Assessing Spatial Dependency under Non-Standard Sampling

Raquel Menezes da Mota Leite

July 2005

Departamento de Estatística e I.O.

Universidad de Santiago de Compostela

Vendo-os assim tão pertinho
A Galiza mai'lo Minho.
São como dois namorados
Que o rio traz separados
Quasi desde o seu nascimento.
Deixa-los, pois namorar
Já que os paes para casar
Lhes não dão consentimento

João Verde
Do livro "Ares da Raya"

Acknowledgments

I would like to thank my advisors Prof.Pilar Garcia-Soidán and Prof.Manuel Febrero-Bande for all their support and direction to this thesis.

From my visits to Lancaster, I thank the stimulating environment and advice provided by Prof.Peter Diggle and Prof.Jonathan Tawn. I also thank Inês, Rose and Susi that, among many others in Lancaster, helped me to find there a comfortable work and leisure environment. My appreciation is also due to Alan, Daniel, Chris and Jamie who struggled to eradicate the worst errors on my non-native english writings.

I would like to acknowledge the combined financial support provided by Universidade do Minho, Prodep grant ref.5.3/N/189.015/01 and a Marie-Curie EU grant that together made this work possible.

Finally I would to thank my Family and Carlos for their permanent support during this research.

Contents

List of Figures

Chapter 1

Introduction

Nowadays, spatial statistics plays an important role, as current technological development helps the derivation of spatial data. Besides the traditional sources of spatial data, such as maps, census material and aerial photography, more sophisticated and reliable data sources have appeared, such as remote sensors aboard satellites. These technologies resulting from the development of fast computers and specific software, such as image processing software and geographic information systems, are providing better tools and new demands from spatial statistics.

Spatial models work with data collected from different spatial locations. These models measure the relationship between observations at various locations. They should reflect the intuitive hypothesis that data items close in space are correlated and that this correlation decreases when distance increases. They should also provide evidence of the existence of spatially correlated errors. The spatial correlation analysis enables us to see how variables, such as pollutant loads, measured at different points in space are related. More important from the practical viewpoint, such relationships can be used for estimating values at sites where no measurements are taken.

A sample of data may consist of observations taken in one, two or three dimensions. Measurements of water quality along a river are taken in one-dimension. On the other hand, rainfall or other meteorological variables are measured at par-

ticular points, but collectively they constitute a two-dimensional random field. Finally, measurements of a specific mineral in ground is sometimes treated as a three-dimensional problem because it can occur over different levels.

The spatial process is a stochastic process. It may be represented as a set of random variables (or vectors) $Z(\mathbf{x})$, indexed by $\mathbf{x}$ which belongs to a set $D \subset I\!R^d$, a d-dimensional euclidean space with $d = 1, 2, 3$. Its usual notation is $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$. Let us suppose one has the spatial locations $\mathbf{x_1}, ..., \mathbf{x_n}$, then $Z(\mathbf{x_1}), ..., Z(\mathbf{x_n})$ identifies observed data at those locations. These observations may be obtained from one or more, discrete or continuous, variables.

According to Cressie (1993), the nature of D allows us to identify three major spatial processes, namely lattice processes, point processes and continuous processes. The latter are commonly referred to as geostatistics (Matheron 1963) and, opposite to lattice processes, between any two spatial points associated to existent observations, there is always another point where the random variable could also be observed.

The developed work falls within the scope of geostatistics and it applies theory of point processes, presuming that spatial locations have been produced by some form of stochastic mechanism. Under *assessing spatial dependency*, we consider two distinct issues, namely the estimation of the spatial dependency structure and the subsequent spatial prediction procedure. Within the realm of geostatistics, it is commonly believed that sample locations are equally spread over the observed region. Furthermore, it is assumed that the point process for data locations does not depend on the data process. So, under *non-standard sampling*, we are considering the failure of one or both of previous assumptions.

In Chapter 2, we introduce notation and definitions related to geostatistical data modelling. We review some well-known facts about the convenient assumption of stationarity of the underlying process. The usage of the variogram as a tool to measure spatial dependence between samples is highlighted.

In Chapter 3, we start focusing the estimation of spatial dependency under

standard sampling. A bibliographic search of current variogram estimators is described. A comparison simulation study is presented, covering different kinds of spatial dependence situations.

In Chapter 4, a motivating example is introduced, the Rongelap island, where the data was collected over a two-stage process of uniform and clustered samples, which may have an impact on conclusions. Centered on the multi-stage case, we assess the effect of clustered and biased sampling on spatial dependency estimation. A new variogram estimator for clustered data is proposed.

In Chapter 5, we proceed with the theoretical study of the proposed estimator. It is shown to enjoy good properties, such as asymptotic unbiasedness and consistency.

In Chapter 6, we introduce the preferential sampling concept, as a formal definition for the dependency of data locations on data values. A flexible class of point processes for preferential sampling, based on log-Gaussian Cox processes, is presented. We then assess the effect of preferential sampling on the classical geostatistical methods typically used for estimation of the spatial dependency and for spatial prediction (kriging).

In Chapter 7, a motivating example of pollution data is introduced, reinforcing the importance of doing parametric model analysis when data is suspected to be preferentially sampled. An intuitive approximate model for preferential sampling is proposed. We proceed with likelihood inference to estimate the parameters of this model and a simulation study is performed to show the benefits of this model-based approach versus the traditional one. We close this Chapter with a short discussion of future work.

Chapter 2

Stationary spatial processes

2.1 Introduction

Suppose that $\{Z(\mathbf{x}) : \mathbf{x} \in D \subset I\!R^d\}$ is a random spatial process, where D is a bounded region with positive d-dimensional volume. The mean of $Z(\mathbf{x})$, $\mathrm{E}[Z(\mathbf{x})]$, is the mean of all possible realizations of the process at points $\mathbf{x}$. Typically, just one realization of the given process is observed, possibly denoted by $z(\mathbf{x})$. The process is said to act over a random field Ω. Additionally, the difference random process $Z(\mathbf{x}) - \mathrm{E}[Z(\mathbf{x})]$ represents departures of the original process from the mean at the points considered. The study of such processes is based on the identification of appropriate characteristics of regularity, which is referred to as stationarity in the context of stochastic processes (e.g. Kottegoda and Rosso 1997).

The process $Z(\mathbf{x})$ is usually defined through the finite-dimensional distribution

$$F_{\mathbf{x_1},...,\mathbf{x_n}}(z_1, ..., z_n) = P\{Z(\mathbf{x_1}) \leq z_1, ..., Z(\mathbf{x_n}) \leq z_n\}, \quad n \geq 1,$$

which must satisfy the Kolmogorov's conditions of symmetry, i.e. remain invariant when z_j and $\mathbf{x_j}$ are subject to the same permutation, and the consistency condition, i.e. $F_{\mathbf{x_1},...,\mathbf{x_{n+m}}}(z_1, ..., z_n, +\infty, ..., +\infty)$ must be equal to $F_{\mathbf{x_1},...,\mathbf{x_n}}(z_1, ..., z_n)$. Furthermore, the following hypothesis for the values of the mean and the variance may be considered.

Strict (or strong) stationarity

A spatial process $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$ is strictly stationary, if for any finite collection $\{\mathbf{x_1}, ..., \mathbf{x_n}\}$ of spatial locations, for any finite collection of real values $\{z_1, ..., z_n\}$ and for any vector $\mathbf{u} \in I\!R^d$ for which $\mathbf{x_i} + \mathbf{u} \in D$ whenever $\mathbf{x_i} \in D$, then:

- $F_{\mathbf{x_1},...,\mathbf{x_n}}(z_1, ..., z_n) = F_{\mathbf{x_1}+\mathbf{u},...,\mathbf{x_n}+\mathbf{u}}(z_1, ..., z_n), \quad \forall\ n, \mathbf{u}$

This means that a stationary process remains invariant when subject to translation transformations of its coordinates.

Second-order (or weak or wide-sense) stationarity

The spatial process $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$ is second-order stationary, if its first moment is a constant and the covariance between two variables is a function of the difference between theirs locations:

- $\mathrm{E}[Z(\mathbf{x})] = \mu(\mathbf{x}) = \mu \quad \forall\ \mathbf{x} \in D$
- $\mathrm{Cov}[Z(\mathbf{x_i}), Z(\mathbf{x_j})] = c(\mathbf{x_i} - \mathbf{x_j}), \quad \forall\ \mathbf{x_i}, \mathbf{x_j} \in D$

The function $c(.)$ is called the *stationary covariance function* or sometimes *covariogram*. The function $\mu(\mathbf{x})$ is known as the *trend* of the process.

Intrinsic stationarity

The spatial process $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$ is intrinsically stationary if its first moment is a constant and the variance of the difference between two variables is a function of the difference between theirs locations:

- $\mathrm{E}[Z(\mathbf{x})] = \mu(\mathbf{x}) = \mu \quad \forall\ \mathbf{x} \in D$
- $\mathrm{Var}[Z(\mathbf{x_i}) - Z(\mathbf{x_j})] = 2\gamma(\mathbf{x_i} - \mathbf{x_j}), \quad \forall\ \mathbf{x_i}, \mathbf{x_j} \in D$

The function $2\gamma(.)$ is called the *variogram* and $\gamma(.)$ the *semivariogram*, but the latter is often also referred as to the variogram to simplify terminology. We shall focus on the importance of this function in geostatistics.

One may note that if a process is strictly stationary, then it is also second-order stationary. Furthermore, if a process is second-order stationary, then it is also intrinsically stationary. Let us confirm this last implication:

$\mathrm{Var}[Z(\mathbf{x_i}) - Z(\mathbf{x_j})] = \mathrm{Var}[Z(\mathbf{x_i})] + \mathrm{Var}[Z(\mathbf{x_j})] - 2\mathrm{Cov}[Z(\mathbf{x_i}), Z(\mathbf{x_j})] =$

$= c(0) + c(0) - 2c(\mathbf{x_i} - \mathbf{x_j}) = 2\gamma(\mathbf{x_i} - \mathbf{x_j})$, being

$$\gamma(.) = c(0) - c(.) \tag{2.1}$$

Strict and second-order stationarity coincide if the spatial process is Gaussian, i.e. if the joint distribution of any finite collection of variables is Gaussian. Second-order and intrinsic stationarity coincide if the variance is finite and it does not depend on $\mathbf{x}$, i.e. $\mathrm{Var}[Z(\mathbf{x})] = c(0) = \sigma^2 < \infty, \quad \forall\ \mathbf{x}$.

Isotropy

The random process $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$ is isotropic if it remains invariant when subject to rotations of coordinates, in contrast to the anisotropic process. For example, the intrinsic random process $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$ is isotropic if $\forall\ \mathbf{x_i}, \mathbf{x_j} \in D$ then:

- $\mathrm{E}[Z(\mathbf{x_i}) - Z(\mathbf{x_j})] = 0$,
- $\mathrm{Var}[Z(\mathbf{x_i}) - Z(\mathbf{x_j})] = 2\gamma(\|\mathbf{x_i} - \mathbf{x_j}\|)$

where $\|.\|$ denotes the Euclidean norm. The variogram here only depends on the distance between the two locations and not on the direction of the difference vector.

The isotropy condition is not so restrictive in practice, since a linear transformation of the coordinates sometimes produces an acceptable approximation to isotropy, or it might be possible to fit a different variogram in different directions in the case of anisotropy. Physically, the former corresponds to a rotation and stretching of the original spatial locations. Algebraically, it means to apply some linear transformation to the space of locations, given some *anisotropy angle* and *anisotropy ratio*.

Ergodicity

A subset of the second-order stationary random processes possesses an important property known as *ergodicity*, which is required for the estimation of the characteristics of a process based on its realizations. This is applicable if the estimates of its moments, taken from the available realizations, converge in probability to the theoretical moments, when the available sample increases. Hence, under ergodicity, one realization will suffice for the estimation of these moments. In practice, this property is normally assumed to hold.

2.2 The mean function, covariance function and variogram

A clear difference between the covariance function and the variogram is that the former is a direct function of the *association* between two variables, whereas the latter measures the *disassociation*. Variograms are more general than covariance functions, and many important properties have been initially established for covariance functions. Gneiting, Sasvári and Schlather (2001) explores the relationship between the two functions and present some analogous results for variograms.

For a second-order stationary and isotropic random process, $\text{Var}[Z(\mathbf{x})] = \sigma^2$ and it is useful to write the covariance function as $c(u) = \sigma^2\rho(u)$, where $\rho(.)$ is the *correlation function* depending on a scalar argument u. As example of an useful correlation functions adopted in geostatistical data modelling, we have the Matérn family (Wackernagel 1998) represented in Figure 2.1 and with algebraic form given by

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}K_{\kappa}(u/\phi), \tag{2.2}$$

where $\kappa > 0$ and $\phi > 0$ are parameters, and $K_{\kappa}(.)$ denotes a Bessel function of order κ. The parameter ϕ determines the rate at which the correlation decays to zero with increasing u. The parameter κ determines the analytic smoothness of $Z(.)$ (see e.g. Ribeiro Jr 2002 for more details).

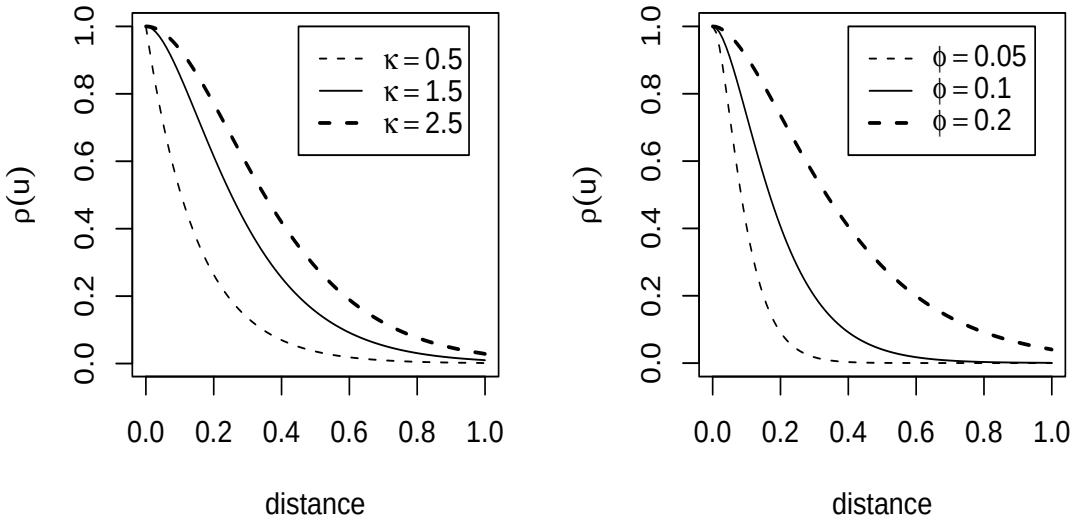


Figure 2.1: Examples of the Matérn correlation function: $\phi = 0.15$, varying κ (left plot); $\kappa = 1.5$, varying ϕ (right plot).

For a process with stationary covariance structure, equation (2.1) shows that the variogram reduces to

$$\gamma(u) = \sigma^2(1 - \rho(u)). \tag{2.3}$$

Typically the variogram approaches a constant value as the separation distance u increases; this value is known as the *sill*. In (2.3) the sill is given by σ^2. When a variogram has a sill, it means that there is a distance beyond which the correlation between variables is zero; this distance is called the *range* or radius of influence.

For many practical applications, it is useful to consider a Gaussian spatial process with a possibly varying mean function but stationary covariance structure. For such processes, $Z(\mathbf{x}) - \mu(\mathbf{x})$ is a stationary Gaussian process with zero mean. A possible solution is to specify $\mu(\mathbf{x})$ as a regression model, with the aim of fitting a smooth surface to values measured over a sample of points. The regression itself provides a summary of the trend as well as a means of predicting the value at any location within the modelled surface. For example, the analysis of rainfall data in

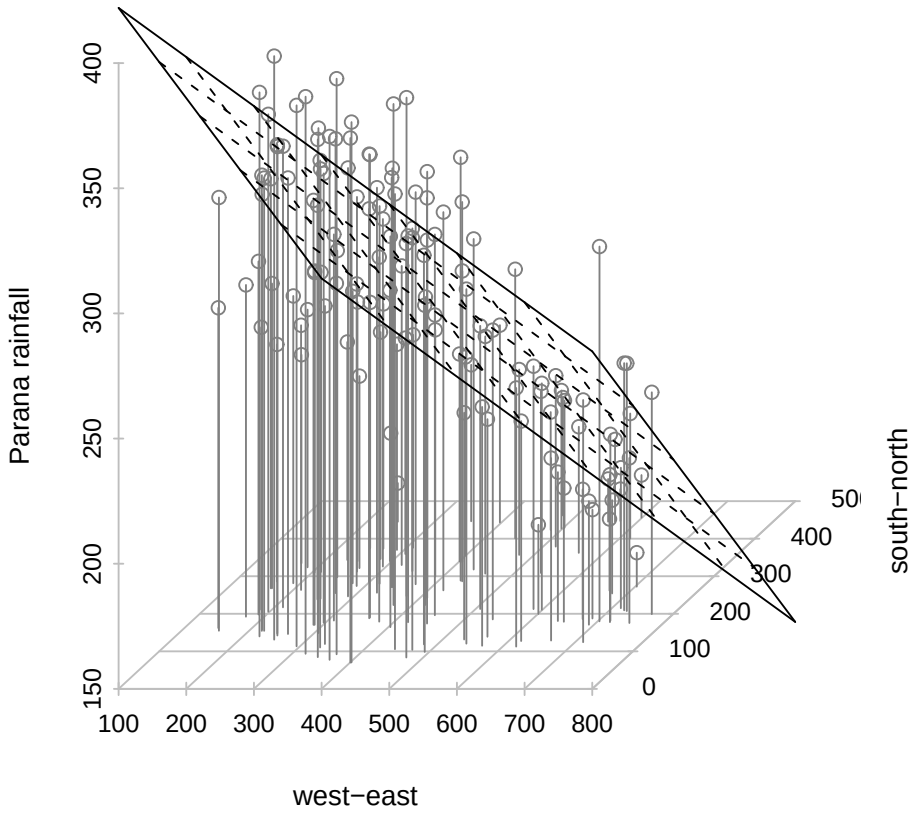


Figure 2.2: Rainfall data - Paraná state in Brazil

Paraná State in Brazil from Diggle and Ribeiro Jr (2002) points to a clear trend from east to west and from south to north which is modelled with a first degree polynomial (see Figure 2.2).

In the presence of scientifically relevant spatially referenced explanatory variables, it is reasonable to include these as covariates in the trend model (Wackernagel 1998). For example, in the analysis of average temperatures over a 24 hour period over the whole of Argentina, it might be important to consider altitude as a covariate. As we shall see in Chapter 6, the procedure for making predictions using a polynomial trend in the coordinate variables is often referred to as *universal*

kriging, whereas using other covariates is referred to as *kriging with a trend model* (Goovaerts 1997).

2.3 Scale of variation

The random spatial process $Z(\mathbf{x})$ may be written as

$$Z(\mathbf{x}) = \mu(\mathbf{x}) + \delta(\mathbf{x}), \quad \mathbf{x} \in D$$

where $\mu(.)$ represents the trend and $\delta(.)$ contains the spatial dependence structure. The trend is a deterministic function, whereas $\delta(\mathbf{x})$ is a random process with zero mean for each $\mathbf{x} \in D$. One can also say that $\mu(\mathbf{x})$ represents the *large-scale variation* and $\delta(\mathbf{x})$ represents the *small-scale variation*. A further decomposition, see e.g. Cressie (1993), can be represented by

$$\delta(\mathbf{x}) = W(\mathbf{x}) + \eta(\mathbf{x}) + \epsilon(\mathbf{x}), \quad \mathbf{x} \in D$$

where:

$W(.)$ is a zero mean L_2-continuous intrinsically stationary process, identifying a smooth small-scale variation.

$\eta(.)$ is a zero mean, intrinsically stationary process, independent of W, whose variogram range exists and is smaller than $min(\|\mathbf{x_i} - \mathbf{x_j}\|)$. It is called micro-scale variation.

$\epsilon(.)$ is a white-noise process, independent of W and η. We call $\epsilon(.)$ the *measurement error*, and denote $\text{Var}[\epsilon(\mathbf{x})] = \tau^2$.

While $\epsilon(.)$ represents the idea that repeated observations at the same spatial location can be significantly unequal, the idea of $\eta(.)$ is to reflect some local effects. In practice, however, $\eta(.)$ and $\epsilon(.)$ might be indistinguishable. For example, in the earliest studies in geostatistics, which had applications to mining, the micro-scale variation was assumed to be caused by the existence of small nuggets of enriched

ore and was approximated by a white noise process. Since then, the term *nugget effect* has been adopted in many other applications, and it is now used to describe, according to context, measurement error, micro-scale variation or a unidentifiable combination of the two.

From the previous statements, an alternative decomposition is

$$Z(\mathbf{x}) = S(\mathbf{x}) + \epsilon(\mathbf{x}), \qquad \mathbf{x} \in D. \tag{2.4}$$

The $S(.)$ process is often referred to as the noiseless version of the $Z(.)$ process. Since the variogram introduced in (2.3) is the one for the noiseless version, the existing noise variance must be added to the variogram of $Z(\mathbf{x})$, giving

$$\gamma(u) = \tau^2 + \sigma^2(1 - \rho(u)).$$

2.4 Further variogram properties

One of the most critical properties characterizing a variogram is that of *conditional negative-definiteness*, i.e. the requirement that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \gamma(\mathbf{x_i} - \mathbf{x_j}) \leq 0$$

for any finite set of spatial locations, $\{\mathbf{x_1}, ..., \mathbf{x_n}\}$, and for any set of real numbers $\{a_1, ..., a_n\}$, such that $\sum_{i=1}^{n} a_i = 0$.

In the absence of this property, the mean square prediction error could be estimated by an absurd negative value. This leads to the impossibility of using some variogram estimators within the inference and prediction context, and we shall classify them as *non valid variograms* from now on. One possible approach to solve this problem is to approximate the variogram by any parametric model which is known to be valid. The idea is to search, among the families of valid variograms, for one that best approximates the underlying spatial dependence of the available sample data. This will be discussed in some detail in Chapter 3.

Some other important properties of the variograms are now stated. Write $D_1 = \{\mathbf{x} = \mathbf{x_1} - \mathbf{x_2} : \ \mathbf{x_1}, \mathbf{x_2} \in D\}$. Then, for all $\mathbf{x} \in D_1$, one has:

- $2\gamma(\mathbf{0}) = 0$, i.e. they are null at the origin;
- $2\gamma(\mathbf{x}) \geq 0$;
- $\lim_{\|\mathbf{x}\| \to 0} \gamma(\mathbf{x}) = \theta_0$, where $\theta_0 \geq 0$ is the nugget effect;
- $2\gamma(\mathbf{x}) = 2\gamma(-\mathbf{x})$, i.e. they are symmetric functions;
- $\lim_{\|\mathbf{x}\| \to \infty} \gamma(\mathbf{x}) / \| \mathbf{x} \|^2 = 0$, i.e. the rate of increase of $\gamma(.)$ should be smaller than $\| \mathbf{x} \|^2$ (Matheron 1971).

If this last property fails, it means that there is no second-order stationarity for the increments, and the presence of a trend structure $\mu(\mathbf{x})$ is then expected.

The first properties confirm that the variogram is not necessarily a continuous function. In theory, at very close distances the disassociation between values of the variable approaches zero. In practice, however, at very small separation distances the variogram can be significantly different from zero, reflecting some local effects or measurement error as previously discussed. The behaviour of the variogram near the origin helps us to define continuity properties of the random process $Z(.)$. Matheron (1971) categorizes the most common types as follows:

i. If $\theta_0 = 0$, then $Z(.)$ is L_2-continuous.

ii. If $\theta_0 \neq 0$, then $Z(.)$ is not even L_2-continuous and is highly irregular.

iii. If $\gamma(.)$ is a positive constant (except the origin where it is zero), then $Z(\mathbf{x_i})$ and $Z(\mathbf{x_j})$ are uncorrelated for any $\mathbf{x_i} \neq \mathbf{x_j}$, regardless of how close they are; $Z(.)$ is often called white noise.

Some isotropic variogram models

Some examples of valid variograms which possess the previous properties are now illustrated in Figure 2.3. These smooth curves are members of some valid parametric family of the type

$$P = \{\gamma : \ \gamma(.) = \gamma(.; \theta), \ \theta \in \Theta\}$$

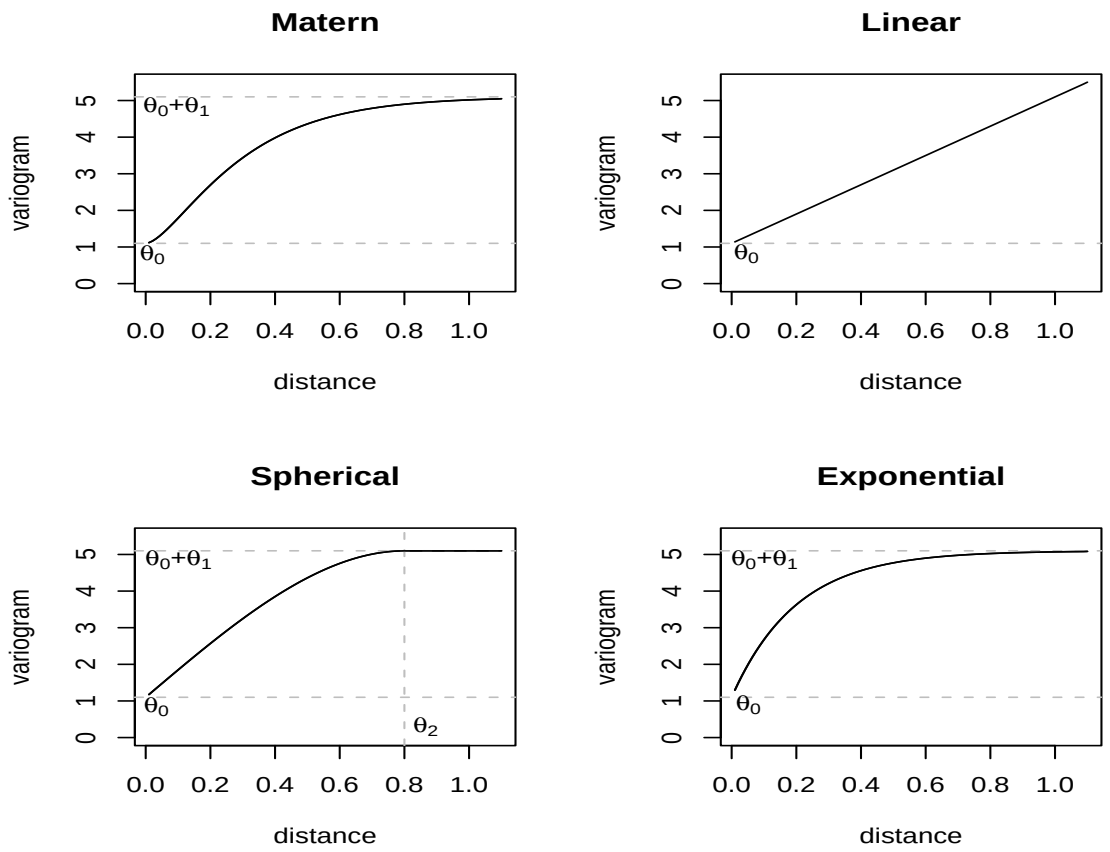


Figure 2.3: Examples of isotropic variogram models: Matérn (with $\kappa = 1.0$), linear, spherical and exponencial models.

where $\gamma(.;\theta)$ is a conditionally negative-definite function depending on the values found in the vector of parameters θ. We consider four basic isotropic models: Matérn, linear, spherical and exponential.

- Matérn model:

$$\gamma(u;\theta) = \theta_0 + \theta_1(1 - \rho(u;\phi = \theta_2, \kappa = \theta_3))$$

  where $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)^t$ and $\rho(u;\phi,\kappa)$ is given in (2.2).

- Linear model:

$$\gamma(u;\theta) = \theta_0 + \theta_1 u$$

where $\theta = (\theta_0, \theta_1)^t$. This function has no sill, $\lim_{u\to\infty} \gamma(u) = \infty$ and, thus, it does not correspond to a stationary process.

- Spherical model:

$$\gamma(u;\theta) = \begin{cases} \theta_0 + \theta_1 \left(\frac{3u}{2\theta_2} - \frac{1}{2} \left(\frac{u}{\theta_2} \right)^3 \right) & , 0 < u \leq \theta_2 \\ \theta_0 + \theta_1 & , u > \theta_2 \end{cases} \tag{2.5}$$

where $\theta = (\theta_0, \theta_1, \theta_2)^t$. In this model, $\theta_0 + \theta_1$ is the sill and θ_2 is the range. It has a linear behaviour near the origin and, in practice, it is one of the most used as it can be easily adjusted to data.

- Exponential model:

$$\gamma(u;\theta) = \theta_0 + \theta_1 \left(1 - exp\left(-\frac{u}{\theta_2} \right) \right), \ u \neq 0 \tag{2.6}$$

where $\theta = (\theta_0, \theta_1, \theta_2)^t$. In this model, $\theta_0 + \theta_1$ is the sill in an asymptotic sense only, while $\sqrt{3}\theta_2$ is the corresponding range. It has a parabolic behaviour near the origin.

Note that, with $\kappa = 0.5$ in (2.2) we get $\rho(u) = exp(-u/\phi)$, so the Matérn model and the exponential model will be the same.

Variogram estimation outline

According to Section 2.1, the variogram of an intrinsic and isotropic spatial process $Z(\mathbf{x})$ reduces to

$$\gamma(\|\mathbf{x_i} - \mathbf{x_j}\|) = \frac{1}{2}\text{Var}[Z(\mathbf{x_i}) - Z(\mathbf{x_j})] = \frac{1}{2}\text{E}[(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2].$$

Consequently, we can estimate the variogram from sample data $\{(\mathbf{x_i}, Z(\mathbf{x_i})),\ i = 1, ..., n\}$ replacing the previous theoretical expectation by the corresponding sample average. That is, for a given lag u, we can average the square differences between pairs of observations $Z(\mathbf{x_i})$ and $Z(\mathbf{x_j})$ whose corresponding locations $\mathbf{x_i}$ and $\mathbf{x_j}$ have $\|\mathbf{x_i} - \mathbf{x_j}\| = u$. This gives an idea of how one might estimate the variogram of a stationary process from the observed data. As we will see in Chapter 3, weights can be included in the averaging procedure to smooth the variogram estimation.

2.5 Trend and outlier identification

A sample variogram can be highly misleading if derived from data with a spatially varying trend. The same happens in the presence of outliers; bear in mind that a single outlying measurement, say $Z(\mathbf{x_k})$, might contribute to the variogram estimation with $n-1$ square differences $(Z(\mathbf{x_k}) - Z(\mathbf{x_i}))^2, \quad i = 1, ...n$ and $i \neq k$. Hence, it is important to consider some techniques of exploratory data analysis to identify possible non stationarity in the mean or isolated outliers.

Returning to the example of rainfall data in Paraná State in Brazil, we first show the rainfall data against each of the coordinates (two top panels in Figure 2.4). These confirm the trend from east to west and from south to north. We then show similar plots but replace the original data by residuals from a linear trend fitted by ordinary least squares (two bottom panels in Figure 2.4). Assuming that the remaining structure seen in the data can be attributed to the random part of the model, these ordinary least square residuals $Z(\mathbf{x_i}) - \widehat{\mu}(\mathbf{x_i})$ can be used for the subsequent variogram estimations. Moreover, this sample variogram could now be used to make a new estimation of the trend, as more reliable estimations of the deterministic and stochastic components of the spatial process depend on each other.

Cressie (1993) summarizes useful methods for exploratory spatial data analysis. If data is located on a regular grid[1], one possible solution is to calculate the sample mean or sample median (as a more robust estimator) across rows and down columns. Plots may then be created, summarizing row and column results. This allows us to identify the existence of a linear trend for mean or median along rows or columns. This type of analysis indicates whether and how the spatial location has influence on the variable values.

The use of median and mean serves another purpose. The comparison of these two variables has the additional function of highlighting rows or columns that may

[1]When the data is not located on a regular grid, a low-resolution grouping of observations into a two-way table, still allows these methods to be carried out.

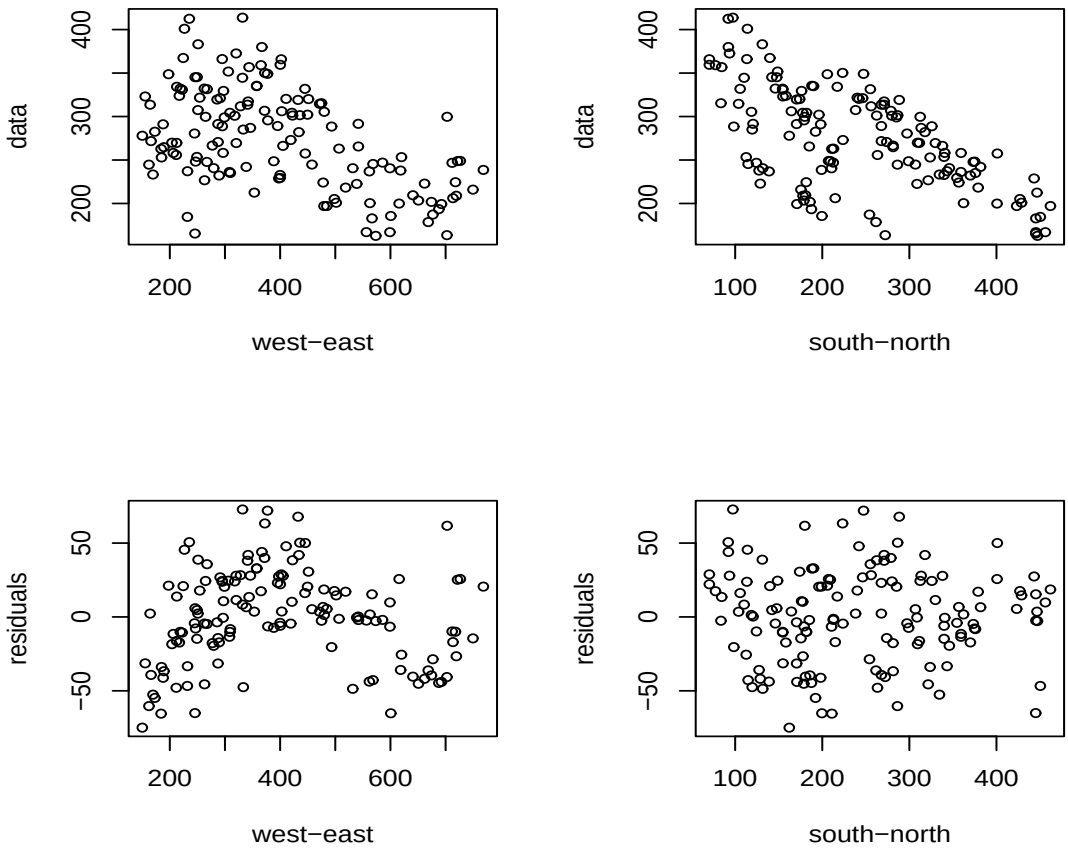


Figure 2.4: Paraná rainfall data shown against each of the coordinates: top two panels display original data and bottom two panels display residuals after fitting linear trend.

contain atypical observations. A high value for $mean - median$ indicates a possible outlier.

Another solution for stationarity analysis is drawing a bivariate plot of $Z(\mathbf{x})$ and $Z(\mathbf{x}+\mathbf{u})$, for a fixed direction $\mathbf{u}$, as $\mathbf{x}$ varies over the data locations. In case of local stationarity, when norm of $\mathbf{u}$ is small enough, this plot points should be near the bisectrix. This is also a way of detecting outliers, which correspond to those isolated values found far from the bisectrix.

The previous methods prove to be useful in detecting gross trends or isolated

outliers. Furthermore, a technique like the one called pocket-plot can allow the identification of localized areas as atypical with respect to a stationary model, i.e. pockets of non stationarity. Once more, this is done by exploiting the spatial nature of the data along rows and columns. These pockets, once discovered, should be removed from variogram estimation; they may be considered afterwards, when modelled and incorporated into final results analysis.

Non-stationary spatial processes

So far, it has been assumed that the data has come from a stationary process, apart from small pockets of non stationarity. This is a convenient working assumption which can be relaxed in various ways.

Sampson and Guttorp (1992) proposes *spatial deformation models* of the form

$$D(\mathbf{x_1}, \mathbf{x_2}) = \gamma(f(\mathbf{x_1}), f(\mathbf{x_2}))$$

with $\gamma(.)$ an isotropic variogram and $f(.)$ a smooth nonlinear map from $I\!R^d$ to $I\!R^{d'}$. In principle one may permit $d' \neq d$ though in most of their work the equality is assumed. The idea is that the map $f(.)$ takes the coordinates from the geographical space into an alternative dispersion space in which stationarity holds.

The transformation of the data itself can also help the non stationarity issue. When the responses $Z(\mathbf{x_i})$ $i = 1, ..., n$ are continuous but the Gaussian model is clearly inappropriate, some additional flexibility is achieved by introducing an extra parameter λ defining a Box-Cox transformation of the response (see e.g. Ribeiro Jr 2002).

Recently a number of methods, including kernel convolutions, deformations and spatially adaptive spectra, have been suggested to allow for non stationarity in the stochastic component of the underlying process. These methods build non stationarity directly into the covariance function (see e.g. Pintore and Holmes 2004 for a brief review). In addition, these authors show how, by working in the spectral domain, one can build non stationary covariance functions which are centred on

popular classes of stationary models such as the Matérn or Gaussian. The resulting non stationary models are defined as "localised" versions of their stationary counterparts. Alternative proposals are found in references like Higdon, Swall and Kern (1999), Fuentes (2002) and Stein (2005).

Bearing in mind the existence of methods to tackle non stationarity, such as previous ones, and the advantages of increased simplicity coming from more restrictive modelling assumptions, we shall keep the assumption of stationarity of the underlying spatial process in the remaining Chapters. A more restrictive assumption can, indeed, help models remain easily interpretable and, in case of complex models fitted to sparse data, can help to avoid issues of poor identifiability of model parameters.

Chapter 3

A comparison of approaches for valid variogram achievement

3.1 Introduction

In spatial prediction, a basic question is that, given a set of n observations of the process $Z(.)$ at points $\mathbf{x_i}$, $i = 1, ..., n$, what is the value taken by the variable at a point, $\mathbf{x_0}$, where data are unavailable? The approach differs from regression in that local features can affect the solution. In principle, all measurements should be considered. Having in mind that some measurements in the vicinity of the point investigated, or sometimes elsewhere, are more closely related than others to the true value at point $\mathbf{x_0}$, the appropriate procedure would be to adopt a *weighted mean*:

$$\widehat{Z}(\mathbf{x_0}) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{x_i}).$$

This linear combination may be considered to be an optimum estimate if the coefficients, or weights, λ_i are such that they sum to one and the estimator is unbiased and has minimum variance[1]. Only data within the radius of influence should be considered.

[1]More precisely, this estimator is classified as BLUE, i.e. *best linear unbiased estimator*.

The estimation of a valid variogram plays here a decisive role, as it is commonly used to find the optimal solution to the values of the weights. The method is called *kriging*, coined this way by Matheron (1963) to honour the mining engineer D.G.Krige.

The optimum solution can be found by using the Lagrange multiplier (see e.g. Kottegoda and Rosso 1997). In an isotropic field with estimated variogram values $\widehat{\gamma}(u_{ij})$ between points $\mathbf{x_i}$ and $\mathbf{x_j}$ at distances u_{ij}, the estimated weights, $\widehat{\lambda}_j,\ j = 1, ..., n$ may be found by solving the following $n+1$ simultaneous equations:

$$\begin{cases} \sum_{j=1}^{n} \lambda_i \widehat{\gamma}(u_{ij}) \ + \ \lambda \ = \ \widehat{\gamma}(u_{i0}), \quad i = 1, ..., n \\ \\ \sum_{i=1}^{n} \lambda_i \ = \ 1 \end{cases}$$

where index "0" relates to the unsampled location and λ is the Lagrange multiplier. Once the weights are estimated, a prediction value can be easily obtained for the process $Z(.)$ at the point with coordinates $\mathbf{x_0}$.

Some more details about kriging methods will be discussed in Chapter 6. In this Chapter our main aim is to stress the contribution of the variogram with respect to inference procedures. Moreover, variogram analysis provides a useful tool for summarizing spatial data and it may be used to measure spatial dependence between samples.

Commonly used variogram estimators

The first proposal, for the presence of stationary processes, for a variogram estimator is due to Matheron (1962). This estimator is based on the method of moments and it is often referred to as the classical estimator:

$$\widehat{\gamma}(u) = \frac{1}{2|N(u)|} \sum_{N(u)} \left(Z(\mathbf{x_i}) - Z(\mathbf{x_j})\right)^2 \tag{3.1}$$

where $N(u) = \{(\mathbf{x_i}, \mathbf{x_j}) : \|\mathbf{x_i} - \mathbf{x_j}\| = u,\ u \in I\!R\}$ and $|N(u)|$ is the total of pairs in $N(u)$. Matheron's estimator is unbiased[2], however it presents some drawbacks

[2]It is unbiased for $\gamma(.)$ when $Z(.)$ is intrinsically stationary (Cressie 1993, page 71).

such as being badly affected by atypical values due to the squared term in the summand of (3.1). In general, its statistical properties are difficult to study. If $Z(.)$ is a Gaussian random process, then $\widehat{\gamma}(u)$ is a linear combination of χ^2 random variables on one freedom degree.

According to Journel and Huijbregts (1978), a minimum of 30 pairs is recommended in $|N(u)|$. When data is not regularly spaced, this estimator can be obtained by considering a tolerance region around u.

Note that the squared term is also used to propose a related variogram, called the variogram cloud. If $\{(Z(\mathbf{x_i}), \mathbf{x_i}) : i = 1, ..., n\}$ is the sample data set, then the scatterplot of the points $\{(u_{ij}, v_{ij}) : \ j > i, \ \ u_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\|, \ \ v_{ij} = \frac{1}{2}(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2\}$ identifies the corresponding variogram cloud. As expected, this estimator is also sensitive to outliers.

Cressie and Hawkins (1980) has minimized this weakness, by working with square-root absolute differences and, under a Gaussianity assumption, have produced the estimator

$$\widehat{\gamma}(u) = \frac{\left\{\frac{1}{2|N(u)|} \ \sum_{N(u)} |Z(\mathbf{x_i}) - Z(\mathbf{x_j})|^{\frac{1}{2}}\right\}^4}{0.457 + \frac{0.494}{|N(u)|}} \tag{3.2}$$

where the term $0.457 + \frac{0.494}{|N(u)|}$ is used to make it unbiased.

Unfortunately, it has been suggested that the estimators in (3.1) and (3.2) should not be used for inference and prediction. The reason for this is that they may fail the conditionally negative-definite property which may lead to absurd negative values for the mean square prediction errors, as proved in Cressie (1993). If this were to occur, then the estimators are deemed to be invalid.

3.2 Traditional three stages

A common approach to achieving a valid variogram estimator is to approximate an empirical variogram by some theoretical model which is known to be valid. The

idea is to select, within the families of valid variograms, a function which captures the underlying spatial dependence of the available data. Traditionally, these type of approaches are accomplished through three distinct stages:

1. *Compute an empirical* variogram (typically non valid);

2. *Choose a theoretical model* among the family of valid parametric or non-parametric variograms;

3. Estimate the variogram by *fitting the theoretical model to the empirical* variogram.

To accomplish these tasks, there are several different approaches strongly defended by their authors. In this Chapter, our work's main purpose was to identify these approaches (Menezes 2002) and compare some of them based on a numerical study, covering different kind of spatial dependence situations. The comparisons are mainly based on the integrated squared errors of the resulting valid estimators. The main contributions of this work appear in Menezes, Garcia-Soidán and Febrero-Bande (2005*a*).

Note that some authors prefer to group these three stages into two parts, *variogram estimation* and *variogram fitting*; the latter part incorporates stages 2 and 3 simultaneously (e.g. Cressie 1993). In contrast, we argue that, when possible, three separate stages allow a better classification of the existing approaches. The output of stage 2 is a vague valid candidate and its complete specification is only obtained from stage 3.

Before giving details about the complete approaches that we examined, we make some generic comments on each of the previously listed stages. We shall point out some references, if we think they introduce a relevant idea for the implementation of these tasks.

3.2.1 Stage 1 – Empirical variogram estimation

The word "empirical" means *based on observation or experiment.* The estimation of the empirical variogram always, unsurprisingly, begins with the observed data, whichever estimator is used. Examples include those estimators introduced in Section 3.1.

Robustness to outliers is normally considered an important characteristic for any estimator. In this regard, some other robust empirical estimators have been proposed in addition to (3.2) (see e.g. Dutter 1996, for a general review, or Gunst and Hartfield 1997, for large data sets). They usually avoid one or both of the following issues associated with the classical estimator:

- the square term, $(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2$, because it induces distortion in data values;
- the mean, because it is not a robust location estimator.

For instance, Armstrong and Delfiner (1980) proposed to use the square of the interquartile range of the differences

$$[\mathrm{UQ}\{Z(\mathbf{x_i}) - Z(\mathbf{x_j})\} - \mathrm{LQ}\{Z(\mathbf{x_i}) - Z(\mathbf{x_j})\}]^2$$

where UQ and LQ stands for upper and lower quartiles, i.e. the 75^{th} and 25^{th} percentiles of the differences $Z(\mathbf{x_i}) - Z(\mathbf{x_j})$. Additionally, they have also considered a sample quantile, the median, of squared differences:

$$\mathrm{med}\{(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2 : (\mathbf{x_i}, \mathbf{x_j}) \in N(u)\}.$$

Note that these approaches need a correct normalization to make them unbiased.

Genton (1998*a*) proposes a variogram estimator based on the highly robust scale estimator of Rousseeuw and Croux (1992,1993), denoted below by Q_{N_u}. Considering $Z(.)$ a intrinsically stationary but not necessarily isotropic process,

they use the theory of M-estimators of scale to derive robustness properties. The resulting estimator is

$$2\widehat{\gamma}(\mathbf{u}) = (Q_{N_u})^2,\ \mathbf{u} \in I\!R^d$$

where Q_{N_u} is defined by

$$Q_{N_u} = 2.2191\{|(Z(\mathbf{x_i}+\mathbf{u}) - Z(\mathbf{x_i})) - (Z(\mathbf{x_j}+\mathbf{u}) - Z(\mathbf{x_j}))| : i < j\}_{(k)}$$

The factor 2.2191 is used for consistency in the Gaussian distribution. N_u is the cardinality of $\{(\mathbf{x_i}, \mathbf{x_j}) : \mathbf{x_i} - \mathbf{x_j} = \mathbf{u},\ \mathbf{u} \in I\!R^d\}$, $k = \binom{[N_u/2]+1}{2}$ and $[N_u/2]$ denotes the integer part of $N_u/2$. This means that they use the k^{th} quantile of all sorted $|.|$ values. One may note that Q_{N_u} does not rely on any location knowledge and is thus said to be location-free, in contrast to Matheron's estimator.

