreline to the baseline. The variance of the predicted values was calculated using an additive model for the errors in (4.2). For each of the three regions, the most conservative estimate for the sill of the empirical variogram was used as an estimate for the variance of the underlying spatial-temporal covariance process. We present the results for three representative transects in the three regions in the figures 4.13, 4.14 and 4.15. The vertical line at time=48 in the three figures is indicative of the time after which we used our model to forecast and compare with the observed values for the distance of the shoreline to the baseline.

The interpretation of 95% confidence here is that 95% of the prediction interval-future value combinations in the population will be successful. In other words 95% of the times the confidence interval for the prediction at a particular period will contain the actual future value.

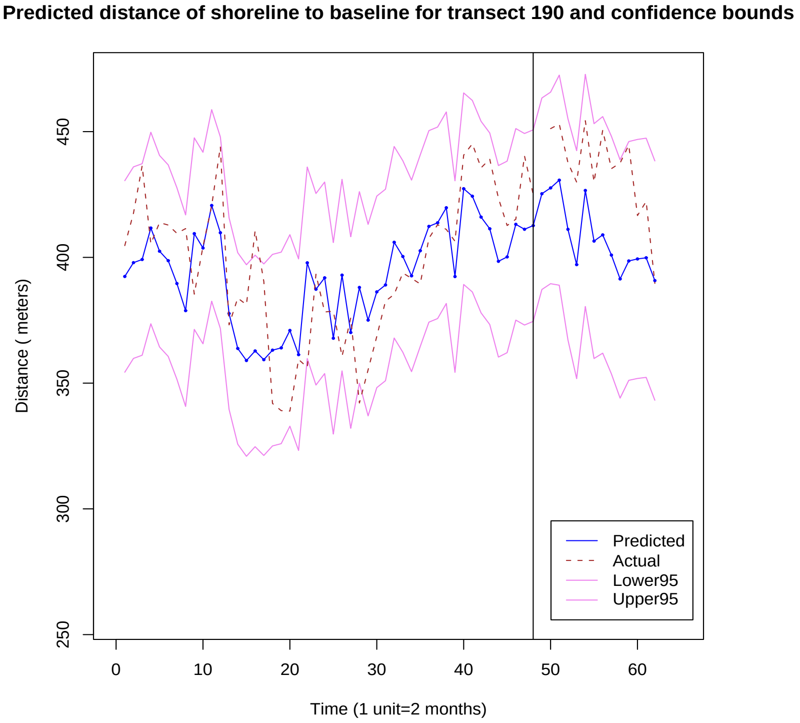


Figure 4.13: Northern Region: Predicted Distance of shoreline to baseline

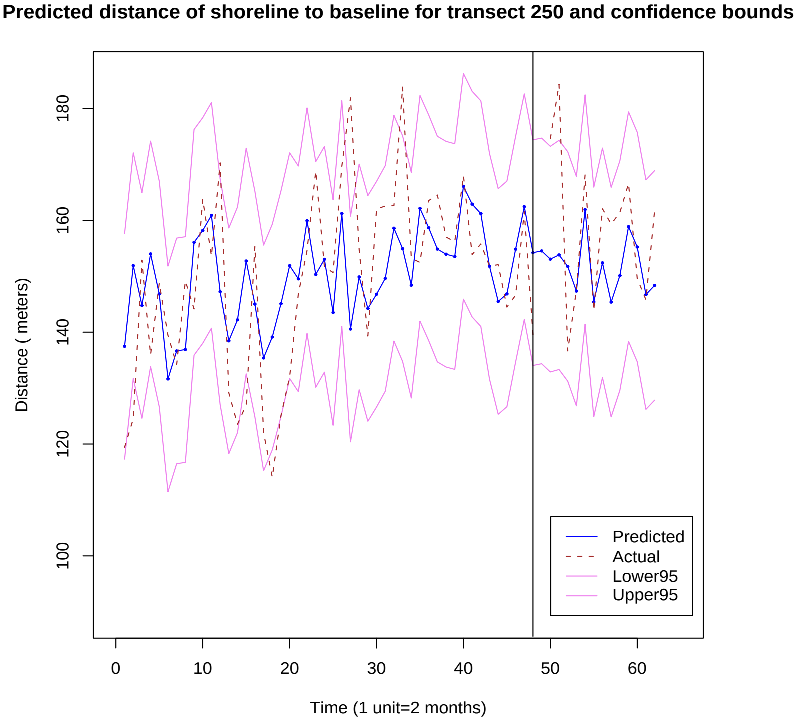


Figure 4.14: Central Region: Predicted Distance of shoreline to baseline

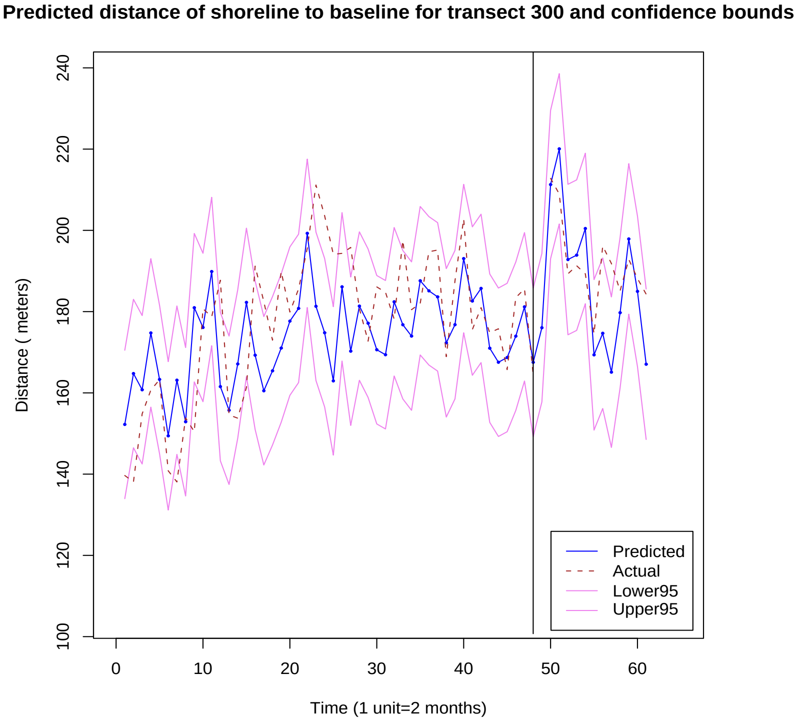


Figure 4.15: Southern Region: Predicted Distance of shoreline to baseline

To be able to forecast the distance of the shoreline to the baseline at any particular transect we need to be able to model the covariates. We simulated values for the covariates using an empirical approach since we did not want to assume a particular distribution structure or time series behaviour for the covariates. First, the residuals were obtained by subtracting the observed value of the distance of shoreline to baseline from the predicted distance of shoreline to baseline using the model given in (4.2). We treated the given observed values of the covariates as one block and reused this block to give us the values of the covariates at future time points. This assumes that the covariates have the same behaviour in the future as that was observed in the past. Next, we constructed blocks of size b from the series of residuals of length n, and we sampled with replacement from this set of blocks. We had values of the residuals for n equal to 48 time points and we sampled with replacement from the set of all blocks of size b equal to 6 rows that could be constructed from the series of