We now introduce a non-parametric approach for the empirical estimation of $\gamma(.)$ that employs a kernel density. The Nadaraya-Watson's kernel estimator is given by the weighted average

$$\widehat{\gamma}_h(u) = \frac{\sum_i \sum_{j \neq i} w_{ij}[Z(\mathbf{x_i}) - Z(\mathbf{x_j})]^2}{\sum_i \sum_{j \neq i} w_{ij}} \quad \text{where} \quad w_{ij} = K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right) \quad (3.3)$$

K is a symmetric, zero-mean and bounded density function, with compact support $[-C, C]$, and h is a bandwidth parameter. In Garcia-Soidán, Febrero-Bande and Gonzalez-Manteiga (2004) several properties of this estimator are studied and an asymptotically optimal bandwidth parameter obtained.

With Matheron's estimator, only pairs $(\mathbf{x_i}, \mathbf{x_j})$ such that $\|\mathbf{x_i} - \mathbf{x_j}\| = u$ are used to compute a specific $\widehat{\gamma}(u)$. If data is not regularly spaced, Matheron's estimator can be adapted to consider a tolerance region around u. For kernel estimator, all pairs are used and they are all given a particular weight: the weights are at their maximum when the distance between two points is close to u, and zero values if $\left|\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right| > C \iff \|\mathbf{x_i} - \mathbf{x_j}\| \notin [u - hC, u + hC]$. Consequently, this kernel estimator offers a smoother estimation of the variogram. In addition, take into account that the Nadaraya-Watson estimator with the uniform kernel, provides

the Matheron's estimator over a tolerance region given by $(u - hC, u + hC)$.

Two other kernel estimators are found in Yu and Mateu (2002) and in Garcia-Soidán, Gonzalez-Manteiga and Febrero-Bande (2003). In expression (3.3), they change the weights to, respectively:

- $w_{ij} = \frac{1}{\delta_0(\|\mathbf{x_i}-\mathbf{x_j}\|)} K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h\delta_0(\|\mathbf{x_i}-\mathbf{x_j}\|)}\right)$,
  where $\delta_0(.) > 0$ is an additional parameter that adapts the amount of smoothing to the local density of distances, originating a *nearest-neighbour* variogram estimator.
- $w_{ij} = K\left(\frac{\|\mathbf{x_i} - \mathbf{x_j}\| - u}{h}\right) \times$
  $\times \quad \sum_{k=1}^{n}\sum_{l=1}^{n} K\left(\frac{\|\mathbf{x_k} - \mathbf{x_l}\| - u}{h}\right)(\|\mathbf{x_k} - \mathbf{x_l}\| - u)(\|\mathbf{x_k} - \mathbf{x_l}\| - \|\mathbf{x_i} - \mathbf{x_j}\|)$,
  originating a *local linear regression* estimator of the variogram. It can be seen as a generalization of the Nadaraya-Watson estimator, as the former is associated to a generic straight line equation, whereas the latter is associated to a constant. This allows a better behaviour near the boundaries.

Bear in mind that, in practice, kernel estimators present some bias, which may not be negligible for small n. However, the latter term tends to zero as n increases and, therefore, they are asymptotically unbiased.

3.2.2 Stage 2 – Valid model selection

The aim of this stage is to find a negative-definite function which, as a measure of spatial dependence, is in some sense closest to the sample data. The notion of "in some sense" is considered in detail at stage 3. At this stage we are concerned with questions such as the choice of spherical vs exponential families, or parametric vs non-parametric estimators.

For the vast majority of identified approaches, this search is restricted to the *space of parametric families*. This is the set of all analytic functions that depend on parameters in such a way that two members of the same family can be distinguished

by their parameters values, and all members are conditionally negative-definite. Examples are all those isotropic models introduced in Section 2.4. The alternative *non-parametric space* will be supported by the spectral representation of positive definite functions (e.g. Stein 1999 or Chilès and Delfiner 1999).

The most common methods used to pick a valid family are based on graphical tools, with model selection reduced to approximating the estimated variogram curve by one from the valid family. In recent years, some alternatives have been suggested.

Maglione and Diblasi (2001) proposes a statistical method for choosing a valid model for the variogram of an isotropic process. The test statistic for their approach is based on smoothed random variables which reflect the underlying spatial variation. The distribution of this test statistic, which is a ratio of quadratic forms, can be approximated by a shifted χ^2 distribution and is used to verify the *distance* between the underlying model for the variogram and the one in the null hypothesis.

Any test statistic to assess a specific valid model for the variogram must take into account the variability of an estimator, possibly the empirical one. This variability can be gathered by a function of the stochastic process of differences $Z(\mathbf{x}') - Z(\mathbf{x})$, which has zero expectation. Functions $(.)^2$ or $|.|^{\frac{1}{2}}$ are usual options.

Their method can then be summarized in this way:

1. $R_k = Z(\mathbf{x_j}) - Z(\mathbf{x_i})$, where k may be obtained from a bijection between the set of all (i, j) for each $(\mathbf{x_i}, \mathbf{x_j})$ and a set of positive integers $\{1, 2, ..., n(n-1)\}$;

2. $S_k = |R_k|^{\frac{1}{2}} - E_0(|R_k|^{\frac{1}{2}})$, where E_0 means expected value under the null hypothesis;

3. $\widetilde{S}_k = \sum_{r=1}^{N} w_r^{(k)} S_r$ are the smoothed variables with weights

$$w_r^{(k)} = \frac{exp\left(-\left(\frac{u_k - u_r}{b}\right)^2\right)}{\sum_{r=1}^{N} exp\left(-\left(\frac{u_k - u_r}{b}\right)^2\right)}$$

where b is a bandwidth parameter that controls the degree of smoothing on S_r, u_k and u_r are distances, and N is the total number of distances in the region;

4. Finally, the proposed test statistic becomes

$$T = \frac{\sum_{k=1}^{N} \left(S_k - (2\gamma(u_k;\theta_0))^{1/4} \, \overline{S} \right)^2 - \sum_{k=1}^{N} \left(S_k - \widetilde{S}_k \right)^2}{\sum_{k=1}^{N} \left(S_k - \widetilde{S}_k \right)^2}$$

One possible test of hypothesis would be

$$H_0 : \ 2\gamma(.) = 2\gamma(.;\theta_0) \quad \text{versus} \quad H_1 : \ 2\gamma(.) \neq 2\gamma(.;\theta_0)$$

If $T > t_{1-\alpha}$, then H_0 should be rejected. According to the usual notation, $t_{1-\alpha}$ is a (1-α)–quantile calculated from the distribution of T under H_0 and α is a given significance level.

Gorsich and Genton (2000) suggests a method for the selection of a valid parametric model via the derivative of a non-parametric variogram estimate, without assuming a prior model. The basic idea of their proposal is to avoid choosing among valid parametric variogram models, as they may look similar, and to choose instead among their derivatives, as they are often quite different. These derivatives should be compared with the one obtained from the non-parametric variogram estimate based on the spectral representation of positive definite functions, as we will see below.

The first non-parametric approach to the selection of a valid model appeared in Shapiro and Botha (1991). A key result behind these approaches is Bochner (1955)'s theorem, which states that a covariance function $c(u)$ is positive definite iff it has the following form (Cressie 1993):

$$c(u) = \int_0^\infty \Omega_d(ut) dF(t)$$

where $\Omega_d(x) = (2/x)^{(d-2)/2}\,\Gamma(d/2)J_{(d-2)/2}(x)$ is a basis for functions in $I\!R^d$, $F(t)$ is a nondecreasing bounded function, Γ is the gamma function, and J_υ is the Bessel function of the first kind of order υ. Some common families are $\Omega_1(x) = cos(x)$, $\Omega_2(x) = J_0(x)$ and $\Omega_3(x) = sin(x)/x$.

This theorem, together with the relation $\gamma(u) = c(0) - c(u)$ in (2.1), are employed to represent the family of non-parametric valid variograms. To allow the numeric evaluation of γ, $F(t)$ should be considered a step function with a finite number m of positive jumps $p_1, ..., p_m$ at points $t_1, ..., t_m$. A valid non-parametric estimator can then be given by

$$\widehat{\gamma}(u_i) = \Sigma_{j=1}^{m}\; p_j\,(1 - \Omega_d(u_i t_j)) \tag{3.4}$$

Returning to the proposal of Gorsich and Genton (2000), the complete specification of expression (3.4) requires the derivation of jumps p_j. They choose to minimize the squared differences between the Matheron's empirical estimations, $\widehat{\gamma}_M(u_i)$, and those from (3.4). This minimization problem can be formulated, in matrix notation, as

$$\min\left\{(\widehat{\gamma}_M - M\mathbf{p})^t\, W^{-1}\,(\widehat{\gamma}_M - M\mathbf{p})\right\} \tag{3.5}$$

where $M_{ij} = 1 - \Omega_d(u_i t_j)$, $\widehat{\gamma}_M = (\widehat{\gamma}_M(u_1), ..., \widehat{\gamma}_M(u_n))^t$, $\mathbf{p} = (p_1, ..., p_m)^t$ and W is a weighting identity matrix. Finally, a classical approach to estimate the derivative of $\widehat{\gamma}$ is to use the estimated jumps $\mathbf{p}$, differentiate the function Ω_d and then consider $M'_{ij} = -\Omega'_d(u_i t_j)$. The result becomes $\widehat{\gamma}' = \mathbf{M'p}$.

Bear in mind that both proposals of Maglione and Diblasi (2001) and Gorsich and Genton (2000) end with the selection of a valid parametric model. Alternatively, Shapiro and Botha (1991) proposes a selection among a broad class of *permissible* variograms, using as we have seen the Bochner's theorem. This results into a valid non-parametric model.

3.2.3 Stage 3 – Model fitting

The classical fit criteria may be used to complete the specification of the final variogram. Possible choices are the *minimum variance* or *norm quadratic unbiased* (MIVQU or MINQU), the *maximum likelihood* (ML) and the *least squares* (LS) criteria. Those based on LS are known as being less limited in scope and require the fewest distributional assumptions about $Z(\mathbf{x})$. In matrix notation, a LS minimizing problem is written as

$$min\left\{(\widehat{\gamma}-\gamma_v)^T W^{-1}(\widehat{\gamma}-\gamma_v)\right\}$$

where $\widehat{\gamma}$ identifies an empirical estimator and γ_v identifies a valid non-parametric or parametric model. In the latter case, $\gamma_v = \gamma_\theta$ whose exact form is known, except for the unknown parameter θ. If W is an identity matrix, then one has the *ordinary least squares* (OLS) criterion. If $W = V$ where V is the variance-covariance matrix whose elements are of type $V_{ij} = Cov[\widehat{\gamma}(u_i), \widehat{\gamma}(u_j)]$, then one has the *generalized least squares* (GLS) criterion. If matrix V is reduced to its diagonal, then the resulting criterion is called *weighted least squares* (WLS).

Cressie (1985) considers WLS as a pragmatic compromise between GLS efficiency and OLS simplicity and suggests $w_j = \frac{|N(u_j)|}{\gamma(u_j)^2}$, where the unknown γ should be approximated by γ_θ through an iterated procedure (weights can start, for instance, equal to 1). Some notes about Cressie's weights are:

- If the cardinality of $N(u_j)$ is larger, then the associated weights w_j are also larger;
- If the value of γ_θ is smaller, then the associated weight is larger. This allows a better characterization near the origin;
- If the variance is larger, then the associated weight is smaller.

Genton (1998*b*) refuses to accept WLS as the solution for GLS complexity and proposes an explicit formula for the covariance structure V, calling the resulting method GLSE. The basic idea is to obtain a generic covariance structure by using, iteratively, the correlation structure of Matheron's estimator in the independent case. The main steps of this algorithm for variogram model fitting are:

1. Determine the matrix $V = V(\theta)$ such that

$$V_{ij} = Cov(2\widehat{\gamma}(u_i), 2\widehat{\gamma}(u_j)) = \frac{Corr(2\widehat{\gamma}(u_i), 2\widehat{\gamma}(u_j))\gamma(u_i;\theta)\gamma(u_j;\theta)}{\sqrt{N(u_i)N(u_j)}}$$

   where $Corr(2\widehat{\gamma}(u_i), 2\widehat{\gamma}(u_j))$ can be approximated by the result obtained for the Matheron's estimator in the independent case;

2. Choose $\theta^{(0)}$ randomly or by using OLS or WLS criteria, and let $i = 0$;

3. Compute matrix $V(\theta^{(i)})$ and determine $\theta^{(i+1)}$ which minimizes

$$G(\theta) = (2\widehat{\gamma} - 2\gamma_\theta)^T V(\theta^{(i)})^{-1} (2\widehat{\gamma} - 2\gamma_\theta)\,;$$

4. Repeat (3) until convergence to obtain $\widehat{\theta}$.

Genton concludes his work by carrying out some simulations to show that the GLSE criterion, combined with a robust variogram estimator, may improve the fit significantly, even in the presence of outliers.

3.2.4 Existing combinations of the previous stages

Next, we shall introduce some existing complete approaches to reach our target: a valid variogram estimator. All of them result from distinct combinations of previous stages, summarized in Table 3.1.

We shall begin with a mandatory reference, **Zimmerman and Zimmerman (1991)**, where seven different approaches are compared through a Monte Carlo simulation study. This comparative study, in spite of being considerably exhaustive,

is somehow restricted in scope as it only involves parametric techniques. In fact, these seven approaches are mainly distinguishable by their third stage, four of which are LS-based, two ML-based and one using a modified MIVQU.

Their main conclusions may be summarized as follows. They confirm that the performance of each estimator improves, in the sense that its distribution becomes less dispersed as the sample size increases. In general terms, the estimators of parameters perform better as the spatial dependence is weaker. However, the standard 95% prediction intervals perform better when the spatial dependence is stronger. Moreover, they conclude that, in some particular situations, the likelihood-based methods can perform a little better than the least squares methods. These are, however, less computationally demanding and are deemed to be perfectly acceptable in most cases.

The paper of **Shapiro and Botha (1991)** is pioneered towards selecting a valid model in a non-parametric space, as we mentioned in Section 3.2.2. They combine the Matheron's estimator at stage 1, a broad class of *permissible variograms* at stage 2 and at the last stage, a WLS fitting criterion where the optimization problem is reduced to a quadratic programming problem. Following Christakos (1984), they define $f(u),\ u \in I\!R^d$ as a *permissible* variogram function, if it is continuous (except possibly at the origin), $f(u) = f(-u),\ \ f(u) \geq 0$ for all u, and $-f(u)$ is conditionally nonnegative definite. The resulting valid variogram estimator then fulfills equation (3.4).

Additionally, requirements such as variogram smoothness, monotonicity or convexity may be incorporated into the fitted variogram γ_v leading to a better approximation. Suppose the empirical $\widehat{\gamma}(u_i)$ are scattered, then γ_v may change rapidly. In this case, it may be important to impose a smoothness condition, forcing this function's derivative to be bounded. It may also be important that the γ_v be monotonically increasing or that it be convex. As expected, the monotonicity condition may be ensured by forcing the derivative to be positive for all $u > 0$,

and the convexity condition by forcing the second derivative to be negative.

This approach was evaluated by Cherry, Banfield and Quimby (1996), where they conclude that this "non-parametric method is faster, easier to use and more objective than parametric methods".

Gribov, Krivoruchko and Ver Hoef (2000) suggests a new method of computing the empirical variogram of Matheron. The squared differences $[Z(\mathbf{x_i}) - Z(\mathbf{x_j})]^2$ are binned into K distinct bins, and point estimations of the semivariogram at K points are obtained. The complicating issue is how to best bin the data. They introduce the notion of logarithmic increases in the size of tolerance regions against the traditional fixed size. This new concept allows better results in estimation near the origin. They also propose to use a kernel method to assign, within a given bin, weighted values depending on how close a value is to the center of the bin. This requires fewer elements per bin than the recommended minimum of 30 pairs suggested from the classic guideline of Journel and Huijbregts (1978), as well as the weights' presence minimizes a possibly existing unequal distribution of lags.

With respect to the *model fitting* stage, they propose a modified WLS[3] procedure. They split this stage into two steps. At step 1, typically with two iterations, they consider logarithmic lag sizes. At step 2, a default lag size obtained from the range estimate in step 1 is used instead.

The last approach included in our survey is the one proposed by **Garcia-Soidán et al. (2004)**. These authors propose the usage of the non-parametric empirical estimator given by equation (3.3), together with the *permissible* function of Shapiro and Botha (1991). The empirical $\widehat{\gamma}_h(u)$ and the theoretical curve are fitted through a re-iterated WLS criterion. The former is shown to have desirable properties, such as asymptotically unbiasedness and consistency.

[3]This algorithm is included into the Geostatistical Analyst extension to GIS ArcInfo/ArcView8.1 (Krivoruchko, 1999).

One should have in mind that an important issue of kernel estimation is the selection of the bandwidth parameter, h. These authors address the problem by asymptotically minimizing the *mean square error* (MSE) or the *mean integrated square error* (MISE), in order to derive the local and the global bandwidth, respectively. Both expressions involve the unknown function $\gamma(u)$. For the purpose of the bandwidth derivation, a simple parametric approach, like the first one presented by Zimmerman and Zimmerman (1991) (see Table 3.1), may be used to estimate $\gamma(u)$. This isolated parametric estimation can even be improved by being incorporated into an iterated non-parametric procedure.

Table 3.1: Taxonomy of existing approaches for valid $\widehat{\gamma}$ achievement. Bold identifies those approaches selected for the comparative study and, between brackets, main equations references are found.

Approaches	Stage 1	Stage 2	Stage 3
Zimmerman	Matheron(3.1)	P model	OLS
and	Cressie-Haw.(3.2)	P model	WLS
Zimmerman	**Matheron**	**P model**	**WLS**
(1991)	Matheron	P model	WLS-Delfiner (1976)
	—	P model	ML
	—	**P model**	**REML**
	Matheron	P model	OLS+MIVQU
Shapiro and Botha (1991)	Matheron	NP function(3.4)	WLS
Gribov *et al.* (2000)	Matheron-modif.	P model	WLS-modified
Garcia-Soidán *et al.* (2004)	**NW kernel**(3.3)	**NP function**(3.4)	**WLS**

Outside the boundary, the bias of the Nadaraya-Watson estimator (3.3) is of the order h^2; however, the latter order amounts to h for distances u close to 0. Then, proceeding as in Kyung-Joon and Shucany (1998), we may denote by $\hat{\gamma}_{q,h}(u)$ the estimator obtained by substituting a boundary kernel H_q for the symmetric one K in (3.3), where $q = \min\{uh^{-1}, C\}$ and

$$H_q(z) = \frac{K(z) - rL(z)}{1-r}, \quad z \in [-C, q]$$

where K and L are symmetric kernel functions, $r = c_{1,K}c_{0,L}(c_{0,K}c_{1,L})^{-1} \neq 1$ and $c_{i,G} = \int_{-C}^{q} z^i G(z)dz$. This particular selection of the boundary kernel H_q produces a variogram estimator $\hat{\gamma}_{q,h}(u)$ that makes it negligible the term of order h in the bias and preserves the same convergence orders for all $u > 0$, as shown in Garcia-Soidán et al. (2004).

3.3 Simulation study

In order to analyze the performance of the previous approaches for valid $\widehat{\gamma}$ achievement, simulations of spatial data in $I\!R^2$ were carried out for different kinds of dependence situations. We considered the spherical and the exponential variogram models given in (2.5) and (2.6), respectively. Additionally, the wave model was also considered, because of its atypical irregular behaviour:

- Wave model: $\gamma_w(u;\theta) = \theta_0 + \theta_1\left[1 - \theta_2\ sin(u/\theta_2)/u\right], \quad u \neq 0.$

Bear in mind that we want to restrict ourselves, in this Chapter, to the estimation of the spatial dependency under standard sampling. Thus, in all cases, a uniform distribution on $[0,1] \times [0,1]$ was assumed for spatial locations $\mathbf{x_i} = (x_{i1}, x_{i2}),\ i = 1,..,n$, where n represents the sample size. Several data sets were generated with Gaussian data, $Z(\mathbf{x_i}),\ i = 1,..,n$, using one of the above variogram models. The parameters of these models were chosen in such a way that the corresponding curves were comparable according to their range. We fix the values for the nugget θ_0 and θ_1 to be 0.25 and 5.0, respectively. The third

parameter was the one chosen depending on the model: exponential, $\theta_2 = 0.167$; spherical, $\theta_2 = 0.5$; and wave, $\theta_2 = 0.113$. With this selection, the theoretical variograms have a sill of 5.25 and a range (referred to as the minimum value for which the variogram reaches either the sill or 95% of the sill, in case that the range is not finite) of 0.5. More precisely, the wave model oscillates around the sill value and, consequently, the 0.5 value identifies the global maximum of the corresponding variogram function.

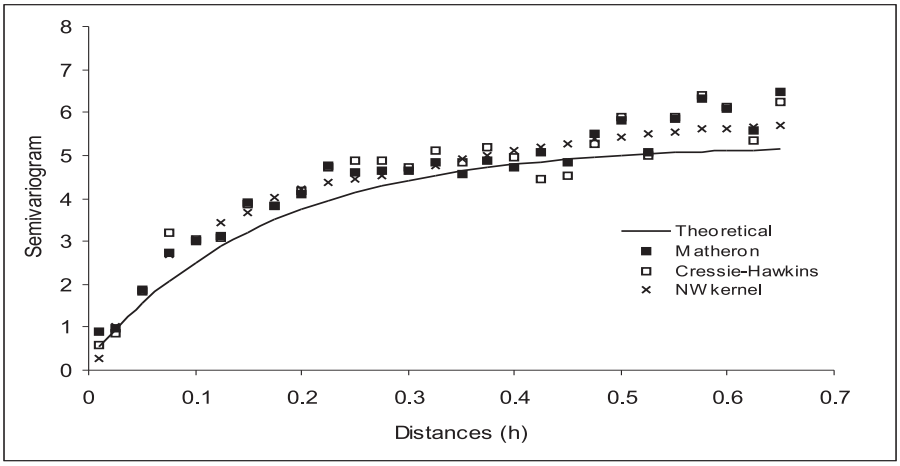
3.3.1 Comparing empirical estimators

The aim of our first numerical study was to compare the three main empirical estimators used at stage 1 for the approaches included in Table 3.1; these are given in expressions (3.1), (3.2) and (3.3). For data generation, we took a sample size of $n = 200$ and we started by selecting the exponential model.

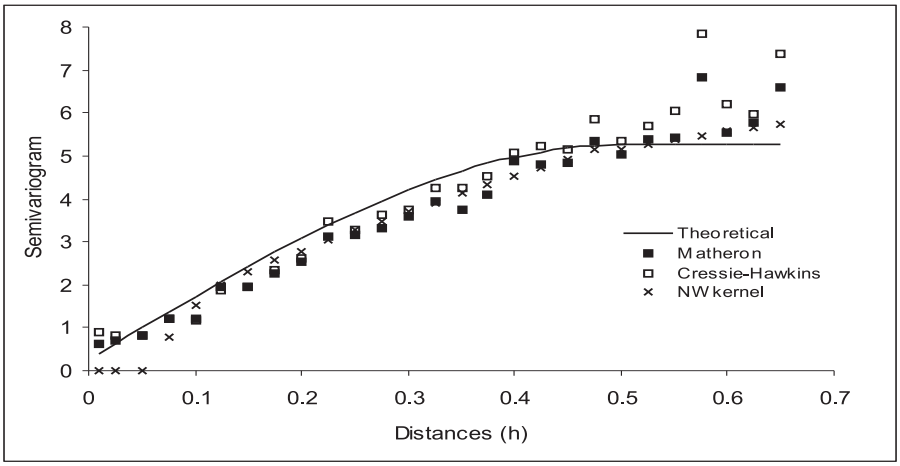
Unusual estimated values were obtained using estimators (3.1) and (3.2) for the largest lags. Additionally, some of them did not have the recommended minimum of 30 pairs. Therefore, in posterior simulations, we have decided to only consider the first 55% of lags. One may note that this guideline is still less *conservative* than the one proposed by Journel and Huijbregts (1978), who specifies that the largest used lag u_k should be less than or equal to half of the largest existent lag. As non-parametric estimation requires more lags than those empirically obtained, we have also decided to consider a larger number of lags, equally spaced, within interval $[\ min(u_k)\ ,\ 0.55 * max(u_k)\ ]$.

Following these considerations, Figure 3.1 shows the obtained data, as well as two more graphs assuming the spherical and the wave models for data generation. All graphics included in this Chapter use the following notation: lines are used to represent a valid estimator; and isolated symbols, e.g. small squares, are used for empirical estimates.

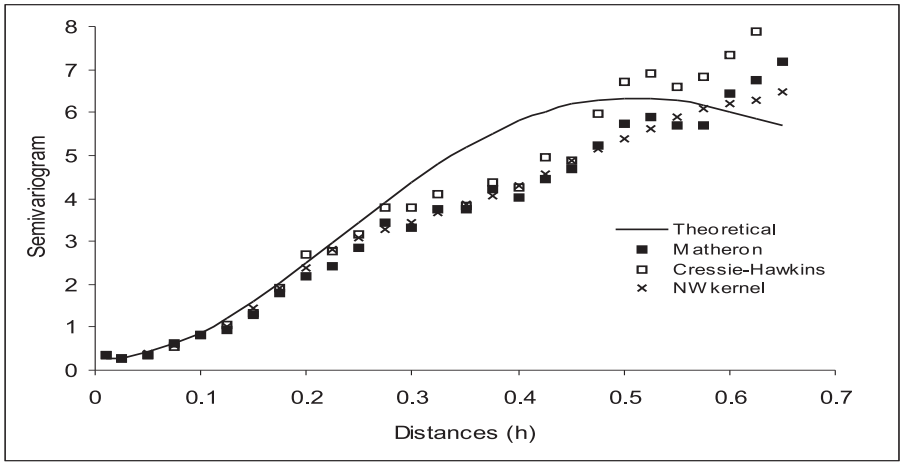
Figure 3.1 demonstrates the behaviour of the estimator when one sample is considered, although it will depend strongly on the sample variability. For this



a) Exponential

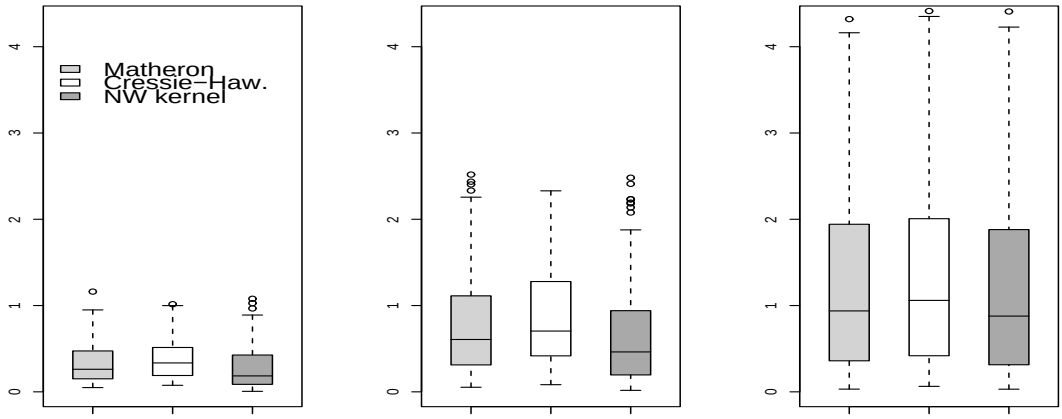


b) Spherical



c) Wave

Figure 3.1: Three empirical estimators and the associated theoretical curve. Data simulated with three distinct models. Sample size equals 200.



a) Exponential b) Spherical c) Wave

Figure 3.2: Boxplot of the evaluated ISE from three empirical estimators, using data simulated from three distinct models. The simulation consisted of 100 replications, each with a sample size of 200.

reason, we include a second study where 100 independent samples are considered. For each one, the *integrated squared error* (ISE) between each of the three empirical estimators and the theoretical variogram, given by

$$ISE = \int \left[\widehat{\gamma}(u) - \gamma(u)\right]^2 du, \tag{3.6}$$

was approximated numerically through the trapezoid rule, $\widehat{\gamma}(u)$ represents an empirical estimator and $\gamma(u)$ represents the theoretical curve. This simulation was repeated for the previous models: exponential, spherical and wave.

The results are summarized in the boxplot in Figure 3.2. If one compares the *median* values associated with the three estimators, then the best performance is clearly achieved by the non-parametric estimator, using the Nadaraya-Watson kernel. Another advantage of this non-parametric estimator is that it is a continuous function. In contrast, estimators (3.1) and (3.2) propose point values of the semivariogram for given distances u, making them discontinuous. Most analyses requires knowledge about estimations in a continuous range of $\gamma(u)$.

We conclude by bringing attention to the different orders of magnitude of the ISE values for each theoretical model, lower values are associated to the exponential curve and the largest ones to the wave curve.

3.3.2 Comparing complete approaches

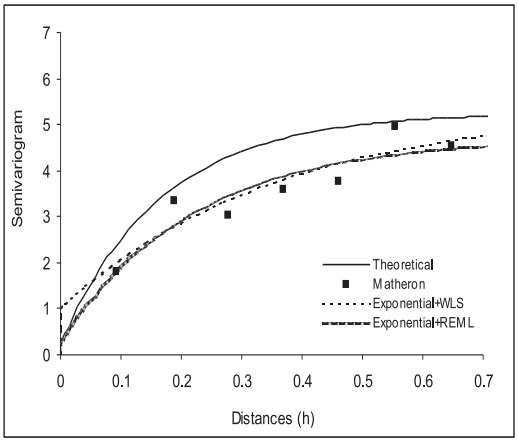
We highlight three approaches (marked in bold) from Table 3.1, which we consider the most representative of the existing alternatives. For two of the approaches, a valid model is chosen within the space of parametric families. Hereafter, they are identified as the *parametric approaches* (P), one of which uses WLS as the fitting criterion and the other REML. The third approach, introduced by Garcia-Soidán et al. (2004), will be referred to as the *non-parametric approach* (NP).

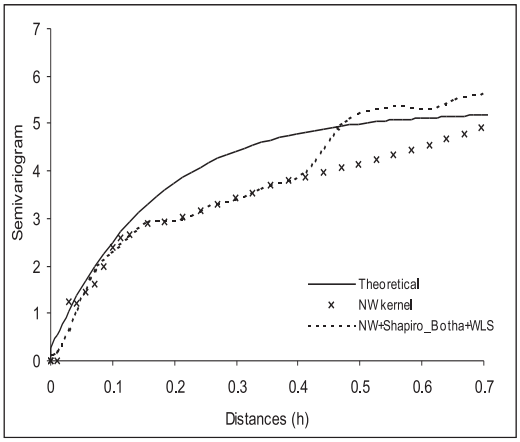
The superior results of the Nadaraya-Watson kernel estimator, when compared to the Matheron's estimator, led us not to include Shapiro and Botha (1991) in our numerical study. Gribov et al. (2000) was also excluded as, under isotropy, their main contribution is reduced to the usage of weights within a given bin. In this case, the kernel estimator does not differ much from their proposal and may be indeed a better choice.

Under the NP approach, we preferred to asymptotically minimize the MSE to derive a local bandwidth parameter. For this purpose, the symmetric Epanechnikov kernel was employed. Additionally, as the bandwidth derivation needs itself an estimation of the variogram, the available WLS parametric estimation was used for this purpose. Near the variogram endpoint 0, a specific asymmetric boundary kernel was constructed from the Epanechnikov kernel and the quartic kernel.

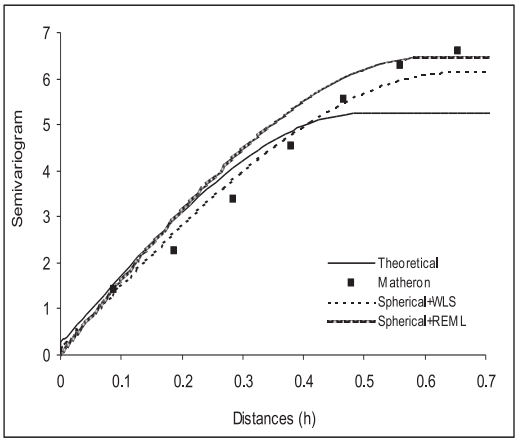
For the implementation of the REML fitting criterion, we used the geoR library from R, which provides several functions for geostatistical analysis as explained in Ribeiro Jr and Diggle (2001). Excluding this particular case, we used Fortran to implement our numerical study.

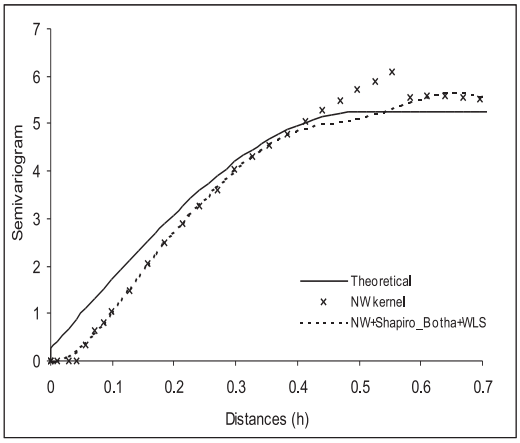
Figure 3.3 shows an example of results obtained with the three selected approaches, when using each of our theoretical models for data simulation and a



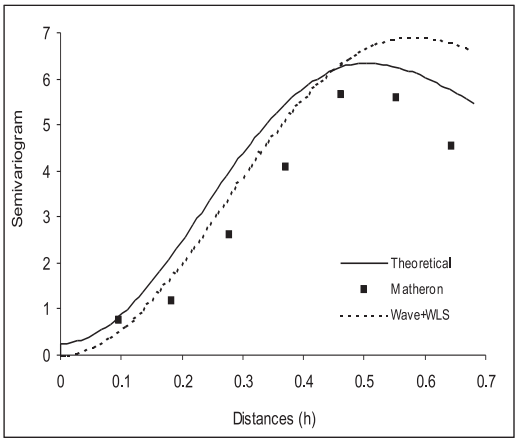


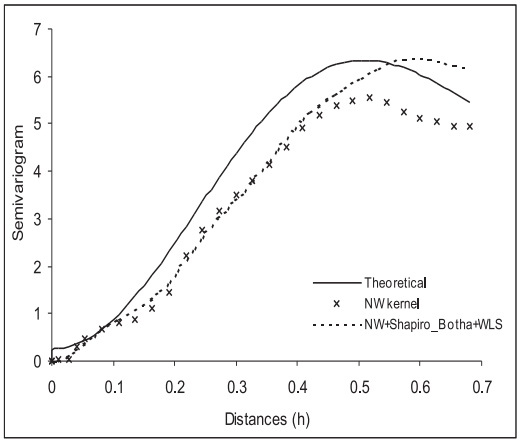
a) Data simulated with an exponential model





b) Data simulated with a spherical model





c) Data simulated with a wave model

Figure 3.3: Approaches to achieving a valid $\widehat{\gamma}(u)$: the 2 parametric approaches are on the left and the non-parametric approach is on the right. Data simulated with three distinct models. Sample size equals 50.

sample size of $n = 50$. The correct specification of the theoretical variogram is considered: if data is generated with a given model then this same model is the one elected at stage 2. On the left side, are the valid estimators resulting from the P approaches. On the right side, the final valid estimator is given by a *permissible* function of Shapiro and Botha when fitted through WLS to the NW kernel's estimations.

The wave variogram, see e.g. Figure 3.3c, causes problems in achieving a valid estimator through REML fitting criterion, because the Cholesky factorization of the variance-covariance matrix was required and this matrix was typically non-positive definite.

Impact of P model selection at stage 2

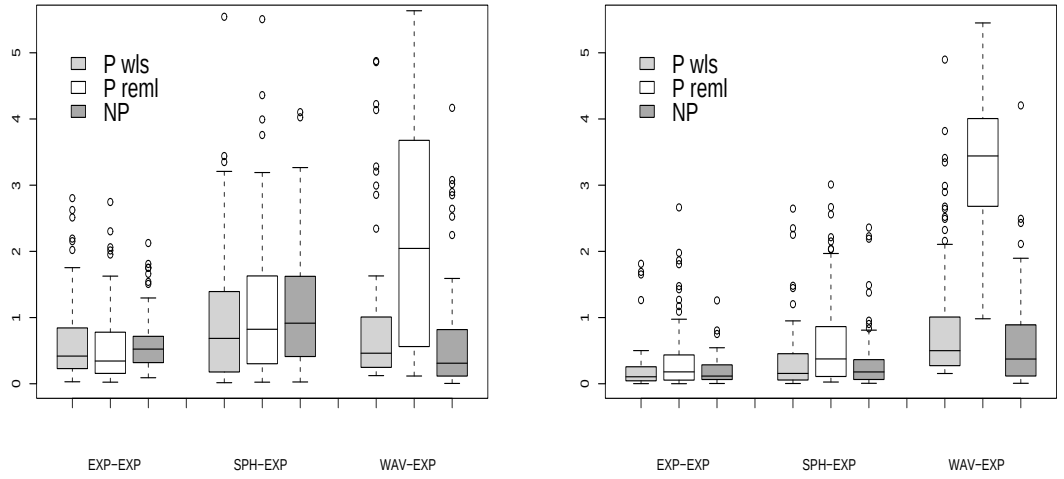
Next, we will cover different kinds of spatial dependence situations. The data were generated using any of our three elected models, exponential, spherical or wave, and we supposed that any of these models could be chosen as the best guess by the user at stage 2. The idea is to analyze how the wrong selection of a parametric model affects our approaches. It is worth noting that even the NP approach is expected to be somehow affected by this error, through the procedure of bandwidth derivation.

Table 3.2 shows the mean values of the evaluated ISE for 100 independent data sets. The errors associated with the P approaches, WLS and REML fitting criteria, are denoted by $ISE_{P_{wls}}$ and $ISE_{P_{reml}}$ respectively, and the errors associated with the NP approach are denoted by ISE_{NP}. Two different sample sizes, $n = 50$ and $n = 200$, were considered.

The NP approach is, in general, preferable, as it provides the smallest mean of ISE values in 55.5% of the cases considered for samples of size $n = 50$ and 100% for $n = 200$. More precisely, the results achieved for the NP approach exceed, at least 5%, those obtained for the second best approach in 44.4% and 77.8% of the total cases for $n = 50$ and $n = 200$, respectively.

		n = 50			**n = 200**		
Theoretical model		Parametric model EXP	SPH	WAV	Parametric model EXP	SPH	WAV
E	Mean($\text{ISE}_{P_{wls}}$)	0.61	**0.58**	0.79	0.21	0.20	0.38
X	Mean($\text{ISE}_{P_{reml}}$)	**0.54**	0.63	1.00	0.37	0.77	1.42 *
P	Mean(ISE_{NP})	0.61	0.63	**0.74**	**0.19**	**0.19**	**0.22**
S	Mean($\text{ISE}_{P_{wls}}$)	**0.96**	**0.85**	0.95	0.34	0.26	0.44
P	Mean($\text{ISE}_{P_{reml}}$)	1.10	1.10	1.12	0.63	0.69	1.60 *
H	Mean(ISE_{NP})	1.14	0.87	**0.88**	**0.32**	**0.25**	**0.24**
W	Mean($\text{ISE}_{P_{wls}}$)	0.84	0.99	1.35	0.89	0.68	0.70
A	Mean($\text{ISE}_{P_{reml}}$)	2.22	2.98	N/A	3.31	4.77	N/A
V	Mean(ISE_{NP})	**0.62**	**0.95**	**1.20**	**0.62**	**0.67**	**0.55**

Table 3.2: Mean values of the obtained ISE for the three approaches (P_{wls}, P_{reml} and NP) chosen to achieve a valid $\widehat{\gamma}(u)$. Data simulated with three theoretical models. Total number of replicas is 100 and each sample size is either 50 or 200. For each combination of one theoretical and one parametric model, bold identifies the lowest mean when comparing the three approaches. * For these two cases about 80 replicas were used, as for the remaining replicas the variance-covariance matrix was non-positive definite, not allowing the Cholesky factorization.



a) Sample size $n = 50$ b) Sample size $n = 200$

Figure 3.4: Boxplot of the evaluated ISE from the three approaches (P_{wls}, P_{reml} and NP) chosen to achieve a valid $\widehat{\gamma}(u)$. Data was simulated with exponential, spherical and wave models, but the exponential model was estimated.

The P_{wls} approach seems competitive with NP (i.e. not more than 5% inferior or even superior) in 44.4% of the observed cases for samples of size $n = 50$ and 22.2% for $n = 200$. As regards the P_{reml} approach, it should be avoided for larger samples, as well as when the wave model is involved on the procedure for valid $\widehat{\gamma}(u)$ achievement. The P_{reml} approach presents the best behaviour when the exponential model is correctly specified and $n = 50$.

The boxplot in Figure 3.4 shows more detailed information about previous ISE values, for one particular situation: the exponential curve was elected as the parametric model. This illustrates a likely situation as this family is one of the most popular, making it a strong candidate for election at stage 2. This boxplot contains three different groups of boxes: the first one, labelled EXP-EXP, stands for data simulated with an exponential model, whereas the second, SPH-EXP, and third, WAV-EXP, represent two cases of wrong specification, as data was simulated with a spherical and wave model, respectively.

In this boxplot, the NP approach shows the lower dispersion, measured in terms of the interquartile range, even when the median value of its evaluated ISE is worst. Another interesting conclusion is that the larger median values and the larger interquartile ranges are normally associated to the smaller sample size, i.e. $n = 50$. The exception is the P_{reml} approach, in the specific case of the wave model, as its median value is degraded by a large sample size.

From the boxplot, it is also evident that the NP approach is the preferred choice in the presence of the wave model. Otherwise, one of the two P approaches might be acceptable.

Estimation of main variogram features

The last comparison of estimates included in our simulation study involves important features typically associated with the variogram function, nugget, sill and range, which were introduced in Sections 2.2 and 2.3.

As stated in Ploner and Dutter (2000), with the spectral representation of the variogram, these features are no longer easily available. This problem can be ameliorated by providing a smooth curve as a variogram estimator. In this sense, some additional constraints on derivatives can be imposed to avoid spurious oscillations of the variogram (issue first raised by Hall, Fisher and Hoffmann 1994). Moreover, Genton and Gorsich (2002) presents a discretization of this spectral representation using Fourier−Bessel matrices, providing smooth and positive definite nonparametric estimators in the continuum.

In case of sill estimation (with null nugget), Cherry (1997) exploits the technique of Shapiro and Botha (1991), where $c(0) = \int_0^\infty dF(t)$ is approximated by $\widehat{c}(0) = \sum_{j=1}^{m} \widehat{p}_j$. As these estimates tended to be biased and highly variable, they propose a penalized fitting algorithm. With W as an identity matrix in (3.5), the problem under consideration is to minimize $\|M\mathbf{p} - \widehat{\gamma}\|$ subject to the constraint $\mathbf{p} \geq 0$. An additional constraint on the sill estimate can be incorporated

by minimizing

$$\left\| \begin{pmatrix} M \\ \sqrt{\lambda}\mathbf{1_p}^t \end{pmatrix} \mathbf{p} - \begin{pmatrix} \widehat{\gamma} \\ 0 \end{pmatrix} \right\|$$

where $\mathbf{1_p}$ is a $p \times 1$ vector of ones, and λ is a non-negative scalar penalty term. Some practical ways of selecting λ are discussed.

An alternative proposal is based on analysing the behaviour of the derivative of $\hat{\gamma}(u)$ to acquire an estimate of the sill approximation and, then, the corresponding estimated range. Table 3.3 summarizes the median values (P_{50}) and the mean square errors (MSE) of our estimates, comparing the outcome results from P_{wls}, P_{reml} and NP approaches. The correct specification of the theoretical model was always considered.

The nugget effect's estimator is given by $\widehat{\theta}_0$, for the P approaches, and by $\widehat{\gamma}(0)$, otherwise. These estimates should be compared with the theoretical value of 0.25. In respect to the remaining features of $\widehat{\gamma}(u)$, according to our empirical method, the derivative of $\widehat{\gamma}(u)$ is then used.

Under a wave model, we compare the global maximum of $\widehat{\gamma}(u)$ obtained from the NP approach, against $\widehat{\theta}_0 + (1 - (\sqrt{2}\pi)^{-1}\sin(\sqrt{2}\pi))\ \widehat{\theta}_1$ obtained from the P approaches. The corresponding range is defined as $\sqrt{2}\pi\widehat{\theta}_2$. Under the exponential and the spherical models, the range's estimators are $3\widehat{\theta}_2$ and $\widehat{\theta}_2$, respectively. All three models share a theoretical range of 0.5. For these last models, the estimated sill approximation is specified as the maximum of $\widehat{\gamma}(u)$ or 95% of this value, when considering a finite number of lags. The P approaches use $\widehat{\theta}_0 + 0.95\widehat{\theta}_1$, for the exponential model, and $\widehat{\theta}_0 + \widehat{\theta}_1$, for the spherical one.

In terms of MSE values, the NP approach offers the best estimators for the variogram's features, as it provides the lowest values in 77.8% of the total evaluated MSE's. The P_{wls} approach always presents the worst MSE values. In terms of median values, however, the best results are not necessarily associated with the NP approach. More precisely, the nugget effect seems to be under-estimated when a spherical or a wave model is used. In addition, this same approach seems to

			$\widehat{Nugget}$		$\widehat{SillApprox}$		$\widehat{RangeApprox}$	
Model	n	Approach	P_{50}	MSE	P_{50}	MSE	P_{50}	MSE
EXP	50	P_{wls}	0.31	0.71	5.58	7.74	0.57	0.98
		P_{reml}	*0.30*	0.29	5.15	4.36	*0.55*	0.20
		NP	0.16	**0.06**	*5.15*	**3.93**	0.39	**0.02**
	200	P_{wls}	0.40	0.58	*5.08*	8.76	*0.49*	1.15
		P_{reml}	0.20	0.03	4.74	**2.60**	0.47	0.06
		NP	*0.20*	**0.02**	7.87	9.31	0.52	**0.01**
SPH	50	P_{wls}	0.23	0.29	5.65	8.46	*0.52*	0.12
		P_{reml}	*0.23*	**0.07**	5.38	7.81	0.54	0.07
		NP	0.00	0.23	*5.30*	**5.70**	0.44	**0.01**
	200	P_{wls}	0.60	0.34	*5.12*	3.33	0.38	0.05
		P_{reml}	*0.32*	**0.04**	4.98	**3.27**	0.33	0.04
		NP	0.17	0.04	7.40	8.41	*0.51*	**0.01**
WAV	50	P_{wls}	*0.37*	0.15	6.64	7.10	*0.49*	0.02
		NP	0.00	**0.05**	*6.30*	**5.14**	0.48	**0.01**
	200	P_{wls}	*0.35*	0.46	*6.43*	8.67	*0.50*	0.03
		NP	0.08	**0.03**	7.24	**6.28**	0.49	**0.01**

Table 3.3: Summary of main features of $\widehat{\gamma}$: nugget, sill approximation and corresponding range. Data simulated with 3 theoretical models. 100 replications and each sample size is 50 or 200. Bold identifies the lowest MSE when comparing chosen approaches, while italic identifies the P_{50} value closest to the theoretical value.

over-estimate the sill approximation when sample size is equal to 200. An aspect also worth mentioning is that, overall, the performance of each estimator improves as sample size increases.

Computational costs

A final remark about the numerical study is related to the computational cost of the three approaches chosen to achieve a valid estimator. The CPU execution times were recorded for each sample, without considering data simulation but just the time needed to implement all existing stages. The results are summarized next in Table 3.4.

	n=50	n=200
P_{wls}	1.3 s	1.5 s
P_{reml}	3.0 s	$\approx$ 30 s
NP	$\approx$ 30 s	$\approx$ 30 s

Table 3.4: Summary of the computational costs of the three approaches chosen for our comparison study.

The lowest computational cost was achieved by the P_{wls} approach, being around 1.3 and 1.5 seconds for $n = 50$ and $n = 200$, respectively. The cost for the P_{reml} approach was around 3 seconds for $n = 50$, this is at least 10 to 15 times greater for $n = 200$. With respect to the NP approach, we have registered CPU times from 27 to 36 seconds for $n = 50$, being the lowest values associated to the spherical data and the greatest to the exponential data. These computational costs have only shown a slight increase when we moved to sample sizes of $n = 200$. Bear in mind that the heavy costs obtained for the NP approach are usually justified by the optimal bandwidth derivation.

3.3.3 Closing remarks

The problem of estimation of the variogram can be analyzed in practice from several points of view. If the aim is just to obtain an approximation of the dependence structure of the spatial data, then the classical and the Nadaraya-Watson kernel provide good estimators that behave better than the robust estimator proposed by Cressie and Hawkins, using as a term of comparison the values estimated for the median and interquartile range of the ISE; however, the robust estimator reduces the range of variation of the ISE.

If we focus on the problem of spatial prediction, we modify the variogram estimators to obtain valid variograms; otherwise, negative mean squared prediction errors may be achieved. From the different alternatives discussed, the valid kernel estimation (referred to as the NP approach) has the best performance for large sample sizes in terms of the values estimated for the ISE, regardless of the parametric model that is considered. In this respect, it is surprising that fitting the correct parametric family does not produce a better fit than the non-parametric method. The misspecification of the parametric family has a second order effect on the kernel estimator, since it affects estimation of values associated to the bandwidth parameter. On the other hand, when considering typical features associated with the variogram (nugget, sill and range), we conclude that the valid kernel estimation provides the lowest MSE values, although the P approaches prove competitive in the estimation of the corresponding median values.

In general, the results presented here show that a valid variogram estimator obtained from a NP approach is a good alternative to those valid estimators obtained from the classic parametric approaches. The NP approach has the additional advantage of avoiding problems associated with using the wrong parametric model, which can occur in many conventional approaches. These advantages become even more evident if sample data underlies an atypical spatial dependence, like the one from the wave model. However, one must be prepared to pay an extra computational cost over the cost associated to a simple P approach like the one that fits a

valid model to some empirical estimations through the WLS criterion.

The P approach using REML as fitting criterion is only able to compete with the other methods in the presence of small datasets and, simultaneously, observed data does not follow a wave-type structure.

Chapter 4

Assessing the effect of clustered and biased multi-stage sampling

4.1 Introduction

The geostatistical methods introduced in Chapter 3 rely on the expected assumption that the sampling design for locations $\mathbf{x_i}$, $i = 1, ..., n$ is deterministic or it is stochastic but independent of the data process, and all analyses are carried out conditionally on $\mathbf{x_i}$ (Diggle, Ribeiro Jr and Christensen 2003). It is then assumed that the sampling points have been chosen independently of the values of the spatial variable. However, dependencies can occur due to the adopted sampling method, such as the favored selection of specific areas that are believed critical (e.g. maximum values search).

Schlather, Ribeiro Jr and Diggle (2004) proposes methods to detect the dependence between marks and locations of marked point processes. As described in Mateu and Ribeiro Jr (1999), the random field (that we have been studying) and the marked point process are two type of spatial processes such that:

- The former is defined in every point of the observed region, and the sample positions can be determined by the scientist himself (example of deterministic

sampling design);

- For the latter, the locations are always given by a stochastic point process, and interactions among the locations and the marks are normally expected. Otherwise, one has the so called *random field model* (marked point process becomes a special class of a random field).

If the data are consistent with a random field model, the point pattern and the marks can be analysed separately using standard techniques for point processes (e.g. Ripley 1981 and Diggle 2003) and for geostatistical data (e.g. many references in Chapter 3). Therefore, this analysis is greatly simplified.

The examination of second-order characteristics, like the variogram, of a spatial process should consider if data come from a random field or a genuine marked point process. Example of references concerned with this subject are Walder and Stoyan (1996), Mateu and Ribeiro Jr (1999) and Schlather (2002).

Schlather et al. (2004) indicates next two likely situations for point and data processes being dependent, and subsequent failure of this important geostatistics assumption. Firstly, if the dependency is an intrinsic property of the data themselves, for example the relative positions of trees impact on their size due to their competition for light and nutrient. This is the case of genuine marked point processes. Alternatively, this dependency can be justified by a prior scientific knowledge of the spatial variable of interest, for example of the expected local level of contamination in air pollution. This can lead to the gathering of samples in areas with atypical values.

Our work concerns the problems resulting from the second situation, that we think of major importance in geostatistics because of its high likelihood of occurrence on actual field measurements, and often either ignored or addressed by generic techniques like declustering ones (e.g. Goovaerts 1997 and Isaaks and Srivastava 1989).

In this Chapter, we are motivated by the application example of the radioactivity data from Rongelap island, where a two-stage data collection was used, leading to the presence of clustered data. So that we start by restricting our attention to multi-stage samples, aiming to assess the presence of multi-stage dependence, or also referred to as *sequential dependence*, where the choice of sampling points is driven by previous measurements.

We propose some data exploratory methods which are intended to detect biased multi-stage collection of spatial data. We investigate corrector models that aim to minimize the impact on variogram estimation due to the adoption of the type of non-standard sampling designs just described. Moreover, we assess the effect of these methods on the Rongelap data.

4.2 Assessing through simulation

We start by using simulated data to develop and study our diagnostic tools for data analysis. It is well known that simulation allows a level of knowledge and control that leads to more robust and defendable solutions. Using simulated data sets, where the characteristics of the data and the sampling designs are controlled and varied, will help the research of the technique's potential, and to assess its performance in specific situations. We can gain insight about what happens when assumptions are violated since the true model is known.

4.2.1 Sample generation algorithm

Typically, when one carries out some study of geostatistical data, the sample locations are uniformly spread over the observation region. Suppose now that one wishes to proceed with a multi-stage collection of data. If the goal is to better characterize the spatial variability for short distances, then one solution is to include some clusters of locations into later stages. Alternatively, suppose the goal is, as exemplified before, to pursue the maximum values of the spatial variable

of interest, then the complete sample data set is expected to be mainly represented by large data values.

All previous situations may condition the sampling design. In our simulation study, we shall then consider four distinct sampling designs:

- complete spatial randomness (CSR);
- just clustered;
- biased but non-clustered;
- and, finally, biased and clustered.

Furthermore, we start with a two-stage approach for sampling collection, with the second stage potentially influenced by the first. We consider spatial locations $\mathbf{x}$ within the unit square $[0,1] \times [0,1]$. Data sets are generated with Gaussian data, $Z(\mathbf{x})$, using some chosen variogram model.

We prefer to propose a more generic algorithm where K clusters can be generated, each one inside a sub-region R_k. For example, if one wishes to produce a biased sample with just one cluster, this can be done by restricting the sampling points from stage 2 to R_1 and around[1] the maximum of measurements from stage 1. The total sample size will be n, with n_1 from stage 1 and n_2 from stage 2. The algorithm may be summarized as follows

1. Sample n_1 points $\mathbf{x_i}$ at random on $[0,1] \times [0,1]$;
2. Generate $\mathbf{Z} = (Z(\mathbf{x_1}), ..., Z(\mathbf{x_{n_1}})) \sim MVN$;
3. For $k = 1, ..., K$
   (a) If biased=TRUE then select $Z(\mathbf{x_{m,k}}) = \max_i \{Z(\mathbf{x_i}) | \mathbf{x_i} \in R_k\}$
   else select $Z(\mathbf{x_{m,k}}) = \texttt{random}_i \{Z(\mathbf{x_i}) | \mathbf{x_i} \in R_k\}$;
   (b) Sample $n_{2,k}$ points at random on $[\mathbf{x_{m,k}} - \delta \ , \ \mathbf{x_{m,k}} + \delta]^2$;

[1] A small square with side length 2δ will be considered. Points must be simultaneously inside of the unit square.

4. Consider $n_2 = \sum_{k=1}^{K} n_{2,k}$;

5. Generate $\mathbf{Z}^* = (Z(\mathbf{x_{n_1+1}}), ..., Z(\mathbf{x_{n_1+n_2}}))$ where

$$\mathbf{Z}^* \mid \mathbf{Z} \sim MVN\left(\Sigma_{12}^T \Sigma_{11}^{-1} \mathbf{Z} , \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12} \right)$$

and $\Sigma_{22} = \text{var}\{\mathbf{Z}^*\}$, $\Sigma_{11} = \text{var}\{\mathbf{Z}\}$, $\Sigma_{12} = \text{cov}\{\mathbf{Z}, \mathbf{Z}^*\}$;

The conditional distribution from step 5 was derived from the joint distribution using properties of the multivariate Gaussian distribution (Anderson 1984).

Under the adopted notation, the sampled points from stage 1 share a common time label t_0, while those generated at stage 2 are assigned a time label t_1. Bear in mind that a completely random sample can be obtained avoiding stage 2, i.e. $n_2 = 0$, or generating the n_2 points uniformly spread over all unit square. Moreover, the cluster effect tends to disappear for a large K.

The two-stage approach reflects more directly the sampling design defined for Rongelap data. The previous algorithm can be easily extended to more than two stages, even though our experience confirmed that similar results are obtained. We also tried a specific multi-stage approach, hereafter termed *serial sampling*, according to which all points and corresponding data values after stage 1 are generated one at a time. So, we will have $\{(\mathbf{x_i}, Z(\mathbf{x_i})) : i = 1, ..., n_1\}$ from stage 1, $(\mathbf{x_{n_1+1}}, Z(\mathbf{x_{n_1+1}}))$ from stage 2, $(\mathbf{x_{n_1+2}}, Z(\mathbf{x_{n_1+2}}))$ from stage 3, ..., and $(\mathbf{x_n}, Z(\mathbf{x_n}))$ from last stage.

Figure 4.1 shows an example of spatial locations derived by our serial sampling algorithm. The latest points, associated to time labels t_i where $i > 0$, are conditioned by the maximum of previous measurements. This algorithm also tends to originate a cluster of biased data, within a neighborhood of length approximately equal to 2δ.

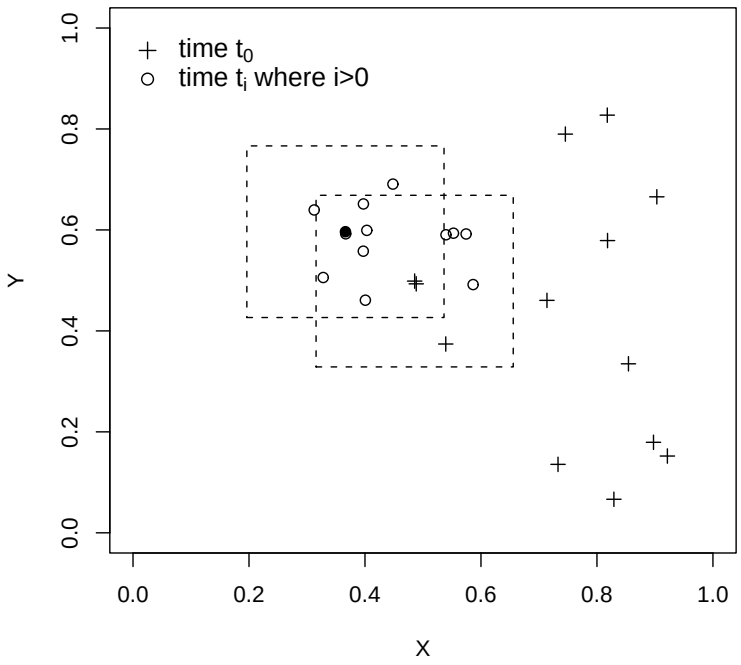


Figure 4.1: Serial sampling algorithm. The center point of each square identifies the maximum value at each stage.

4.2.2 Impact on variogram estimation

We now want to analyse the impact of clustered and biased multi-stage sampling on variogram estimation. We consider two popular empirical estimators, the classical one from Matheron (1962) and the Nadaraya-Watson kernel estimator, given in (3.1) and (3.3), respectively.

Figure 4.2 illustrates an example of degraded behaviour of these estimators under this type of non-standard sample. The resulting estimations are plotted against the theoretical curve of the variogram model chosen for sample generation. In case A, the data set was obtained by random sampling, whereas in case B a serial sampling was considered with 70% of biased clustering[2]. The results of case B are, at least partially, justified by the sample locations not being sufficiently

[2]In our serial sampling algorithm, we have specified a total $n = 200$ and $n_1 = 60$ for stage 1.

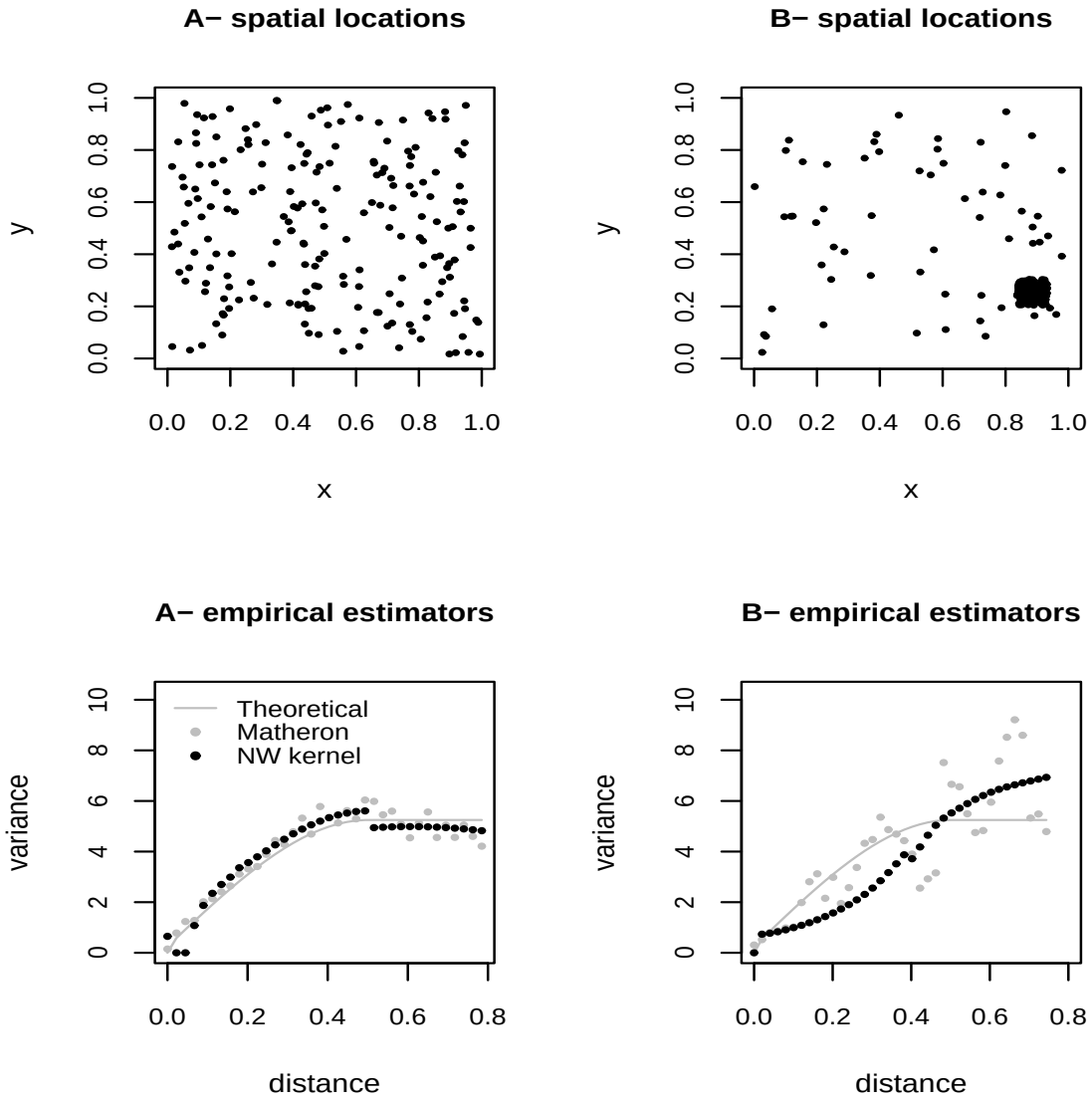


Figure 4.2: Behaviour of $\widehat{\gamma}$ under random sampling (case A) against clustered and biased sequential sampling (case B).

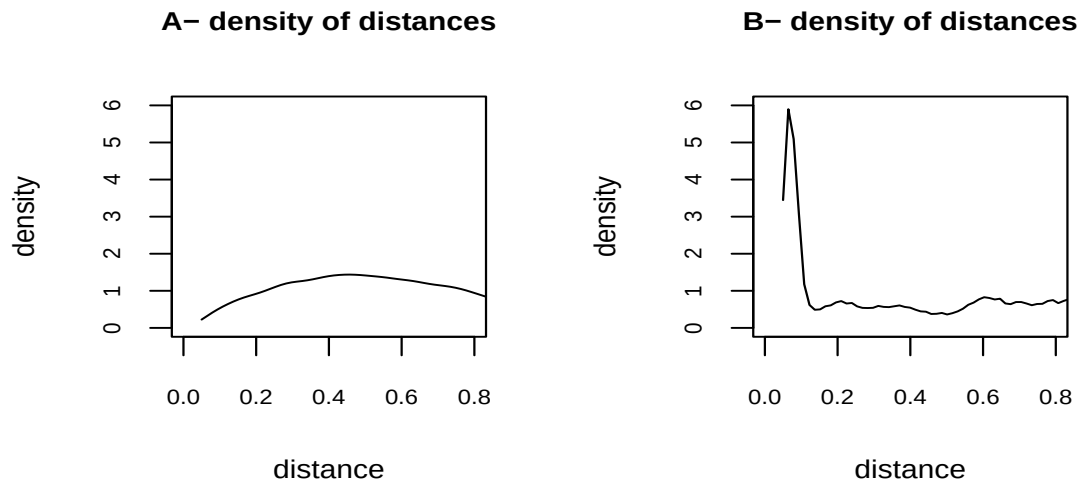


Figure 4.3: Density of the distances between sample locations for cases A and B of Figure 4.2.

spatially representative of the overall data, as they are irregularly distributed over the observation region. Additionally, note that to better exemplify the impact of biased clustering, we chose as case B an *extreme* situation, where the density of the distances between locations presents a strong mode for small lags, damaging the sample average in expressions (3.1) and (3.3) (see Figure 4.3).

Again, a more conclusive analysis of these estimators' behaviour must be based on results from several independent cases. We then generate a total of 100 independent data sets and, for each one, derive the integrated square error[3] (ISE) between the estimator and the theoretical variogram. A Matérn model with order $\kappa = 1$, a range $\phi = 0.2$ and a partial sill equals to 2.25 was chosen to model the spatial dependency.

In Table 4.1, we summarize the mean values of the ISE, considering the 4 possible combinations of biased two-stage sampling and clustered sampling. To generate a biased but not strongly clustered sample, we split the sample grid

[3]The ISE, given in (3.6), was approximated numerically through the trapezoid rule.

Sampling Design	$\widehat{\gamma}(u)$	$u \leq 0.6$	$u \leq 0.3$	$u \leq 0.2$	$u \leq 0.1$
Random	Matheron	0.608	0.239	0.123	0.032
	NW kernel	0.580	0.275	0.161	0.043
	RobClust	0.598	0.234	0.121	0.031
	NewRobClust	0.571	0.265	0.154	0.042
	Pooled	0.574	0.264	0.153	0.041
Just Clustered	Matheron	0.957	0.453	0.315	0.070
	NW kernel	0.567	0.248	0.159	0.063
	RobClust	0.645	0.298	0.188	0.058
	NewRobClust	0.512	0.260	0.164	0.059
	Pooled	0.472	0.238	0.152	0.059
Just Biased	Matheron	0.685	0.350	0.246	0.111
	NW kernel	0.507	0.268	0.164	0.048
	RobClust	0.680	0.346	0.243	0.111
	NewRobClust	0.496	0.255	0.154	0.044
	Pooled	0.352	0.177	0.105	0.027
Biased and Clustered	Matheron	2.989	0.766	0.336	0.090
	NW kernel	1.882	0.338	0.176	0.071
	RobClust	1.308	0.430	0.226	0.069
	NewRobClust	1.102	0.402	0.218	0.068
	Pooled	0.415	0.212	0.142	0.061

Table 4.1: Comparison of five distinct variogram estimators, through the mean values of the evaluated ISE. Four distinct sampling designs were considered, from simultaneously biased and clustered sample to a completely random sample. Total of replicas equals to 100 and each replica total sample size equals to 200.

into 25 sub-areas. We start to generate randomly 75 values and locations in the total area and then generate 5 more points clustered around the maximum of previous measurements in each sub-area. A final sample size of 200 is obtained. For standardization reasons, in the remaining cases, n_1 equals to 75 and n_2 equals to 125 is also chosen.

The results from Table 4.1, for Matheron and NW kernel estimators, confirm the poor performance found previously of both estimators under biased clustering. Note that Table 4.1 also includes the results of three other variogram estimators, *RobClust*, *NewRobClust* and *Pooled*, that will be introduced in later Sections of this Chapter. We now want to restrict our attention to the Matheron and NW kernel estimators.

One may observe that there is a larger degradation for the Matheron's estimator. In fact, when all lags less than or equal to 0.6 are considered, this estimator and the kernel one grow worse 4.9 and 3.2 times, respectively.

The worst results normally associated to Matheron's estimator tend to be not so obvious for smaller lags. This should be an indirect consequence of the typical less satisfactory behaviour of kernel estimators in boundaries. In any case, with respect to larger lags under *just biased* or *just clustered* sampling designs, note that the NW kernel estimator performs quite well.

Finally, from Table 4.1, the clustering issue seems to have a larger impact on variogram estimates than the sequential dependence issue.

This same experience was repeated for a multi-stage approach, applying our serial sampling algorithm. The results, with a very similar interpretation, can be observed in boxplots from Section 4.5.3.

4.3 Data exploratory methods

We have shown that the non-standard sampling designs described in Section 4.2.1 are responsible for a more difficult estimation of the spatial dependency structure. This suggests the need for detecting biased multi-stage sampling and, ideally, for

correcting solutions.

Still using simulated data, we now investigate data exploratory tools to reveal hidden dependency patterns in a given sample data set. More precisely, the practical part of this research is concerned with the exploratory analysis of sampled data in order to understand if it is reasonable to assume dependency between data values and locations and if this dependency is indeed sequential.

4.3.1 Detection of dependence

In a context of marked point processes, Schlather et al. (2004) investigates marks (M) and locations (L) interactions, by introducing two functions of the inter-point distance u, under the assumption of stationary and isotropy. These functions denote the conditional expectation and the conditional variance of a mark, given that there is a further point of the process a distance u away. Writing $\Phi = \bigcup_i \mathbf{x_i}$ for the corresponding unmarked point process, the functions may be represented, respectively, in this way:

- $E(u) = \mathrm{E}\left[Z(\mathbf{x}) \mid \mathbf{x}, \mathbf{x}' \in \Phi,\ ||\mathbf{x} - \mathbf{x}'|| = u\right]$
- $V(u) = \mathrm{E}\left[(Z(\mathbf{x}) - E(u))^2 \mid \mathbf{x}, \mathbf{x}' \in \Phi, \|\mathbf{x} - \mathbf{x}'\| = u\right]$

Schlather et al. (2004) presents tests based on E and V for the hypothesis of dependency between the values of the marks and their locations. If this dependence does not occur, one has $[M, L] = [M][L]$ with [.] meaning "the distribution of", and classical geostatistical methods can then be applied to model $[M]$. Otherwise, some open issues are the classification of the existing dependency and how to model the conditional distribution $[L|M]$. One of the purposes of this Chapter is to address the former issue, while the latter will be addressed in Chapter 6.

4.3.2 Detection of sequential dependence

In order to decide if existing dependency is sequential, we propose a new version for the conditional expectation function, denoted by $E_{seq}(u)$, restricted to the *latest*

values of the spatial variable. Under the notation we have been using, this new function can be defined as:

- $E_{seq}(u) = \mathrm{E}\left[Z(\mathbf{x}) \mid \mathbf{x}, \mathbf{x}' \in D,\ ||\mathbf{x} - \mathbf{x}'|| = u,\ t(\mathbf{x}) > t(\mathbf{x}')\right]$ (4.1)

where $t(.)$ identifies the stage when data were collected, i.e. a time label. It is then assumed that the analyst is aware of how, or if one prefers when, the collection of the sample data occurred, making time labels available. Other functions that can be defined are:

- $V_{seq}(u)$, the corresponding centred second order moment
- $V^*_{seq}(u)$, second order moment about $E_{seq}(u)$ using both data values
- $E(Z)$, the overall expectation (not dependent on inter-point distance)

Estimators for E_{seq}, V_{seq} and V^*_{seq}

Both conditional expectation functions, $E(u)$ and $E_{seq}(u)$, can be approximated through a sample average, but the second considers a sub set of the total data values considered for the first. Considering a tolerance region for lag u in (4.1), i.e. $||\mathbf{x} - \mathbf{x}'|| \simeq u$, our estimator can be defined as

$$N_u^{-1} \sum_{\substack{|\|\mathbf{x_i}-\mathbf{x_j}\|-u| \le \frac{\varepsilon}{2} \\ t(\mathbf{x_i}) > t(\mathbf{x_j})}} f(Z(\mathbf{x_i}), Z(\mathbf{x_j})) \tag{4.2}$$

where $\varepsilon > 0$ is a fixed bin-width and N_u is the number of pairs $(\mathbf{x_i}, \mathbf{x_j})$ for which $|\|\mathbf{x_i} - \mathbf{x_j}\| - u| \le \frac{\varepsilon}{2}$ and $t(\mathbf{x_i}) > t(\mathbf{x_j})$. Additionally, function $f(.)$ is as follows.

- For $\widehat{E}_{seq}(u)$, $f(Z(\mathbf{x_1}), Z(\mathbf{x_2})) = Z(\mathbf{x_1})$.
- For $\widehat{V}_{seq}(u)$, $f(Z(\mathbf{x_1}), Z(\mathbf{x_2})) = (Z(\mathbf{x_1}) - \widehat{E}_{seq}(u))^2$.
- For $\widehat{V}^*_{seq}(u)$, $f(Z(\mathbf{x_1}), Z(\mathbf{x_2})) = \frac{1}{2}\{(Z(\mathbf{x_1}) - \widehat{E}_{seq}(u))^2 + (Z(\mathbf{x_2}) - \widehat{E}_{seq}(u))^2\}$.

Let us compare estimations for $E(u)$ and $E_{seq}(u)$ through a simple example. For a given inter-point distance u_1, suppose one has three pairs of variables

$[Z(\mathbf{x_1}), Z(\mathbf{x_6})]$, $[Z(\mathbf{x_2}), Z(\mathbf{x_3})]$ and $[Z(\mathbf{x_7}), Z(\mathbf{x_9})]$ and that the time labels are t_0, t_3, t_0, t_0, t_4 and t_6, respectively. Then, the estimated value for $E(u_1)$ will be a direct average of the total six data values, while the estimated value for $E_{seq}(u_1)$ just considers $Z(\mathbf{x_6})$ and $Z(\mathbf{x_9})$. The data values $Z(\mathbf{x_2})$ and $Z(\mathbf{x_3})$ are disregarded, as they share the same time label.

According to this example, an estimate for $V_{seq}(u_1)$ would be obtained using terms $[Z(\mathbf{x_6}) - \widehat{E}_{seq}(u_1)]^2$ and $[Z(\mathbf{x_9}) - \widehat{E}_{seq}(u_1)]^2$, whereas the one for $V^*_{seq}(u_1)$ would use the two additional terms $[Z(\mathbf{x_1}) - \widehat{E}_{seq}(u_1)]^2$ and $[Z(\mathbf{x_7}) - \widehat{E}_{seq}(u_1)]^2$.

4.3.3 Impact of sample designs on E_{seq}, V_{seq} and V^*_{seq}

We investigate the influence of sequential biased and clustered sampling on the behaviour of the conditional expectation functions. The simulation study considers 1000 independent data sets and the same features chosen for our previous study: same sampling designs and spatial dependency structure.

In Figures 4.4 and 4.5, we plot the mean of 1000 estimated conditional expectation functions, given by $\widehat{E}_{seq}(u) - \widehat{E}(u)$, $\widehat{E}_{seq}(u) - \widehat{E}(Z)$ and $\widehat{E}(u) - \widehat{E}(Z)$. The confidence intervals (CI) for the sampling distribution of differences constructed from 1000 samples were added to check the variability of these estimations.

We conclude that under the absence of a sequential biased sample, and just in this case, the difference functions $\widehat{E}_{seq}(u) - \widehat{E}(u)$ and $\widehat{E}_{seq}(u) - \widehat{E}(Z)$ are approximately zero. The corresponding CIs embrace the theoretical $E_{seq}(u) - E(u) = E_{seq}(u) - E(Z) = 0$. Otherwise, with or without the presence of strong clustering, our two difference functions are clearly non-zero, reflecting the existence of bias in *latest* data points (higher values in our simulation).

The plots illustrate the dependency pattern of a biased multi-stage collection of sample data

$$E_{seq}(u) - E(u) \neq 0. \tag{4.3}$$

Note that in the case of *just clustered*, the three difference functions share a

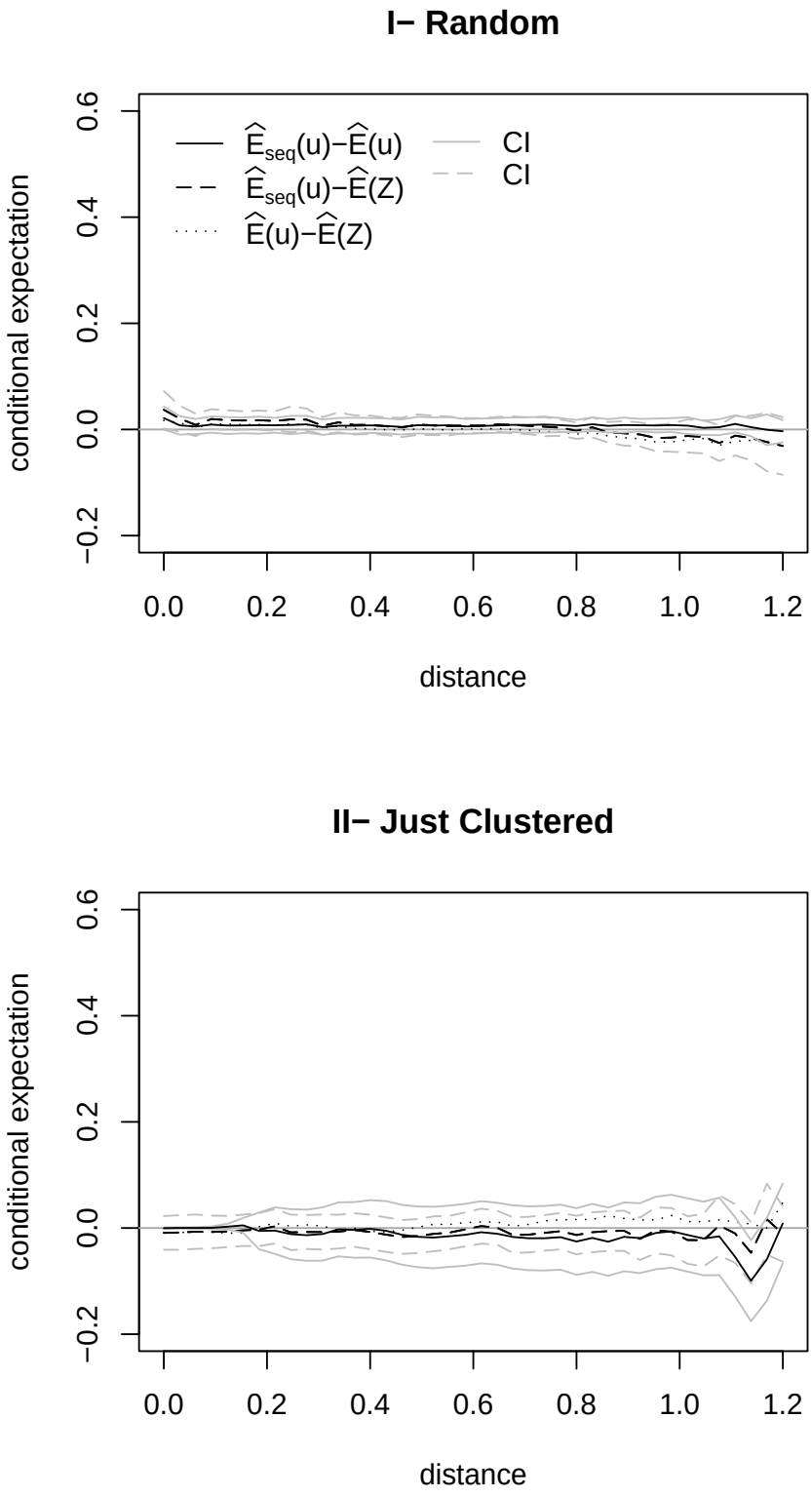


Figure 4.4: Part I - mean values of estimated conditional expectation functions. Total of replicas equals to 1000.

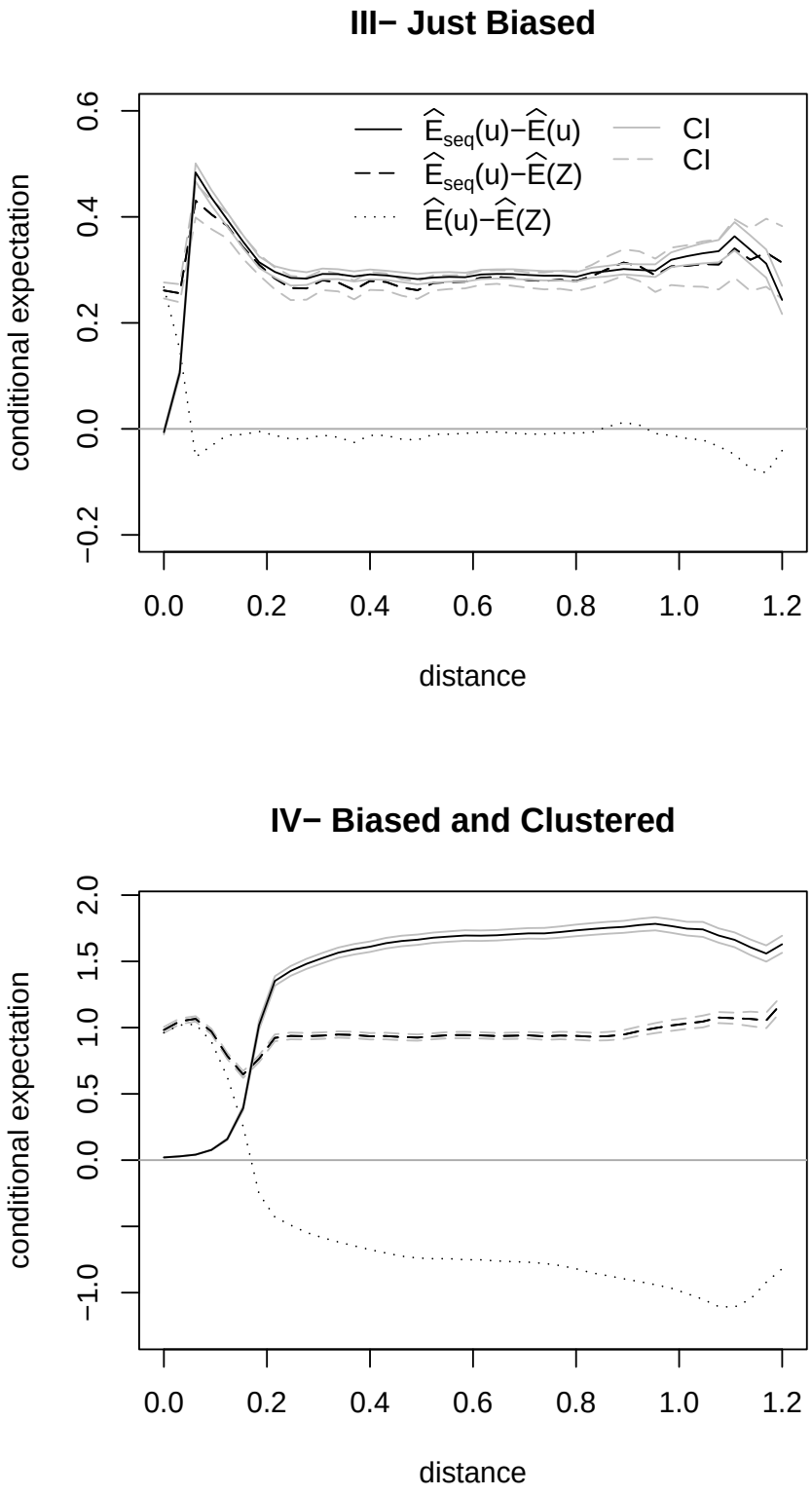


Figure 4.5: Part II - mean values of estimated conditional expectation functions. Total of replicas equals to 1000.

similar behaviour, as the choice of *latest* data points is randomly affected by previous measurements (not always the maximum). In this case, the CIs are slightly larger than in other cases, maybe because of the random selection of the cluster position.

The analysis of the corresponding estimated standard deviation functions[4], given by $\sqrt{\widehat{V}_{seq}(u)}$, $\sqrt{\widehat{V}^*_{seq}(u)}$ and $\sqrt{\widehat{V}(u)}$, was not so conclusive (see Figure 4.6). According to Schlather et al. (2004), we would expect an approximately constant variance when there is no dependence between data values and data locations, i.e. under *random* and *just clustered* sampling designs. However, this only happens when we estimate the second order moment about the theoretical overall expectation, the $E(Z)$ chosen for our simulation study. This means making $f(Z(\mathbf{x_1}), Z(\mathbf{x_2})) = (Z(\mathbf{x_1}) - E(Z))^2$ in the estimator defined in (4.2).

Actually, the main pattern found in the variance's plots is imposed by the clustering issue, responsible for smaller estimates of the variance for smaller lags[5]. Our function $\widehat{V}_{seq}(u)$ seems to under-estimate the theoretical variance, because it is restricted to data values of stage 2, typically with less variability. We decided then to focus on the conditional expectation functions.

4.4 Monte Carlo tests

The widely used Monte Carlo significance testing was originally proposed by Barnard (1963) and its basic idea is as follows. Suppose H_0 is the null hypothesis about the model which generates $Y = \{(\mathbf{x_i}, Z(\mathbf{x_i})) : i = 1, ..., n\}$, and r_1 is an observed value of a real valued statistic $R = h(Y)$, which has a distribution function F, possibly mathematically intractable. Moreover, suppose we agree to reject H_0 for a *large* value of r_1.

Hence, we can use pseudo-random numbers to simulate a random sample

[4]We decide not to plot the CIs, because they would not add new interpretation hints.

[5]Note that our simulation study involved a cluster of diameter approximately equal to 0.1.

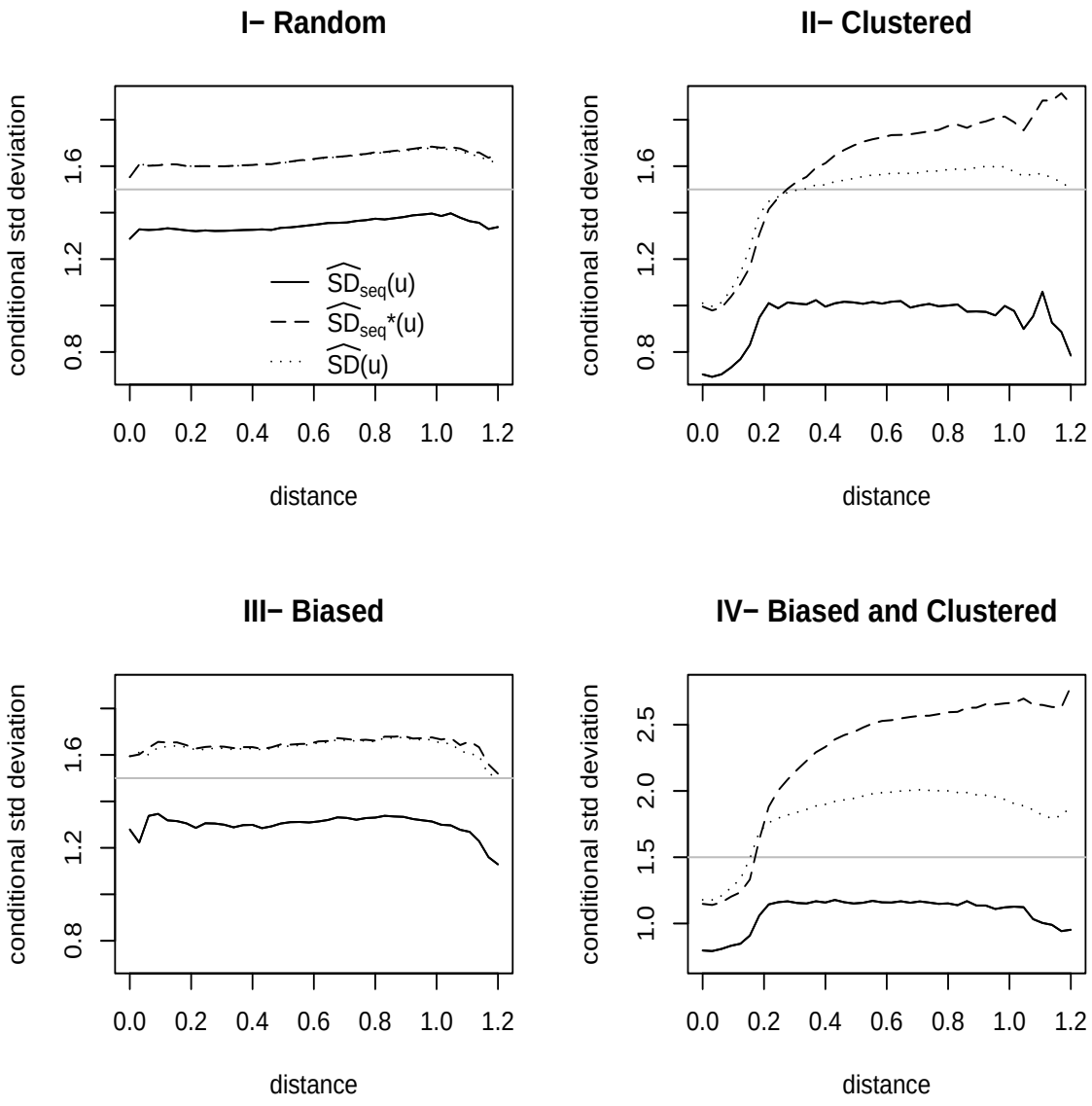


Figure 4.6: Mean values of estimated conditional standard deviation functions (theoretical stddev is displayed in grey). Total of replicas equals to 1000.

$r_2, ..., r_m$ of $m-1$ observations from distribution F and to construct a test by comparing these simulated values with r_1. If F is continuous and $k = 1 + \#\{j : j = 2, ..., m \; and \; r_1 > r_j\}$, then H_0 will be rejected at the k/m attained significance level, since the rank of r_1 is uniformly distributed on the integers $1, ..., m$ when H_0 is true. See Besag and Diggle (1977) for a general discussion of Monte Carlo tests. Note that the parametric bootstrap techniques work in a similar way to those described in here (see e.g. Gentle 2002 or Hall 1991).

In our work, we are interest in a test for the hypothesis that a given data set does not incorporate sequential biasing, so that we shall define

$$H_0 : \; E_{seq}(u) - E(u) = 0. \tag{4.4}$$

Under this hypothesis, the spatial process can be generated by sampling a random field Z at the given locations $\mathbf{x_i}$, $i = 1, ..., n$, with no sequential dependence. In this way, we can simulate $m-1$ further data sets under H_0, and define r_j to be a measure of discrepancy between $\widehat{E}^j_{seq}(u)$ and $\widehat{E}^j(u)$ over the whole range of u. For example, our test statistic can be given by the integrated squared difference

$$r_j = \int \{\widehat{E}^j_{seq}(u) - \widehat{E}^j(u)\}^2 du. \tag{4.5}$$

We can then proceed to a formal test based on the rank of r_1 amongst r_j, because under H_0 all ranking of r_1 are equiprobable. Bear in mind that m is rather smaller than might perhaps be expected, in contrast with the much larger sample which would be needed for accurate estimation of F, the distribution function of R. According to Hope (1968), for a one-sided test at the conventional 5% level of significance, $m = 100$ is suitable.

One may not wish to proceed directly to formal testing. A preliminary rough visual guide to address the problem being investigated can be provided by means of the well-known "simulation envelopes" plot. Testing involves comparing an

observed test statistic with samples from the model under consideration. Consequently, this visual approach is based directly on the variation in estimates obtained from data generated from the model. The maximum and minimum of the total $m-1$ independent simulations allow the definition of *upper* and *lower envelopes*. See Figure 4.7, for an example.

Diggle (2003) emphasises the use of such a plot as a visual aid to interpretation. Comparison of the observed[6] curve $\widehat{E}^1_{seq}(u) - \widehat{E}^1(u)$ with that expected from a random arrangement of $\widehat{E}^j_{seq}(u) - \widehat{E}^j(u)$, $j = 2, ..., m$ allows an assessment of the overall degree of coverage. If the observed curve lies between the two envelopes, this suggests the acceptance of hypothesis H_0 given in (4.4). If the observed curve exceeds the envelopes for some distances u, this is an initial and informal indication of the possibility of H_0 rejection. Anyway, in our case, we prefer to deepen analysis and to proceed with a formal Monte Carlo test.