residuals, for a particular transect. The number of blocks that we resample using the resampling technique depends upon the length of the series that we need to construct. To obtain the simulated value of the distance of the shoreline to the baseline we used the appropriate regression model developed earlier (depending upon the region that the transect that we are considering, lies in) and added to it the series of residuals that we constructed using the resampling technique. Using this method we simulated values for the distance of the shoreline to the baseline for 120 consecutive time points and constructed empirical 95% confidence intervals using the percentile method. Figures 4.16, 4.17 and 4.18 give plots of the simulated values and confidence intervals for the distance of the shoreline to the baseline using this approach.

Another approach used currently by scientists to obtain forecasts is the Empirical Simulation Technique. This was proposed by Borgman and Scheffner (Scheffner, Clausner, Militello, Borgman, Edge & Grace 1999). The EST is a procedure based on a resampling with replacement, nearest neighbour interpolation technique in which random sampling of a finite length database is used to generate a larger database. It assumes that future events will be statistically similar in magnitude and frequency to past events. The EST begins with the development of a database of storm parameters, then uses these parameters as a basis for simulating multiple repetitions of multiple years of storm activity.

To compare the results from our approach and the EST we present the yearly mean of the distance of the shoreline to the baseline per year and also the yearly maximum distance of the shoreline to the baseline per year. Figures 4.19, 4.20 and 4.21 give plots of the median of the yearly mean of the simulated values and empirical confidence intervals using our approach and the EST. Overlayed on the above three plots are also the yearly mean of the observed values for the distance of the shoreline to the baseline. Figures 4.22, 4.23 and 4.24 give plots of the median of the yearly maximum of the simulated values and empirical confidence intervals using our approach and the EST. The yearly maximum of the observed values for the distance of the shoreline to the baseline are also given. The confidence intervals were calculated using the percentile method.