In Sections 4.4.1 and 4.4.2, we take for granted some model assumptions. In Section 4.4.3, we examine an alternative non-parametric approach, applying some randomization test ideas.

4.4.1 Example of a simulated data set

We first emphasize the distinction between our proposal and the one described in Schlather et al. (2004), through a simple simulated example. We generate a sample data set $\{(\mathbf{x_i}, Z(\mathbf{x_i})) : i = 1, ..., n\}$, incorporating some sequential bias but not forming any obvious cluster of spatial locations. This is our *observed* data set.

We need now to simulate 99 further data sets under H_0 in (4.4). The idea is to fix the sample locations $\mathbf{x_i}$ and to generate Gaussian data on them. Bear in mind that this normally requires the estimation of the spatial dependency structure. The accuracy of this estimation may strongly impact the results of Monte Carlo testing, as we shall see in the remainder of this Chapter.

In the left panel of Figure 4.7, we plot the observed $\widehat{E}_{seq}(u) - \widehat{E}(u)$ against

[6]The one derived from the observed data set $\{(\mathbf{x_i}, Z(\mathbf{x_i})) : i = 1, ..., n\}$.

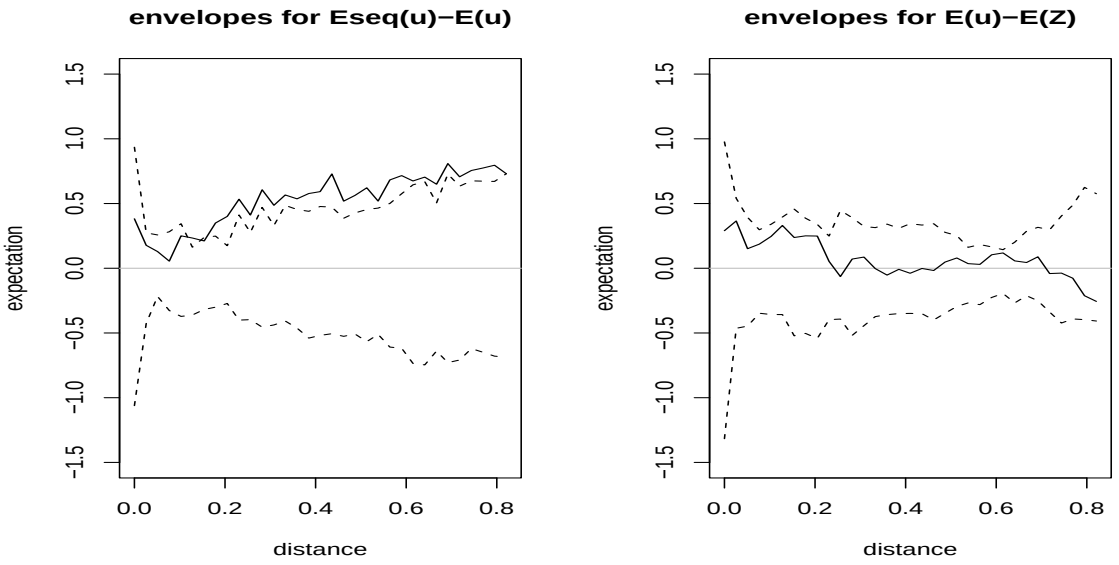


Figure 4.7: Simulation envelopes of $E_{seq}(u) - E(u)$ and $E(u) - E(Z)$ for a simulated data set, with sequential bias but not clustered: data (solid curve); upper and lower envelopes from 99 simulations of a random field (dashed curves).

the corresponding 99 simulations of a random field, which points to a possible rejection of H_0. This rejection was confirmed with a formal test based on (4.5) at the 5% level of significance. However, the simulation envelopes of $E(u) - E(Z)$ plotted in the right panel of Figure 4.7 could lead to the acceptance of H_0. It is an example of just one simulated data set, but it suggests some caution when using the conditional expectation functions to detect dependency and, whenever applicable, we should take advantage of the possibly available multistage information[7].

4.4.2 Rongelap island's data

We now describe the application of our methods to the data set from Rongelap island, whose sampling design has inspired part of the research work described in this Chapter.

This island is located in the Pacific Ocean approximately 4000 kilometres south-

[7]Information about when the collection of the sample data occurred, namely time labels.

west of Hawaii. The data were collected for the analysis of current levels of radioactivity contamination that resulted from a nuclear weapons testing programme during the 1950s. The scientific problem has been the estimation of the maximum level of radioactivity over the island, as part of a wider investigation to decide whether Rongelap can safely be resettled. See Diggle, Tawn and Moyeed (1998) or Diggle, Harper and Simon (1997) for more detail on these data.

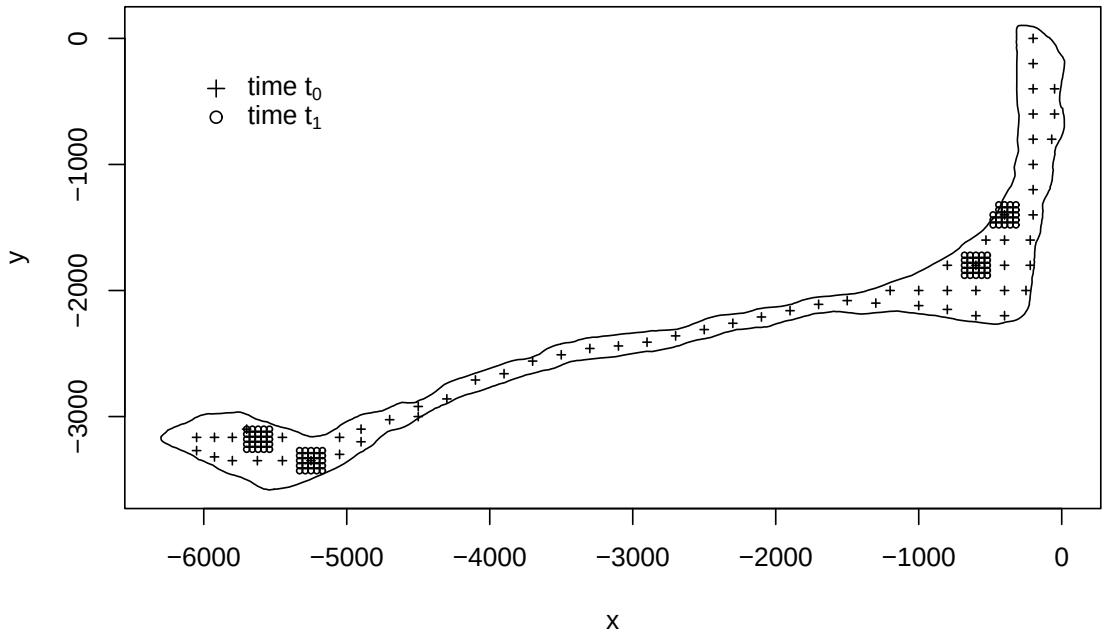


Figure 4.8: Rongelap's island: two-stage strategy of uniform and clustered samples.

The sampling design defined for data collection is illustrated in Figure 4.8. It started with a coarse grid of 63 locations and ended up with 98 additional measurements within four fine grids. These locations are identified by time label t_0 and t_1, respectively. As this process involved two-stage of uniform and clustered samples, we wonder about the impact on conclusions from a standard analysis that does not account for either of these features. To proceed our analysis, the methods are applied to transformed data with a constant variance as described next.

Our variables of interest from Rongelap data set are the spatial coordinates $\mathbf{x_i}$, the counts Y_i of radioactive emissions at each location, the length l_i of time over which the counts are recorded, and the stage in sampling measurements were made. The total sample size is $n = 161$. Note that the Y_i are treated as realizations of mutually independent Poisson random variables with expectations $l_i\lambda(\mathbf{x_i})$, where $\lambda(\mathbf{x})$ measures the local radioactivity at location $\mathbf{x}$. We chose the data transformation $Z_i = \sqrt{Y_i/l_i}$ to make the variability more consistent and more Gaussian[8].

It was found convenient to start with the maximum likelihood estimation of the spatial dependency structure. So, we consider a variogram estimator, obtained by using the coarse data and, derived through restricted maximum likelihood (REML). In Figure 4.9, we present the results of our Monte Carlo test. We generate 99 simulations of a random field over the total 157 distinct[9] locations of Rongelap's island. From this plot, we would not reject the null hypothesis, as confirmed through a formal test. So, we would tend to refuse the existence of sequential bias. However, when one replaces this variogram estimator for one of those proposed in Section 4.5, this tendency is not that clear.

4.4.3 Randomization tests

The previous approach requires model assumptions, like Gaussianity of data. If one wishes to avoid it, an alternative Monte Carlo method can be supported by the theory of randomization tests. The basic idea is to calculate a test statistic from the observed data, and then reshuffle the data a large number of times, recalculating the test statistic for each iteration. These statistics are used as before to generate a distribution of values. The observed value can be compared to the distribution to see whether the observed case is a tail value, i.e. an event that is unlikely to occur through chance.

[8] According to Delta method used to estimate a variance of a transformed parameter, one has $\mathrm{Var}[G(T)] \simeq \mathrm{Var}[T] \times (G'(\mu))^2 = const$, where $T = Y/l$, $\mathrm{E}[T] = \mathrm{Var}[T] = \mu$ and $G(T) = \sqrt{T}$.

[9] Four locations were overlapped in fine and coarse grids.

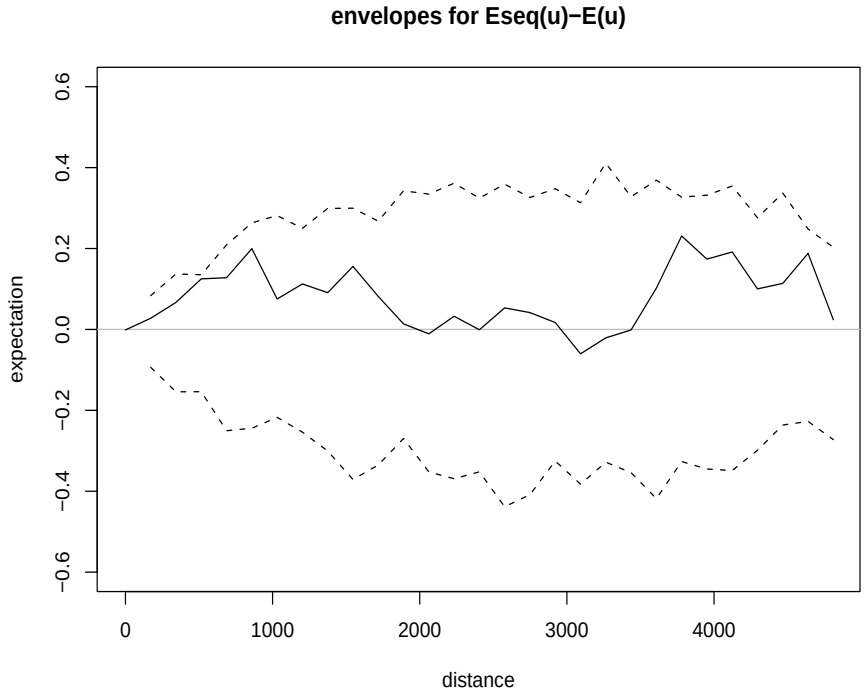


Figure 4.9: Simulation envelopes of $E_{seq}(u) - E(u)$ for Rongelap island's data, with $\widehat{\gamma}$ obtained through REML: data (solid curve); upper and lower envelopes from 99 simulations of a random field (dashed curves).

Often in comparisons between groups, like in some examples of biology applications (Manly 1991), only the group memberships are randomized while the same set of measurements are maintained. The latter tests are sometimes referred to as permutation ones, because the randomization can be done by reordering the positions of elements in an array. Monte Carlo tests evaluated by permutation are also quite applied to geostatistical data. Examples are the Mantel's correlation test of two association matrices, that may refer to distances, and the Moran's I test based on an empirical spatial autocorrelation coefficient. See Cliff and Ord (1981) for more detail on Moran's I test.

A non-parametric approach

In geostatistics, a natural permutation test can be derived when the actual data values are maintained, but they are randomly permuted in order to obtain the distribution of the test statistic. Exactly how they are permuted depends on the null hypothesis to be tested.

In our case, for testing H_0 given in (4.4) on Rongelap data, we propose the following non-parametric approach. Suppose the locations and values for the first stage of the sampling were fixed a priori, then we can assess the variation in the test statistic over randomisation of the second stage sampling. In here, to keep avoiding the assumption of a model for the spatial process, we can select at random over all the locations from the two stages and using the observed values at these selected sites, this would avoid the need for a model.

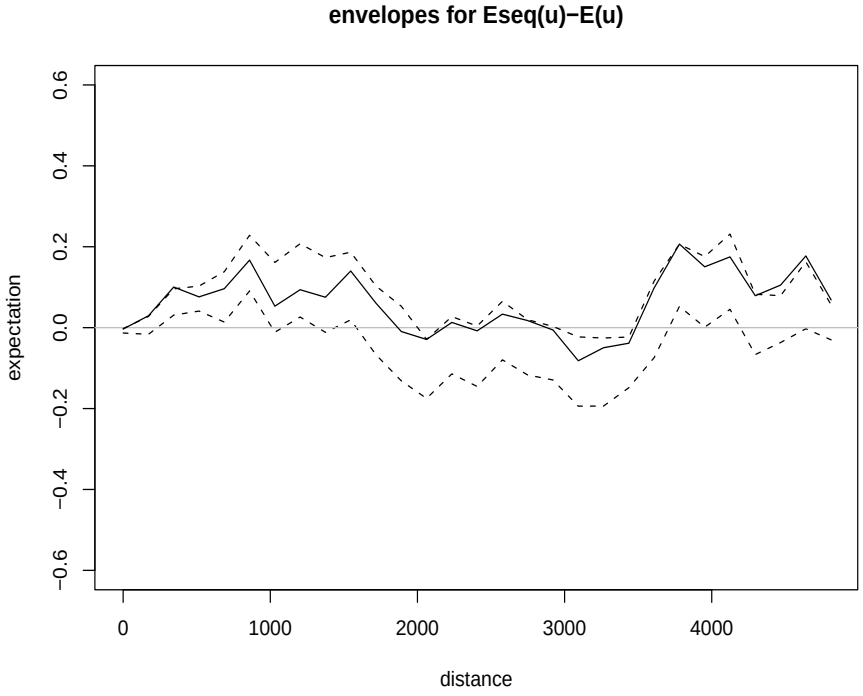


Figure 4.10: Simulation envelopes of $E_{seq}(u) - E(u)$ for Rongelap island's data, using a non-parametric approach: data (solid curve); upper and lower envelopes from 99 simulations of a random field (dashed curves).

The results plotted in Figure 4.10 were derived following this type of approach. We fixed as true the 63 sampled locations and values from the first stage

$$\{(\mathbf{x_i}, Z(\mathbf{x_i})) : \ i = 1, ..., 161, \ \ t(\mathbf{x_i}) = t_0\}.$$

For each simulation, we chose randomly 98 extra data points $\mathbf{x_k}$, among the total 161 available and we got a new data set representative of the second stage

$$\{(\mathbf{x_k}, Z(\mathbf{x_k})) : t(\mathbf{x_k}) = t_0 \ \text{ or } \ t(\mathbf{x_k}) = t_1\}.$$

We could then derive the conditional expectation functions $\widehat{E}^j_{seq}(u)$ and $\widehat{E}^j(u)$ for $j = 1, ..., 99$.

According to Figure 4.10, we realize that this non-parametric approach gives a narrow envelope interval when compared to the one obtained through REML in Figure 4.9, probably because of the smaller variability associated to a permutation test. These simulation envelopes actually suggest a possible rejection of H_0 given in (4.4). However, this rejection was not confirmed with the formal test based on (4.5). The observed test statistic r_1 was the 92^{th} largest of all values r_j, so H_0 should be accepted with an attained significance level of 0.92.

4.5 Some non-standard sampling correctors

In Section 4.2.2, when analysing the Matheron's and Nadaraya-Watson kernel variogram estimators under different sampling designs, we have concluded that they may produce poor estimates of the spatial variability. This can happen when the sampling strategy causes later samples to be located in areas with atypical, usually high or low, data values. While, it is likely that these samples give good information on the spatial variance within the clusters, they are not representative of the remaining area. The naive approach of discarding clustered biased data would force us to lose useful information, as well as, it may not always be possible to identify those that should be kept and those that should be discarded. To obtain a good estimate of the global spatial variance, one may claim a method of

weighting individual samples and clustered ones, in such a way the latter do not have an undue influence on the estimate.

4.5.1 Method to adjust for clustering

Our first concern is then the clustering issue. We propose to modify the NW kernel estimator trying to adjust for clustering of samples and minimize the negative impact on variogram estimation.

Consequently, a compensation for the unpopulated areas is proposed, by suggesting an inverse weight to a given neighbourhood density and, simultaneously, joining the benefits outcome from a kernel estimator. A possible way to extend $w_{ij}(u) = K\left((u - \|\mathbf{x_i} - \mathbf{x_j}\|)/h\right)$ in equation (3.3) to adjust for clustering is to use

$$w_{ij}(u) = \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right), \qquad n_i = \sum_k I_{\{\|\mathbf{x_i} - \mathbf{x_k}\| \leq \delta\}} \tag{4.6}$$

where n_i represents the number of points that fall within the circle of radius δ and center $\mathbf{x_i}$, as $I_{\{A\}}$ denotes the indicator function of the set A. Hereafter, this estimator is tagged *NewRobClust*.

A correction for over-populated areas is also suggested in Reilly and Gelman (2004). The direct effect of introducing a kernel estimator into this declustering weight process is *up-weighting* each lag $\|\mathbf{x_i} - \mathbf{x_j}\|$ according to its proximity to lag u under estimation. Moreover, the *down-weight* component $(n_i \times n_j)^{-\frac{1}{2}}$ allows a correction for high density areas, as they might not be sufficiently spatially representative of the overall data. Some preliminary results of this new estimator are presented in Menezes and Tawn (2003).

4.5.2 Sequential biased corrector

The second concern is about the bias possibly present in the final sample data set, when some type of multi-stage sampling design is adopted. Here, we propose a very naive approach. If the exploratory analysis from previous Sections suggests the presence of sequential bias, we propose to slightly modify the adjust for clustering's

method in such a way that those data values included and those not included into a region regarded as sequential biased would not be mixed. Once more, the implementation of this proposal assumes certainly that one keeps track of time labels associated to each data value, i.e., that knowledge about how the multi-stage collection of data occurred is made available. Then, the weight expression (4.6) changes to

$$w_{ij}(u) = \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right) \times I_{\{t(\mathbf{x_i})=t(\mathbf{x_j})\}}, \quad n_i = \sum_k I_{\{\|\mathbf{x_i}-\mathbf{x_k}\|\leq\delta\}}$$

Hereafter, this estimator is tagged *Pooled*. Under biased clustered samples, the resulting *pooled* variogram estimator can be roughly described as using latest data values for small lags' estimations and the remaining data values otherwise. Under the absence of sequential bias, this estimator should produce very similar results to estimator (4.6).

4.5.3 Results

The comparison study of variogram estimators described in Section 4.2.2 can now be concluded. In Table 4.1, we include the results achieved by the variogram estimators proposed in the two previous Sections. Remember that, for this simulation study, we have considered a two-stage approach for sampling collection, with the second stage possibly influenced by the first, suggesting four possible combinations of biased sampling and clustered sampling.

In this Table, one also finds the performance results of a distinct robust to clusters estimator, which does not include a kernel component. This estimator is the one identified by tag *RobClust* and its corresponding weight is simplified to $w_{ij}(u) = (n_i \times n_j)^{-\frac{1}{2}}$.

Under random sampling, the five estimators present similar results, with just a slightly better performance for the three kernel estimators. The best improvement accomplished by the *Pooled* estimator occurs under simultaneously biased

and clustered sampling, when the errors decrease 4.5 times and 7.2 times, when compared to NW kernel and Matheron estimators, respectively. Under just clustered sampling, these same values decrease 1.2 and 2.0 times, respectively. For the last combination, bias but no strong clustering, the *Pooled* estimator origins values 1.9 and 1.4 smaller than NW kernel and Matheron estimators.

As expected, the new estimators *NewRobClust* and *Pooled* present very similar ISE values under the absence of sequential sampling. However, under simultaneously biased and clustered sampling, there is some practically relevant improvement of our naive *Pooled* estimator over existing methods. This suggests to consider time labels in the estimates whenever possible.

The direct comparison of *NewRobClust* and *RobClust* confirms the importance of the kernel component in our proposal.

As previously mentioned, this same numerical study was repeated for a more than two stages approach, applying the serial sampling generation algorithm introduced in Section 4.2.1. We aim to reinforce the analogy of results when two or more stages are considered for data collection. In boxplots from Figures 4.11-4.14, we compare the estimates given by the five variogram estimators, giving us now some additional information about the dispersion of ISE values, measured in terms of the interquartile range.

The analysis of these boxplots emphasises the weak results of NW kernel estimator near the endpoint 0, that have a direct impact on the proposed estimators. Consequently, for smaller lags, the described gains are not so evident and the new estimators can even produce slightly worse results than Matheron. Anyway, one should have in mind that the small lags' behaviour has a small relative contribution to the global behaviour. In fact, the ISE values found for $u \leq 0.6$ are about 50 times larger than those found for $u \leq 0.1$.

Figure 4.11: Behaviour of $\widehat{\gamma}$ under random sampling.

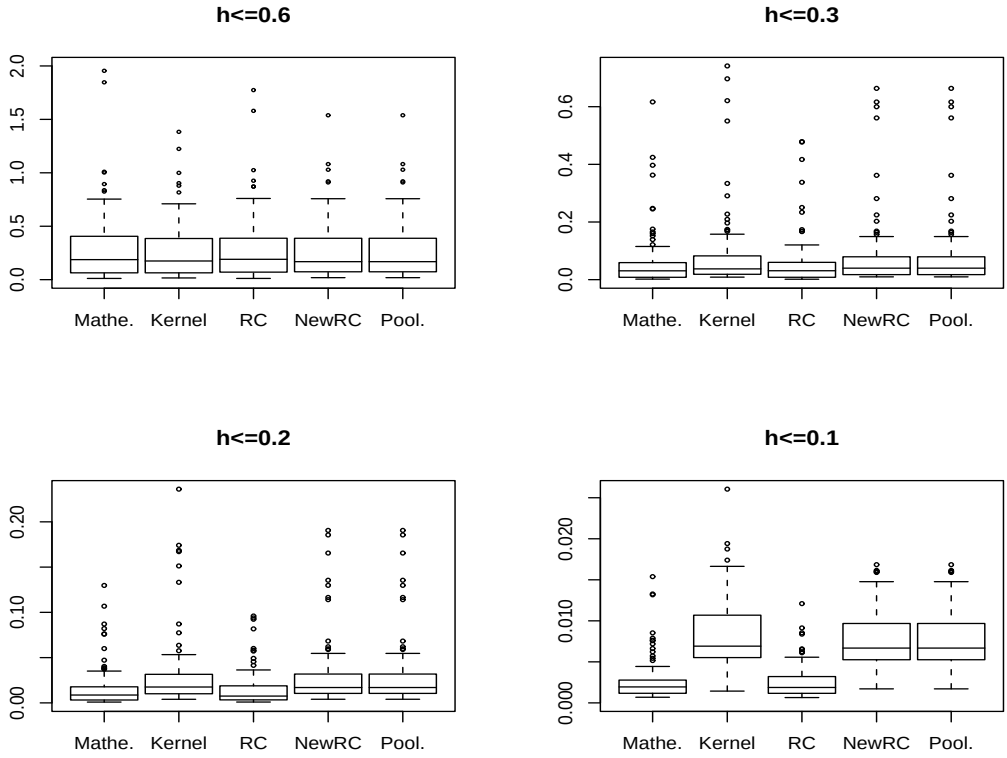


Figure 4.12: Behaviour of $\widehat{\gamma}$ under clustered and sequential biased sampling.

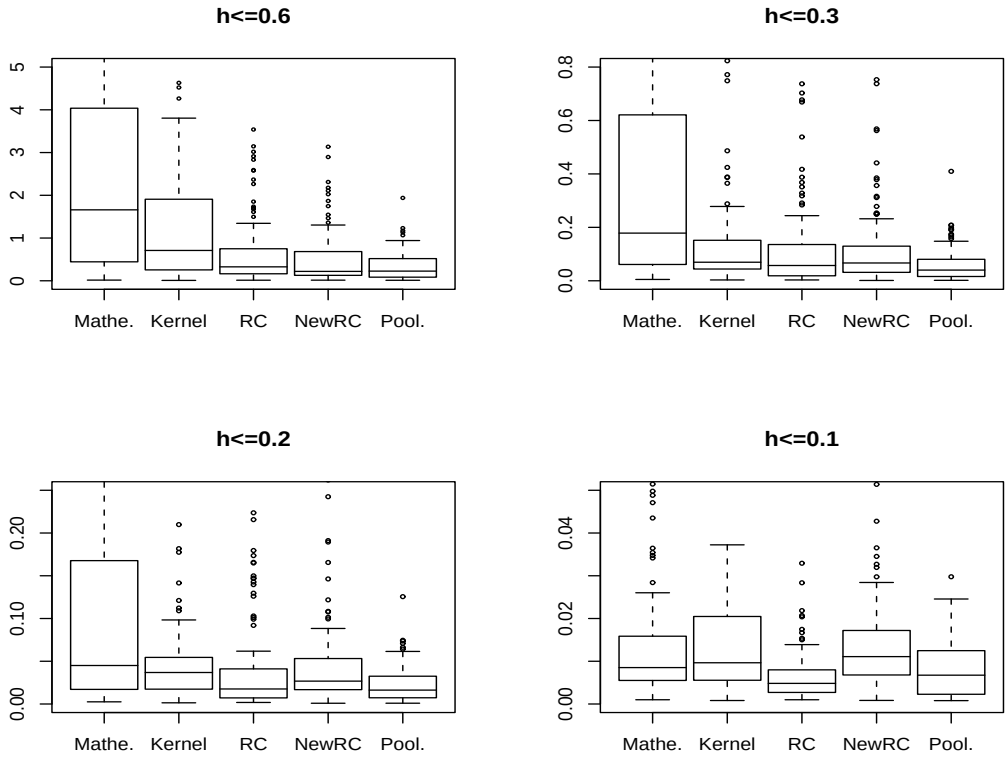


Figure 4.13: Behaviour of $\widehat{\gamma}$ under clustered sampling (no sequential biased).

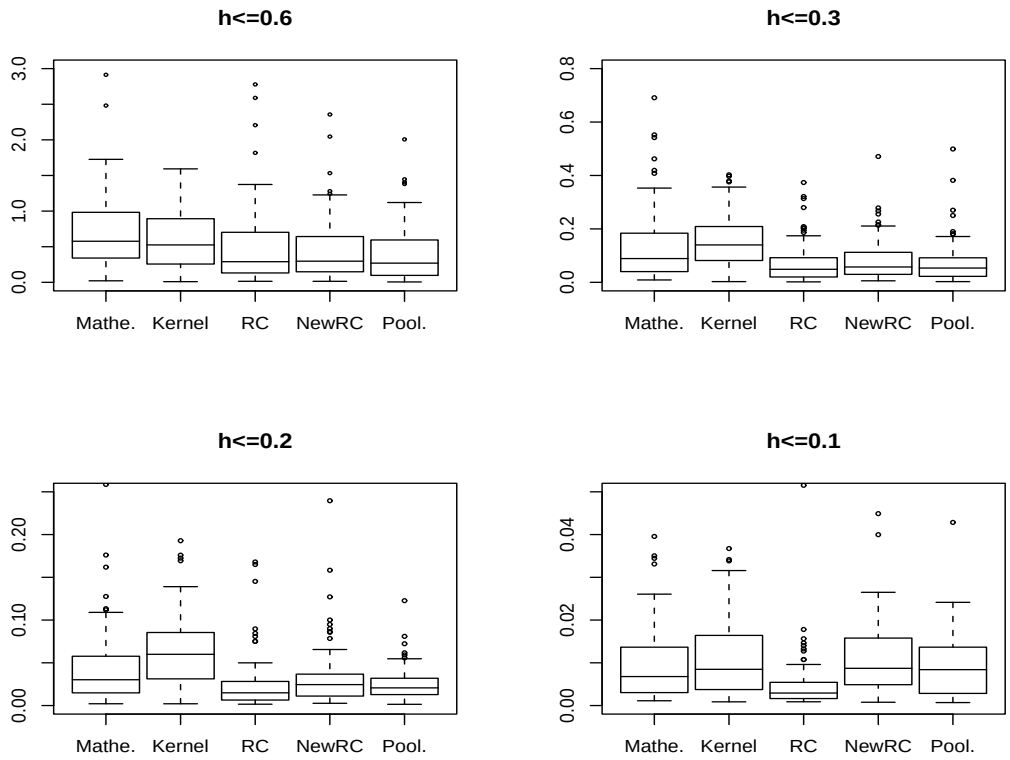
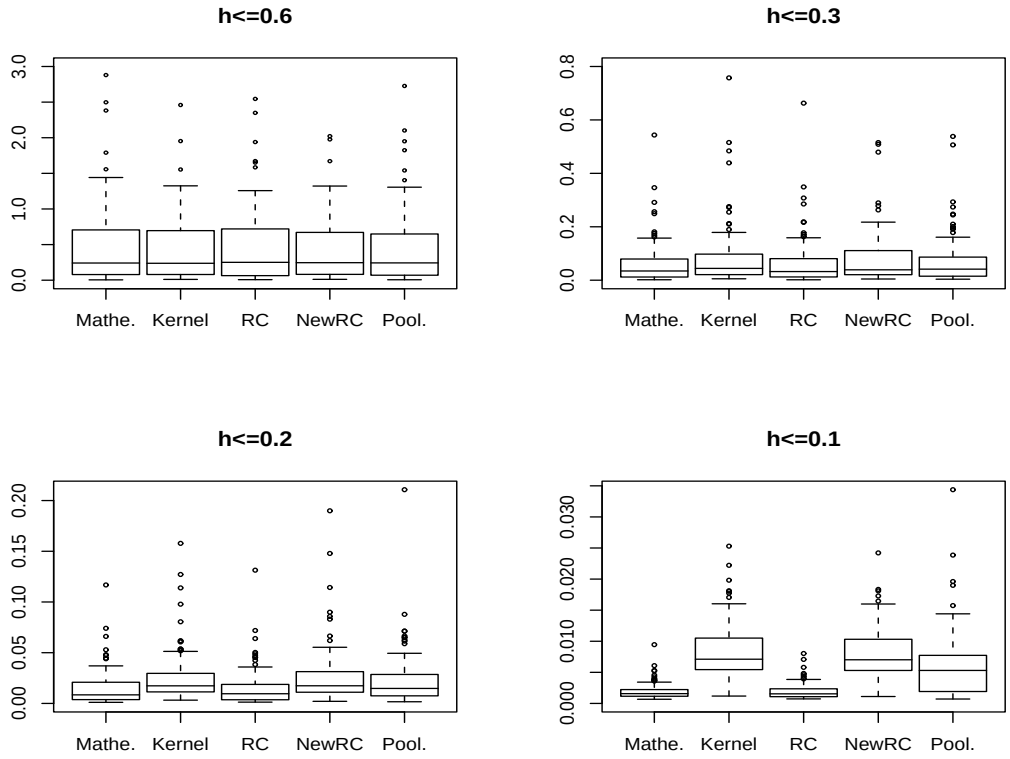


Figure 4.14: Behaviour of $\widehat{\gamma}$ under sequential biased sampling (no clustered).



4.5.4 Rongelap island's data

We conclude this Chapter by proceeding with the assessment of the proposed variogram estimators on Rongelap data, when testing for the presence of sequential dependency.

In Section 4.4.2, the Monte Carlo test suggested for the Rongelap data has employed a variogram estimator derived through maximum likelihood. This estimation was required to model the spatial dependency and to generate 99 simulations of a random field according to the null hypothesis H_0 in (4.4).

We now investigate the influence of adopting the new variogram estimators instead, bearing in mind that we must use their corresponding valid versions. These are obtained by fitting the empirical estimators introduced in Sections 4.5.1 and 4.5.2 to a permissible variogram given by Bochner's theorem in equation (3.4). Both estimators, with similar outcomes, were applied to all data from coarse and fine grids.

In Figure 4.15, we choose to illustrate the Monte Carlo test related to the valid *Pooled* estimator. The observed curve $\widehat{E}_{seq}(u) - \widehat{E}(u)$ is outside the simulation envelopes for small and large lags. According to these results, the presence of sequential dependency in the samples should not be totally excluded. Actually, the rejection of H_0 in (4.4) was confirmed with a formal test based on (4.5) at the 5% level of significance.

As a last note, we highlight that this data set underlies some characteristics, like a low spatial variance and locations almost forming a straight line due to the island's layout, requiring a careful estimation. Consequently, corrector methods like the ones proposed are advised.

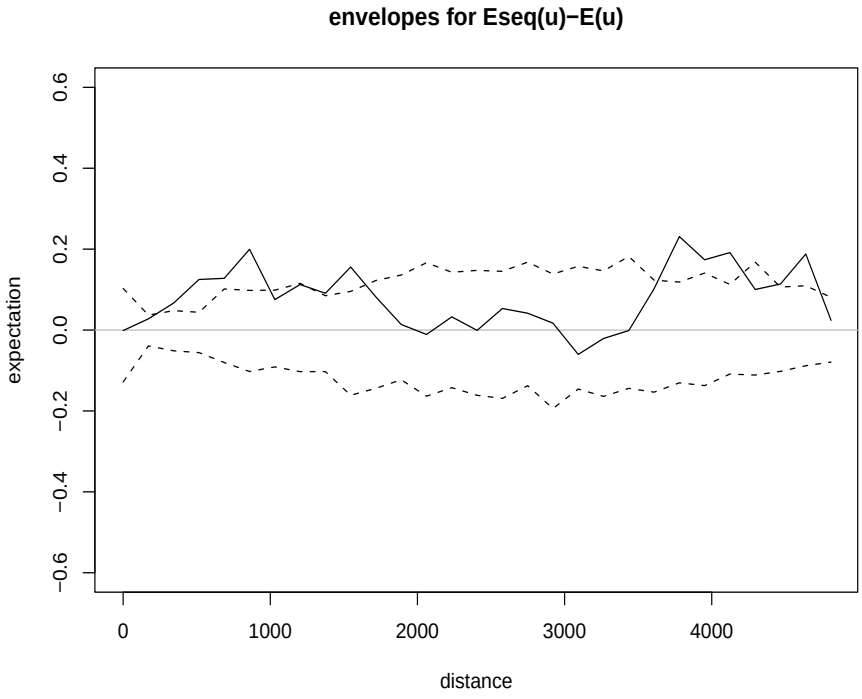


Figure 4.15: Simulation envelopes of $E_{seq}(u) - E(u)$ for Rongelap island's data, using the valid *Pooled* variogram estimator: data (solid curve); upper and lower envelopes from 99 simulations of a random field (dashed curves).

Chapter 5

Properties of a kernel variogram estimator for clustered data

5.1 Introduction

This Chapter is dedicated to the theoretical study of the new kernel variogram estimator proposed in Section 4.5.1 for clustered data, proving its asymptotic unbiasedness and consistency. Additionally, we shall propose optimal values for its unknown *smoothing parameters*, two user-adjustable quantities that affect the estimator's performance.

Bearing in mind that $\{Z(\mathbf{x}) : \mathbf{x} \in D \subset I\!R^d\}$ is an intrinsic and isotropic random process and denoting by $Z(\mathbf{x_1}), ..., Z(\mathbf{x_n})$ the values of the process observed at spatial locations $\mathbf{x_1}, ..., \mathbf{x_n}$, the suggested variogram estimator is defined as follows:

$$\widehat{\gamma}(u) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right)[Z(\mathbf{x_i}) - Z(\mathbf{x_j})]^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right)}, \quad u \geq 0, \qquad (5.1)$$

where $n_i = \sum_k I_{\{\|\mathbf{x_i}-\mathbf{x_k}\| \leq \delta\}}$ and $n_j = \sum_k I_{\{\|\mathbf{x_j}-\mathbf{x_k}\| \leq \delta\}}$; h and δ represent the bandwidth and neighbourhood radius selectors, respectively.

Declustering methods are quite intuitive, and their need is well recognized in the spatial statistics literature to estimate spatially representative mean trends for clustered data (see e.g. Goovaerts (1997) and Isaaks and Srivastava (1989); or Dubois and Saisana (2002) for a comparison of classical declustering methods). In contrast, the corresponding need for the reliable estimation of the second-order spatial structures is not normally considered. The presence of clustered sample data is, however, not negligible at all as shown in Chapter 4. See for example Figure 4.2 or Table 4.1, which exhibit the decay of traditional variogram estimators under unequal samples density.

Some of the main reasons for clustering of sample locations are:

- External factors, like selection of locations conditional on specific geographic or demographic spots.
- The need to better characterize short-range variability, requiring a denser sampling, but sometimes too costly to cover the whole observation region.
- Adoption of a denser sampling in areas that are deemed critical. For example, the search of maximum values based on some prior knowledge.

The recent paper from Kovitz and Christakos (2004) concerns the clustering issue and the estimation of the second-order structure. These authors suggest a modified form of Matheron's estimator that also incorporates some declustering weights, but based on zones of proximity. Each zone of any data point is defined by the area of the Voronoi polygon that contains all points closer to that interior data point than to any other data point. The performance of this modified estimator of the variogram is analysed in terms of a numerical application.

In our case, we prove that our variogram estimator enjoys good asymptotic properties. A short version of the preliminary theoretical result is found in Menezes, Garcia-Soidán and Febrero-Bande (2004). As this estimator requires the selection of the bandwidth h and the radius δ, we recommend: the first will be treated via the MSE, i.e. the minimum square error; and the latter will result from the

analysis of the density estimation derived on the observation region. The main contributions of this work are described in the extended version Menezes, Garcia-Soidán and Febrero-Bande (2005*b*).

The remainder of this Chapter is organized as follows. We first introduce additional notation and summarize the main assumptions considered in our asymptotic study. We then include comments about the neighbourhood radius selector. Next, the fundamental properties of the proposed non-parametric variogram estimator are established and corresponding proofs are developed. The results derived for bias and variance are used for the optimal bandwidth selection. We end with some numerical studies and implementation details about the proposed estimator.

5.2 Assumptions

To ensure estimation consistency, we follow the strategy proposed by Hall et al. (1994), and recently, adopted by Garcia-Soidán et al. (2004), according to which the observation region is considered to be increasing. Then,

(A1) We start by assuming $D = D_n = \lambda D_0$ where $\lambda = \lambda_n$ may diverge to $+\infty$ and $D_0 \subset I\!R^d$ is a bounded and fixed region.

(A2) Additionally, a random design is assumed for spatial locations $\mathbf{x_i} = \lambda \mathbf{v_i}$, $i = 1...n$, where $\mathbf{v_i}$ is a realization of a random sample $\mathbf{V_i}$ from f_0, the density function defined on D_0.

(A3) For all $\mathbf{v} \in D_0$ and for some positive constants d_1 and d_2, one has $d_1 \leq f_0(\mathbf{v}) \leq d_2$. [1]

(A4) $\gamma(.)$ admits three continuous derivatives in a neighbourhood of u, for all $u > 0$.

[1] This allow us to guarantee that $0 < \int f_0(\mathbf{x})^i d\mathbf{x} < +\infty,\ i = 2, 3, 4$.

(A5) There is a bounded and continuously differentiable function $g : I\!R^{3d} \to I\!R$ satisfying that $\text{Cov}\left[(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2, (Z(\mathbf{x_k}) - Z(\mathbf{x_l}))^2\right] = g(\mathbf{x_i} - \mathbf{x_j}, \mathbf{x_i} - \mathbf{x_k}, \mathbf{x_i} - \mathbf{x_l})$. We assume

$$\lim_{\|\mathbf{x_2}\| \geq r \bigvee \|\mathbf{x_3}\| \geq r} |g(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3})| = 0, \quad \text{where } 0 < r < +\infty$$

(A6) With respect to convergence rates, it is assumed that

$$\lim_{n \to \infty} \{h + \lambda^{-1} + \lambda^d n^{-1} + (nh)^{-1}\} = 0$$

(A7) Take $\delta = \lambda a$, where δ is the neighbourhood radius in D space and $a > 0$ is the equivalent in D_0. We assume that a has an upper bound.

Bear in mind that, in the context of a Gaussian process, one has

$$\text{Cov}\left[(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2, (Z(\mathbf{x_k}) - Z(\mathbf{x_l}))^2\right] =$$

$$= 2\left[\gamma(\|\mathbf{x_i} - \mathbf{x_k}\|) + \gamma(\|\mathbf{x_j} - \mathbf{x_l}\|) - \gamma(\|\mathbf{x_i} - \mathbf{x_l}\|) - \gamma(\|\mathbf{x_j} - \mathbf{x_k}\|)\right]^2$$

and, afterwards, one may take

$$g(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}) = 2\left[\gamma(\|\mathbf{x_2}\|) + \gamma(\|\mathbf{x_3} - \mathbf{x_1}\|) - \gamma(\|\mathbf{x_3}\|) - \gamma(\|\mathbf{x_2} - \mathbf{x_1}\|)\right]^2$$

so that condition **(A5)** is satisfied provided that the variogram is bounded and has an asymptotic range. Thus a model with no finite range, such as the exponential, is acceptable. We are not considering unbounded variograms, such as the linear, but in real data applications we think it reasonable to restrict to a bounded spatial correlation.

5.3 Neighbourhood radius selector

We first apply standard techniques of exploratory data analysis to gain a better understanding on an advisable value for δ. Namely, we used some elementary

theory of spatial point patterns to detect the presence of clusters (Diggle 2003). In this context, two useful functions are the cumulative distribution functions of point-to-point and origin-to-point nearest neighbour distances, G and F respectively.

Suppose a spatial point pattern dataset with n points. Let d_i denote the distance from the i^{th} point to the closest of the other $n-1$ points. For a grid of k sampling origins, let e_i denote the distance from the i^{th} origin to the closest of the n points. Then, these functions may be derived as

$$\widehat{G}(u) = n^{-1} \sum_i I_{\{d_i \leq u\}} \qquad \text{and} \qquad \widehat{F}(u) = k^{-1} \sum_i I_{\{e_i \leq u\}}$$

where $I_{\{.\}}$ is the indicator function.

The estimates of G or F can be used for formal inference purposes about the pattern, when compared to the true value of G or F for a completely random (Poisson) point process, which are

$$G(u) = F(u) = 1 - exp(-\lambda \pi u^2)$$

where λ is the intensity (expected number of points per unit area).

In Figure 5.1, we exemplify a graphic diagnostic with three distinct spatial models. The first one represents the example of the complete spatial randomness (CSR), where the locations within the unit square were obtained from an uniform distribution. The second and third models were obtained from a mixture of uniform and beta distributions, such that one or two strong clusters were achieved. For each of these models, we plot the estimated G function, as well as, $\widehat{G}$ and $\widehat{F}$ against each other. As expected, the estimates of G and F present similar values under the first model.

In our exploratory analysis, it was found convenient to consider the observation region to be defined in such way that edge effects can safely be ignored. In any case, if one is analysing some clustered area, it is reasonable to presume the cluster itself is not too close to the borders.

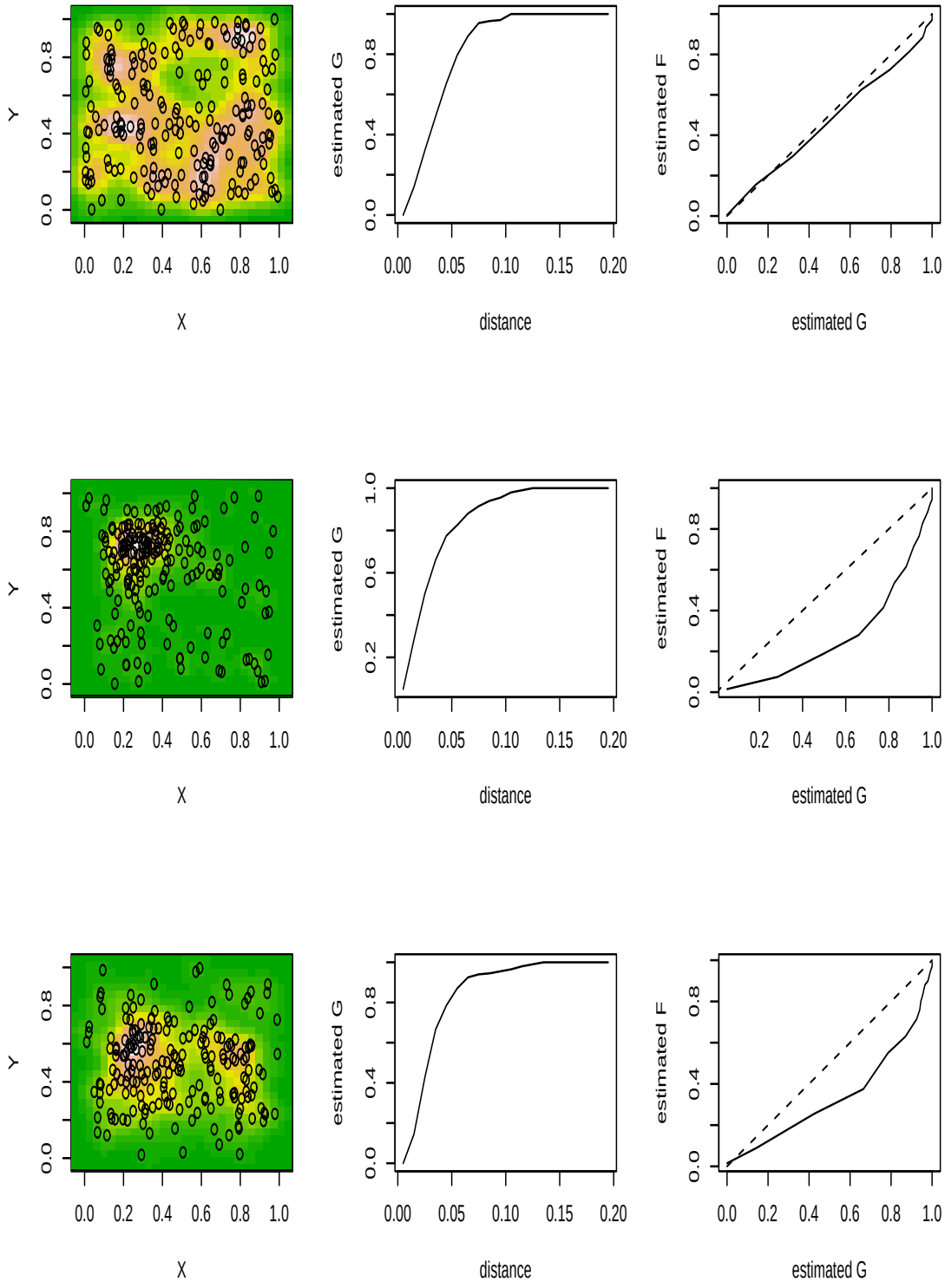


Figure 5.1: $\widehat{F}$ and $\widehat{G}$ under three distinct models: CSR, 1 cluster and 2 clusters.

Another useful empirical function to summarise an observed pattern is the K function (variously called "Ripley's K-function" or the "reduced second moment function") of a stationary point process. This is defined as the expected number of additional random points within a distance u of a typical random point, divided by the overall intensity of the points. The K function is determined by the second order moment properties of the point process. Under CSR, one expects

$$K(u) = \pi u^2.$$

For any of the previous empirical functions, Monte Carlo simulations can be used to test hypotheses and construct confidence intervals. Testing involves comparing an observed test statistic with samples from the model under consideration, in this case a CSR model. Confidence intervals can be based directly on the variation in estimates observed in data generated from the model. The maximum and minimum of the total independent simulations allow the definition of *upper* and *lower envelopes*. The observed test statistic may be compared against these simulation envelopes (Diggle 2003). In Figure 5.2, one finds examples of K estimates. The dashed curve, from the graphics in the middle side panels, identifies the πu^2 curve. The graphics, found in the right side panels, represent the corresponding Monte Carlo tests for the expression $\sqrt{\frac{K(u)}{\pi}} - u = 0$. The upper and lower envelopes (dashed curves) were derived from 99 CSR simulations. As expected, the second and third models suggest a rejection of the hypothesis "cluster absence". Note that all these plots were produced using the R package Splancs, presented in Rowlingson and Diggle (1993).

Some alternative methodologies for cluster analysis are directly motivated from techniques for density estimation (examples are Wong and Lane 1983, Silverman 1986, Cuevas, Febrero and Fraiman 2001). All of them are based on the natural idea of *clusters correspond to modes or peaks in the underlying density function* f *on* $\mathbb{R}^d$. Very often the unknown theoretical function f is replaced by a non-

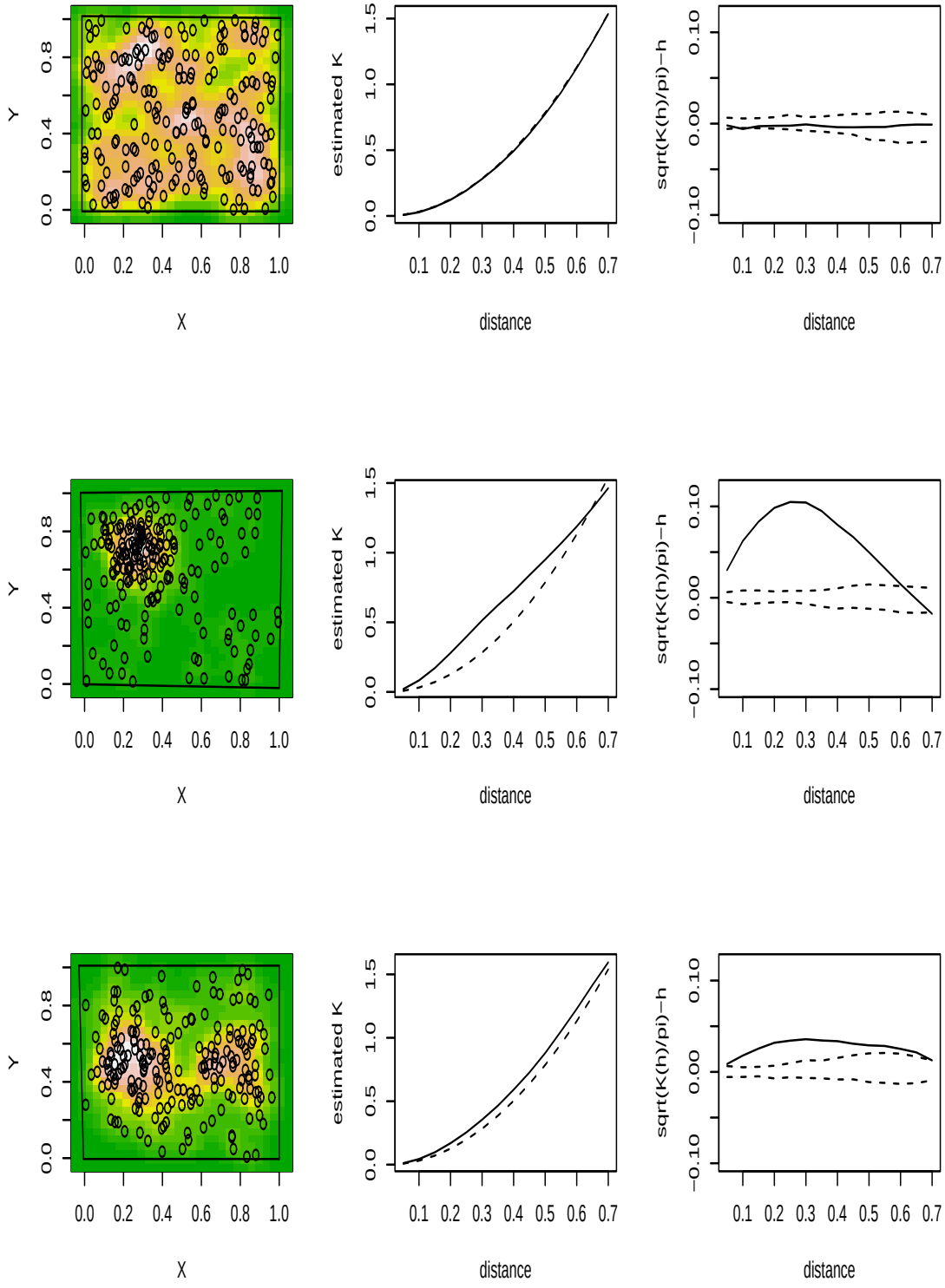


Figure 5.2: $\widehat{K}$ under three distinct models: CSR, 1 cluster and 2 clusters.

parametric density estimator of kernel-type

$$\widehat{f}(\mathbf{x}; h) = \frac{1}{nh^d} \sum_{i=1}^{n} K_d \left(\frac{\mathbf{x} - \mathbf{x_i}}{h} \right) \tag{5.2}$$

for some d-variate kernel density K_d.

Bear in mind that we aim to select a value for the neighbourhood radius δ in (5.1). Thus, we are more interested in the density estimation of the distances between locations than on the density estimation of the locations themselves. Additionally, one should be aware that we do not pretend to derive the *exact size* of each potential cluster, but just propose a reasonable neighbourhood radius for the down-weight operation in (4.6). Note that if this proposal is too small this new variogram estimator will tend to be equal to the Nadaraya-Watson one in (3.3). The other extreme of too large a value is not interesting either, as all locations will be treated in the same way and we aim to up-weight areas with low density.

We start by trying two different approaches for density estimation of the distances. Suppose $\{\mathbf{x_i}\}_{i=1}^{n}$ are locations in $I\!R^d$, then define $d_j = \|\mathbf{x_i} - \mathbf{x_k}\|$, $j = 1, ..., \frac{n(n-1)}{2}$. First, a kernel estimation is applied on equispaced distances, ranging from the lowest to the largest sampled distance d_j. In the second case, the estimation is applied on the sampled distances themselves. The δ quantity may then be derived from the maximum of these functions or even from, for instance, the 10% highest values. The final results from these two approaches tend to be very similar.

Another possible approach for δ derivation is based on *counts of distances*. For a given point and for a list of equispaced distances, one must count how the remaining $n-1$ points are spread within that list of distances. After repeating this for all n points, the partial sum organized by the distances, gives us the distance for the maximum count, i.e the proposal for δ value.

All previous approaches tend to return equivalent results, as illustrated in Figure 5.3, where the same set of spatial models were considered. Under the CSR model, we limited the value of the radius neighbourhood to 0.4. Please note that

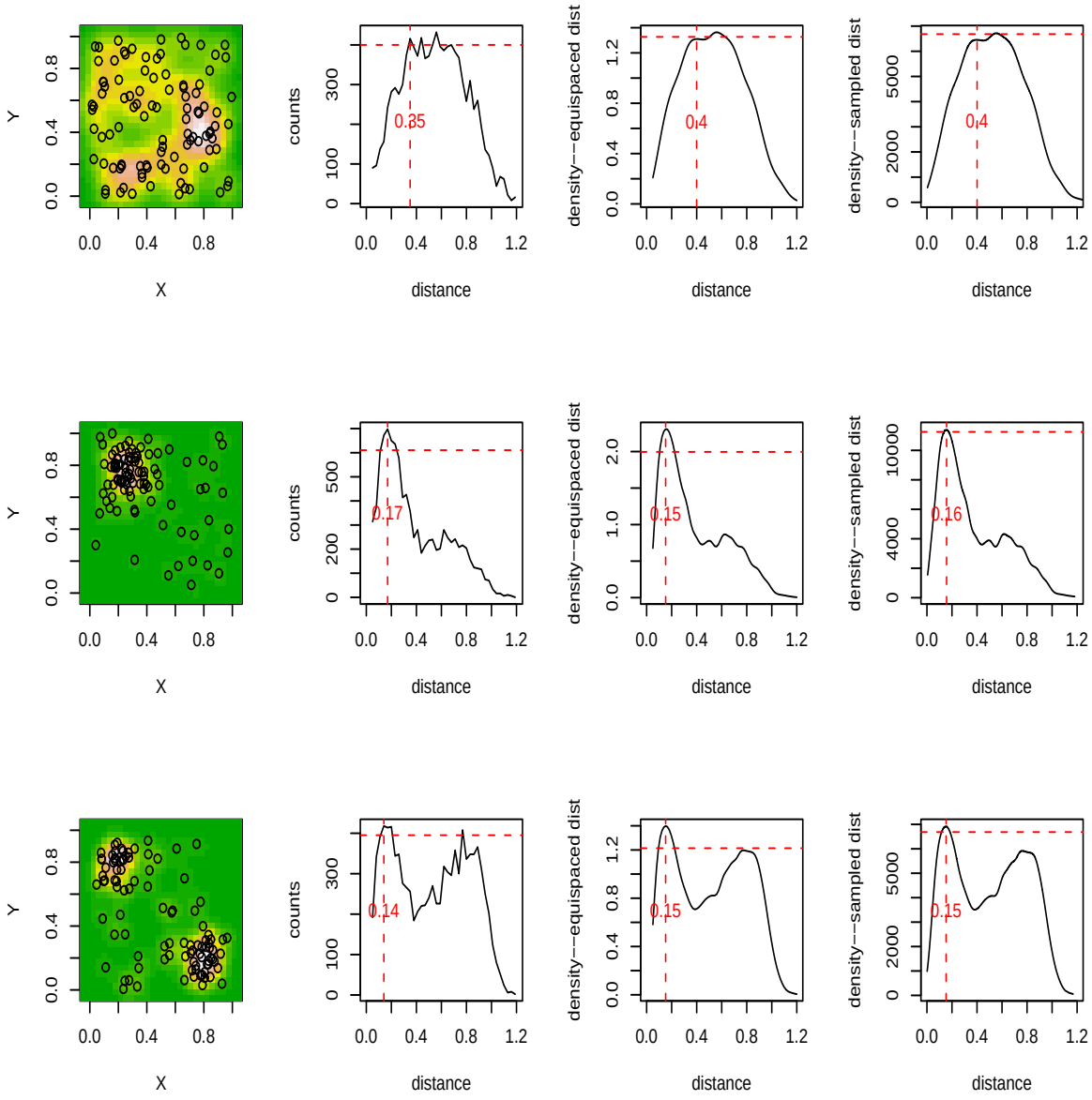


Figure 5.3: δ derivation under three distinct approaches and spatial models: CSR, 1 cluster and 2 clusters.

in the left panels of Figure 5.3, we plot the sample locations and the corresponding density estimation of the locations given by (5.2).

Moreover, our simulations seem to suggest that:

- If there is just 1 cluster, then one should expect only 1 strong mode;
- If there are 2 clusters, then one should expect 2 modes; This may be not so evident, if the clusters are too close;
- The distance between the 1^{st} mode and a possible existent 2^{nd} mode may be used to identify the distance between 2 clusters.

Finally, one should be aware that δ is the required neighbourhood radius in D space. We shall need the equivalent value, $a > 0$, in D_0 and these two are related through the expression $\delta = \lambda a$, given in assumption **(A7)**.

5.4 Bias of $\widehat{\gamma}(.)$ robust to clusters

We now describe the derivation of the fundamental asymptotic results, such as those for bias and variance, for the proposed variogram estimator. This derivation requires the assumptions introduced in Section 5.2, leading us to a desirable consistent estimation. Additionally, one should bear in mind that:

- under isotropy, the variogram domain is restricted to non-negative values;
- the kernel function operates on the distances $\|\mathbf{x_i} - \mathbf{x_j}\| \in [u - Ch, u + Ch]$;
- it is assumed that interval $[u - Ch, u + Ch]$ is wholly contained within the domain of $\gamma(.)$.

In this way, the following results are attained on $u \geq Ch$.

Theorem 5.1 *Assume that conditions* **(A1)-(A4)** *are satisfied. Additionally, suppose the convergence rates stated in* **(A6)**. *Then, for* $u \geq Ch$, *one has*

$$\mathrm{E}\left[\hat{\gamma}(u)\right] - \gamma(u) = \frac{1}{2} c_K \gamma''(u) h^2 + o\left(h^2\right)$$

with $c_K = \int z^2 K(z)dz$, *showing that the proposed estimator is asymptotically unbiased.*

Remark 5.2 *According to Theorem 5.1, the bias of* $\hat{\gamma}(u)$ *is of the exact order* h^2, *for* $u \geq Ch$; *however, near the endpoint 0,* $u < Ch$, *an order* h *rather than* h^2 *is expected, due to the boundary effect. As suggested in Garcia-Soidán et al. (2004), the adoption of a specific combination of boundary kernels is a possible solution to keep the same rate of convergence. Although, Theorem 5.1 would remain valid in practice for any* $u > 0$ *and large* n, *since the bandwidth parameter* h *tends to 0 as* n *increases.*

Lemma 5.3 *Let* $\{X_n\}$ *be a sequence of uniformly bounded random variables such that* $X_n = o(1)$ *a.s. Then,* $\mathrm{E}[X_n^r] = o(1)$, *for all* r.

According to the previous Lemma, as

$$\mathrm{Bias}\left[\hat{\gamma}(u)\right] = \mathrm{E_{Z,P}}\left[\hat{\gamma}(u) - \gamma(u)\right] = \mathrm{E_P}\left[\mathrm{E_Z}\left[\hat{\gamma}(u) - \gamma(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}\right]\right]$$

with P identifying the random process for the spatial locations, then our target becomes the derivation of the order of $\mathrm{E_Z}\left[\hat{\gamma}(u) - \gamma(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}\right]$. Future references to $\mathrm{E_Z}\left[\hat{\gamma}(u)|\mathbf{V_1}, ..., \mathbf{V_n}\right]$ will be simplified to $\mathrm{E}\left[\hat{\gamma}(u)|\mathbf{V_1}, ..., \mathbf{V_n}\right]$.

$$\mathrm{E}[\widehat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}] = \mathrm{E}\left[\frac{\sum w_{ij}(u)[Z(\mathbf{x_i}) - Z(\mathbf{x_j})]^2}{2\sum w_{ij}(u)} \mid \mathbf{V_1}, ..., \mathbf{V_n}\right] =$$

$$= \frac{\sum w_{ij}(u)\gamma(\|\mathbf{x_i} - \mathbf{x_j}\|)}{\sum w_{ij}(u)}$$

where w_{ij} is given in (4.6).

Write $a_1(u) = \sum w_{ij}(u)$ and $a_2(u) = \sum w_{ij}(u)\gamma\left(\|\mathbf{x_i} - \mathbf{x_j}\|\right)$. The spatial locations have been taken as $\mathbf{x_i} = \lambda\,\mathbf{v_i}$, $1 \leq i \leq n$; see conditions **(A1)** and **(A2)**. Then,

$$\mathrm{E}\left[\hat{\gamma}(u)\,|\mathbf{V_1},...,\mathbf{V_n}\right]-\gamma(u)=\frac{a_2(u)-a_1(u)\gamma(u)}{a_1(u)} \tag{5.3}$$

In Sections 5.4.1 and 5.4.2, we shall show that the following orders hold for $u \geq Ch$:

$$a_1(u) = u^{d-1}A_d\ n^2\lambda^{-d}\ h \quad \int f_0(\mathbf{w_1})^2\ \sum_{k=0}^{n-2}\frac{\binom{n-2}{k}H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-2-k}}{k+2}\ d\mathbf{w_1}\ + \\ +\ o(n^2\lambda^{-d}h\ a_n) \quad \text{a.s.}$$

$$a_2(u)-a_1(u)\gamma(u) = \tfrac{1}{2}c_K\gamma''(u)\ u^{d-1}A_d\ n^2\lambda^{-d}\ h^3. \\ .\int f_0(\mathbf{w_1})^2\ \sum_{k=0}^{n-2}\frac{\binom{n-2}{k}H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-2-k}}{k+2}\ d\mathbf{w_1}+ o(n^2\lambda^{-d}h^3\ a_n) \quad \text{a.s.}$$

for some bounded sequence a_n, as described in Section 5.7.2.

Consequently, by considering the latter relations and applying Lemma 5.3 to expression (5.3), one obtains

$$\mathrm{E}\left[\hat{\gamma}(u)\right]-\gamma(u)=\mathrm{E}\left[\frac{a_2(u)-a_1(u)\gamma(u)}{a_1(u)}\right]=\frac{1}{2}c_K\gamma''(u)h^2+o\left(h^2\right)$$

which would allow one to conclude Theorem 5.1 is valid.

5.4.1 Order of $a_1(u)$ for $u \geq Ch$

For $u \geq Ch$, as the kernel function K is compactly supported, one has

$$a_1(u)=\sum_{i\neq j}\frac{K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right)}{\sqrt{n_in_j}}+nK\left(\frac{u}{h}\right)=\sum_{i\neq j}\frac{K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right)}{\sqrt{n_in_j}}$$

Proceeding in a similar way as in the proof of Theorem 3.1 of Garcia-Soidán et al. (2004), the dominant term of $a_1(u)$ will be given by $n^2\alpha$, where

$$\alpha=\mathrm{E}\left[\frac{K\left(\frac{u-\lambda\|\mathbf{V_1}-\mathbf{V_2}\|}{h}\right)}{\sqrt{\sum_{k_1,k_2}I_{\{\lambda\|\mathbf{V_1}-\mathbf{V_{k_1}}\|\leq\delta,\lambda\|\mathbf{V_2}-\mathbf{V_{k_2}}\|\leq\delta\}}}}\right]$$

Consider the new random variables $\mathbf{W_2} = \mathbf{V_1} - \mathbf{V_2}$, ..., $\mathbf{W_n} = \mathbf{V_1} - \mathbf{V_n}$.

Keep also in mind that a realization of $\mathbf{W_2}$ obeys to $\lim_{\lambda\to\infty} \|\mathbf{w_2}\| = 0$. This happens since K is compactly supported, i.e. $K(z) = 0$ if $|z| > C$, meaning that $\lambda^{-1}(u - Ch) \leq \|\mathbf{w_2}\| \leq \lambda^{-1}(u + Ch)$. Then

$$\alpha = \int ... \int \frac{K\left(\frac{u-\lambda\|\mathbf{w_2}\|}{h}\right)\ f_{n-1}(\mathbf{w_2}, ..., \mathbf{w_n})}{\sqrt{1 + 2\sum_{k_1\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq\frac{\delta}{\lambda}\}} + \sum_{k_1,k_2\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq\frac{\delta}{\lambda}, \|\mathbf{w_{k_2}}-\mathbf{w_2}\|\leq\frac{\delta}{\lambda}\}}}}\, d\mathbf{w_2}...d\mathbf{w_n}$$

As a marginal distribution, $f_{n-1}(\mathbf{w_2}, ..., \mathbf{w_n})$ can be written as $\int f_n(\mathbf{w_1}, ..., \mathbf{w_n})d\mathbf{w_1}$ $= \int f_0(\mathbf{w_1})f_0(\mathbf{w_1} - \mathbf{w_2})...f_0(\mathbf{w_1} - \mathbf{w_n})d\mathbf{w_1}$ and, consequently,

$$\alpha = \int ... \int \frac{K\left(\frac{u-\lambda\|\mathbf{w_2}\|}{h}\right) f_0(\mathbf{w_1})f_0(\mathbf{w_1} - \mathbf{w_2})...f_0(\mathbf{w_1} - \mathbf{w_n})}{\sqrt{\left(1 + 2\sum_{k_1\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}}\right) + \left(\sum_{\substack{k_1\geq 2\\ k_2\geq 2}} I_{\{\|\mathbf{w_{k_1}}\|\leq a, \|\mathbf{w_{k_2}}-\mathbf{w_2}\|\leq a\}}\right)}}\, d\mathbf{w_1}...d\mathbf{w_n}$$

The expression under the square root of α may be simplified as follows

$$\left(1 + 2\sum_{k_1\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}}\right) + \left(\sum_{k_1,k_2\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a, \|\mathbf{w_{k_2}}-\mathbf{w_2}\|\leq a\}}\right) =$$

$$= \left(3 + 2\sum_{k_1\geq 3} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}}\right) + \left(\sum_{\substack{k_1=k_2\\ k_1\geq 2}} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{\substack{k_1\neq k_2\\ k_1,k_2\geq 2}} I_{\{\|\mathbf{w_{k_1}}\|\leq a, \|\mathbf{w_{k_2}}\|\leq a\}}\right) =$$

$$= \left(3 + 2\sum_{k_1\geq 3} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}}\right) + \left(1 + 3\sum_{k_1\geq 3} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{\substack{k_1\neq k_2\\ k_1,k_2\geq 3}} I_{\{\|\mathbf{w_{k_1}}\|\leq a\|\}} I_{\{\|\mathbf{w_{k_2}}\|\leq a\}}\right) =$$

$$= 4 + 5\sum_{k_1\geq 3} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{\substack{k_1\neq k_2\\ k_1,k_2\geq 3}} I_{\{\|\mathbf{w_{k_1}}\|\leq a\|\}} I_{\{\|\mathbf{w_{k_2}}\|\leq a\}}$$

Then

$$\alpha = \int ... \int \frac{K\left(\frac{u-\lambda\|\mathbf{w_2}\|}{h}\right) f_0(\mathbf{w_1})^2 f_0(\mathbf{w_1}-\mathbf{w_3})...f_0(\mathbf{w_1}-\mathbf{w_n})}{\sqrt{4+5\sum_{k_1\geq 3} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{\substack{k_1,k_2\geq 3 \\ k_1\neq k_2}} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} I_{\{\|\mathbf{w_{k_2}}\|\leq a\}}}} d\mathbf{w_1}...d\mathbf{w_n}$$

Now, convert $\mathbf{w_2} = (w^{(1)}, ..., w^{(d)})$ to spherical polar coordinates with the transformation

$$w^{(i)} = r\cos\theta_i \prod_{j=0}^{i-1} \sin\theta_j$$

where $\sin\theta_0 = \cos\theta_d = 1$, $0 \leq \theta_{d-1} < 2\pi$ and $0 \leq \theta_i < \pi$, for $i = 1, ..., d-2$. The corresponding Jacobian transformation is given by

$$r^{d-1} J_d(\theta_1, ..., \theta_{d-1}) = r^{d-1}(\sin\theta_1)^{d-2}(\sin\theta_2)^{d-3}... \ \sin\theta_{d-2}.$$

Furthermore, suppose $\sum_{k_1\geq 3} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} = k$ and apply some basic combinatory rules to obtain

$$\alpha = \left(\int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} \int_0^{m_0} r^{d-1} J_d(\theta_1, ..., \theta_{d-1}) \ K\left(\frac{u-\lambda r}{h}\right) dr d\theta_1 ... d\theta_{d-1} \right) .$$

$$\cdot \sum_{k=0}^{n-2} \frac{\binom{n-2}{k} \ \int f_0(\mathbf{w_1})^2 H(a, \mathbf{w_1})^k (1 - H(a, \mathbf{w_1}))^{n-2-k} d\mathbf{w_1}}{\sqrt{4+5k+k(k-1)}}$$

where $m_0 = sup\{\|x\| : \ x \in D_0\}$ and $H(a, \mathbf{w_1}) = \int_{\|\mathbf{w}\|\leq a} f_0(\mathbf{w_1} - \mathbf{w}) d\mathbf{w}$.

Finally, with the following change of variable

$$t = h^{-1}(u - \lambda r) \quad \Rightarrow \quad r = \lambda^{-1}(u - th) \quad \Rightarrow \quad dr = -\lambda^{-1} h \ dt,$$

and as K is compactly supported, the dominant term in α becomes

$$\left(\int ... \int J_d\left(\theta_1, ..., \theta_{d-1}\right) d\theta_1 ... d\theta_{d-1} \right) \left(\int_{\frac{u-m_0\lambda}{h}}^{\frac{u}{h}} \left(\lambda^{-1}(u - th)\right)^{d-1} K(t) \ \lambda^{-1} h \ dt \right) .$$

$$\cdot \sum_{k=0}^{n-2} \frac{\binom{n-2}{k} \ \int f_0(\mathbf{w_1})^2 H(a, \mathbf{w_1})^k (1 - H(a, \mathbf{w_1}))^{n-2-k} d\mathbf{w_1}}{k+2} =$$

$$= u^{d-1} A_d \ \lambda^{-d} \ h \quad \int f_0(\mathbf{w_1})^2 \ \sum_{k=0}^{n-2} \ \frac{\binom{n-2}{k} H(a, \mathbf{w_1})^k (1 - H(a, \mathbf{w_1}))^{n-2-k}}{k+2} \ d\mathbf{w_1}$$

where

$$A_d = \int ... \int J_d \left(\theta_1, ..., \theta_{d-1}\right) d\theta_1 ... d\theta_{d-1} \tag{5.4}$$

In the final expression of α, bear in mind that the integral performed over variable $\mathbf{w_1}$ is bounded due to condition **(A3)**.

5.4.2 Order of $a_2(u) - a_1(u)\gamma(u)$ for $u \geq Ch$

Similarly, for $u \geq Ch$, one has

$$a_2(u) - a_1(u)\gamma(u) = \sum_{i \neq j} \frac{K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right)}{\sqrt{n_i n_j}} \ (\gamma(\|\mathbf{x_i} - \mathbf{x_j}\|) - \gamma(u))$$

and one may prove that the dominant term in here is given by $n^2\beta$, where

$$\beta = \mathrm{E}\left[\frac{K\left(\frac{u - \lambda\|\mathbf{V_1} - \mathbf{V_2}\|}{h}\right)}{\sqrt{\sum_{k_1,k_2} I_{\{\lambda\|\mathbf{V_1} - \mathbf{V_{k_1}}\| \leq \delta, \lambda\|\mathbf{V_2} - \mathbf{V_{k_2}}\| \leq \delta\}}}} \ (\gamma(\lambda\|\mathbf{V_1} - \mathbf{V_2}\|) - \gamma(u)) \right]$$

Let us again convert $\mathbf{w_2} = (w^{(1)}, ..., w^{(d)})$ to spherical polar coordinates and perform the change of variable $t = h^{-1}(u - \lambda r)$, to obtain that

$$\beta = A_d \left(\int_{\frac{u - m_0\lambda}{h}}^{\frac{u}{h}} \left(\lambda^{-1}(u - th)\right)^{d-1} K(t) \ (\gamma(u - th) - \gamma(u)) \ \lambda^{-1} h \ dt \right) .$$
$$\cdot \sum_{k=0}^{n-2} \frac{\binom{n-2}{k} \ \int f_0(\mathbf{w_1})^2 H(a, \mathbf{w_1})^k (1 - H(a, \mathbf{w_1}))^{n-2-k} d\mathbf{w_1}}{k+2}$$

Asymptotically, by using condition **(A4)**, the expression $(\gamma(u - th) - \gamma(u))$ may be reduced to the second term of its Taylor expansion, i.e. $\frac{\gamma''(u)}{2}(-th)^2$. Then, the dominant term in β becomes

$$\frac{1}{2} c_K \gamma''(u) \ u^{d-1} A_d \ \lambda^{-d} \ h^3 \quad \int f_0(\mathbf{w_1})^2 \ \sum_{k=0}^{n-2} \ \frac{\binom{n-2}{k} H(a, \mathbf{w_1})^k (1 - H(a, \mathbf{w_1}))^{n-2-k}}{k+2} \ d\mathbf{w_1}$$

where $c_K = \int z^2 K(z) dz$.

5.5 Variance of $\widehat{\gamma}(.)$ robust to clusters

For the analysis of the asymptotic efficiency, it is important now to proceed with the derivation of variance for the proposed variogram estimator. A decreasing variance estimate means a growing efficiency of the estimator, as it will tend to be more accurate.

Theorem 5.4 *Assume the hypotheses required in Theorem 5.1. Additionally, suppose that assumptions* **(A5)** *and* **(A7)** *are satisfied. Then, for* $u \geq Ch$*, one has*

$$\mathrm{Var}\left[\hat{\gamma}(u)\right] = \frac{B_d(u)\ d_K}{2u^{d-1}A_d^2}\ E_d(n,a)\ n^{-2}\lambda^d h^{-1}\ +\ \frac{C_d(u)}{A_d^2}\ F_d(n,a)\ n^{-1}\ +$$

$$+\ \frac{D_d(u)}{4A_d^2}\ G_d(n,a)\ \lambda^{-d}\ +\ o(n^{-2}\lambda^d h^{-1}\ +\ n^{-1}\ +\ \lambda^{-d} + h^4)$$

where $d_K = \int (K(z))^2 dz$ *and* A_d*,* $B_d(u)$*,* $C_d(u)$*,* $D_d(u)$*,* $E_d(n,a)$*,* $F_d(n,a)$ *and* $G_d(n,a)$ *are as given in (5.4), (5.9), (5.10), (5.11), (5.6), (5.7) and (5.8), respectively.*

Remark 5.5 *In a similar way as in Theorem 5.1, if we assume n sufficiently large then Theorem 5.4 holds for any* $u > 0$*.*

Let us start by considering that

$$\mathrm{Var}\left[\hat{\gamma}(u)\right] = \mathrm{Var}\left[\mathrm{E}\left[\hat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}\right]\right] + \mathrm{E}\left[\mathrm{Var}\left[\hat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}\right]\right] \tag{5.5}$$

By using Theorem 5.1 and Lemma 5.3, it is straightforward to see that for $u \geq Ch$

$$\mathrm{Var}\left[\mathrm{E}\left[\hat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}\right]\right] = o(h^4).$$

We need now to check that $\mathrm{Var}\left[\hat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}\right] = O(n^{-2}\lambda^d h^{-1} + n^{-1} + \lambda^{-d})$ and, again by Lemma 5.3, it will lead us to the convergence rate of

$$\mathrm{E}\left[\mathrm{Var}\left[\hat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}\right]\right].$$

Consequently, the convergence rate stated in Theorem 5.4 will be proved to be valid.

About the detailed expression obtained for the conditional variance, we have that

$$\begin{aligned}
&\mathrm{Var}\left[\hat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}\right] = \mathrm{E}\left[(\widehat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n} - \mathrm{E}[\widehat{\gamma}(u) \mid \mathbf{V_1}, ..., \mathbf{V_n}])^2\right] = \\
&= \mathrm{E}\left[\left(\frac{\sum_{i\neq j} w_{ij}(u)\left((Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2 - \mathrm{E}[(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2]\right)}{2\sum_{i\neq j} w_{ij}(u)}\right)\cdot\right. \\
&\left.\cdot\left(\frac{\sum_{k\neq l} w_{kl}(u)\left((Z(\mathbf{x_k}) - Z(\mathbf{x_l}))^2 - \mathrm{E}[(Z(\mathbf{x_k}) - Z(\mathbf{x_l}))^2]\right)}{2\sum_{k\neq l} w_{kl}(u)}\right)\right] = \\
&= \frac{\sum_{i\neq j} w_{ij}(u)\ \sum_{k\neq l} w_{kl}(u)\ \mathrm{Cov}\left[(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2, (Z(\mathbf{x_k}) - Z(\mathbf{x_l}))^2\right]}{\left(2\sum_{i\neq j} w_{ij}(u)\right)^2} = \\
&= (2a_1(u))^{-2} \sum_{\substack{i\neq j \\ k\neq l}} K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right)\ K\left(\frac{u - \|\mathbf{x_k} - \mathbf{x_l}\|}{h}\right). \\
&\cdot\frac{1}{\sqrt{n_i n_j}}\frac{1}{\sqrt{n_k n_l}}\ g\left(\mathbf{x_i} - \mathbf{x_j}, \mathbf{x_i} - \mathbf{x_k}, \mathbf{x_i} - \mathbf{x_l}\right) = \frac{2e_1(u) + 4e_2(u) + e_3(u)}{4(a_1(u))^2}
\end{aligned}$$

where

$$e_1(u) = \sum_{i\neq j} \frac{K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right)^2\ g(\mathbf{x_i} - \mathbf{x_j}, 0, \mathbf{x_i} - \mathbf{x_j})}{n_i n_j} \qquad \Leftarrow\ (i = k \wedge j = l)$$

$$e_2(u) = \sum_{\substack{i\neq j \\ j\neq l}} \frac{K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right) K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_l}\|}{h}\right)\ g(\mathbf{x_i} - \mathbf{x_j}, 0, \mathbf{x_i} - \mathbf{x_l})}{\sqrt{n_i n_j}\sqrt{n_i n_l}} \qquad \Leftarrow\ (i = k)$$

$$e_3(u) = \sum_{\substack{i\neq j,k,l \\ j\neq k,l \\ k\neq l}} \frac{K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right) K\left(\frac{u-\|\mathbf{x_k}-\mathbf{x_l}\|}{h}\right)\ g(\mathbf{x_i} - \mathbf{x_j}, \mathbf{x_i} - \mathbf{x_k}, \mathbf{x_i} - \mathbf{x_l})}{\sqrt{n_i n_j}\sqrt{n_k n_l}}$$

Then, according to the results from Section 5.4.1 about $a_1(u)$, and those from Sections 5.5.1, 5.5.2 and 5.5.3 about $e_1(u)$, $e_2(u)$ and $e_3(u)$, respectively, we obtain:

- $$\frac{e_1(u)}{2(a_1(u))^2} = \frac{B_d(u)\ d_K}{2u^{d-1}A_d^2}\ E_d(n,a)\ n^{-2}\lambda^d h^{-1}\ +\ o(n^{-2}\lambda^d h^{-1}) \quad \text{a.}s.$$

where

$$E_d(n,a) = \frac{\int f_0(\mathbf{w_1})^2\ \sum_{k=0}^{n-2} \frac{\binom{n-2}{k}\ H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-2-k}}{(k+2)^2}\ d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2\ \sum_{k=0}^{n-2} \frac{\binom{n-2}{k}\ H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-2-k}}{k+2}\ d\mathbf{w_1}\right)^2} \tag{5.6}$$

- $$\frac{e_2(u)}{(a_1(u))^2} = \frac{C_d(u)}{A_d^2}\ F_d(n,a)\ n^{-1}\ +\ o(n^{-1}) \quad \text{a.}s.$$

where

$$F_d(n,a) = \frac{\int f_0(\mathbf{w_1})^3\ \sum_{k=0}^{n-3} \frac{\binom{n-3}{k}\ H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-3-k}}{(k+3)^2}\ d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2\ \sum_{k=0}^{n-2} \frac{\binom{n-2}{k}\ H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-2-k}}{k+2}\ d\mathbf{w_1}\right)^2} \tag{5.7}$$

- $$\frac{e_3(u)}{4(a_1(u))^2} = \frac{D_d(u)}{4A_d^2}\ G_d(n,a)\ \lambda^{-d}\ +\ o(\lambda^{-d}) \quad \text{a.}s.$$

where

$$G_d(n,a) = \frac{\int f_0(\mathbf{w_1})^4\ \sum_{k=0}^{n-4} \frac{\binom{n-4}{k}\ H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-4-k}}{(k+4)^2}\ d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2\ \sum_{k=0}^{n-2} \frac{\binom{n-2}{k}\ H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-2-k}}{k+2}\ d\mathbf{w_1}\right)^2} \tag{5.8}$$

The validity of the latter expressions demand that E_d, F_d and G_d are bounded, even for large n, which is proved numerically in Section 5.7.2.

For the specific case of $I\!R^2$, a supplemental simulation study for the dependency analysis of E_2, F_2 and G_2 on n was included in Section 5.7.3.

5.5.1 Order of $e_1(u)$ for $u \geq Ch$

The dominant term of $e_1(u)$ is given by $n^2\alpha_1$, where

$$\alpha_1 = \mathrm{E}\left[\frac{K\left(\frac{u-\lambda\|\mathbf{V_1}-\mathbf{V_2}\|}{h}\right)^2 \; g(\lambda(\mathbf{V_1}-\mathbf{V_2}),0,\lambda(\mathbf{V_1}-\mathbf{V_2}))}{\sum_{k_1,k_2} I_{\{\|\mathbf{V_1}-\mathbf{V_{k_1}}\|\leq a,\|\mathbf{V_2}-\mathbf{V_{k_2}}\|\leq a\}}}\right] =$$

$$= \int\cdots\int \frac{K\left(\frac{u-\lambda\|\mathbf{w_2}\|}{h}\right)^2 g(\lambda\mathbf{w_2},0,\lambda\mathbf{w_2})\; f_0(\mathbf{w_1})^2 f_0(\mathbf{w_1}-\mathbf{w_3})...f_0(\mathbf{w_1}-\mathbf{w_n})}{4+5\sum_{k_1\geq 3} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{\substack{k_1,k_2\geq 3\\ k_1\neq k_2}} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} I_{\{\|\mathbf{w_{k_2}}\|\leq a\}}} \, d\mathbf{w_1}...d\mathbf{w_n}$$

As in Section 5.4.1, we may convert $\mathbf{w_2}$ to spherical polar coordinates and make a change of variable to obtain

$$\alpha_1 = \int_{\frac{u-m_0\lambda}{h}}^{\frac{u}{h}} \int_0^\pi ... \int_0^\pi \int_0^{2\pi} J_d\left(\theta_1,...,\theta_{d-1}\right)\left(\lambda^{-1}(u-th)\right)^{d-1} K(t)^2 \; \lambda^{-1}h.$$

$$.g\left((u-th)(\cos\theta_1,...,\prod_{j=0}^{d-1}\sin\theta_j),\;0,\;(u-th)(\cos\theta_1,...,\prod_{j=0}^{d-1}\sin\theta_j)\right)\; dt\; d\theta_1...d\theta_{d-1}.$$

$$.\sum_{k=0}^{n-2} \frac{\binom{n-2}{k} \quad \int f_0(\mathbf{w_1})^2 H(a,\mathbf{w_1})^k (1-H(a,\mathbf{w_1}))^{n-2-k} d\mathbf{w_1}}{(k+2)^2}$$

The dominant term will be given by

$$\alpha_1 = u^{d-1} B_d(u)\; d_K\; \lambda^{-d}\; h \;\; \int f_0(\mathbf{w_1})^2 \sum_{k=0}^{n-2} \frac{\binom{n-2}{k} H(a,\mathbf{w_1})^k (1-H(a,\mathbf{w_1}))^{n-2-k}}{(k+2)^2} \, d\mathbf{w_1}$$

where $d_K = \int (K(z))^2 dz$ and

$$B_d(u) = \int_0^\pi ... \int_0^\pi \int_0^{2\pi} J_d(\theta_1,...,\theta_{d-1}).$$

$$.g\left(u(\cos\theta_1,...,\prod_{j=0}^{d-1}\sin\theta_j),\;0,\;u(\cos\theta_1,...,\prod_{j=0}^{d-1}\sin\theta_j)\right) d\theta_1...d\theta_{d-1} \qquad (5.9)$$

5.5.2 Order of $e_2(u)$ for $u \geq Ch$

In here, three distinct indices i, j, l, are involved, thus the dominant term of $e_2(u)$ will be given by $n^3\alpha_2$, where

$$\alpha_2 = \mathrm{E}\left[\frac{K\left(\frac{u-\lambda\|\mathbf{V_1}-\mathbf{V_2}\|}{h}\right)K\left(\frac{u-\lambda\|\mathbf{V_1}-\mathbf{V_3}\|}{h}\right)\ g(\lambda(\mathbf{V_1}-\mathbf{V_2}),\ 0,\ \lambda(\mathbf{V_1}-\mathbf{V_3}))}{\sqrt{\sum_{k_1,k_2} I_{\{\|\mathbf{V_1}-\mathbf{V_{k_1}}\|\leq a,\|\mathbf{V_2}-\mathbf{V_{k_2}}\|\leq a\}}}\ \sqrt{\sum_{k_1,k_2} I_{\{\|\mathbf{V_1}-\mathbf{V_{k_1}}\|\leq a,\|\mathbf{V_3}-\mathbf{V_{k_2}}\|\leq a\}}}}\right]$$

For random variables $\mathbf{W_i} = \mathbf{V_1} - \mathbf{V_i},\ i = 2, 3,$ as K is compactly supported, we shall show that the dominant term of the expectation above can be reduced to those values $\|\mathbf{w_2}\|$ and $\|\mathbf{w_3}\|$ tending to 0. Then, it becomes

$$\alpha_2 = \int ... \int \frac{K\left(\frac{u-\lambda\|\mathbf{w_2}\|}{h}\right)K\left(\frac{u-\lambda\|\mathbf{w_3}\|}{h}\right)\ g(\lambda\mathbf{w_2},\ 0,\ \lambda\mathbf{w_3})}{\sqrt{1+2\sum_{k_1\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{k_1,k_2\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a,\|\mathbf{w_{k_2}}-\mathbf{w_2}\|\leq a\}}}}\cdot$$

$$\cdot\frac{f_0(\mathbf{w_1})f_0(\mathbf{w_1}-\mathbf{w_2})...f_0(\mathbf{w_1}-\mathbf{w_n})}{\sqrt{1+2\sum_{k_1\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{k_1,k_2\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a,\|\mathbf{w_{k_2}}-\mathbf{w_3}\|\leq a\}}}}\ d\mathbf{w_1}...d\mathbf{w_n}$$

The square root can actually be eliminated, as

$$\sum_{k_1,k_2\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a,\|\mathbf{w_{k_2}}-\mathbf{w_3}\|\leq a\}} = \sum_{k_1,k_2\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a,\|\mathbf{w_{k_2}}-\mathbf{w_2}\|\leq a\}} = \sum_{k_1,k_2\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a,\|\mathbf{w_{k_2}}\|\leq a\}}$$

Furthermore

$$1+2\sum_{k_1\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{k_1,k_2\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a,\|\mathbf{w_{k_2}}\|\leq a\}} =$$

$$= \left(1+2\sum_{k_1\geq 2} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{\substack{k_1=k_2\\ k_1\geq 2}} I_{\{\|\mathbf{w_{k_1}}\|\leq a,\|\mathbf{w_{k_1}}\|\leq a\}}\right) + \left(\sum_{\substack{k_1\neq k_2\\ k_1,k_2\geq 2}} I_{\{\|\mathbf{w_{k_1}}\|\leq a,\|\mathbf{w_{k_2}}\|\leq a\}}\right) =$$

$$= \left(7+3\sum_{k_1\geq 4} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}}\right) + \left(2+4\sum_{k_1\geq 4} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{\substack{k_1,k_2\geq 4\\ k_1\neq}} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} I_{\{\|\mathbf{w_{k_2}}\|\leq a\}}\right)$$

According to previous results

$$\alpha_2 = \int ... \int \frac{K\left(\frac{u-\lambda\|\mathbf{w_2}\|}{h}\right) K\left(\frac{u-\lambda\|\mathbf{w_3}\|}{h}\right)\ g(\lambda\mathbf{w_2},\ 0,\ \lambda\mathbf{w_3})}{9+7\sum_{k_1\geq 4} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} + \sum_{k_1,k_2\geq 4, k_1\neq k_2} I_{\{\|\mathbf{w_{k_1}}\|\leq a\}} I_{\{\|\mathbf{w_{k_2}}\|\leq a\}}}.$$
$$.f_0(\mathbf{w_1})^3 f_0(\mathbf{w_1}-\mathbf{w_4})...f_0(\mathbf{w_1}-\mathbf{w_n})\ d\mathbf{w_1}...d\mathbf{w_n}$$

Now, convert $\mathbf{w_2} = (w^{(1,1)}, ..., w^{(d,1)})$ and $\mathbf{w_3} = (w^{(1,2)}, ..., w^{(d,2)})$ to spherical polar coordinates with the transformation

$$w^{(i,1)} = r_1 \cos\theta_{i,1} \prod_{j=0}^{i-1} \sin\theta_{j,1} \qquad and \qquad w^{(i,2)} = r_2 \cos\theta_{i,2} \prod_{j=0}^{i-1} \sin\theta_{j,2}$$

where, for $k = 1, 2$, $\sin\theta_{0,k} = \cos\theta_{d,k} = 1$, $0 \leq \theta_{d-1,k} < 2\pi$ and $0 \leq \theta_{i,k} < \pi$, for $i = 1, ..., d-2$. The corresponding Jacobian transformations are given by

$$r_k^{d-1} J_d(\theta_{1,k}, ..., \theta_{d-1,2}) = r_k^{d-1} (\sin\theta_{1,k})^{d-2} (\sin\theta_{2,k})^{d-3} ...\ \sin\theta_{d-2,k}.$$

In this way,

$$\alpha_2 = \left(\int_0^{m_0} \int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} \int_0^{m_0} \int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} r_1^{d-1} J_d\left(\theta_{1,1}, ..., \theta_{d-1,1}\right) r_2^{d-1}\ .\right.$$
$$.J_d\left(\theta_{1,2}...\theta_{d-1,2}\right)\ g\left(\lambda r_1(\cos\theta_{1,1}, ..., \prod_{j=0}^{d-1}\sin\theta_{j,1}),\ 0,\ \lambda r_2(\cos\theta_{1,2}, ..., \prod_{j=0}^{d-1}\sin\theta_{j,2})\right).$$
$$\left. .\ K\left(\frac{u-\lambda r_1}{h}\right)\ K\left(\frac{u-\lambda r_2}{h}\right)\ dr_1 dr_2 d\theta_{1,1}...d\theta_{d-1,1} d\theta_{1,2}...d\theta_{d-1,2} \right).$$
$$.\sum_{k=0}^{n-3} \frac{\binom{n-3}{k} \int f_0(\mathbf{w_1})^3 H(a,\mathbf{w_1})^k (1-H(a,\mathbf{w_1}))^{n-3-k} d\mathbf{w_1}}{9+7k+k(k-1)}$$