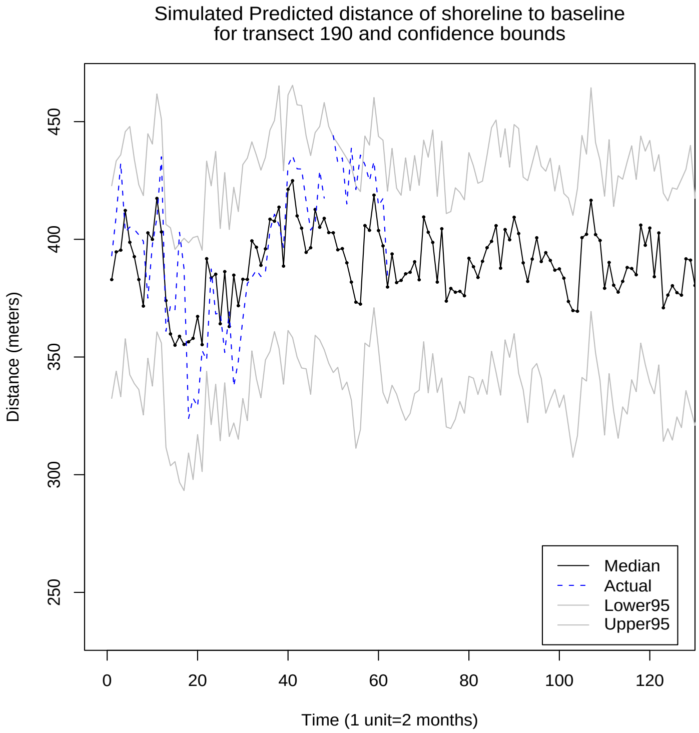


Figure 4.16: Northern Region: Simulated Distance of shoreline to baseline

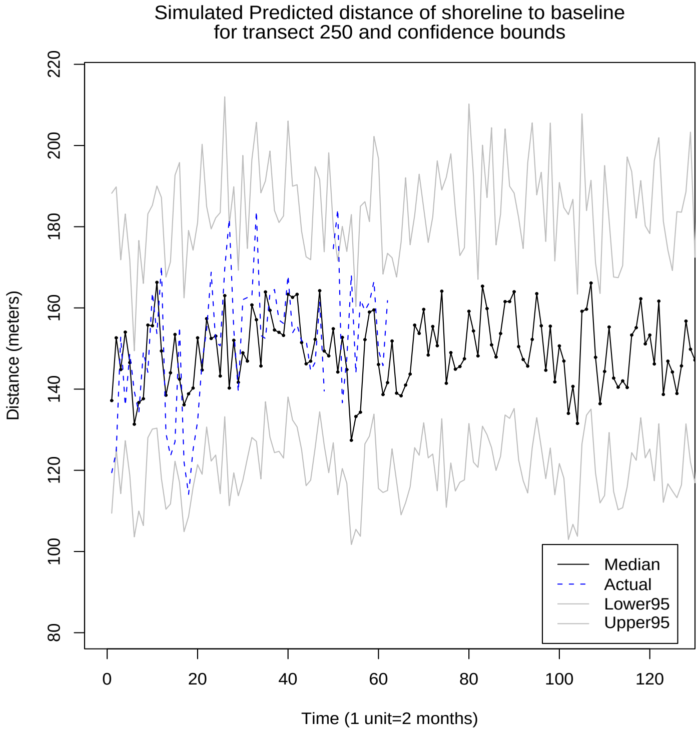


Figure 4.17: Central Region: Simulated Distance of shoreline to baseline

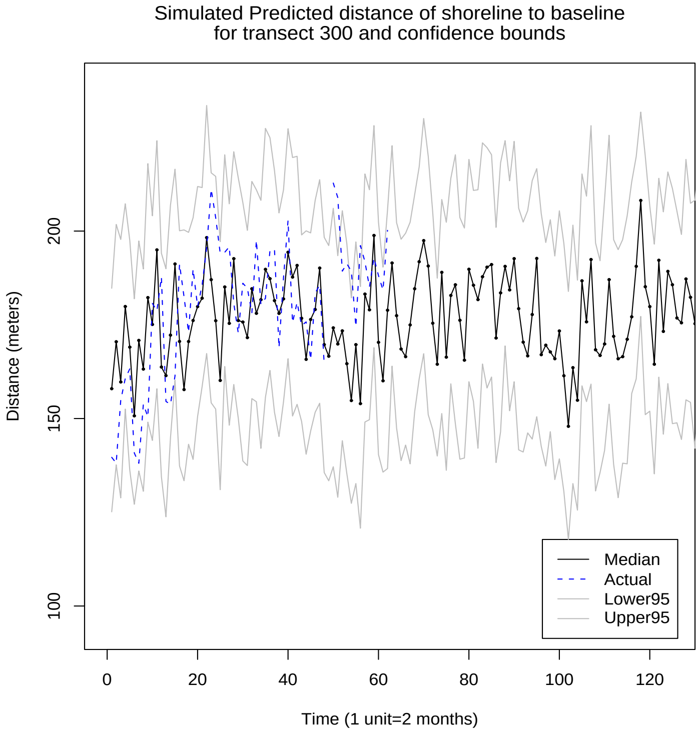


Figure 4.18: Southern Region: Simulated Distance of shoreline to baseline

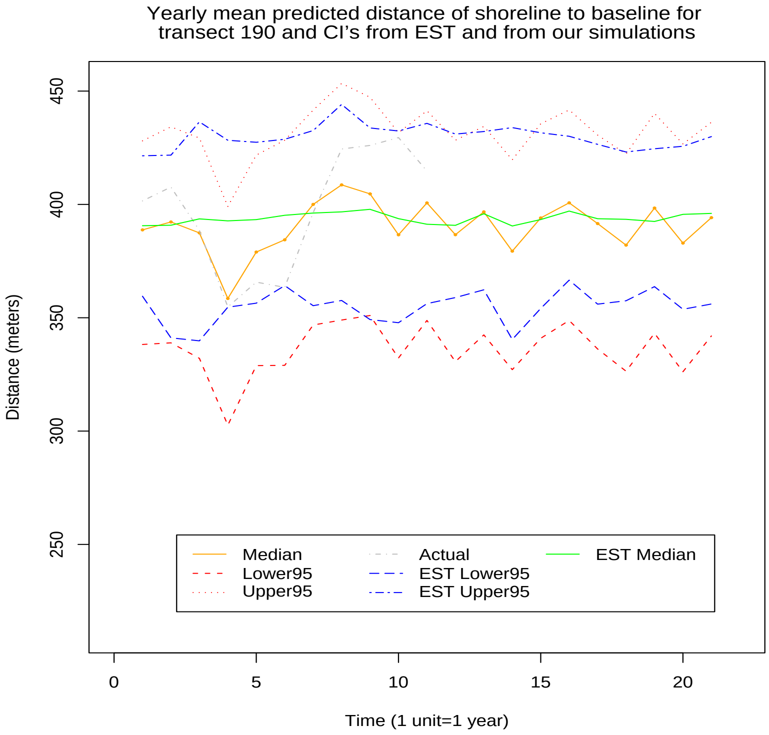


Figure 4.19: Northern Region: Comparison of Our approach and EST

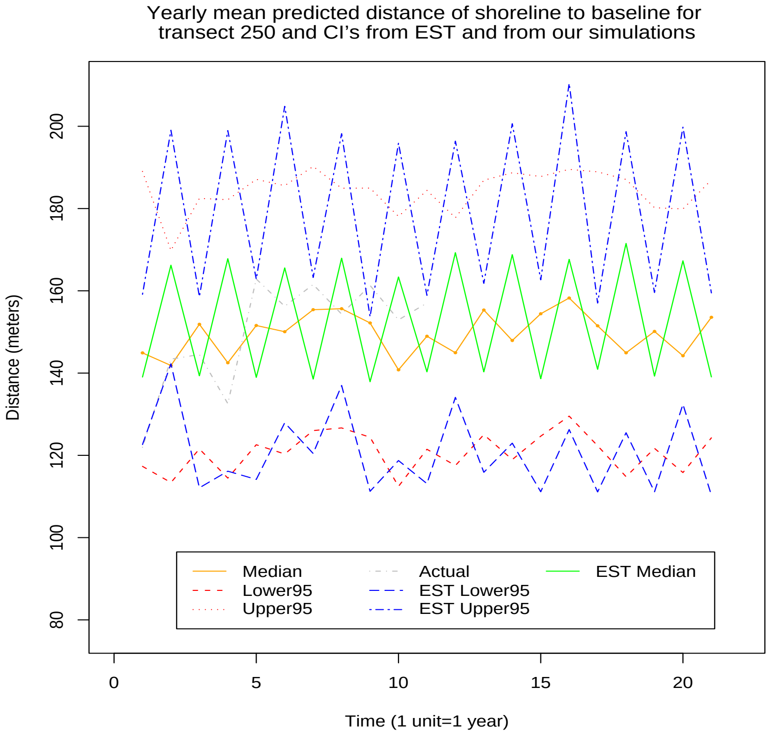


Figure 4.20: Central Region: Comparison of Our approach and EST

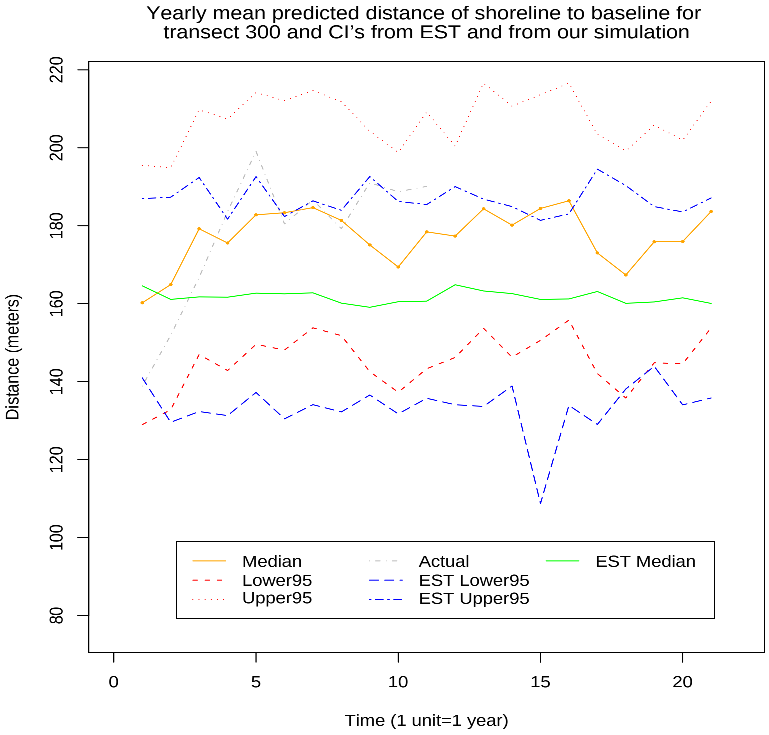


Figure 4.21: Southern Region: Comparison of Our approach and EST

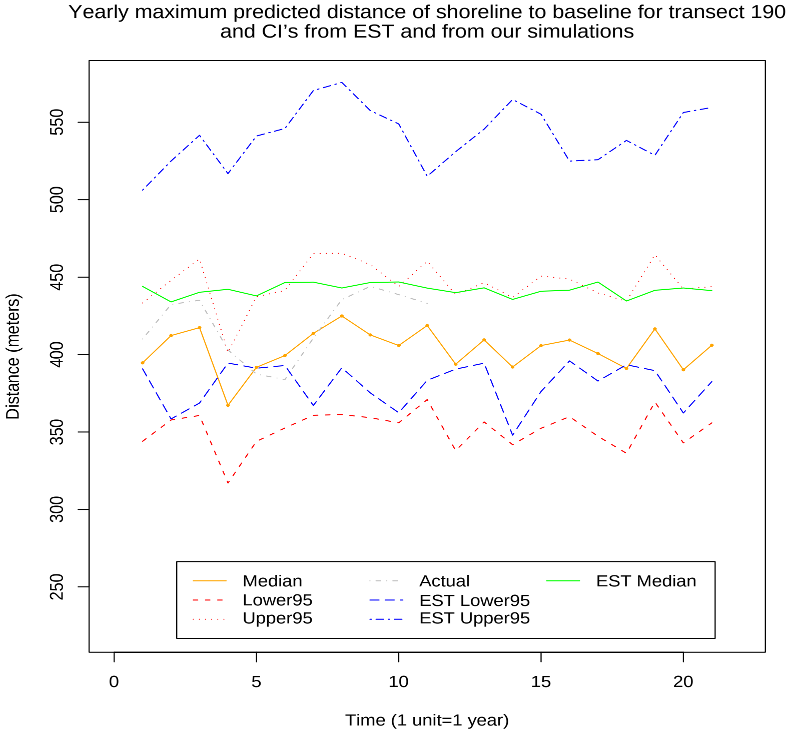


Figure 4.22: Northern Region: Comparison of Our approach and EST

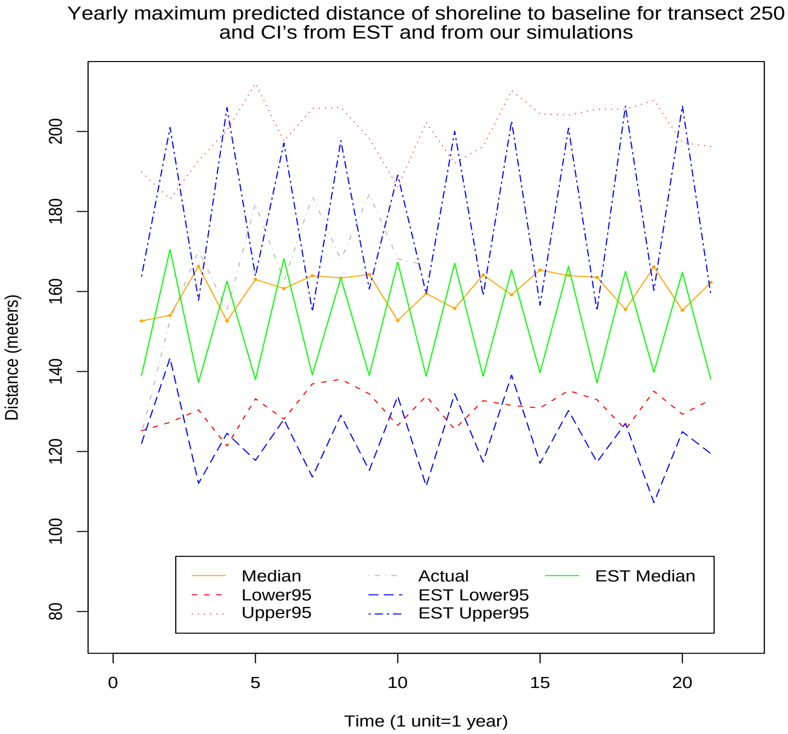


Figure 4.23: Central Region: Comparison of Our approach and EST

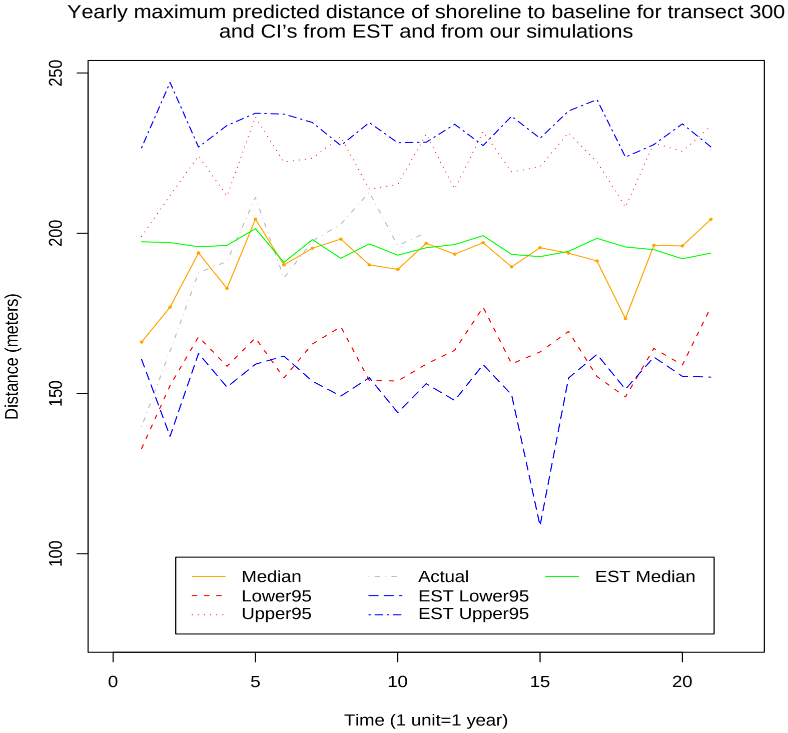


Figure 4.24: Southern Region: Comparison of Our approach and EST

4.6 Discussion

Our method performs better than the EST as it is able to capture the behaviour of the series over time. This can be seen by looking at Figures 4.19 through 4.24. Our approach has some advantages over the EST as we are able to model the behaviour over time. We are also able to model the covariance structure of the process which is not an option using the Empirical Simulation Technique.

To obtain predictions using our model it would be beneficial to model the covariates and thus forecast them at time points in the future. This would enable us to provide predictions that would take into account the behaviour of the storms without having to assume that future storms would be similar to those that have been observed in the past.

A simple method of modeling the shoreline change process would be to use a simple linear regression model with time as the only explanatory variable. However, this approach has some significant disadvantages. This model can only capture the linear trend over time and can not model any cyclical effects that might be present in the data. Further, it ignores the presence of any covariance structure in the residuals and assumes that the residuals are independent and identically distributed. As illustrated in chapter 1, not taking correlation into consideration would lead to narrower confidence intervals. Using our approach we are able to model the periodicities in the data and also take into account any correlation that may be present in the residuals. This enables our model to capture short term behaviour and makes our model more informative.

In the table below we give the simulated values for the distance of the shoreline to the baseline using our approach and also using the simple linear regression on time, for time 50 years in the future. Figures 4.25, 4.26 and 4.27 give plots of the simulated distance of the shoreline to the baseline using our approach and also using the simple linear regression on time as well as the corresponding 95% confidence intervals for the predictions for 10 years in the future.

Our approach **Simple Linear Regression**

Transect No.	Distance	Empirical 95 % CI		Distance	95 % CI	
		Lower	Upper		Lower	Upper
190	367	313	410	593	467	719
250	136	105	165	241	177	305
300	172	137	204	324	259	390