Finally, with the following changes of variable, for $k = 1, 2$

$$t_k = h^{-1}(u-\lambda r_k) \quad \Rightarrow \quad r_k = \lambda^{-1}(u-t_k h) \quad \Rightarrow \quad dr_k = -\lambda^{-1} h\ dt_k$$

the dominant term becomes

$$\alpha_2 = \lambda^{-2}h^2 \int_0^{\pi}...\int_0^{\pi}\int_0^{2\pi}\int_0^{\pi}...\int_0^{\pi}\int_0^{2\pi} J_d\left(\theta_{1,1},...,\theta_{d-1,1}\right) J_d\left(\theta_{1,2}...\theta_{d-1,2}\right).$$

$$.g\left(u(\cos\theta_{1,1},...,\prod_{j=0}^{d-1}\sin\theta_{j,1}),\ 0,\ u(\cos\theta_{1,2},...,\prod_{j=0}^{d-1}\sin\theta_{j,2})\right) d\theta_{1,1}...d\theta_{d-1,1}d\theta_{1,2}...d\theta_{d-1,2}.$$

$$.\left(\int_{\frac{u-\lambda m_0}{h}}^{\frac{u}{h}}\int_{\frac{u-\lambda m_0}{h}}^{\frac{u}{h}}(\lambda^{-1}(u-t_1h))^{d-1}(\lambda^{-1}(u-t_2h))^{d-1}\ K(t_1)K(t_2)\ dt_1dt_2\right).$$

$$.\sum_{k=0}^{n-3}\frac{\binom{n-3}{k}\int f_0(\mathbf{w_1})^3H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-3-k}d\mathbf{w_1}}{9+7k+k(k-1)} =$$

$$= u^{2(d-1)}C_d(u)\ \lambda^{-2d}\ h^2 \int f_0(\mathbf{w_1})^3\sum_{k=0}^{n-3}\frac{\binom{n-3}{k}H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-3-k}}{(k+3)^2}\ d\mathbf{w_1}$$

where the integral over variable $\mathbf{w_1}$ is bounded due to condition **(A3)**, and

$$C_d(u) = \int_0^{\pi}...\int_0^{\pi}\int_0^{2\pi}\int_0^{\pi}...\int_0^{\pi}\int_0^{2\pi} J_d\left(\theta_{1,1},...,\theta_{d-1,1}\right) J_d\left(\theta_{1,2}...\theta_{d-1,2}\right).$$

$$.g\left(u(\cos\theta_{1,1},...,\prod_{j=0}^{d-1}\sin\theta_{j,1}),\ 0,\ u(\cos\theta_{1,2},...,\prod_{j=0}^{d-1}\sin\theta_{j,2})\right).$$

$$.d\theta_{1,1}...d\theta_{d-1,1}d\theta_{1,2}...d\theta_{d-1,2} \tag{5.10}$$

5.5.3 Order of $e_3(u)$ for $u \geq Ch$

In here, four distinct indices i, j, k, l, are involved, thus the dominant term of $e_3(u)$ will be given by $n^4\alpha_3$, where α_3 is equal to

$$\mathrm{E}\left[\frac{K\left(\frac{u-\lambda\|\mathbf{V_1}-\mathbf{V_2}\|}{h}\right)K\left(\frac{u-\lambda\|\mathbf{V_3}-\mathbf{V_4}\|}{h}\right)\ g(\lambda(\mathbf{V_1}-\mathbf{V_2}),\ \lambda(\mathbf{V_1}-\mathbf{V_3}),\ \lambda(\mathbf{V_1}-\mathbf{V_4}))}{\sqrt{\sum_{k_1,k_2}I_{\{\|\mathbf{V_1}-\mathbf{V_{k_1}}\|\leq a,\|\mathbf{V_2}-\mathbf{V_{k_2}}\|\leq a\}}}\ \sqrt{\sum_{k_1,k_2}I_{\{\|\mathbf{V_3}-\mathbf{V_{k_1}}\|\leq a,\|\mathbf{V_4}-\mathbf{V_{k_2}}\|\leq a\}}}}\right]$$

One will have $\lim_{\lambda\to\infty}\|\mathbf{w_2}\|=0$ and $\lim_{\lambda\to\infty}\|\mathbf{w_4}-\mathbf{w_3}\|=0$, as K is compactly supported. The second convergence rate allow us to define $A=I_{\{\|\mathbf{w_4}\|\leq a\}}=I_{\{\|\mathbf{w_3}\|\leq a\}}$. Then, it becomes

$$\alpha_3=\int\dots\int\frac{K\left(\frac{u-\lambda\|\mathbf{w_2}\|}{h}\right)K\left(\frac{u-\lambda\|\mathbf{w_4}-\mathbf{w_3}\|}{h}\right)\ g(\lambda\mathbf{w_2},\ \lambda\mathbf{w_3},\ \lambda\mathbf{w_4})}{\sqrt{4+12A+(5+4A)\sum_{k_1\geq 5}I_{\{\|\mathbf{w_{k_1}}\|\leq a\}}+\sum_{\substack{k_1,k_2\geq 5\\ k_1\neq k_2}}I_{\{\|\mathbf{w_{k_1}}\|\leq a\}}I_{\{\|\mathbf{w_{k_2}}\|\leq a\}}}}\cdot$$

$$\cdot\frac{f_0(\mathbf{w_1})^2f_0(\mathbf{w_1}-\mathbf{w_4})^2f_0(\mathbf{w_1}-\mathbf{w_5})\dots f_0(\mathbf{w_1}-\mathbf{w_n})\ d\mathbf{w_1}d\mathbf{w_2}\dots d\mathbf{w_n}}{\sqrt{4+12A+(5+4A)\sum_{k_1\geq 5}I_{\{\|\mathbf{w_{k_1}}-\mathbf{w_4}\|\leq a\}}+\sum_{\substack{k_1,k_2\geq 5\\ k_1\neq k_2}}I_{\{\|\mathbf{w_{k_1}}-\mathbf{w_4}\|\leq a\}}I_{\{\|\mathbf{w_{k_2}}-\mathbf{w_4}\|\leq a\}}}}$$

Before proceeding, we propose an additional assumption to allow us to consider the previous two expressions under the square roots as asymptotically equivalent. In fact, the third argument of the $g(.)$ function is $\lambda\mathbf{w_4}$ and, according to hypothesis (**A5**), it is reasonable to assume that $\lambda\|\mathbf{w_4}\|<const$, where $const>r$. Then, $\|\mathbf{w_4}\|<\lambda^{-1}const$ and $\lim_{\lambda\to\infty}\|\mathbf{w_4}\|=0$. Asymptotically, this allows us to eliminate the square root and to consider $f_0(\mathbf{w_1}-\mathbf{w_4})=f_0(\mathbf{w_1})$

The usual conversion to spherical polar coordinates can now be applied

$$\begin{aligned}
&\mathbf{w_2}=(w^{(1,1)},\dots,w^{(d,1)}) && \text{where} \quad w^{(i,1)}=r_1\cos\theta_{i,1}\textstyle\prod_{j=0}^{i-1}\sin\theta_{j,1}\\
&\mathbf{w_4}-\mathbf{w_3}=(w^{(1,2)},\dots,w^{(d,2)}) && \text{where} \quad w^{(i,2)}=r_2\cos\theta_{i,2}\textstyle\prod_{j=0}^{i-1}\sin\theta_{j,2}\\
&\mathbf{w_4}=(w^{(1,3)},\dots,w^{(d,3)}) && \text{where} \quad w^{(i,3)}=r_3\cos\theta_{i,3}\textstyle\prod_{j=0}^{i-1}\sin\theta_{j,3}
\end{aligned}$$

and, consequently, $\mathbf{w_3}=\mathbf{w_4}-(\mathbf{w_4}-\mathbf{w_3})$.

Bearing in mind that $I_{\{\|\mathbf{w_4}\|\leq a\}}=\begin{cases}1 & \text{if } 0\leq\|\mathbf{w_4}\|\leq a\\ 0 & \text{if } \|\mathbf{w_4}\|>a,\end{cases}$

where a is a bounded value (see assumption **(A7)**), and $g(.)$ is asymptotically equal to zero for $\|\mathbf{w_4}\|>a$ according to assumption **(A5)**, then the dominant term in α_3 is given by

$$\int_0^a\int_0^\pi\dots\int_0^\pi\int_0^{2\pi}\int_0^{m_0}\int_0^\pi\dots\int_0^\pi\int_0^{2\pi}\int_0^{m_0}\int_0^\pi\dots\int_0^\pi\int_0^{2\pi}r_1^{d-1}J_d\left(\theta_{1,1},\dots,\theta_{d-1,1}\right).$$

$$.r_2^{d-1} J_d\left(\theta_{1,2}...\theta_{d-1,2}\right)\ r_3^{d-1} J_d\left(\theta_{1,3}...\theta_{d-1,3}\right)\ K\left(\frac{u-\lambda r_1}{h}\right).$$

$$.\ K\left(\frac{u-\lambda r_2}{h}\right)\ g\left(\lambda r_1(\cos\theta_{1,1},...,\prod_{j=0}^{d-1}\sin\theta_{j,1}),\right.$$

$$\lambda(r_3\cos\theta_{1,3}-r_2\cos\theta_{1,2},...,r_3\prod_{j=0}^{d-1}\sin\theta_{j,3}-r_2\prod_{j=0}^{d-1}\sin\theta_{j,2}),$$

$$\left.\lambda r_3(\cos\theta_{1,3},...,\prod_{j=0}^{d-1}\sin\theta_{j,3})\right)\ dr_1dr_2dr_3d\theta_{1,1}...d\theta_{d-1,1}d\theta_{1,2}...d\theta_{d-1,2}\theta_{1,3}...d\theta_{d-1,3}.$$

$$.\int f_0(\mathbf{w_1})^2 f_0(\mathbf{w_1})^2\ \sum_{k=0}^{n-4}\frac{\binom{n-4}{k}\ H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-4-k}}{16+9k+k(k-1)}\ d\mathbf{w_1}$$

Consequently, by using the following changes of variable

$$\begin{array}{lllll} t_1=\dfrac{u-\lambda r_1}{h} & \Rightarrow & r_1=\lambda^{-1}(u-t_1h) & \Rightarrow & dr_1=-\lambda^{-1}h\ dt_1 \\ t_2=\dfrac{u-\lambda r_2}{h} & \Rightarrow & r_2=\lambda^{-1}(u-t_2h) & \Rightarrow & dr_2=-\lambda^{-1}h\ dt_2 \\ t_3=\lambda r_3 & \Rightarrow & r_3=\lambda^{-1}t_3 & \Rightarrow & dr_3=\lambda^{-1}dt_3, \end{array}$$

we may obtain that α_3 is equal to

$$\lambda^{-3}h^2\int_0^{\lambda a}\int_0^{\pi}...\int_0^{\pi}\int_0^{2\pi}\int_{\frac{u-\lambda m_0}{h}}^{\frac{u}{h}}\int_0^{\pi}...\int_0^{\pi}\int_0^{2\pi}\int_{\frac{u-\lambda m_0}{h}}^{\frac{u}{h}}\int_0^{\pi}...\int_0^{\pi}\int_0^{2\pi}(\lambda^{-1}t_3)^{d-1}.$$

$$.(\lambda^{-1}(u-t_1h))^{d-1}(\lambda^{-1}(u-t_2h))^{d-1}J_d\left(\theta_{1,1},...,\theta_{d-1,1}\right).$$

$$.J_d\left(\theta_{1,2}...\theta_{d-1,2}\right)J_d\left(\theta_{1,3}...\theta_{d-1,3}\right)K(t_1)K(t_2).$$

$$.g\left(u(\cos\theta_{1,1},...,\prod_{j=0}^{d-1}\sin\theta_{j,1}),\ (t_3\cos\theta_{1,3}-u\cos\theta_{1,2},...,t_3\prod_{j=0}^{d-1}\sin\theta_{j,3}-u\prod_{j=0}^{d-1}\sin\theta_{j,2}),\right.$$

$$\left.t_3(\cos\theta_{1,3},...,\prod_{j=0}^{d-1}\sin\theta_{j,3})\right)\ dt_1dt_2dt_3d\theta_{1,1}...d\theta_{d-1,1}d\theta_{1,2}...d\theta_{d-1,2}\theta_{1,3}...d\theta_{d-1,3}.$$

$$.\int f_0(\mathbf{w_1})^4\sum_{k=0}^{n-4}\frac{\binom{n-4}{k}\ H(a,\mathbf{w_1})^k(1-H(a,\mathbf{w_1}))^{n-4-k}}{(k+4)^2}\ d\mathbf{w_1}$$

Finally, the dominant term in α_3 becomes

$$u^{2(d-1)} D_d(u)\ \lambda^{-3d}\ h^2 \quad \int f_0(\mathbf{w_1})^4 \sum_{k=0}^{n-4} \frac{\binom{n-4}{k} H(a,\mathbf{w_1})^k (1-H(a,\mathbf{w_1}))^{n-4-k}}{(k+4)^2}\ d\mathbf{w_1}$$

where the integral over variable $\mathbf{w_1}$ is bounded due to condition **(A3)**, and

$$D_d(u) = \int_0^{\delta}\int_0^{\pi} ... \int_0^{\pi}\int_0^{2\pi}\int_0^{\pi} ... \int_0^{\pi}\int_0^{2\pi}\int_0^{\pi} ... \int_0^{\pi}\int_0^{2\pi} J_d\left(\theta_{1,1},...,\theta_{d-1,1}\right).$$

$$.J_d\left(\theta_{1,2}...\theta_{d-1,2}\right) J_d\left(\theta_{1,3}...\theta_{d-1,3}\right)\ t^{d-1}\ g\left(u(\cos\theta_{1,1},...,\prod_{j=0}^{d-1}\sin\theta_{j,1}),\right.$$

$$\left.(t\cos\theta_{1,3} - u\cos\theta_{1,2},...,t\prod_{j=0}^{d-1}\sin\theta_{j,3} - u\prod_{j=0}^{d-1}\sin\theta_{j,2}),\ t(\cos\theta_{1,3},...,\prod_{j=0}^{d-1}\sin\theta_{j,3})\right).$$

$$.dtd\theta_{1,1}...d\theta_{d-1,1}d\theta_{1,2}...d\theta_{d-1,2}d\theta_{1,3}...d\theta_{d-1,3} \tag{5.11}$$

5.6 Kernel bandwidth selector

Our goal here is to use the information available in the sampled data for optimal approximation of the kernel bandwidth h, in the sense of providing values of $\widehat{\gamma}$ close to the true values of γ. We need, then, some definition of being "close". One of the most common measures is Mean Square Error, or MSE, which is defined as

$$\mathrm{MSE}[\hat{\gamma}(u)] = \mathrm{E}\left[(\widehat{\gamma}(u) - \gamma(u))^2\right].$$

Hence, we propose the selection of the optimal kernel bandwidth, as the value that minimizes the MSE function, originating a local bandwidth selector.

Then, according to previous results, one has

$$\mathrm{MSE}\left[\hat{\gamma}(u)\right] = \left(\mathrm{Bias}\left[\hat{\gamma}(u)\right]\right)^2 \ +\ \mathrm{Var}\left[\hat{\gamma}(u)\right] \ \simeq\ \frac{c_K^2 \gamma''(u)^2}{4} h^4 \ +$$

$$+\ \frac{B_d(u)\ d_K\ E_d(n,a)}{2u^{d-1}A_d^2}\ n^{-2}\lambda^d h^{-1} \ +\ \frac{C_d(u)\ F_d(n,a)}{A_d^2}\ n^{-1} \ +\ \frac{D_d(u)\ G_d(n,a)}{4A_d^2}\ \lambda^{-d}$$

From here, for $u \geq Ch$, the bandwidth parameter that asymptotically minimizes the MSE $[\hat{\gamma}(u)]$ becomes

$$h_{opt}(u) = \left[\frac{B_d(u)\ d_K\ E_d(n,a)}{2\ u^{d-1} A_d^2\ c_K^2 \gamma''(u)^2} \right]^{1/5} n^{-2/5} \lambda^{d/5}$$

Remark 5.6 *Alternatively, one might deal with a global bandwidth parameter, by minimizing the Mean Integrated Square Error, or MISE, defined as*

$$\text{MISE}[\hat{\gamma}(u)] = \int_R (\text{MSE}[\hat{\gamma}(u)])\ du = \int_R (\text{Bias}\,[\hat{\gamma}(u)])^2\ du\ +\ \int_R \text{Var}\,[\hat{\gamma}(u)]\ du$$

for some $R \subset [0, +\infty)$. *For instance, we may take* $R = [m_0, m]$, *where* $m = \sup\{\|\mathbf{x_i} - \mathbf{x_j}\| : \mathbf{x_i}, \mathbf{x_j} \in D\}$ *and some constant* m_0, $0 < m_0 < m$. *The resulting optimal bandwidth would be*

$$h_{opt} = \left[\frac{\int_R \frac{B_d(u)}{u^{d-1}} du\ d_K\ E_d(n,a)}{2\ A_d^2\ c_K^2 \int_R \gamma''(u)^2 du} \right]^{1/5} n^{-2/5} \lambda^{d/5}$$

where $c_K = \int z^2 K(z) dz$, $d_K = \int (K(z))^2 dz$ *and* A_d, $B_d(u)$ *and* $E_d(n,a)$ *are as given in (5.4), (5.9) and (5.6), respectively.*

Both derived local and global bandwidth expressions involve the unknown function $\gamma(u)$. For this purpose, a simple parametric approach, like one of those described in Chapter 3 (for example, first entry of Table 3.1), may be used to estimate $\gamma(u)$. This parametric estimation can be improved by being incorporated into an iterated non-parametric procedure.

The fundamental idea behind our asymptotic study is that the observation region may be considered expansible, in $D_n = \lambda_n D_0$ (see assumption **(A1)**). As the region D_n grows, more locations are expected. As a consequence of stationarity, the spatial dependency, described by the theoretical γ, is kept unchanged. The

estimate of γ may suffer some impact resulting from a larger sample size, due to the fact that new distances may be used in the estimation process.

This reasoning of expansion must occur in a controlled way. A possible negative effect of adopting a too large expansion rate is to make the "radius of influence", known as the range, into a negligible value when compared to the maximum distance on D_n. This would lead us to a meaningless spatial dependency on the observation region.

With respect to the convergence rates stated in **(A6)**, furthermore, we can take

$$\lambda^d = c_1 n^{c_0} + o(n^{c_0}) \tag{5.12}$$

for some constants $c_0 > 0$ and $c_1 > 0$. The expansion rate is established by constant c_0, i.e. a smaller value for c_0 means a slower expansion for D_n.

From here, the bandwidth parameter that asymptotically minimizes the MSE $[\hat{\gamma}(u)]$ becomes

$$h_{opt}(u) = \left[c_1 \frac{B_d(u)\ d_K\ E_d(n,a)}{2\ u^{d-1} A_d^2\ c_K^2 \gamma''(u)^2} \right]^{1/5}\ n^{-(2-c_0)/5} \tag{5.13}$$

With this selection of the bandwidth parameter, it follows that

$$\text{MSE}\,[\hat{\gamma}(u)] = O(n^{-4(2-c_0)/5}\ I_{\{c_0 > \frac{8}{9}\}} + n^{-c_0}\ I_{\{c_0 \leq \frac{8}{9}\}})$$

and the minimum order is achieved for $c_0 = \frac{8}{9}$.

5.6.1 Order of variance

Bear in mind the relation (5.12), $\lambda^d = O(n^{c_0})$, and suppose O_1, O_2 and O_3 identify the convergence rates of the first, second and third terms of the conditional variance, respectively. Then

$$\begin{aligned} O_1 &= O(n^{-2}\lambda^d h^{-1}) = O(n^{-(2-c_0)} h^{-1}) \\ O_2 &= O(n^{-1}) \\ O_3 &= O(\lambda^{-d}) = O(n^{-c_0}). \end{aligned}$$

As a result from **(A6)**, one has $0 < c_0 < 1$, which conveys that O_2 is of a less order than O_1 and O_3 and, therefore, the contribution of O_2 to the variance is asymptotically negligible. As a result of this, an alternative expression for the variance may be written as below.

Corollary 5.7 *Assume the hypotheses required in Theorem 5.4. Additionally, suppose the convergence rate announced in (5.12) is satisfied. Then, for $u \geq Ch$, one has*

$$\mathrm{Var}\,[\hat{\gamma}(u)] = \frac{B_d(u)\ d_K}{2u^{d-1}A_d^2}\ E_d(n,a)\ n^{-2}\lambda^d h^{-1}\ +\ \frac{D_d(u)}{4A_d^2}\ G_d(n,a)\ \lambda^{-d}\ +$$

$$+\ o\left(n^{-2}\lambda^d h^{-1}\ +\ \lambda^{-d}\ +\ h^4\right)$$

where $d_K = \int (K(z))^2 dz$ and A_d, $B_d(u)$, $D_d(u)$, $E_d(n,a)$ and $G_d(n,a)$ are as given in (5.4), (5.9), (5.11), (5.6) and (5.8), respectively.

5.7 Numerical studies

We end this Chapter describing three simulation experiments related to the study of properties of the proposed $\widehat{\gamma}(.)$ for clustered data. The first simulation aims mainly to illustrate how to apply the new estimator, while checking its performance. Implementation details about the local bandwidth derivation are given.

The other two simulation studies were required at some point of the course of our proofs, to show numerically the given expression is bounded.

5.7.1 Performance of $\widehat{\gamma}(.)$ robust to clusters

In order to analyse the performance of our estimator simulations of spatial data in $I\!R^2$ were carried out. Gaussian data were generated on the observation region $D \subset I\!R^2$ by selecting a theoretical variogram model to specify the spatial correlation.

The region D is assumed to be equal to λD_0, where D_0 is the bounded and fixed square unit. The new estimator is compared against the estimator of Matheron and the one using the Nadaraya-Watson kernel, given in (3.1) and (3.3), respectively. The symmetric Epanechnikov kernel was employed in the two previous kernel-type estimators.

To obtain the optimal local bandwidth, we considered the optimal $c_0 = \frac{8}{9}$, and $c_1 = 1$. Then, this local bandwidth can be derived as a function of lag u, like

$$h_{opt}(u) = \left[\frac{B_2(u)\ d_K\ E_2(n,a)}{2\ uA_2^2\ c_K^2 \gamma''(u)^2} \right]^{1/5} n^{-2/9}.$$

The corresponding scale factor, from (5.12), is given by $\lambda = (n^{8/9})^{1/2}$. We considered a sample size $n = 100$ and a theoretical exponential variogram with a nugget effect of 0.6, a sill of 1.336 and the corresponding range equal to 5.0.

About the bandwidth derivation, note that c_K identifies the variance of the Epanechnikov kernel and d_K identifies the integral of this squared kernel. As we are in IR^2, the A_2 expression in (5.4) is reduced to 2π. For the Gaussian case, $B_2(u)$ given in (5.9) can be approximated by $8A_2\gamma(u)^2$. To estimate $E_2(n,a)$, given in (5.6), the existing integrals were numerically approximated to a sample average, as

$$\int g(x)dx = \mathrm{E}\left[\frac{g(X)}{f(X)}\right], \quad \text{where } f(x) \text{ is the density function of } X.$$

Section 5.7.3 provides more detail about $E_2(n,a)$ estimation. Additionally, as the bandwidth derivation needs itself an estimation of the variogram, a rough parametric estimation was used for this purpose.

To proceed with our simulation study, we generate a total of 100 independent data sets and, for each one, derive the integrated square error (ISE) between the estimator and the theoretical variogram. The ISE, defined as $\int_\alpha^\beta [\widehat{\gamma}(u) - \gamma(u)]^2\, du$, was approximated numerically through the trapezoid rule. In Table 5.1, the mean values of the resulting ISEs are compared for two distinct sampling designs:

- A CSR model, where points are uniformly distributed on D;

		$u \leq 0.6\lambda$	$u \leq 0.3\lambda$	$u \leq 0.2\lambda$	$u \leq 0.1\lambda$
CSR	Matheron	1.270	0.943	0.819	0.763
	NW kernel	0.527	0.314	0.276	0.291
	RobCluster	0.500	0.307	0.276	0.298
CLUSTER	Matheron	1.519	1.141	0.889	0.568
	NW kernel	0.582	0.525	0.488	0.400
	RobCluster	0.392	0.294	0.243	0.245

Table 5.1: Mean values of the standardized ISEs, from the empirical estimators. The total number of replicas is 100 and in each replica the total sample size is 100.

- A clustered model, where 40% of the total points are gathered together into one sub-region of D.

As the observation region D depends on λ, we decided to group the mean values of the ISEs into four classes of lags: $(0, 0.6\lambda)$, $(0, 0.3\lambda)$, $(0, 0.2\lambda)$ and $(0, 0.1\lambda)$. To easily compare columns, all ISE values were standardized by dividing them by the corresponding integral interval, $\beta - \alpha$.

According to Table 5.1, the new empirical estimator, named "RobCluster", offers a better performance in the presence of clustered data. Under a CSR model, the Nadaraya-Watson kernel estimator and the new estimator present similar results, and better than those from Matheron's proposal.

We repeat this same experiment with the corresponding valid versions of the previous three empirical estimators, after fitting them to a permissible variogram defined by *Bochner's theorem* in equation (3.4) (see Chapter 3). In Table 5.2, we summarize the mean values of the obtained ISE. Once more, one may confirm the better behaviour of the proposed variogram under clustered data.

Alternatively, one might work with a global bandwidth (see Remark 5.6). In this case, the optimal expression for bandwidth h does not depend on lag u, as it depends instead on some integrals of u. Bear in mind, a global bandwidth is

		$u \leq 0.6\lambda$	$u \leq 0.3\lambda$	$u \leq 0.2\lambda$	$u \leq 0.1\lambda$
CSR	Matheron	0.861	0.672	0.609	0.577
	NW kernel	0.504	0.285	0.223	0.201
	RobCluster	0.479	0.279	0.231	0.214
CLUSTER	Matheron	0.925	0.848	0.764	0.519
	NW kernel	0.490	0.475	0.473	0.385
	RobCluster	0.364	0.278	0.229	0.217

Table 5.2: Mean values of the standardized ISE, from valid estimators. The total number of replicas is 100 and for each replica the total sample size is 100.

expected to lead to faster simulations when compared to a local one, as it avoids a specific estimation for each lag u. The natural drawback is that it proposes a less accurate solution.

5.7.2 Analysis of $E_d(n,a)$, $F_d(n,a)$ and $G_d(n,a)$ for large n

The goal of the current simulation study is to understand how E_d, F_d and G_d, given in (5.6), (5.7) and (5.8), conduct themselves under a large sample size n. These three expressions share the common denominator $\left(\int f_0(\mathbf{w_1})^2\ S\ d\mathbf{w_1}\right)^2$, where

$$S = \sum_{k=0}^{n-2} \frac{\binom{n-2}{k}\ H(a,\mathbf{w_1})^k (1-H(a,\mathbf{w_1}))^{n-2-k}}{k+2} \tag{5.14}$$

As $H(a,\mathbf{w_1}) = \int_{\|\mathbf{w}\|\leq a} f_0(\mathbf{w_1}-\mathbf{w})d\mathbf{w}$, we shall replace $H(a,\mathbf{w_1})$ by H in S, with $0 < H < 1$, and analyse the dependency of S on n and H. The results from this dependency analysis are summarized in Table 5.3. We wish to emphasize that the exact value of H loses importance with increasing sample size n. In fact, the value derived for the standard deviation decreases when n increases. The latter conveys that $S = O(a_n)$, for some bounded sequence a_n.

n / H	0.1	0.3	0.5	0.7	Std Dev
100	9.09E-02	3.26E-02	1.98E-02	1.42E-02	3.52E-02
500	1.96E-02	6.64E-03	3.99E-03	2.85E-03	7.74E-03
1000	9.91E-03	3.33E-03	2.00E-03	1.43E-03	3.91E-03
5000	2.00E-03	6.66E-04	4.00E-04	2.86E-04	7.89E-04
10000	9.99E-04	3.33E-04	2.00E-04	1.43E-04	3.95E-04
50000	2.00E-04	6.66E-05	4.00E-05	2.86E-05	7.89E-05

Table 5.3: Values obtained for S in (5.14), when given n and H. In the last column, are values for the corresponding standard deviation of H, when given n.

Let us now consider the following three quotients, for $i = 2, 3, 4$:

$$Q_i = \frac{\sum_{k=0}^{n-i} \binom{n-i}{k} \frac{H^k (1-H)^{n-i-k}}{(k+i)^2}}{S^2} \tag{5.15}$$

Table 5.4 presents the values obtained for Q_2, Q_3 and Q_4, for the same previous values of H. One notes that these quotients tend to 1 with increasing sample size. This tendency may be observed, for any chosen probability H.

Expressions (5.6), (5.7) and (5.8) may now be re-written as

$$E_d(n,a) = \frac{\int f_0(\mathbf{w_1})^2 \; Q_2 \; S^2 \; d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2 \; S \; d\mathbf{w_1}\right)^2} \simeq \frac{\int f_0(\mathbf{w_1})^2 \; O(a_n)^2 \; d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2 \; O(a_n) \; d\mathbf{w_1}\right)^2} =$$

$$= O\left(\frac{\int f_0(\mathbf{w_1})^2 \; d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2 \; d\mathbf{w_1}\right)^2}\right)$$

$$F_d(n,a) = \frac{\int f_0(\mathbf{w_1})^3 \; Q_3 \; S^2 \; d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2 \; S \; d\mathbf{w_1}\right)^2} \simeq \frac{\int f_0(\mathbf{w_1})^2 \; O(a_n)^2 \; d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^3 \; O(a_n) \; d\mathbf{w_1}\right)^2} =$$

$$= O\left(\frac{\int f_0(\mathbf{w_1})^3 \; d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2 \; d\mathbf{w_1}\right)^2}\right)$$

$$G_d(n,a) = \frac{\int f_0(\mathbf{w_1})^4 \; Q_4 \; S^2 \; d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2 \; S \; d\mathbf{w_1}\right)^2} \simeq \frac{\int f_0(\mathbf{w_1})^2 \; O(a_n)^2 \; d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^4 \; O(a_n) \; d\mathbf{w_1}\right)^2} =$$

Q_2	n / H	0.1	0.3	0.5	0.7
	100	1.08588	1.02326	1.00999	1.00428
	500	1.01798	1.00467	1.00200	1.00086
	1000	1.00900	1.00233	1.00100	1.00044
	5000	1.00180	1.00047	1.00020	1.00009
	10000	1.00090	1.00023	1.00010	1.00004
	50000	1.00018	1.00005	1.00002	1.00001
Q_3	n / H	0.1	0.3	0.5	0.7
	100	0.90084	0.97543	0.98960	0.99556
	500	0.98157	0.99529	0.99798	0.99914
	1000	0.99089	0.99765	0.99900	0.99957
	5000	0.99820	0.99953	0.99980	0.99991
	10000	0.99910	0.99977	0.99990	0.99996
	50000	0.99982	0.99995	0.99998	0.99999
Q_4	n / H	0.1	0.3	0.5	0.7
	100	0.76283	0.93098	0.96982	0.98696
	500	0.94713	0.98603	0.99399	0.99742
	1000	0.97328	0.99301	0.99700	0.99871
	5000	0.99461	0.99860	0.99940	0.99974
	10000	0.99730	0.99930	0.99970	0.99987
	50000	0.99946	0.99986	0.99994	0.99997

Table 5.4: Values obtained for Q_2, Q_3 and Q_4 in (5.15), when given n and H.

$$= O\left(\frac{\int f_0(\mathbf{w_1})^4 \, d\mathbf{w_1}}{\left(\int f_0(\mathbf{w_1})^2 \, d\mathbf{w_1}\right)^2}\right)$$

Assumption **(A3)** allows us to guarantee that $0 < \int f_0(\mathbf{w_1})^i d\mathbf{w_1} < +\infty$, $i = 2, 3, 4$. Consequently, the approximations derived above for E_d, F_d and G_d are of the exact order $O(1)$ and, therefore, they are bounded.

5.7.3 Estimates of $E_2(n, a)$, $F_2(n, a)$ and $G_2(n, a)$

For the specific case of $I\!R^2$, we now describe a supplemental simulation study for the dependency analysis of E_2, F_2 and G_2 on n. We also suggest a numeric approximation for the expressions introduced in (5.6), (5.7) and (5.8). Bear in mind that these are defined in the region $D_0 \subset I\!R^2$, which must be a bounded and fixed region. We have selected D_0 to be the unit square, $[0, 1] \times [0, 1]$. The density function for the spatial locations on D_0 is f_0.

To estimate E_2, F_2 and G_2, the corresponding integrals were numerically approximated to a sample average, as

$$\int g(x)dx = \mathrm{E}\left[\frac{g(X)}{f(X)}\right], \quad \text{where } f(x) \text{ is the density function of } X.$$

For instance, $\widehat{E}_2$ may be derived, as follows:

$$\widehat{E}_2(n, a) = \frac{\frac{1}{N}\sum_{i=1}^{N} \widehat{f}_0(\mathbf{w_i}) \ \sum_{k=0}^{n-2} \frac{\binom{n-2}{k} \ \widehat{H}(a,\mathbf{w_i})^k (1-\widehat{H}(a,\mathbf{w_i}))^{n-2-k}}{(k+2)^2}}{\left(\frac{1}{N}\sum_{i=1}^{N} \widehat{f}_0(\mathbf{w_i}) \ \sum_{k=0}^{n-2} \frac{\binom{n-2}{k} \ \widehat{H}(a,\mathbf{w_i})^k (1-\widehat{H}(a,\mathbf{w_i}))^{n-2-k}}{k+2}\right)^2}$$

where

- n is the number of original sampled points; we chose $n = 100, 200, 400$;
- N is is the number of extra points generated from density f_0 and needed for the integral approximation; we chose $N = 5000$;
- $\widehat{H}(a, \mathbf{w_i}) = \frac{n_i}{n}$, being n_i the number of original sampled points within the circle of center $\mathbf{w_i}$ and radius a;

CSR	n	$\widehat{E}_2$	$\widehat{F}_2$	$\widehat{G}_2$
	100	2.142(0.102)	0.927(0.009)	0.497(0.031)
	200	1.945(0.056)	0.980(0.006)	0.560(0.019)
	400	1.831(0.028)	1.005(0.003)	0.610(0.011)

CLUSTER	n	$\widehat{E}_2$	$\widehat{F}_2$	$\widehat{G}_2$
	100	1.971(1.166)	1.206(0.098)	2.571(1.463)
	200	2.096(0.273)	0.892(0.042)	1.010(0.231)
	400	2.111(0.097)	0.901(0.012)	0.607(0.054)

Table 5.5: Mean values of $\widehat{E}_2$, $\widehat{F}_2$ and $\widehat{G}_2$, obtained from a total of 100 independent samples. The corresponding standard deviations are given between brackets.

- $\widehat{f}_0$ results from a non parametric density estimation of the spatial locations in D_0; we adopted a bivariate kernel-type estimator;

The other two estimates, $\widehat{F}_2$ and $\widehat{G}_2$, may be obtained in a very similar way.

We started with a complete spatial randomness (CSR) design. So, we generated n locations uniformly distributed on D_0. This procedure was repeated to obtain 100 independent samples. Table 5.5 presents the average of those 100 replicas and the corresponding standard deviation.

The next simulation included one clustered area on D_0, where we forced a minimum of 60 points to be restricted to a small square, with area equal to 0.16×0.16 instead of the original 1×1, and a center randomly chosen. The results were also included in Table 5.5.

The main conclusion from both simulations appears to be the absence of an obvious tendency with increasing of sample size. In any case, for any of the three approximations of E_2, F_2 and G_2, the standard deviation clearly decreases with increasing of sample size, so that the mean value provides a good estimate of the unknown term.

Chapter 6

Assessing the effect of preferential sampling

6.1 Introduction

As stated before, in geostatistics, in both prediction and inference contexts, it is commonly assumed that the selection of the sampling locations does not depend on the values of the spatial variable (Diggle et al. 2003). Additionally, most techniques are based on the assumption, possibly tacit, of sampling locations being uniformly distributed over the observed region.

In Chapter 4, we assess the effect of the failure of the earlier assumptions concerning the estimation of the correlation structure in the specific case of multi-stage collection of spatial data. The appraisal of biased data in later stages, conditional on data values from earlier stages, is considered. As the presence of clusters is a natural consequence of non-uniform locations distribution, we propose a kernel estimator robust to clusters. Then, in Chapter 5, we proceed with the theoretical study of the suggested estimator.

We now intend to introduce a formal definition directly related to the failure of the *independency* assumption, and not restricted to multi-stage sample collection. Suppose that, in the nature of the sampling process, involving as it does

Unobserved field S Observed data $(\mathbf{x_i}, y_i)$

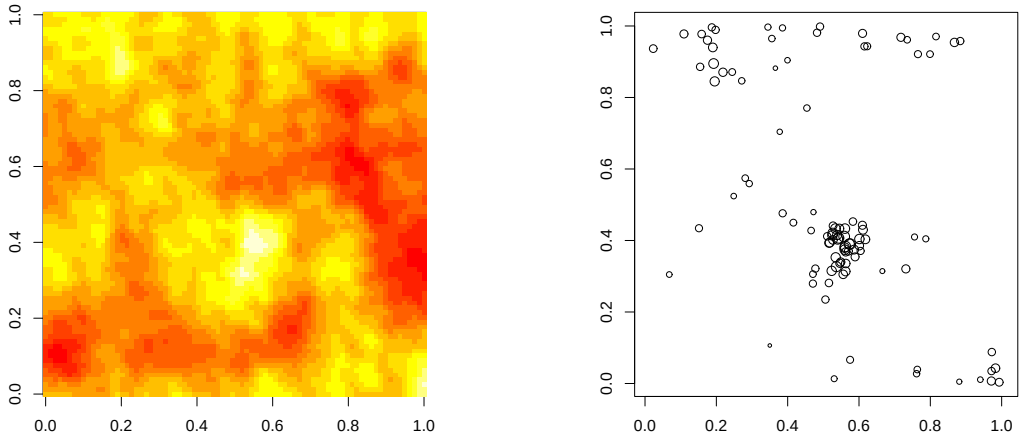


Figure 6.1: Example of an unobserved field process (highest values represented by lightest colors) and the corresponding observed sample data set (highest values of y_i represented by largest bullets).

the search for maximum values, there exists an underlying stochastic relationship between data and locations, then one has *preferential sampling*.

Following the notation used by Diggle et al. (2003), we shall consider that the data for analysis are of the form $(\mathbf{x_i}, y_i) : i = 1, ..., n$, where $\mathbf{x_1}, ..., \mathbf{x_n}$ are locations within an observation region $D \subset I\!R^2$ and $y_1, ..., y_n$ are measurements associated with these locations. The $\{\mathbf{x_i} : i = 1, ..., n\}$ is the *sampling design* and y_i is assumed to be a realization of $Y_i = Y(\mathbf{x_i})$, where $\{Y(\mathbf{x}) : \mathbf{x} \in D\}$ is the *measurement process*. We also assume the existence of an unobserved field process $\{S(\mathbf{x}) : \mathbf{x} \in D\}$, usually regarded as our goal of prediction. Often, Y_i can be considered as a noisy version of the underlying random variable $S(\mathbf{x_i})$, the value at location $\mathbf{x_i}$ of process $S(.)$.

Figure 6.1, in the left hand panel, illustrates an example of a true field S and, in the right hand panel, the corresponding sample data set $(\mathbf{x_i}, y_i)$.

Remark 6.1 *Under* ***preferential sampling****, the sampling design process is assumed to be stochastically dependent of the field process* $S(.)$*. Consequently, the corresponding geostatistical model (specified by the joint distribution of the processes involved) must take into account the conditional distribution of the sampling design.*

Considering the presence of preferential sampling under Gaussian assumptions, we shall propose a model-based approach for spatial prediction. This new parametric model will be founded on a flexible class of log-Gaussian Cox processes to be introduced in the next Section. The remainder of this Chapter is devoted primarily to the analysis of the consequences of ignoring preferential sampling and adopting the classic geostatistical methods. In Chapter 7, we proceed with the likelihood inference to estimate our model parameters.

To terminate this introductory Section, we want to emphasize the distinction between clustered and preferential sampling. As already mentioned, the clustering of locations may be due to the existence of specific geographic or demographic spots, or they may even be used to describe short-range variability better. These are good examples showing that clustered sampling may not imply preferential sampling. On the contrary, the opposite implication tends to occur, as preferred sample locations normally occur in concentrated areas. For example, some prior scientific knowledge about $S(.)$, such as the expected local ore grade in mine exploration, may cause the concentration of samples in areas with atypically large values.

6.2 A class of log-Gaussian Cox processes

The sample locations $\mathbf{x}$ are nothing more than realizations of a point process P. Under complete spatial randomness, the point process modelling is typically based on some homogeneous Poisson process. In our case, however, we are not

interested in a constant intensity function. We need a class of point processes where the constant intensity λ of the Poisson process is replaced by a spatially varying intensity function, $\lambda(\mathbf{x})$. More precisely, we wish to model aggregated spatial point patterns where the aggregation is due to some stochastic heterogeneity. This leads us to a class of inhomogeneous Poisson processes, P, with stochastic intensity functions, called the Cox processes. As described in Diggle (2003), we have:

- $\{\Lambda(\mathbf{x}) : \mathbf{x} \in I\!R^2\}$ is a non-negative-valued stochastic process.
- Conditional on $\{\Lambda(\mathbf{x}) = \lambda(\mathbf{x}) : \mathbf{x} \in I\!R^2\}$, the events form an inhomogeneous Poisson process with intensity function $\lambda(\mathbf{x})$.

Additionally, we must have in mind our intention to model the dependency of point process P on field process S. Assuming Gaussian data, we shall then consider log-Gaussian Cox processes for P, i.e. Cox processes where the logarithm of the intensity surface is a Gaussian process. In Moller, Syversveen and Waagepetersen (1998), the class of stationary log-Gaussian Cox processes is shown to possess various appealing properties. These authors further note that these point processes are flexible models for clustering and easy to simulate. Their analysis is, however, restricted to unmarked point processes.

Remark 6.2 *A **geostatistical model for preferential sampling** is a specification of the joint distribution of the field process, the point process and the measurement process of the form* $[S(.), P(.), Y(.)] = [S(.)]\ [P(.)|S(.)]\ [Y(.)|S(.)]$, *where* [.] *means "the distribution of".*

So, the proposed model for preferential sampling might be exemplified by:

- $S \sim \text{SGP}(\mu, \sigma^2, \rho(.))$ – S is a stationary Gaussian process with mean μ, standard deviation σ and spatial correlation function $\rho(.)$.

- $P|S \sim \text{Poisson}\left(\exp\{\alpha + \beta S(\mathbf{x})\}\right)$ – α and β are real numbers, and $|\beta|$ identifies the **degree of preferability**.

- $Y(\mathbf{x_i}) = S(\mathbf{x_i}) + Z_i,\ i = 1, ..., n,$ where $Z_i \sim \text{N}(0, \tau^2)$ and $\mathbf{x_i}$ are realizations of $P|S$.

According to this model, processes Y and S can be directly related to the same spatial variable, as in the previous example of ore grade in mine exploration, and Z_i basically identifies the measurement error or a generic *nugget effect* (see the decomposition of scale variation given in (2.4), in Chapter 2). Alternatively, suppose one wishes to proceed with some prediction about soil fertility in a given region. This type of model could then be useful, by considering the measurement process Y as the height of specific trees and the process S as the fertility. Moreover, suppose our goal of prediction S is the air pollution in a certain area, then Y can describe the concentration of heavy metals, e.g. nickel, measured at some sample locations.

The sample locations, if taken from areas where S is expected to present larger values, could be considered realizations of a point process $P|S$ with a positive β. On the contrary, a negative β would be an indicator of a negative association between P and S, in the sense that more sample locations should be collected from areas where S is expected to be smaller.

As expected, a null value for β leads to the *classical geostatistical model* and, in this case, one has $[S, P, Y] = [S][Y|S][P] = [S, Y][P]$. Hence, the corresponding log-likelihood function can be given by $l(\boldsymbol{\theta}|Y, S, P) = l_1(\boldsymbol{\theta}|S, Y) + l_2(\boldsymbol{\theta}|P)$, where $\boldsymbol{\theta}$ is the vector of model parameters. Note that the second expression can be ignored, without the model losing validity. So, for the classical geostatistical model, the distribution of the point process can be regarded as irrelevant.

As to the term α from the intensity function of the Poisson process, it could be allowed to vary and as a function of $\mathbf{x}$, so originating a distribution for sample points depending on their spatial locations. This could provide a means by which

to assign different *priorities* to different areas within the observation region.

For the purpose of simulating inhomogeneous Poisson processes, an algorithm based on rejection sampling can be used (see e.g. Lewis and Shedler 1979). This consists in simulating a Poisson process on the observation region with intensity λ_0 equal to the maximum value of $\lambda(\mathbf{x})$ within this region, and retaining an event at $\mathbf{x}$ with probability $\lambda(\mathbf{x})/\lambda_0$. Please note that a constant value chosen for α means using a uniform distribution to pick up the locations using the previous algorithm. We shall consider α to be equal to 0, as its exact value is not really meaningful for the corresponding density function

$$f(\mathbf{x}) = \frac{\lambda(\mathbf{x})}{\int_D \lambda(\mathbf{x})d\mathbf{x}} = \frac{\exp(\alpha + \beta S(\mathbf{x}))}{\int_D \exp(\alpha + \beta S(\mathbf{x}))d\mathbf{x}} = \frac{\exp(\beta S(\mathbf{x}))}{\int_D \exp(\beta S(\mathbf{x}))d\mathbf{x}}$$

In Figure 6.2, we plot a blocking-experiment, where rows correspond to two independent realizations of a Gaussian field S and columns correspond to three distinct values for β (0, 1 and 2). The results allow us to show clearly how β works as a degree of preferability, since a larger value of β indicates a stronger aggregation of sample points around larger values of S.