The underlying covariance structure for the Northern region is significantly different from that for the Central and Southern regions. We fit a spherical model to the empirical estimates of the variograms to all three regions. The estimate for the sill of the process in the Northern region is significantly greater than the sill of the process in the other two regions. The sill, which measures the variance of the process, indicates that there is a lot of variability in the Northern region. The range of the process does not vary much across the three regions.

An important thing to remember is that the decomposition in 4.1 is not unique. Cressie talks about this issue in section 3.1 of his book (Cressie 1993). This decomposition is largely operational in nature. It is possible that different scientists may reach different conclusions on the same set of data, depending on how much variation they attribute to the different components. One person's mean structure may be another person's correlated error structure.

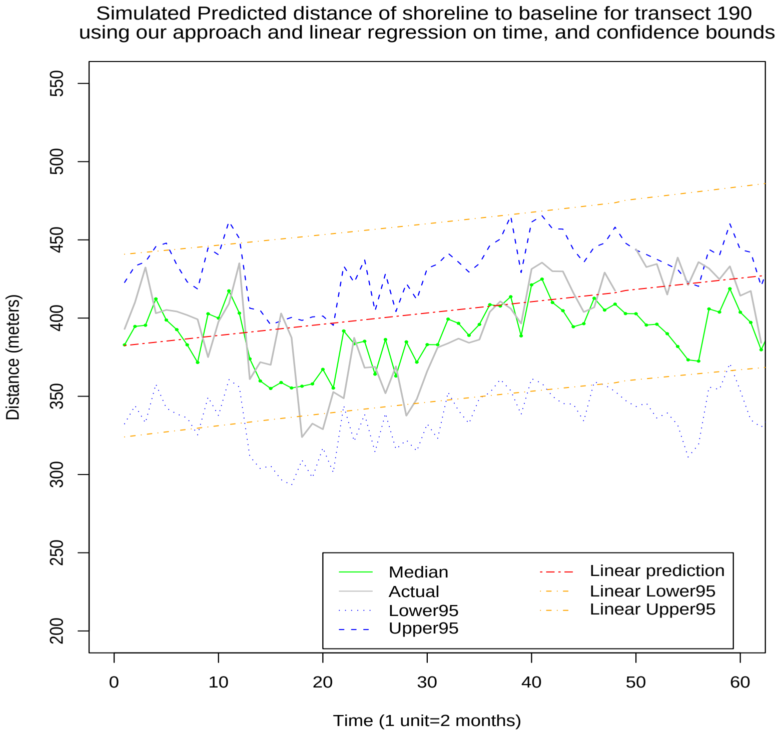


Figure 4.25: Northern Region: Comparison of Our approach and simple linear regression on time