6.3 Effect on variogram estimation

We now analyse the effect of preferential sampling on the estimation of spatial dependency, when this is specified through the variogram function. As already discussed, the variogram of a spatial stochastic process $Y(.)$ can be obtained from its second-moment structure

$$\frac{1}{2}\mathrm{E}[\{Y(\mathbf{x}) - Y(\mathbf{x}')\}^2] = \frac{1}{2}\mathrm{Var}[Y(\mathbf{x}) - Y(\mathbf{x}')] \quad \forall\, \mathbf{x}, \mathbf{x}' \in D.$$

However, as $\mathbf{x}$ and $\mathbf{x}'$ are obtained as realizations of $P|S$, the foregoing expression may be used to derive only an *empirical mark variogram*, here denoted by $\widehat{\gamma}_M(\|\mathbf{x}-\mathbf{x}'\|)$. The empirical variogram estimators presented in Chapter 3, or the one we proposed and studied in Chapter 5, can be used as $\widehat{\gamma}_M(.)$. The *theoretical mark*

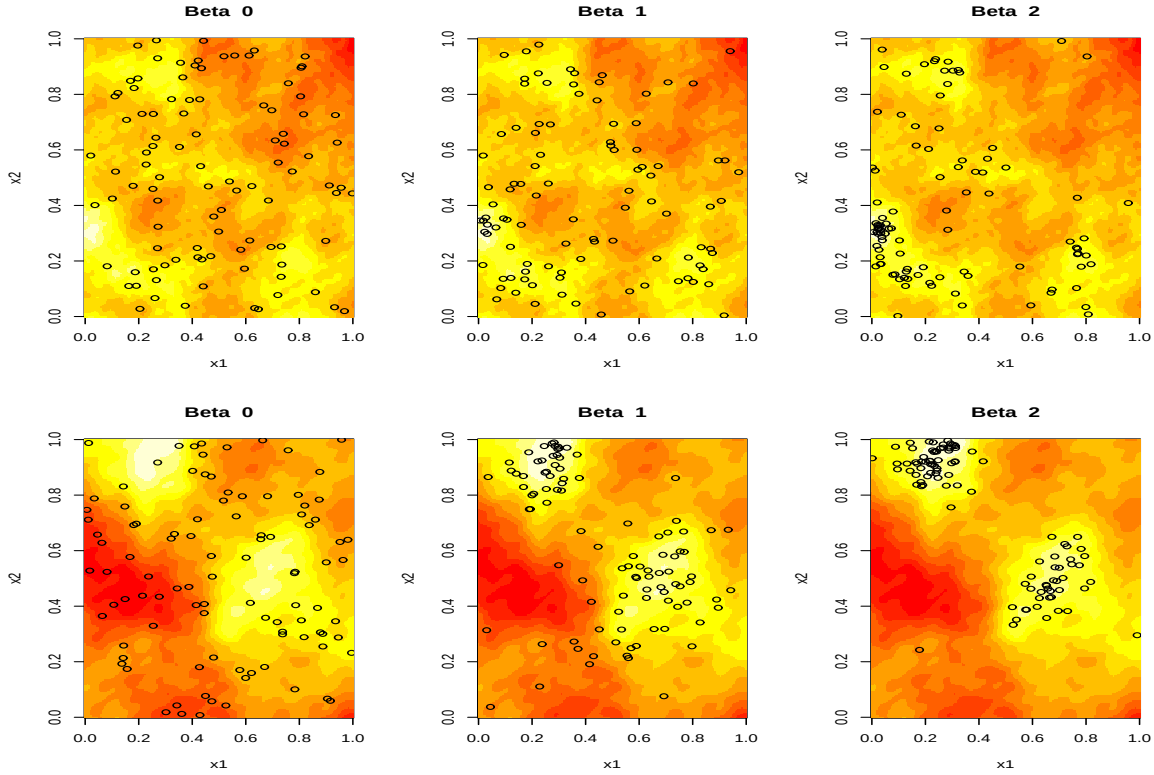


Figure 6.2: Influence of β on sample locations, given two independent Gaussian fields S (rows). The values chosen for β are 0, 1 and 2 (columns).

variogram $\gamma_M(.)$ is not known[1]. However, for our present study, its analytical expression may be considered not relevant. An approximation can be achieved from a sample mean of several independent datasets, by averaging all $\widehat{\gamma}_M(.)$. From now on, we shall denote this Monte Carlo approximation by $\widetilde{\gamma}_M(.)$.

For the Gaussian model introduced in Section 6.2, the *theoretical variogram* function can be stated as

$$\gamma(u) = \tau^2 + \sigma^2(1 - \rho(u))$$

where u represents the distance between spatial locations. The basic structural covariance parameters of our model are the nugget variance τ^2, the total sill given by $\mathrm{Var}[Y(.)] = \tau^2 + \sigma^2$ and the range ϕ (parameter from correlation function $\rho(.)$).

[1]Some brief comments on marked point processes were included in Section 4.1.

Bear in mind that in classical geostatistics, corresponding to $\beta = 0$, $\widetilde{\gamma}_M(.)$ and $\gamma(.)$ are expected to be equal, meaning that the Monte Carlo approximation to the theoretical mark variogram is unbiased.

6.3.1 Some simulation details

The Gaussian model depends on the correlation function $\rho(u) = Corr[S(\mathbf{x}), S(\mathbf{x}')]$, whose specification determines the smoothness of the resulting process $S(.)$. In the following simulation studies, we assume $\rho(.)$ to be a member of the *Matérn* family, which we recall from expression (2.2) to be

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}K_{\kappa}(u/\phi).$$

As a starting point for our simulations, we chose $\mu = 4$, $\sigma^2 = 0.64$, $\tau^2 = 0.01$, $\kappa = 1$ and $\phi = 0.4$, and we took the unit square as our observation region. The value chosen for the sample size of each realization was 100.

The Monte Carlo simulations require the generation of a large number of realizations of the stationary Gaussian process $S(.)$ on a discrete mesh of sample points. This means that a fine grid of data must be prepared. The degree of *fineness* should take into account the value chosen for the range parameter ϕ. The conventional method of *direct matrix decomposition*[2] adopted in previous Chapters becomes inadequate here, as it is limited to about 1000 points, i.e. a 31x31 grid in two dimensions (at least for R packages, see e.g. Schlather 2001). We then adopted the *circulant embedding* method proposed in Chan and Wood (1997) and pointed out as a fast and accurate algorithm for this kind of situation.

6.3.2 Influence of β on bias and variance

For distinct values of β, we intend to measure the *bias* of the mark variogram approximation $\widetilde{\gamma}_M(.)$, when compared to the theoretical variogram $\gamma(.)$. It is also

[2]This method for simulating a Random Field is based on the well-known method for simulating any multivariate Gaussian distribution, using the square root of the covariance matrix. This implementation can use the Cholesky decomposition and the singular value decomposition.

important to measure the variability imposed by this approximation. In Figure 6.3, two important functions of distance u are being plotted, $bias(u)$ and $stddev(u)$, given by $\gamma(u) - \widetilde{\gamma}_M(u)$ and $\sqrt{\widehat{\text{Var}}[\widetilde{\gamma}_M(u)]}$, respectively. The corresponding standard error is also plotted. These results (together with the corresponding Confidence Interval) suggest that, for $\beta = 0$, the Monte Carlo approximation tends to be an unbiased estimation. It is also interesting to observe the direct association between bias enlargement and an increasing β value.

On the contrary, the standard deviation and the standard error do not look as if they depend on β. Note that, for those chosen values of β, the existing bias might not seem very meaningful when compared with the *imprecision degree* (specified by the standard deviation). However, for larger values of β, the relative importance of bias is expected to increase. Furthermore, suppose we have a larger sample size, then the standard deviation should decrease and these two curves should also become closer, an indicator of a more meaningful bias.

We repeated the same experiment for a total of 500 and 1000 replicas. As we obtained similar results, subsequent experiments are limited to 500 independent realizations. This experiment was also repeated with new values for our model parameters. The main conclusions are summarized as follows:

- We confirmed that the value of μ does not affect the results, since the variogram function does not depend on the process expectation.

- The sign of β does not change the results of the variogram estimation. This may be justified by the Gaussian fields properties, making $S(\mathbf{x})$ and $-S(\mathbf{x})$ equivalent.

- We tried distinct values for the range parameter ϕ (0.1, 0.15, 0.2 and 0.4) and for the smoothness parameter κ (0.5, 1 and 2). The impact of β on the variogram estimation, illustrated in Figure 6.3, was unchanged.

- Finally regarding the variance parameters, if the ratio τ^2/σ^2 increases (or

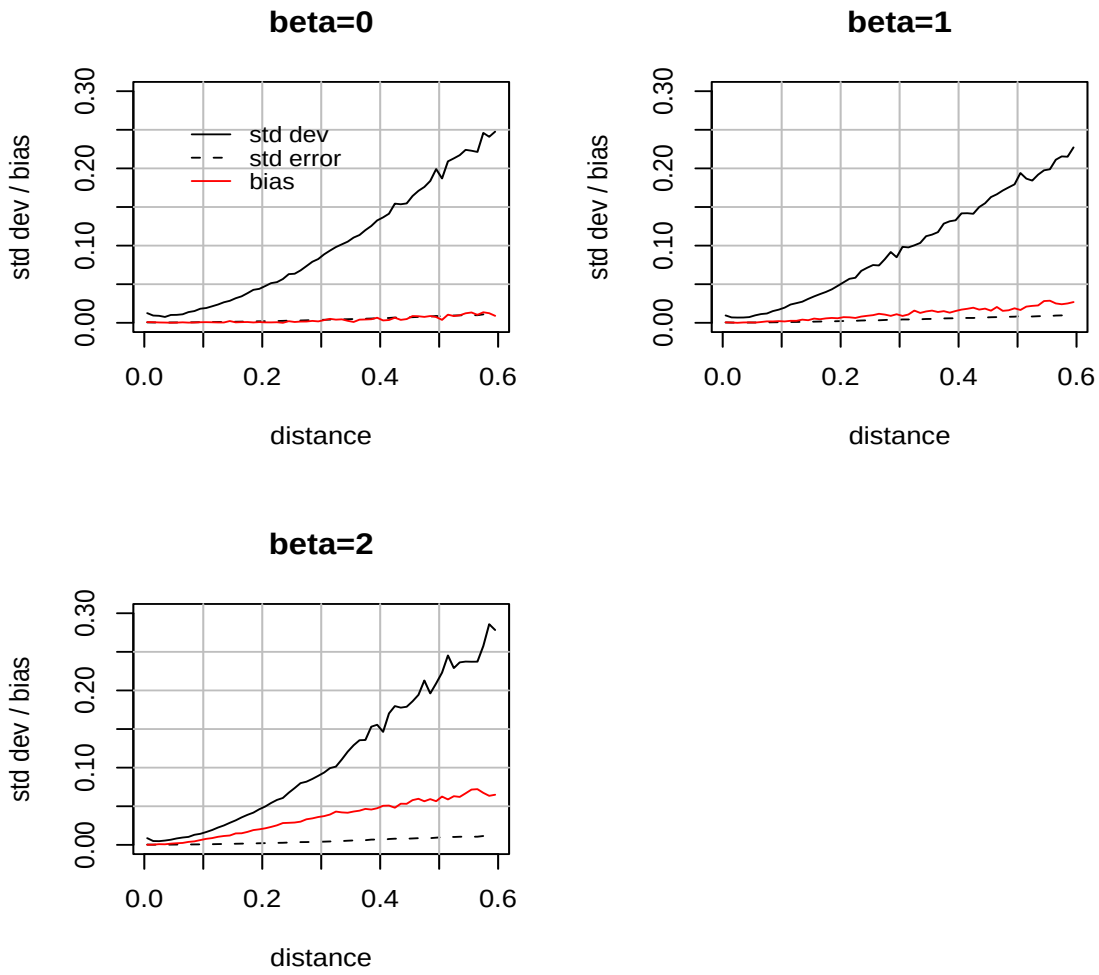


Figure 6.3: Influence of β on variogram estimation. Comparison of the estimation bias and the corresponding approximation variability, given by the standard deviation and the standard error. The simulation consisted of 500 replications, each with a sample size of 100.

$\sigma^2 \to 0$), then $\gamma(.)$ and $\widetilde{\gamma}_M(.)$ become similar. Consequently, the degree of preferability β seems to lose importance under too much white noise, even though the clustering effect persists.

6.3.3 Clustered versus preferential

We now aim to carry out a numerical study to measure the impact of clustered sampling on variogram estimation, both with and without, being also preferential. Three distinct sampling designs are to be considered: *completely spatial randomness* (CSR), *preferential* and *just clustered*. As expected, the CSR sample is obtained for β equal to 0. A pairwise sample generation is adopted, otherwise: we first generate a preferential sampling data set for β equal to 2; we then keep previous locations and generate new multivariate Gaussian data on them to obtain a clustered but non-preferential data set.

Remember that $\widetilde{\gamma}_M(.)$ is derived from a Monte Carlo average of several independent $\widehat{\gamma}_M(.)$. For all results included in the foregoing Section, we choose as $\widehat{\gamma}_M(.)$ the classic estimator from Matheron, given in (3.1). Here, we intend to compare again the performance of three next variogram estimators: the classic, the NW kernel and the robust to clusters (the two latter are given in (3.3) and (5.1), respectively).

In the following simulations, the variance parameters were changed to more significant values, becoming $\tau^2 = 0.25$ and $\sigma^2 = 2.25$. The simulation consisted of 500 independent replicas, each with a sample size of 100.

For each independent data set, we first derive the integrated squared error (ISE) between each empirical estimator and the theoretical variogram, as defined in (3.6). The results are summarized in the boxplots in Figure 6.4, through the quartiles of the ISE values found, when taking all lags smaller than 0.6. These boxplots confirm the results already observed in Chapter 4 (see e.g. Table 4.1), exhibiting the positive contribution of the proposed estimator when sampling is clustered, independently of whether preferential or not.

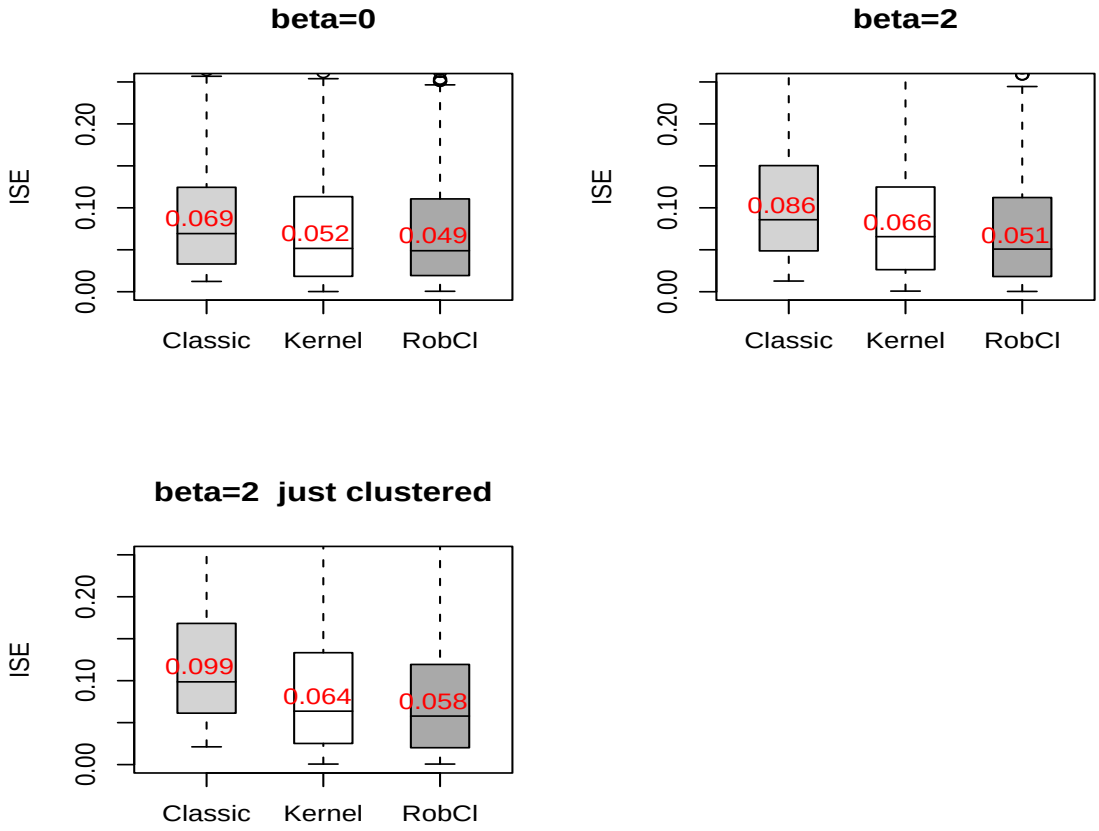


Figure 6.4: Boxplot of the evaluated ISE from three empirical estimators: Classic, NW kernel and Robust to Clusters. Three sampling designs are considered: CSR ($\beta = 0$), preferential ($\beta = 2$) and clustered (non-preferential).

To gain more information concerning the behaviour of these estimators, we then proceed with some efficiency assessment, by comparing their variances and their mean squared errors (MSE). Recall that the latter is defined as

$$\mathrm{MSE}\left[\hat{\gamma}_M(u)\right] = \left(\mathrm{Bias}\left[\hat{\gamma}_M(u)\right]\right)^2 \;+\; \mathrm{Var}\left[\hat{\gamma}_M(u)\right].$$

In Figure 6.5, in the left hand panels, we plot the square roots of variances, including biases, for our three estimators. In the right hand panels the corresponding square roots of the MSE's are likewise plotted.

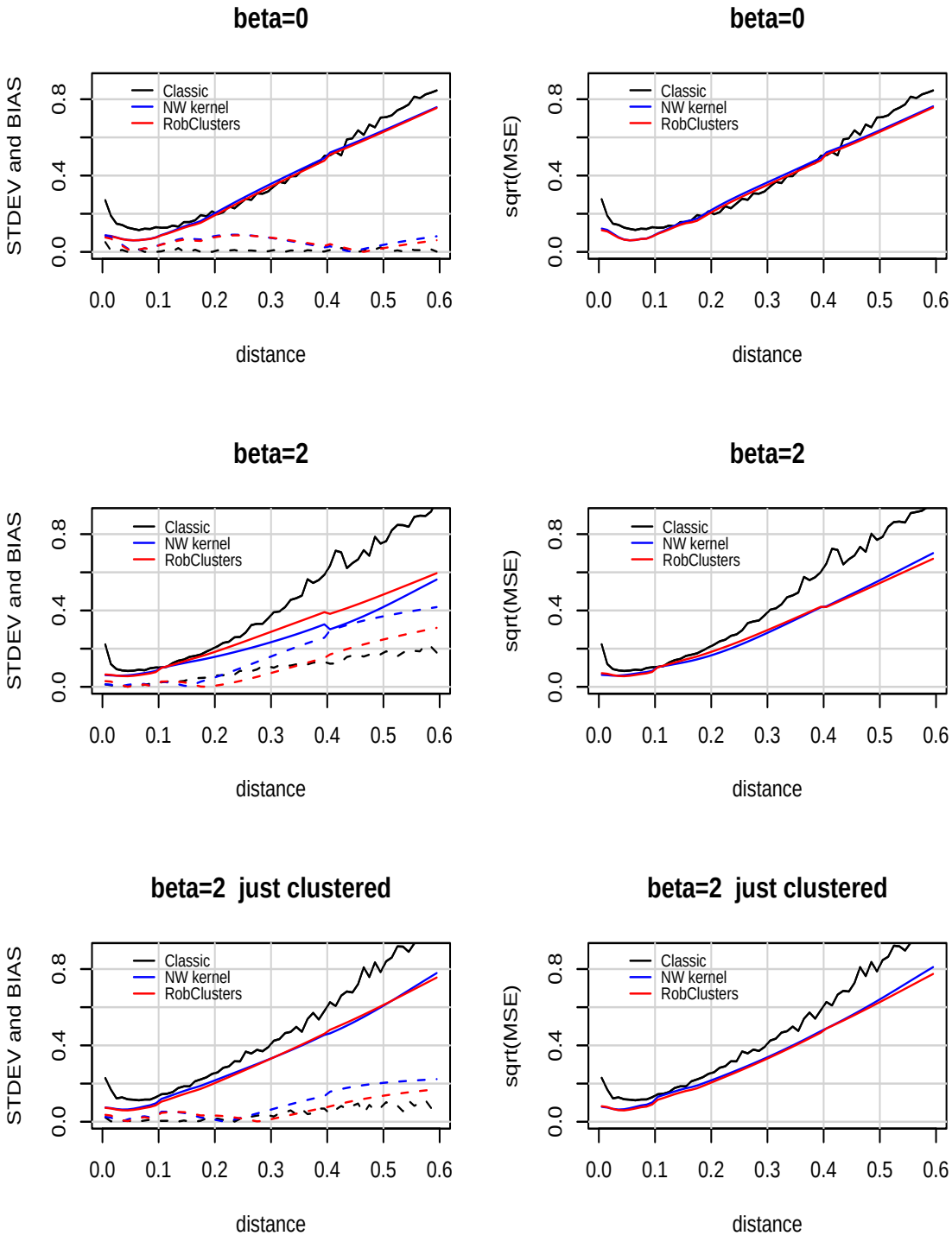


Figure 6.5: Comparison of variogram estimators through their efficiency (measure in terms of $\sqrt{Var}$ and $\sqrt{Mse}$) when sampling is clustered, both with and without, being also preferential. The estimators bias is also plotted (dashed-lines). It is being considered $\tau^2 = 0.25$ and $\sigma^2 = 2.25$.

The analysis of these graphics leads us to the following results. Under CSR, the performance of the two kernel estimators is confirmed to be only slightly better than under the classic one. Actually, the latter presents a little more variance for nugget estimation, and for $\gamma(.)$ estimation regarding those lags larger than the range value, i.e. 0.4. The top-left panel in Figure 6.5 also reveals a minor bias, nearly meaningless in the case of the classic estimator, that it is sometimes referred to as the *smoothing bias*. Here it is of interest to recall that, in classical geostatistics[3], the bias term goes to zero as n increases, as kernel estimators are asymptotically unbiased.

Under clustered sampling, whether or not also preferential, the two kernel estimators always perform significantly better than the classic one. Note that the smaller variance of these estimators, when sampling is also preferential, is an indirect consequence of the smaller variance of the data values under this sampling design. Comparing the two kernel estimators through the MSE and, through the ISE, that robust to clusters always performs at least slightly better. This occurs even when this estimator presents a larger variance, as it is always associated with a smaller bias.

We conclude that, if the sample locations are not homogeneously scattered over the observed region, then corrector methods such as that proposed by downweight clustering, are advisable. Remember that the simulation studies included in Chapter 4 pointed to the same conclusion. Where efficiency is concerned, we have now proved that the benefit should not be disregarded.

Bear in mind, however, that the proposed estimator (like the other estimators) does not consider the preferability issue. Therefore, under preferential sampling, all of them present a non-negligible bias and the efficiency issue consequently becomes less relevant.

[3]That corresponds to first and third rows of panels in Figure 6.5.

Effect of white noise

The last simulation study included in this Section is related to the analysis of the effect of noise. As has already been stated, the $Y(.)$ process can be considered the noisy version of $S(.)$. Suppose we have more noise variance τ^2, but a fixed value for the total variance $\tau^2 + \sigma^2$. We intend to observe how the variogram estimation for the $Y(.)$ process would be affected. To simplify notation, let us represent this estimator by $\widehat{V}_Y$, and that associated with $S(.)$ by $\widehat{V}_S$. The two are related through the expression $\widehat{V}_S \equiv \widehat{V}_Y - \widehat{\tau}^2$. An estimate of τ^2 is, however, normally difficult to obtain.

We changed τ^2 from 0.25 to 0.81. The value for σ^2 was chosen in such a way that the total variance, i.e. sill, remains equal to 2.5. In Figure 6.6, the efficiency results are plotted. The same three sampling designs and the same three empirical estimators are taken into account. Besides efficiency, other performance indicators, such as the ISE, were derived. The main conclusions can be summarized as follows.

Under CSR, the classic estimator is no longer a reasonable alternative to either kernel estimator. The presence of strong noise really degrades the performance of this estimator, in terms of variance and bias, for small lags. The two kernel estimators seem considerably resistant to the presence of noise. This may be justified by the adoption of a specific asymmetric boundary kernel, near the variogram endpoint 0, as briefly explained in Chapter 3 (more detail to be found in Garcia-Soidán et al. 2004).

Under non-CSR, the white noise continues to affect the classical estimator but not the kernel ones. Curiously, these two last estimators present a little less bias than in the previous case study.

Finally, if the two kernel estimators are compared in terms of efficiency, one may say that the gain from using the proposed kernel estimator is almost negligible compared with that from the NW kernel. We highlight the fact that the main contribution of the former continues to be some bias reduction.

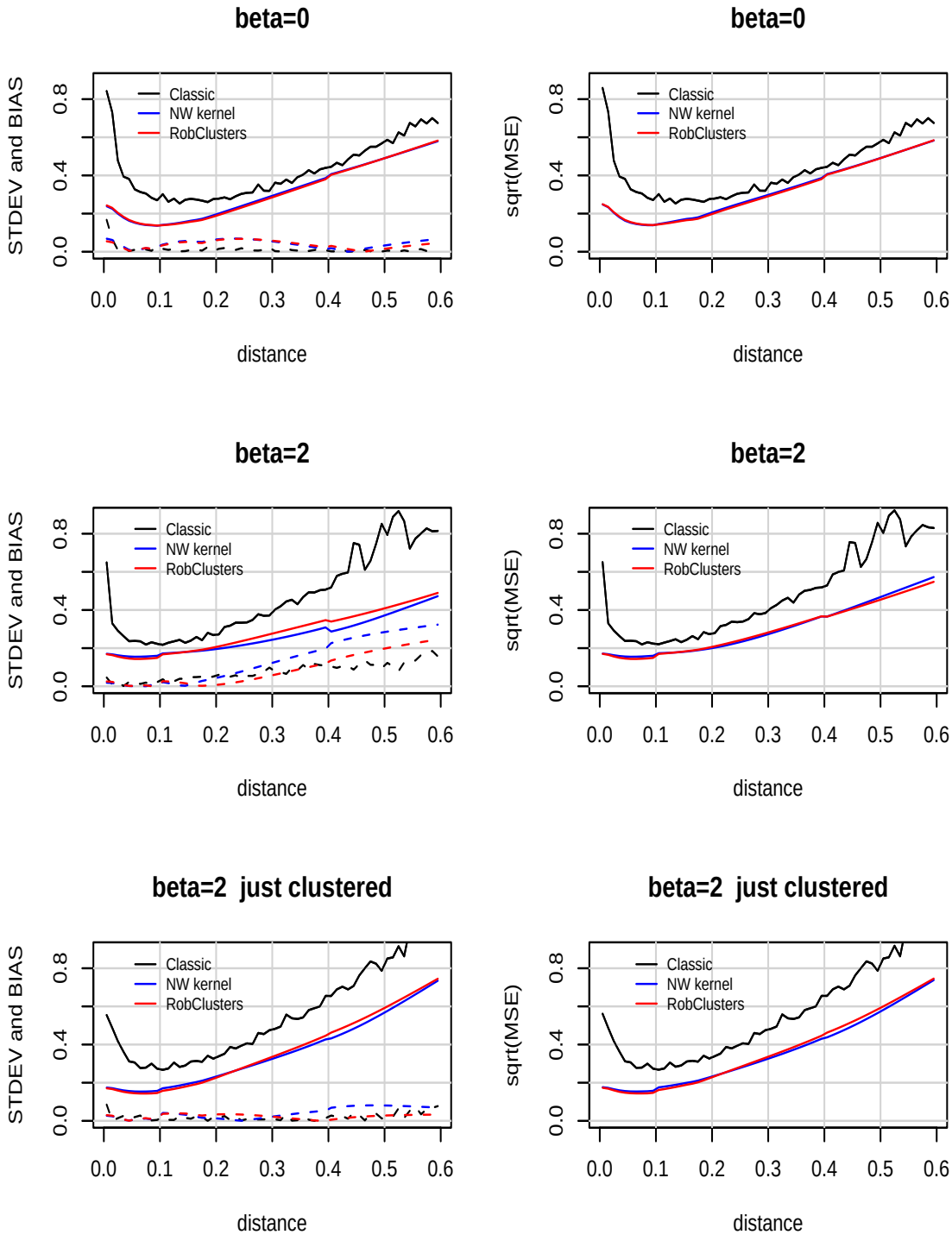


Figure 6.6: Comparison of variogram estimators through their efficiency (measure in terms of $\sqrt{Var}$ and $\sqrt{Mse}$) when sampling is clustered, in both the preferential and non-preferential sub-cases. The estimation bias is also plotted (dashed-lines). It is being considered $\tau^2 = 0.81$ and $\sigma^2 = 1.69$.

6.4 Impact on prediction

Optimal prediction, i.e kriging, typically depends on knowledge of spatial dependency, as it is supposed that some measurements in the vicinity of the point investigated should be more closely related than others to the true unknown value. It is then supported by the second-order properties of the spatial process. One point of interest to be addressed is the analysis of the influence of earlier variogram estimators on the consequences for prediction

Kriging is indeed considered to provide the best linear unbiased estimator (abbreviated as BLUE) of the unknown characteristic studied. *Linear* because its estimates result from a weighted linear combination of sample data; *unbiased* because the mean of prediction errors (deviation between true value and predicted value) is expected to be null; *best* because the variance of the prediction errors (prediction variance) is at its minimum.

For simplicity, suppose now that interest lies in the process $Y(.)$ and not in its noiseless version $S(.)$. Our target for prediction becomes $Y(\mathbf{x_0})$ the value of process $Y(.)$ at a generic location $\mathbf{x_0}$, given sample data $(\mathbf{x_i}, y_i)$, $i = 1, 2, ..., n$. The prediction problem may be formalized by invoking the conditional distribution of $Y(.)$ given the observed data $\mathbf{y}$. $\mathrm{E}[Y(\mathbf{x_0})|\mathbf{y}] = \widehat{Y}(\mathbf{x_0})$ specifies the predicted value and $\mathrm{Var}[Y(\mathbf{x_0})|\mathbf{y}] = \mathrm{E}[(Y(\mathbf{x_0}) - \widehat{Y}(\mathbf{x_0}))^2]$ specifies the prediction variance (or prediction mean square error).

The optimal predictor $\widehat{Y}(\mathbf{x_0})$ minimises $MSE[\widehat{Y}(\mathbf{x_0})] = \mathrm{E}[(Y(\mathbf{x_0}) - \widehat{Y}(\mathbf{x_0}))^2]$, where the expectation is specified by the joint distribution of $Y(\mathbf{x_0})$ and Y. As discussed in Cressie (1993), this *best* predictor is not always linear in $\mathbf{y}$ (under non-Gaussianity). Therefore, one should additionally require

$$\widehat{Y}(\mathbf{x_0}) = \sum_{i=1}^{n} \lambda_i y_i + \lambda.$$

We now aim to minimize (over coefficients $\lambda_1, ..., \lambda_n, \lambda$)

$$\mathrm{E}\left[\left(Y(\mathbf{x_0}) - \sum_{i=1}^{n} \lambda_i y_i - \lambda\right)^2\right] = \mathrm{Var}\left[Y(\mathbf{x_0}) - \sum_{i=1}^{n} \lambda_i y_i\right] + \left(\mu(\mathbf{x_0}) - \sum_{i=1}^{n} \lambda_i \mu(\mathbf{x_i}) - \lambda\right)^2,$$

where $\mu(\mathbf{x}) = \mathrm{E}[Y(\mathbf{x})], \mathbf{x} \in D$. The derived solution is $\lambda = \mu(\mathbf{x_0}) - \sum_{i=1}^{n} \lambda_i \mu(\mathbf{x_i})$ and

$$(\lambda_1, ..., \lambda_n)^t = \mathbf{c}^t \mathbf{\Sigma}^{-1},$$

where $\mathbf{c} = (C(\mathbf{x_0}, \mathbf{x_1}), ..., C(\mathbf{x_0}, \mathbf{x_n}))^t$, $\mathbf{\Sigma}$ is a $n \times n$ matrix with $(i,j)^{th}$ element $C(\mathbf{x_i}, \mathbf{x_j})$ and $C(.)$ is the covariance function. The optimal linear predictor becomes

$$\widehat{Y}(\mathbf{x_0}) = \mathbf{c}^t \mathbf{\Sigma}^{-1}(\mathbf{y} - \mu) + \mu(\mathbf{x_0}),$$

where $\mu = (\mu(\mathbf{x_1}), ..., \mu(\mathbf{x_n}))^t$. The minimized mean-squared prediction error is

$$\mathrm{Var}[Y(\mathbf{x_0})|\mathbf{y}] = C(\mathbf{x_0}, \mathbf{x_0}) - \mathbf{c}^t \mathbf{\Sigma}^{-1} \mathbf{c}.$$

Such a sampling prediction technique, assuming a known mean function $\mu(\mathbf{x})$, was named *simple kriging* by Matheron (1971). If $\mu(\mathbf{x})$ is equal to an unknown constant μ, one is in the presence of *ordinary kriging*. Alternatively, *universal kriging* is used when $\mu(\mathbf{x})$ is linear in a fixed number of unknown parameters. All these kriging techniques produce optimal linear predictors.

6.4.1 Gaussian data

Suppose that we disregard the possibility of data being *preferentially sampled* and we adopt previous standard kriging techniques. Under a Gaussian model like the one proposed in Section 6.2, we would have

$$Y(\mathbf{x}) = S(\mathbf{x}) + \mathrm{N}(0, \tau^2),$$

where $Y(.)$ is the measurement process and $S(.)$ is the target of prediction. So, interest is now in process $S(.)$, given the observed data $\mathbf{y}$.

$[S(\mathbf{x_0}), Y]$ is supposed to be multivariate Gaussian with mean vector $(\mu, \mu\mathbf{1})^t$ and covariance matrix

$$\begin{pmatrix} \sigma^2 & \sigma^2 \mathbf{r}^t \\ \sigma^2 \mathbf{r} & \tau^2 \mathbf{I} + \sigma^2 \mathbf{R} \end{pmatrix}$$

where $\mathbf{r}$ is a vector with elements $r_i = \rho(\|\mathbf{x_0} - \mathbf{x_i}\|) : i = 1, ..., n$, $\mathbf{R}$ is a $n \times n$ matrix with $(i, j)^{th}$ element $\rho(\|\mathbf{x_i} - \mathbf{x_j}\|)$. The $\mathbf{1}$ is a $n-$length vector of ones and $\mathbf{I}$ is the $n \times n$ identity matrix.

As described in Diggle et al. (2003), the minimum mean square error predictor becomes

$$\widehat{S}(\mathbf{x_0}) = \sigma^2 \mathbf{r}^t (\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1} (\mathbf{y} - \mu \mathbf{1}) + \mu \tag{6.1}$$

and with prediction variance

$$\text{Var}[S(x_0)|\mathbf{y}] = \sigma^2 - \sigma^2 \mathbf{r}^t (\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1} \sigma^2 \mathbf{r} \tag{6.2}$$

Consequently, the prediction variance depends on the correlation model, on the spatial configuration of the data and on the prediction location, but it does not directly[4] depend on the actual data.

An extension to this Gaussian model may be taken into account by dealing with a non-constant mean value surface $\mathbf{x}$. Typically, it may be useful to consider

$$\mu(\mathbf{x}) = \Sigma_{j=1}^{p} \beta_j f_j(\mathbf{x}),$$

where $f_1(\mathbf{x}), ..., f_p(\mathbf{x})$ are observed functions of location $\mathbf{x}$, or functions of observed covariates, leading to *universal kriging* or *kriging with a trend model* (Wackernagel 1998). An estimator for the unknown $\beta = (\beta_1, ..., \beta_p)^t$ may be derived by maximum likelihood, yielding

$$\widehat{\beta} = (\mathbf{F}^t \mathbf{V}^{-1} \mathbf{F})^{-1} \mathbf{F}^t \mathbf{V}^{-1} \mathbf{y}, \tag{6.3}$$

where $\mathbf{F}$ is a $n \times p$ matrix with $(i, j)^{th}$ element $f_j(\mathbf{x_i})$ and $\mathbf{V} = \mathbf{R} + \tau^2/\sigma^2 \mathbf{I}$. In this case, the expression (6.1) for minimum mean-squared predictor would change slightly to give

$$\widehat{S}(\mathbf{x_0}) = \sigma^2 \mathbf{r}^t (\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1} (\mathbf{y} - \mathbf{F}\widehat{\beta}) + \mathbf{F_0}\widehat{\beta}$$

where $\mathbf{F_0}$ is a $1 \times p$ matrix with $(1, j)^{th}$ element $f_j(\mathbf{x_0})$.

[4]Note that one may say that the prediction variance indirectly depends on data $\mathbf{y}$, through the estimation of the parameters.

Finally, consider the case of *ordinary kriging* ($\mu(\mathbf{x}) = \mu$, unknown) which occurs when $p = 1$ and $\mathbf{F}$ equals a vector of one's. Expression (6.3) will then return a real value $\widehat{\beta}$ as an estimator of $\widehat{\mu}$.

6.4.2 Simulation study

We now wish to assess how misleading it might be using standard kriging methodology for preferential sampling. In Figure 6.7, we simulate two different sample data sets from the same Gaussian field $S(.)$. A non-null degree of preferability is represented in the second row. The spatial dependency was estimated starting with the empirical variogram estimator proposed by Matheron and, then, fitting to it a Matérn model by ordinary least squares. Remember that only a valid model (i.e. one obeying the conditionally negative-definite property) must be used for prediction, otherwise absurd negative values for the mean square prediction errors may be obtained (Cressie 1993).

Afterwards, an ordinary kriging was performed on a grid of 21×21 prediction locations in $I\!R^2$. The two columns on the right side of Figure 6.7 show the mean square error predictors and the square roots of prediction variances, given by expressions (6.1) and (6.2), respectively. These plots already illustrate our worries about the effect of preferential sampling. A rough visual inspection of the kriging estimates seems to point to *worse* estimates of the underlying Gaussian field, for $\beta = 2$. This is not surprising where the absence of sample data does not allow the correct interpolation that is expected for kriging. However, for instance, on the right bottom corner, there is no obvious reason for worse estimates.

Furthermore, we have larger values (represented by lighter colors) for prediction standard deviations in those areas where there are fewer sample points. The dependency of the latter on the spatial configuration of the data and on the estimation of the spatial correlation model also contributes to these outcomes.

Similar results were obtained when spatial dependency is estimated through maximum likelihood.

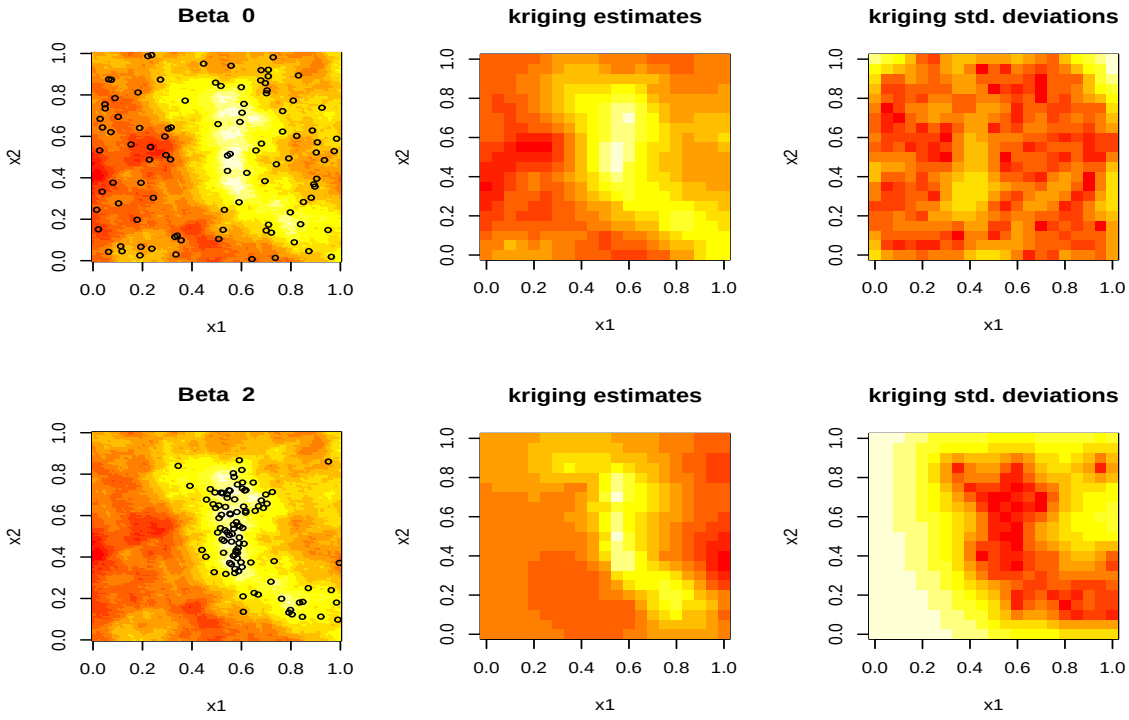


Figure 6.7: Kriging estimates and standard deviations for 2 sample data sets: CSR and preferential sampling. The highest values are represented by lightest colors.

Monte Carlo experiments

Having in mind that kriging provides a BLUE estimator, we shall proceed with two distinct Monte Carlo experiments. First, we examine the unbiasedness issue. It can be anticipated that estimation of the expected prediction error should yield a value close to zero, that is $\mathrm{E}[\widehat{S}(\mathbf{x_0}) - S(\mathbf{x_0})]$.

Second, we check whether estimates of the variance of prediction errors and the prediction variance are approximately equal, given by $\mathrm{Var}[\widehat{S}(\mathbf{x_0}) - S(\mathbf{x_0})]$ and $\mathrm{Var}[S(\mathbf{x_0})|\mathbf{y}]$, respectively. Assuming the absence of bias, these two should be approximately equal to the estimate of prediction mean square error, $\mathrm{E}[(\widehat{S}(\mathbf{x_0}) - S(\mathbf{x_0}))^2]$.

Once more, we consider the Gaussian model $Y(\mathbf{x}) = S(\mathbf{x}) + \mathrm{N}(0, \tau^2 = 0.5^2)$,

Table 6.1: Non-null nugget effect – impact of preferential sampling on prediction. Monte Carlo approximations for: *expectation of prediction errors* (PE), *variance of prediction errors*, *prediction variance* (PV) and *prediction mean square error*. The approximations consisted of 500 realizations.

	$\widetilde{\boldsymbol{\theta}}$	**CSR**	**Clustered**	**Preferential**
$\widehat{\mathrm{E}}[\mathbf{PE}]$	$\widehat{\boldsymbol{\theta}}$	$(-0.081, 0.059)$	$(-0.082, 0.186)$	$(1.290, 1.578)$
	$\boldsymbol{\theta}$	$(-0.088, 0.054)$	$(-0.080, 0.188)$	$(1.153, 1.447)$
$\widehat{\mathrm{Var}}[\mathbf{PE}]$	$\widehat{\boldsymbol{\theta}}$	$(0.282, 0.363)$	$(0.996, 1.137)$	$(1.150, 1.471)$
	$\boldsymbol{\theta}$	$(0.284, 0.365)$	$(1.004, 1.280)$	$(1.190, 1.532)$
$\widehat{\mathrm{E}}[\mathbf{PV}]$	$\widehat{\boldsymbol{\theta}}$	$(0.543, 0.589)$	$(1.078, 1.238)$	$(0.921, 1.097)$
	$\boldsymbol{\theta}$	$(0.558, 0.594)$	$(1.205, 1.369)$	$(1.205, 1.369)$
$\widehat{\mathrm{E}}[\mathbf{PE^2}]$	$\widehat{\boldsymbol{\theta}}$	$(0.268, 0.354)$	$(0.948, 1.300)$	$(2.967, 3.729)$
	$\boldsymbol{\theta}$	$(0.269, 0.359)$	$(0.952, 1.304)$	$(2.635, 3.429)$

where $\log(\Lambda(\mathbf{x})) = \beta S(\mathbf{x})$, $\mathrm{E}[S(.)] = \mu = 4$, $\mathrm{Var}[S(.)] = \sigma^2 = 1.5^2$, $\rho(.)$ is a Matérn function with $\phi = 0.15$ and $\kappa = 1$. We continue to compare three sampling designs: CSR, just clustered and preferential. In each case, 500 realizations in all were used, each with a sample size of 100 (non-fixed point locations). As we are assuming stationarity, we choose just one prediction point, that for which $\mathbf{x_0} = (0.5, 0.5)$.

For each replica $j = 1...500$, we derive the corresponding prediction error (PE) and prediction variance (PV), with $\boldsymbol{\theta} = (\mu, \sigma^2, \tau^2, \phi, \beta)^t$ and $\widetilde{\boldsymbol{\theta}}$ equals $\boldsymbol{\theta}$ or $\widehat{\boldsymbol{\theta}}$:

- $PE_j = \widehat{S}_j(\mathbf{x_0}) - S_j(\mathbf{x_0}) = \mathrm{E}[S_j(\mathbf{x_0})|\mathbf{y}, \widetilde{\boldsymbol{\theta}}\,] - S_j(\mathbf{x_0})$
- and $PV_j = \mathrm{Var}[S_j(\mathbf{x_0})|\mathbf{y}, \widetilde{\boldsymbol{\theta}}\,]$

In Table 6.1, we summarize the Monte Carlo approximations obtained for $\mathrm{E}[\widehat{S}(\mathbf{x_0}) - S(\mathbf{x_0})]$, $\mathrm{Var}[\widehat{S}(\mathbf{x_0}) - S(\mathbf{x_0})]$, $\mathrm{Var}[S(\mathbf{x_0})|\mathbf{y}]$ and $\mathrm{E}[(\widehat{S}(\mathbf{x_0}) - S(\mathbf{x_0}))^2]$, of

which we want to highlight the first two.

Trying to distinguish the side effects caused by two different issues, parameter estimation and spatial prediction, we repeat the prediction process using the true $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$.

The most evident conclusion from the analysis of Table 6.1 is that the bias, represented by $\widehat{\mathrm{E}}[PE]$, considerably increases under preferential sampling. This is because $\widehat{S}(x_0)$ tends to be over-estimated, as we are forcing a higher density of sample locations close to maximum values of the underlying field $S(.)$.

Additionally, it is quite evident, that $\widehat{\mathrm{Var}}[PE]$ increases under preferential or just clustered sampling designs, higher values being found in the former case. Bear in mind that this variance represents an empirical prediction variance[5], which depends on sampling design $\mathbf{x_i}$ and on the estimation of model parameters[6], and that the latter is affected by a non-random design as was shown earlier.

These results emphasize how misleading it would be to adopt classical kriging methodology for preferential sampling. They support the need for an alternative solution, which will be discussed in Chapter 7.

According to the results obtained for the true $\boldsymbol{\theta}$, where bias is concerned, a more *precise* estimation of $\boldsymbol{\theta}$ would seem to be more important under a preferential sampling design. In this case, there is a significant reduction in the prediction errors. Otherwise, the impact of $\boldsymbol{\theta}$ estimation hardly appears to be relevant. Actually, looking at the results obtained for $\widehat{\mathrm{E}}[PV]$ under clustered sampling (including both the preferential and non-preferential sub-cases), we are surprised to observe higher variability when $\boldsymbol{\theta}$ is true.