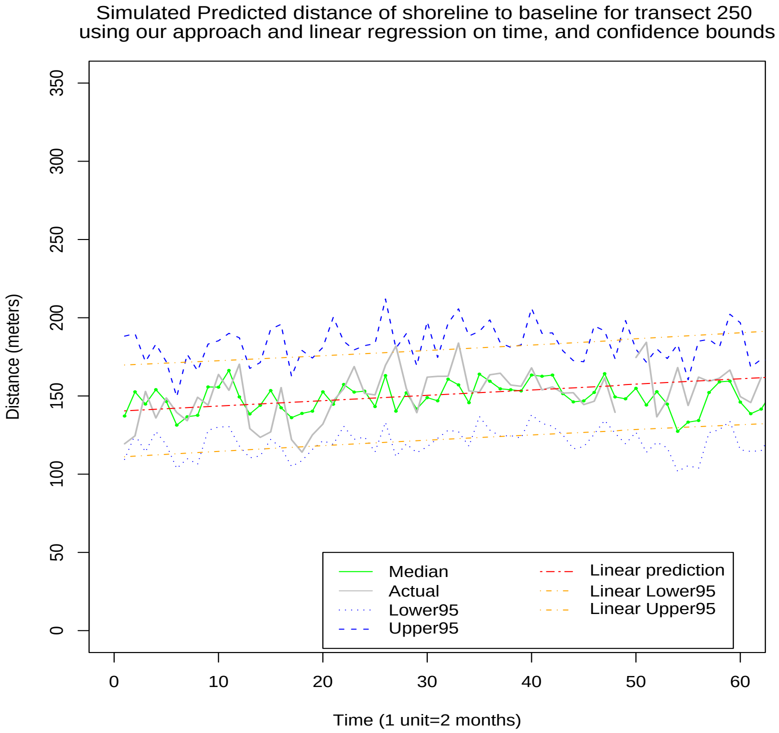


Figure 4.26: Central Region: Comparison of Our approach and simple linear regression on time

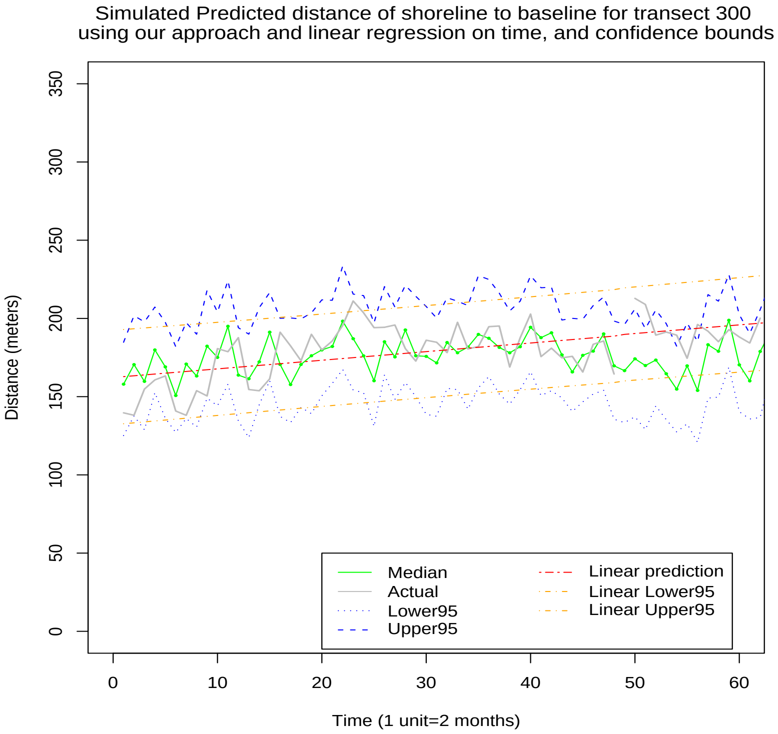


Figure 4.27: Southern Region: Comparison of Our approach and simple linear regression on time

Bibliography

Billingsley, P. (1995). *Probability and Measure (Third Edition).* Wiley.

Brieman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123–140.

Carlstein, E. (1986). The use of sub-series values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics* **14**, 1171–1179.

Casella, G. & Berger, R. L. (1990). *Statistical Inference.* Duxbury Press.

Cressie, N. A. C. (1993). *Statistics for Spatial Data (Revised Edition).* Wiley-Interscience.

Crowell, M., Douglas, B. & Leatherman, S. (1997). On forecasting future U.S. shoreline positions: A test of algorithms. *Journal of Coastal Research* **13**(4), 1245–1255.

Davis, B. M. & Borgman, L. E. (1982). A note on the asymptotic distribution of the sample semi-variogram. *Mathematical Geology* **14**(2), 1– 2.

Dolan, R., Hayden, B., May, P. & Suzette, M. (1980). The reliability of shoreline change measurements from aerial photographs. *Shore and Beach* **48**, 22–29.

Douglas, B. & Crowell, M. (2000). Long term shoreline position prediction and error prediction. *Journal of Coastal Research* **16**(1), 145–152.

Douglas, B., Crowell, M. & Honeycutt, M. (2002). Discussion of Fenster, M.S.; Dolan, R., and Morton, R.A., 2001. Coastal storms and shoreline change:

Signal or noise? Journal of Coastal Research, 17(3),714-720. *Journal of Coastal Research* **18**(2), 388–390.

Douglas, B., Crowell, M. & Leatherman, S. (1998). Considerations for shoreline position prediction. *Journal of Coastal Research* **14**(3), 1025–1033.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. Chapman and Hall.

Fenster, M., Dolan, R. & Elder, J. (1993). A new method for predicting shoreline positions from historical data. *Journal of Coastal Research* **9**(1), 147–171.

Fenster, M., Dolan, R. & Morton, R. (2001). Coastal storms and shoreline change: Signal or noise?. *Journal of Coastal Research* **17**(3), 714–720.

Friedman, J. & Hall, P. (2002). On bagging and nonlinear estimation. *Preprint*.

Fuentes, M. (2001*a*). Fixed-domain asymptotics for variograms using subsampling. *Mathematical Geology* **33**(6), 679–691.

Fuentes, M. (2001*b*). A new high frequency approach for nonstationary environmental processes. *Envirometrics* **12**, 469–483.

Fuentes, M., Overton, M. & Kim, H. (n.d.). Spatial temporal analysis of the shoreline at oregon inlet terminal groin.

Hall, P. (1985). Resampling a coverage pattern. *Stochastic Processes and their applications* **20**, 231–246.

Hall, P. (1998). On confidence intervals for spatial parameters estimated from non-replicated data. *Biometrics* **44**, 271–277.

Hall, P., Horowitz, J. & Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika* **82**, 561–574.

Honeycutt, M., Crowell, M. & Douglas, B. (2001). Shoreline position forecasting: Impact of storms, rate-calculation methodologies, and temporal scales. *Journal of Coastal Research* **17**(3), 721–730.

Kriebel, D., Dalrymple, R., Pratt, A. & Sakovich, V. (1996). A shoreline risk index for northeasters. *in* 'Proc. Conf. on Natural Disaster Reduction, Washington'. ASCE. pp. 251–252.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* **17**, 1217–1241.

Leadbetter, R., Lindgren, G. & Rootźen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag Inc.

Leatherman, S. (1983). Shoreline mapping: A comparison of techniques. *Shore and Beach* **51**, 28–33.

Leatherman, S., Douglas, B. & Crowell, M. (1997). Beach erosion trends and shoreline forecasting. *Journal of Coastal Research* **13**(4), iii–iv.

Liu, R. Y. & Singh, K. (1992). Moving blocks and bootstrap capture weak dependence. *in* 'Exploring the Limits of Bootstrap'. Wiley (New York). p. 00000.

Pajak, M. J. & Leatherman, S. (2002). The high water line as shoreline indicator. *Journal of Coastal Research* **18**, 329–337.

Politis, D. N. & Romano, J. P. (1992). A circular block-resampling procedure for stationary data. *in* 'Exploring the Limits of Bootstrap'. Wiley (New York). pp. 263–270.

Politis, D. N. & Romano, J. P. (1993). Nonparametric resampling for homogeneous strong mixing random fields. *Journal of Multivariate Analysis* **47**, 301–328.

Politis, D. N. & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association* **89**, 1303–1313.

Resnick, S. (1983). *Extreme values, Regular Variation, and Point Processes*. Springer-Verlag Inc.

Rosén, B. (1968). A note on the asymptotic normality of sums of higher dimensionally indexed random variables. *Arkiv för matematik* **8**, 33–43.

Sallenger, A. H. J. (2000). Storm impact scale for barrier islands. *U.S. Geological Survey* **16**(3), 890–895.

Scheffner, N., Borgman, L. & Mark, D. (1996). Empirical simulation technique based storm surge frequency analysis. *Journal of waterway, port,coastal and ocean engineering* **122**(2), 93–101.

Scheffner, N., Clausner, J., Militello, A., Borgman, L., Edge, B. & Grace, P. (1999). Use and application of the empirical simulation technique:user's guide. *Technical Report CHL-99-21,US Army Corps of Engineers.*

Shao, J. & Tu, D. (1995). *The jackknife and bootstrap.* Springer-Verlag Inc.

Stein, M. L. (1995). Predicting integrals of stochastic processes. *The Annals of Probability* **5**, 158–170.

Tikhomirov, A. (1980). On the convergence rate in the central limit theorem for weakly dependent random variables. *Theory of Probability and its applications* **25**, 790–809.

Zhang, K., Douglas, C. & Leatherman, S. (2001). Beach erosion potential for severe nor'easters. *Journal of Coastal Research* **17**, 309–321.

Zhang, K., Huang, W., Douglas, B. & Leatherman, S. (2002). Shoreline position variability and long-term trend analysis. *Shore and Beach* **70**, 31–35.