[5]The nominal prediction variance is given in (6.2)

[6]Note that, if $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$, the prediction variance given by $\widehat{\mathrm{E}}[PV]$ is exactly the same for clustered sampling and preferential sampling, because we have adopted a pairwise sample generation and, for each realization, they share the same spatial configuration.

Table 6.2: Null nugget effect – impact of preferential sampling on prediction. Similar information to Table 6.1.

	$\widetilde{\boldsymbol{\theta}}$	**CSR**	**Clustered**	**Preferential**
$\widehat{\mathrm{E}}[\mathbf{PE}]$	$\widehat{\boldsymbol{\theta}}$	$(-0.010, 0.062)$	$(-0.018, 0.150)$	$(1.169, 1.371)$
	$\boldsymbol{\theta}$	$(-0.008, 0.064)$	$(-0.018, 0.154)$	$(1.129, 1.339)$
$\widehat{\mathrm{Var}}[\mathbf{PE}]$	$\widehat{\boldsymbol{\theta}}$	$(0.143, 0.184)$	$(0.788, 1.010)$	$(1.131, 1.450)$
	$\boldsymbol{\theta}$	$(0.145, 0.186)$	$(0.812, 1.041)$	$(1.217, 1.561)$
$\widehat{\mathrm{E}}[\mathbf{PV}]$	$\widehat{\boldsymbol{\theta}}$	$(0.184, 0.208)$	$(0.758, 0.864)$	$(0.614, 0.724)$
	$\boldsymbol{\theta}$	$(0.196, 0.220)$	$(0.898, 1.018)$	$(0.898, 1.018)$
$\widehat{\mathrm{E}}[\mathbf{PE^2}]$	$\widehat{\boldsymbol{\theta}}$	$(0.136, 0.188)$	$(0.738, 1.044)$	$(2.526, 3.248)$
	$\boldsymbol{\theta}$	$(0.138, 0.190)$	$(0.763, 1.075)$	$(2.511, 3.277)$

Additional results

We repeated the foregoing simulation study, by keeping the spatial configuration of the data sampled points fixed for all realizations. According to the rejection sampling algorithm described in Section 6.2, this new study could not be implemented in the case of preferential sampling. For the two remaining sampling types, CSR and just clustered, the results seem too heavily dependent on the initial design (the one randomly chosen for the first realization). Consequently, we decided not to include the results here.

Finally, concerning Table 6.1, please remember that the approximations for $\mathrm{Var}[\widehat{S}(\mathbf{x_0}) - S(\mathbf{x_0})]$, $\mathrm{Var}[S(\mathbf{x_0})|\mathbf{y}]$ and $\mathrm{E}[(\widehat{S}(\mathbf{x_0}) - S(\mathbf{x_0}))^2]$ (three last rows) were expected to be more similar, except in the case of preferential sampling. Trying to understand whether the reason could be related to the presence of a large noise effect, we repeated the simulation study, now with $\tau^2 = 0$. In fact, the new results, summarized in Table 6.2, match our expectations more closely. All the other main conclusions, from the analysis of the most recent Table, are confirmed for the null-nugget effect case.

Chapter 7

Moss data and a model-based approach

This closing Chapter is split into two main parts. The first part describes an application to real data, allowing us to exemplify the application of several methods of spatial statistics which were discussed in the course of the thesis. Namely, this data set is used to exemplify the usefulness of the non-parametric variogram robust to clusters which was studied in Chapter 5.

Air pollution intensities based on data from Galicia (NW Spain) are analysed. Two distinct datasets are available, as data were collected in two distinct years: 1995 and 2000. In the later year, as more funding was obtained for the project, a grid with more points was used for the sampling design. With respect to the 1995 data, we suspect there is a clear rationale for the sampling being preferential, as most of the data seem to have been sampled close to pollution sources for reasons of cost. The results reinforce the importance of doing parametric model analysis under preferential sampling.

In the second part of Chapter 7, we then proceed with the analysis of a model-based approach to preferential sampling. An intuitive model is proposed, incorporating into the traditional model a correction term for the preferability issue. We describe an algorithm for the direct Monte Carlo approximation to the likelihood

function. We then use likelihood inference to estimate the parameters of the proposed model. A simulation study is performed to show the benefits of this model over the traditional one. Our model analysis suggests that this intuitive model should take into account the presence of some bias. So, we close this Chapter, by including a short discussion on future research objectives related to the evaluation of an *unbiased* model for preferential sampling.

7.1 Application to real data

Air quality can be monitored either by measuring the pollutants directly in the air or in deposition, or by using biomonitors. Direct measurements provide objective information about the level of pollutants, but they are expensive and there is a risk of contamination when determining low concentrations. Biomonitoring is an alternative method, fast and inexpensive, used to screen the intensity and the distribution of the pollutants emission, over surrounding regions of probable pollution sources. This cost-effective method is based on the high bioconcentration of heavy metals in land mosses, specifically the scleropodium purum moss. The Ecotoxicology Group from Santiago de Compostela University presents the first results of the biomonitoring methods for the analysis of pollution data in Galicia region in Fernández, Rey and Carballeira (2000).

Some details are here given concerning the two sample data sets. For the year 1995, the moss was planted in 64 locations, which can be observed on the map in Figure 7.1. Bear in mind that the main towns in Galicia are in the coastal area and that these towns are typically surrounded by industrial areas. In the northern part of Galicia is concentrated considerable manufacturing activity, including a metallurgy industry site and two thermal power stations. In the interest of cost reduction, it was decided to plant a larger quantity of moss in the north. For the year 2000, the sampling grid comprised 118 sampling points spread over the entire region of Galicia (about 28000 km^2), located, roughly speaking, at the vertices of a 15×15 km UTM grid as described in Figure 7.1 (see Aboal, Real, Fernández and

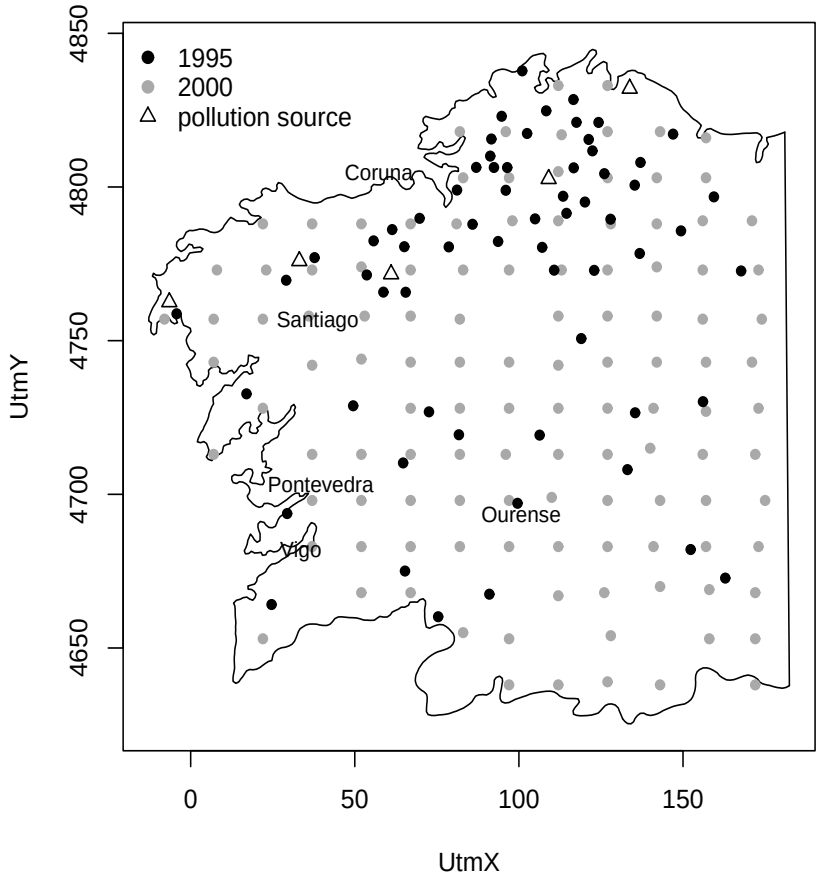


Figure 7.1: Moss data locations – Galicia region in Spain.

Carballeira 2005, for more details). The typical procedure is to plant the moss and some time later to collect it to be analysed and the concentration of heavy metals to be measured. In both years, the sampling was carried out at the beginning of summer. In this work we decide to focus our attention on three heavy metals: chromium, nickel and lead. The corresponding concentration measurements are given in parts per million (ppm).

The results of the preliminary analysis of samples collected in 1995 and 2000 are summarised in Table 7.1. Concentrations of Cr and Ni recorded during the sampling in 2000 were significantly higher than those found in 1995. As far as

	1995 (n=64)					2000 (n=118)				
	Mean	S.D.	Med	Min	Max	Mean	S.D.	Med	Min	Max
Cr	1.41	0.85	1.14	0.18	4.77	15.9	35.0	5.44	0.29	265.2
Ni	1.82	0.78	1.74	0.54	4.50	8.27	15.8	3.27	0.49	115.0
Pb	5.68	8.25	3.73	0.04	60.4	2.11	1.23	1.77	0.29	8.65

Table 7.1: Average, S.D., median, minimum and maximum concentrations of Cr, Ni and Pb in moss samples collected in Galicia in 1995 and 2000.

we know, no meaningful sources of pollution were created during these five years. This growth can be considered normal, and it is possibly due to greater activity on the part of the pollution sources already present. Furthermore, if we observe the spatial distribution of data collected in 2000 represented in Figure 7.2, we conclude that some of the largest concentrations of Cr and Ni are found in the south and east of Galicia, in zones not sampled in 1995. These zones are crossed by important highways linking Galicia to other regions of Spain and, in Castilla-León close to the east border of Galicia, there is a thermal power station. Additionally, note that both zones are classified as being more barren, where the erosion processes are more pronounced than in the rest of Galicia. Concentrations of Ni and, more significantly, of Cr are associated with the soil lithology, so possibly accounting for higher values of these elements in these zones.

With respect to the values of Pb found in Table 7.1, one observes that those recorded during the 1995 sampling survey were higher than those found in 2000, due possibly to the use of unleaded gasoline becoming more popular and to prohibition of the sale of leaded gasoline.

In order to examine further the concentrations of the aforementioned heavy metals, the log transformation was adopted to allow data normalization. The histograms obtained for the standardized log-measurements of Cr, Ni and Pb are given in Figure 7.3. We then proceed to our exploratory analysis with the va-

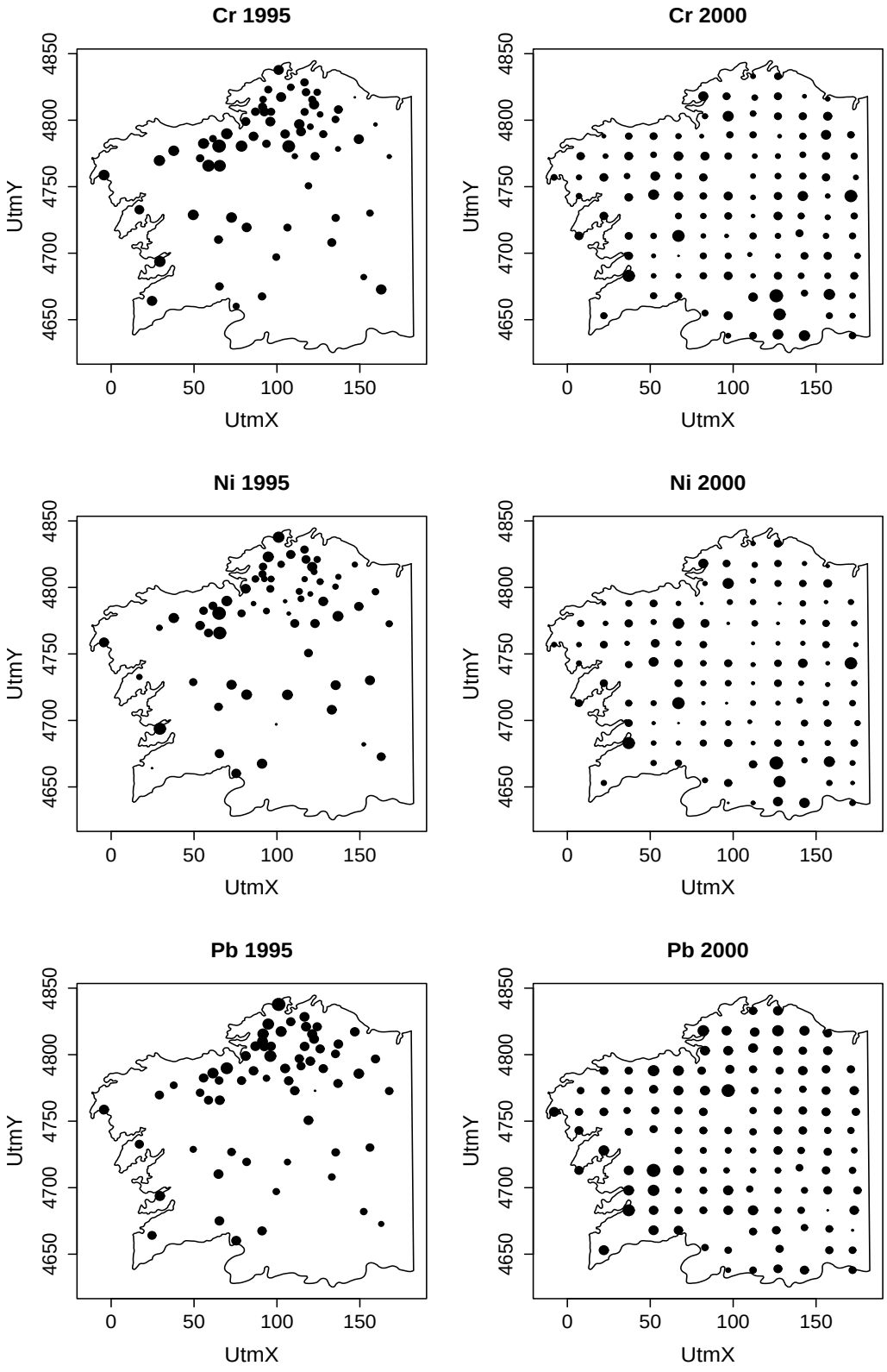


Figure 7.2: Spatial distribution of Cr, Ni and Pb in 1995 and 2000. The radius of each circle is proportional to the concentration of each heavy metal.

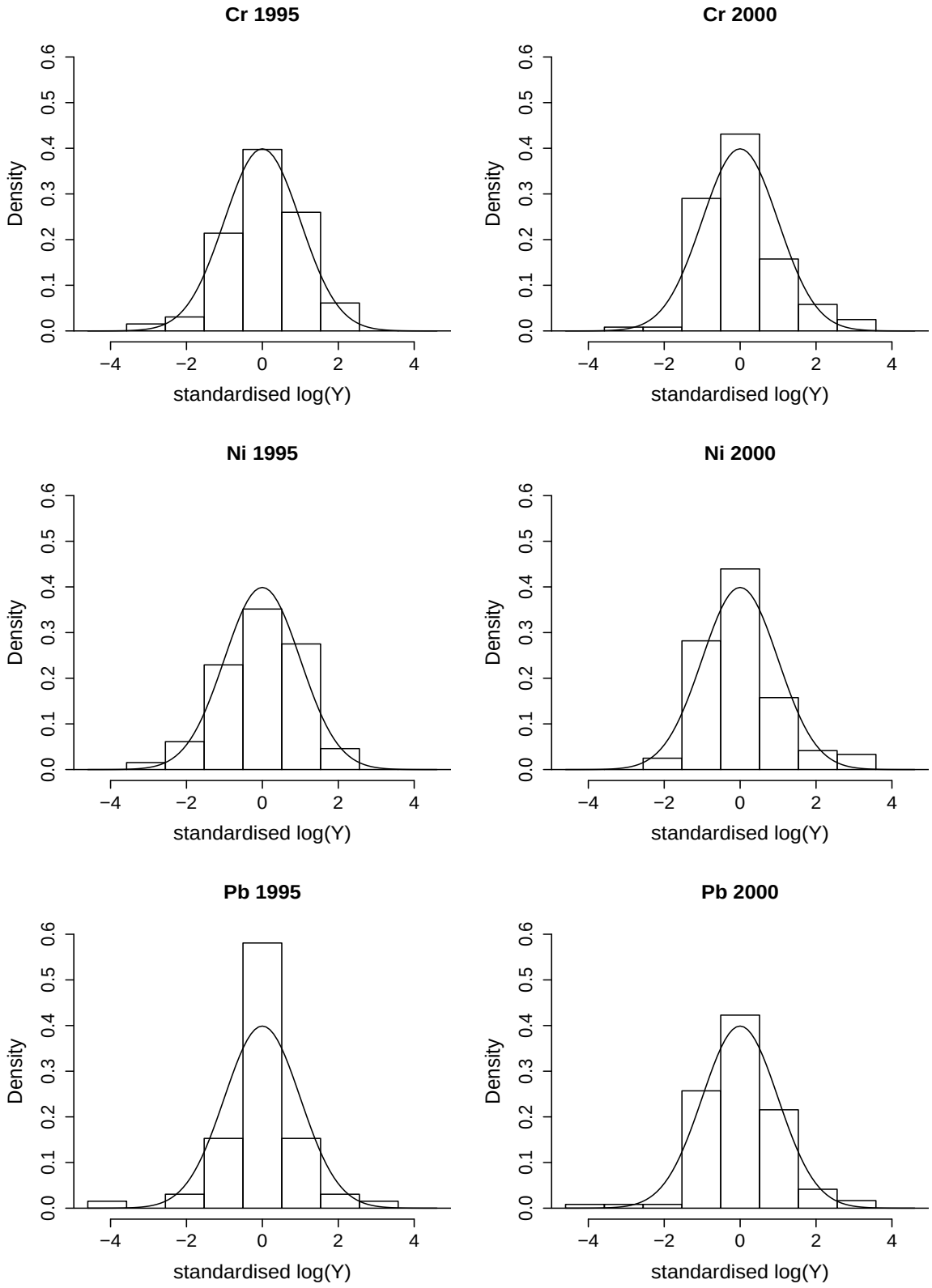


Figure 7.3: Histograms of the standardized data measurements of Cr, Ni and Pb in 1995 and 2000.

riogram estimation in order to check the spatial dependency associated with each of these elements (see Figure 7.4). We first consider the empirical estimation given by Matheron's proposal as defined in equation (3.1). Bearing in mind that the presence of clustering in the 1995 data affects the results of this classic estimator, we then consider the kernel variogram estimator proposed in (5.1), which incorporates an inverse weighting to a given neighbourhood density. The valid, i.e. conditionally negative-definite, version of the robust to clusters estimator is obtained through Bochner's theorem as described in (3.4). This non-parametric estimator, tagged *RobCluster* in Figure 7.4, helps us to find an adequate parametric model. The Matérn family defined in Section 2.4 seems suitable in all cases, considering that parameter κ is used to determine the correct analytic smoothness of each measurement process. Alternative variogram models were compared through the likelihood and the result, tagged *MaxLik*, was also included in Figure 7.4.

In the case of year 2000, as expected, the robust to clusters and the maximum likelihood estimators present similar results, mainly for Pb data. Note that the two other heavy metals present some exceptional high values (see Table 7.1), possibly explaining why the non-parametric estimator fits the empirical estimates slightly better.

In the case of year 1995, the results of the parametric estimator seem too different from the empirical estimates. In contrast, the robust to clusters estimator seems to be more consistent with the empirical estimator, even though presenting a conveniently smoother shape than the latter, since a correction for high density areas was adopted. Figure 7.4 then reveals possible benefits from the proposed non parametric estimator.

7.1.1 Test if sample is preferential

Our interest now lies in deciding for which heavy metals it is reasonable to assume preferential sampling. A typical sign is related to values not expected for the

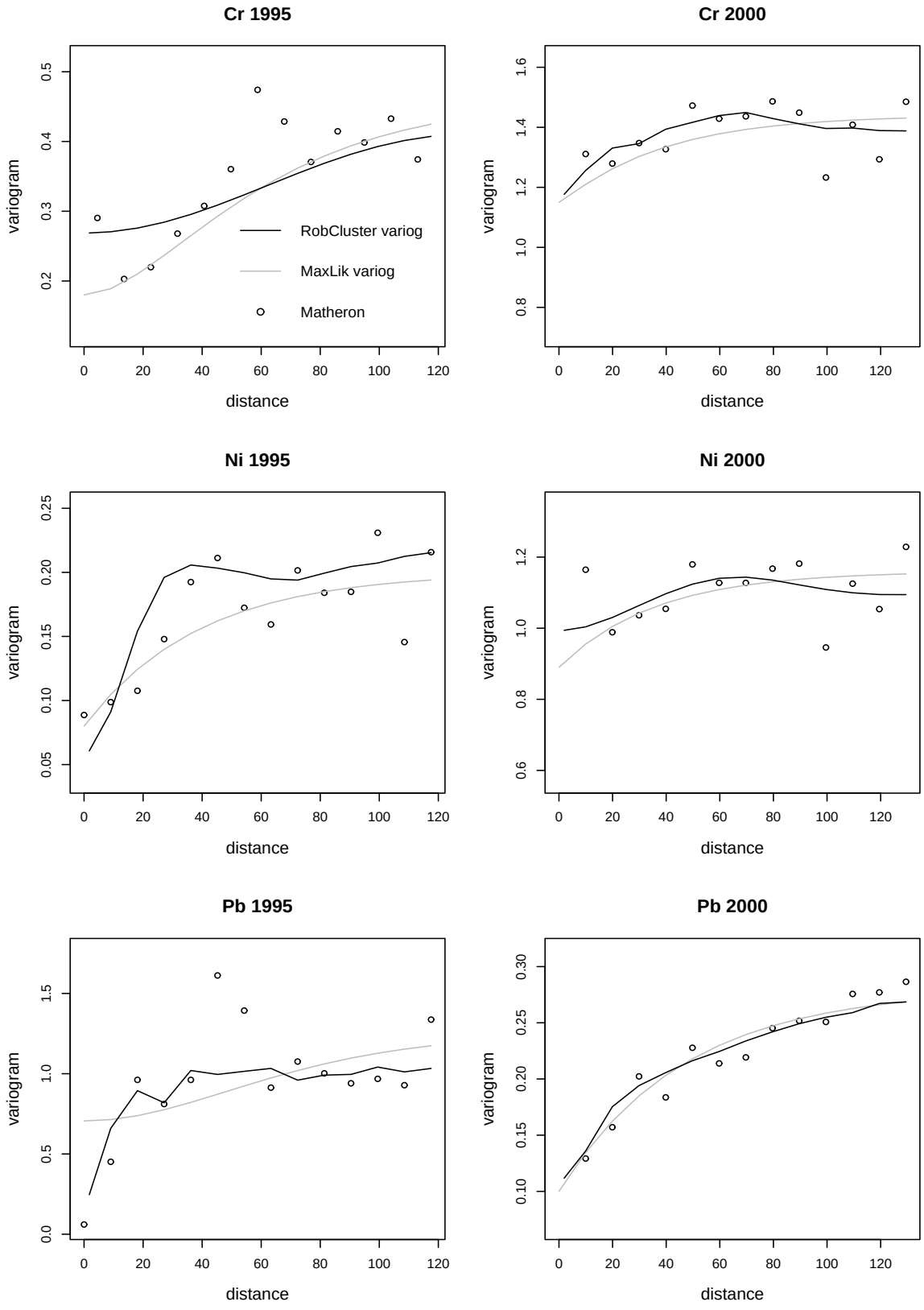


Figure 7.4: Variogram estimation for Cr, Ni and Pb data.

sample average, together with a smaller sample standard deviation than would be expected. For example in the case of Ni concentrations, a negative degree of preferability β in 1995 would explain a smaller average and a smaller S.D. than for the 2000 data. In the case of Pb, a positive degree of preferability β in 1995 would explain an average larger than that for the 2000 data (if together with a smaller S.D., which is not the case). However, the foregoing interpretations require caution, as characteristics specific to a given year, such as the presence of atypical values in zones not sampled in both years, may influence conclusions.

A more exact sign of the preferability issue, given a sample data set Y, can be obtained from a *rough* estimation of the degree of preferability β. Suppose that $Y = \{(\mathbf{x_i}, y_i) : i = 1, ..., n\}$, where the data value y_i is related to the concentration measurement of some heavy metal and $\mathbf{x_i}$ is the corresponding location. A preliminary β_0 can be obtained through a simple algorithm such as: first, use a kernel-type intensity estimator of the locations to derive $\widehat{\lambda}(\mathbf{x})$; and, then, choose β_0 such that $\log \widehat{\lambda}(\mathbf{x}) \simeq const + \beta_0 \mathbf{y}$. With respect to the 1995 data, we have obtained β_0 equals to -0.005, -0.4 and 0.5 for the measurements of Cr, Ni and Pb, respectively. The last two values suggest a further analysis.

Alternatively, we could have considered the log-likelihood function for $\lambda(.)$ based on data $\{\mathbf{x_1}, ..., \mathbf{x_n}\}$: $L(\lambda) = \Sigma_{i=1}^{n} \log \lambda(\mathbf{x_i}) - \int_D \lambda(\mathbf{x}) d\mathbf{x}$, where the integral term represents the mean of the Poisson distribution over the observation region D (see Diggle 2003). The preliminary β_0 could then be obtained as its maximimum likelihood estimate. Note that a finite sum approximation to the integral term is required; a discussion of this problem is found in Baddeley and Turner (2000).

The formal way to decide if one can assume preferential sampling is to proceed with Monte Carlo significance testing, as described in Schlather et al. (2004). Some ideas concerning these and similar hypothesis tests have already been discussed in Chapter 4. Suppose, then, that H_0 is the null hypothesis about the model which generates $Y = \{(\mathbf{x_i}, y_i) : i = 1, ..., 64\}$. H_0 will define the absence of the

preferability issue, t_1 is the observed value of a real valued statistic $T = h(Y)$ and we shall reject H_0 for a large value of t_1. The chosen test statistic is based on the conditional expectation $E(u)$ presented in Schlather et al. (2004). So, consider

- $E(u) = \mathrm{E}\left[Y(\mathbf{x}) \mid \mathbf{x}, \mathbf{x}' \in D,\ ||\mathbf{x} - \mathbf{x}'|| = u\right]$;
- $E(Y)$, the overall expectation (not dependent on inter-point distance).

We shall simulate 99 further data sets under H_0, and for $j = 1, ..., 100$ define t_j to be a measure of discrepancy between $\widehat{E}^j(u)$ and $\widehat{E}(Y)$ over the whole range of u. Considering a tolerance region for lag u, an estimator for $E(u)$ is obtained from the following sample average

$$N_u^{-1} \sum_{|\|\mathbf{x_i}-\mathbf{x_j}\|-u| \leq \frac{\varepsilon}{2}} (y_i + y_j)$$

where $\varepsilon > 0$ is a fixed bin-width and N_u is the number of pairs (y_i, y_j) for which $|\|\mathbf{x_i} - \mathbf{x_j}\| - u| \leq \frac{\varepsilon}{2}$. We choose to define our test statistic as the integrated squared difference

$$t_j = \int \{\widehat{E}^j(u) - \widehat{E}(Y)\}^2 du. \tag{7.1}$$

As described in Section 4.4, we can proceed to a formal test based on the rank of t_1 amongst t_j, because under H_0 all rankings of t_1 are equiprobable.

Assuming H_0, a similar pattern for the data process should be expected when interchanging the data values randomly over the different locations. Thus, we start with a randomization test. We reshuffle the data a large number of times and recalculate the test statistic for each iteration. Alternatively, because we are assuming Gaussian data, we can simulate 99 further data sets under H_0 by fixing the locations and generating Gaussian data on them. The parameters of the corresponding Gaussian model need to be estimated, which task can be carried out by ignoring the preferability issue and proceeding with classic maximum likelihood. Another option is to proceed with a non-parametric estimation of the variance-covariance matrix. Both, parametric and non-parametric, estimators of the spatial dependency structure are represented in Figure 7.4.

We conclude that, in the specific case of Cr, H_0 should not be rejected; such result, together with the foregoing degree of preferability $\beta_0 \simeq 0$, supports the idea that there was no preferential sampling for chromium. However, in the case of Ni and Pb, the rejection of H_0 is confirmed with a formal test based on (7.1) at the 5% level of significance. The foregoing non-null estimates of β_0 and, now, the results of these MC tests support the idea that some preferential sampling has occurred for nickel and lead.

7.1.2 Kriging and cross validation

In order to complement the analysis included in Chapter 6 concerning the effect of preferential sampling, we now want to evaluate the two sampling models from 1995 and 2000 through their performance in actual kriging situations. Thus we shall adopt the technique of cross-validation for the analysis of moss data.

We start by deriving the surface of the ordinary kriging prediction, and the surface of the corresponding standard error, for each of the three heavy metals. To achieve that we perform spatial prediction for a fixed variance structure using global neighbourhood. The spatial dependency structures considered are those specified through the variograms of Figure 7.4. The resulting parametric and non-parametric kriging estimates are represented in Figures 7.5 and 7.6, respectively. The overall appreciation of the P and NP kriging surfaces points to similar results. Perhaps the NP approach is typically associated with smoother transitions of the kriging estimates. For the 1995 data, it is worth noting that higher values of Pb seem to be found in a clustered zone, reflecting a possible positive degree of preferability. Likewise, smaller values of Ni seem to occur in a clustered zone, reflecting a possible negative degree of preferability. As expected no specific clustering pattern is found for the values of Cr.

With respect to the comparison between the kriging estimates of 1995 and 2000, more similar surfaces would be expected, even though the sample size has doubled. This issue is more prominent in the case of Ni, in which the exceptionally

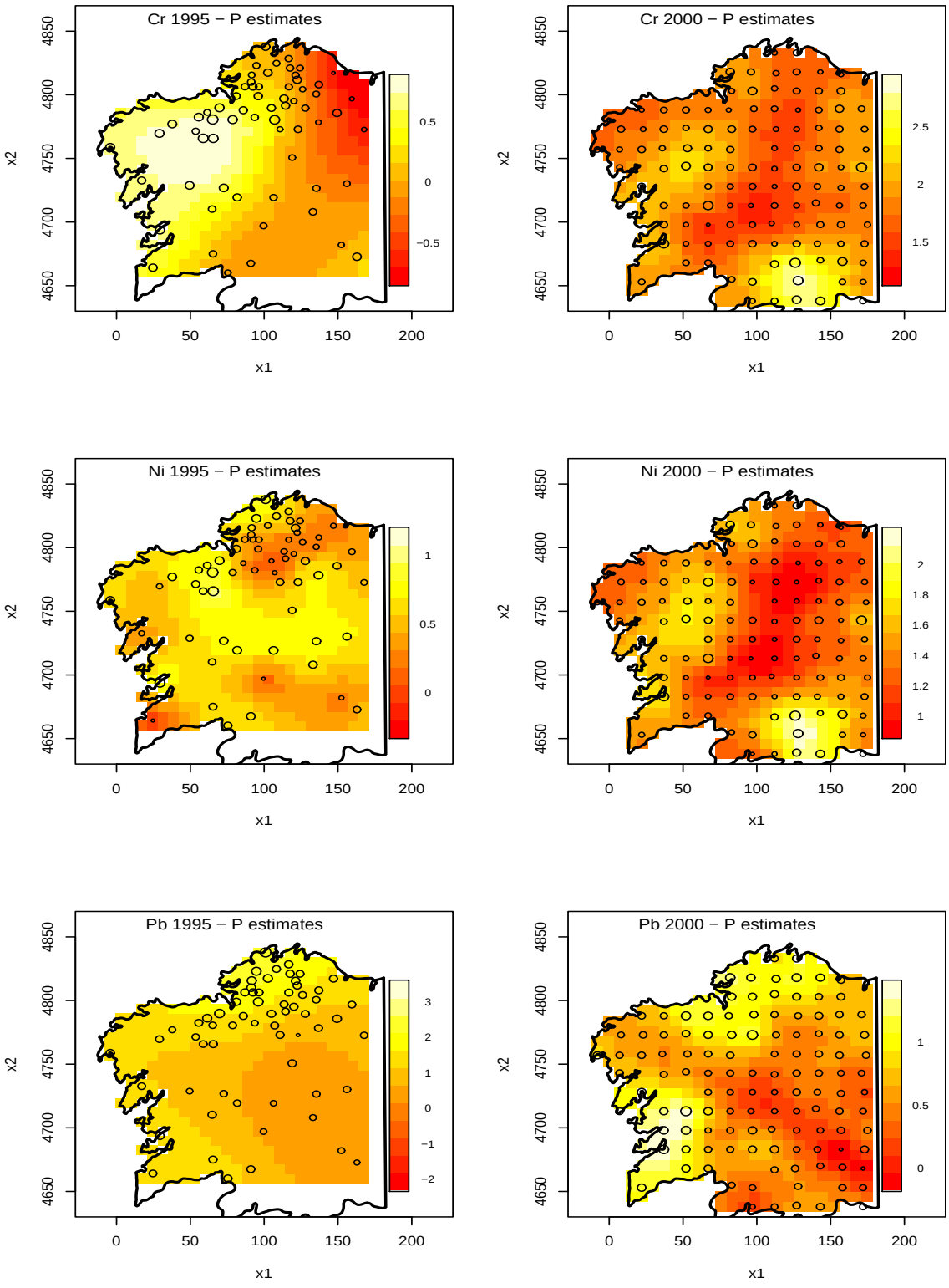


Figure 7.5: Parametric kriging estimates for Cr, Ni and Pb in moss samples collected in Galicia in 1995 and 2000.

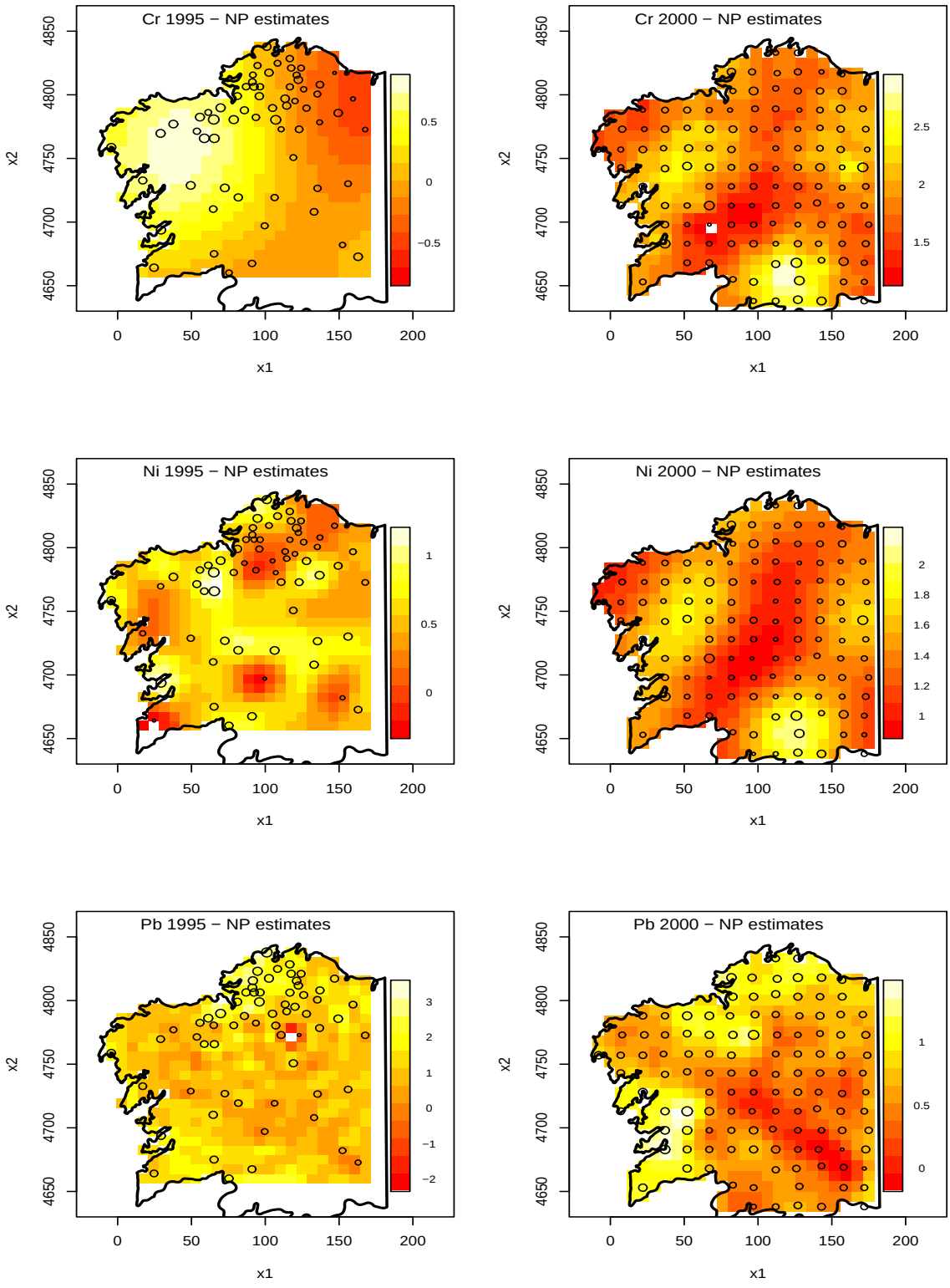


Figure 7.6: Non-parametric kriging estimates for Cr, Ni and Pb in moss samples collected in Galicia in 1995 and 2000.

high values found in the south of Galicia in 2000 seem to strongly affect the global results. In the case of Pb, it is possible to identify the spatial pattern of 1995 five years later; i.e. the area around of Coruña continues to be pointed out as a zone with high values. Furthermore, the concentrations of Pb measured in 2000 allow a better characterization of the area surrounding Vigo and Pontevedra, where a second zone of high values is found. Bear in mind that the main source of Pb emission, apart from oil refineries, is road traffic. Finally, in the case of Cr, once again the exceptionally high concentrations found in the south of Galicia in 2000 seem to affect the global results, making it difficult to maintain the same spatial pattern in 1995 and 2000; note, however, that the area surrounding Santiago continues to present high values of Cr in 2000.

We now proceed with cross-validation. The fundamental idea behind this technique, also called "the leave-one-out method", is to estimate the concentration measurement $Y(\mathbf{x})$ at each sample point $\mathbf{x_i}$ from neighbouring data $Y_j = Y(\mathbf{x_j}),\ j \neq i,$ as if $Y_i = Y(\mathbf{x_i})$ were unknown. In this way at every sample point $\mathbf{x_i}$ we get a kriging estimate $\widehat{Y}_{-i} = \widehat{Y}(\mathbf{x_i})$ and the associated kriging variance σ^2_{-i}. Bearing in mind that the true value Y_i is known, we can compute the prediction error $PE_i = Y_i - \widehat{Y}_{-i}$. If $\gamma(u)$ is the theoretical variogram, PE_i is a random variable with mean zero and variance σ^2_{-i} (Chilès and Delfiner 1999). Moreover, the standardized error $e_i = PE_i/\sigma_{-i}$ is a zero-mean unit-variance random variable. Typically, comparing the results of two cross-validations performed under different conditions is useful in helping one to decide between two models. In our case, we wish to compare the parametric and the non-parametric variogram models, and to inspect the effect of the sampling procedure. In Table 7.2, we include the results for the mean square error and the mean standardized squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{-i})^2 \qquad \text{and} \qquad MSSE = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{Y}_{-i})^2}{\widehat{\sigma^2_{-i}}},$$

respectively. Note that the value of MSE is often essentially due to the contribution of a few points, data poorly "explained" by their neighbors, which possibly explains

		MSE			MSSE		
		Cr	Ni	Pb	Cr	Ni	Pb
1995	P	0.265	0.187	0.798	1.167	1.292	0.996
	NP	0.247	0.172	0.801	0.837	1.305	1.032
2000	P	1.738	1.331	0.217	1.036	1.028	1.012
	NP	1.672	1.276	0.217	0.974	0.978	0.986

Table 7.2: The mean square error and the mean standardized squared error for Cr, Ni and Pb in 1995 and 2000.

higher values for Cr and Ni in 2000 than in 1995.

In an overall appreciation, and with respect to the MSE, we want to highlight the smaller values normally associated with the NP approach, reinforcing the advantages of the NP variogram. Note that these gains are slightly larger in 2000, because we have have a larger sample size; moreover, note that this comparison is being made in a logarithm scale, otherwise the differences would be slightly more important. With respect to the MSSE, we want to highlight that these values are closer to 1 in 2000 than in 1995. Alternatively, one could analyse the histogram of the standardized errors and confirm that they would fit a standard normal better in 2000 than in 1995. We think such results, mainly those for Ni, support the need for a model-based approach to preferential sampling.

7.2 A model-based approach

The term *model-based geostatistics* was first used by Diggle et al. (1998) to describe the application of formal statistical models and likelihood-based methods of inference to geostatistical problems. We suggest a solution for the preferability issue following this approach. We aim to apply an explicit parametric stochastic model to preferential sampling, and proceed with likelihood inference to estimate

the parameters of the proposed model to allow for spatial prediction.

The current work was developed under Gaussian assumptions. Additionally, we continue to assume the stationarity of the underlying spatial process.

7.2.1 Related work

Our first concern is to decide, given a sample of spatial data, if it is reasonable to assume preferential sampling. An informed decision may be based on the Monte Carlo tests proposed by Schlather et al. (2004) to detect dependence between marks and locations of marked point processes. Actually, in the example of an application to moss data presented above, we adopt and recommend the conditional expectation described in Schlather et al. (2004) to test whether a sample is preferential.

Díaz and Ayala (1999) also proposes a Monte Carlo test for the hypothesis of no dependency between marks and locations, motivated, however, by the study of corneal endothelium morphology. This morphology, typically showing a very regular polygonal pattern, can be modified by stress situations such as cataract surgery or corneal transplantation. The location is defined as the centroid of each cell, which is marked by the area of its corresponding cell.

Assuming the hypothesis that there is no relationship between location and observed area, a similarly marked point pattern should be expected when interchanging the areas randomly over the different locations. This suggests a randomization test. The authors then develop an index of spatial homogeneity, using the mark variogram, that quantifies the variability of cell areas taking into account their spatial arrangement. The new index is the $p-$value of the test validating the hypothesis.

Another work, in this case concerning the stochastic dependence between two spatial processes, is described in Foxall and Baddeley (2002). These authors define

non-parametric measures of dependence, using distance methods[1], in bivariate spatial processes (P, Y) where P is a point process and Y is any random set. Note that there are some well known proposals for assessing dependencies between point processes (see e.g. Cox 1972, Baddeley, Moller and Waagepetersen 2000; or Diggle 2003 for further references), but most of them are considered not to generalize easily to the case where the latter is a random set.

Their non-parametric approach makes it possible to proceed with conditional inference about P given Y. In accordance with an example which they provide, suppose P represents the point pattern of gold deposits and Y some geological fault pattern. Here, the aim would be to predict gold deposits from the more easily observable geological faults.

Bear in mind that, in the case of preferential sampling, we wish to model the dependency of a point process and some Gaussian field, as described in Moller et al. (1998). Moreover, our goal of prediction is quite different, as it might be defined as the value of this Gaussian field over the observed region, given some measurements of Gaussian data and their corresponding locations.

Our literature review, now focused on our prediction goal when considering data in some way preferentially located, led us to generic solutions such as *declustering*, *detrending data*, *detecting outliers* or *using the so-called non-ergodic estimators*. These topics are typically included in any classic geostatistical literature, such as, for example Goovaerts (1997) or Isaaks and Srivastava (1989).

In Curriero, Hohn, Liebhold and Lele (2002), they describe a so-called non-ergodic approach as a solution when sampled data are preferentially located and exhibit a skewed frequency distribution. The idea is, before applying kriging techniques, to use a non-ergodic covariogram estimator[2] instead of using the traditional

[1] Analogous to the use of the summary functions F and G for univariate point patterns, described in Chapter 5.

[2] $\widehat{C}(u) = \frac{1}{|N(u)|} \sum_{N(u)} (Y(\mathbf{x_i}) - \overline{Y}(u_i))(Y(\mathbf{x_j}) - \overline{Y}(u_j))$, where statistic $\overline{Y}(u_i)$ represents the

one. The difference between the two is the subtraction of *local* means versus the subtraction of a global mean.

Although this use of different data for each lag may be considered the advantage of the non-ergodic approach, it has a negative impact on estimation at the crucial shorter lags, where the number of available data pairs are often sparse. The authors argue that an alternative solution lies in a detrending of the data followed by robust estimation of the residual variogram. In spite of their contributions, both solutions neglect the interaction between the point process and the underlying field.

To our knowledge, in geostatistics, the preferability issue[3] has always been either ignored or addressed by already familiar generic techniques.

7.3 Likelihood inference

Taking into account that we propose a parametric model for preferential sampling, the natural way to proceed with parameter estimation is to use the likelihood function of the sample data. Loosely speaking, the likelihood of a set of data is the probability of obtaining that specific set of data, given the chosen probability distribution model. This expression implies the existence of model parameters, although these are unknown. The values of these parameters that would maximize the sample likelihood are the *maximum likelihood estimates* or sometimes referred to as the MLE's.

The sample data are of the form $(\mathbf{x_i}, y_i),\ i = 1, ..., n$. When restricted to an observation region $D \subset I\!R^2$, the set of locations $\{\mathbf{x_i} : i = 1, ..., n\}$ identifies the *preferential sampling design*. The set $\{y_i : i = 1, ..., n\}$ represents the measurements associated with these locations, where y_i are realizations of $Y_i = Y(\mathbf{x_i})$ and $Y(.)$ is the measurement process. Under Gaussian assumptions, we shall consider

$$\mathbf{Y} = (Y_1, ..., Y_n)^t \ \sim \ \text{MVN}(\mu_Y \mathbf{1}, \mathbf{\Sigma_Y}),$$

mean of all observations appearing as $Y(\mathbf{x_i})$ in set $N(u)$.

[3] As defined in Remark 6.1.

where **1** denotes the n-element vector of ones. It is also assumed that

$$Y_i = S(\mathbf{x_i}) + Z_i,\ i = 1, ..., n,$$

where $S(.)$ is an unobserved stochastic process, target of prediction, with mean μ, variance σ^2 and correlation function $\rho(u; \phi)$, being Y stochastically dependent on S. The $Z_1, ..., Z_n$ are mutually independent, identically distributed with $Z_i \sim \mathrm{N}(0, \tau^2), i = 1, ..., n$. So, we have

$$\mu_Y = \mu \qquad \text{and} \qquad \mathbf{\Sigma_Y} = \sigma^2 \mathbf{R_Y}(\phi) + \tau^2 \mathbf{I_n}$$

where $\mathbf{I_n}$ is the $n \times n$ identity matrix and $\mathbf{R_Y}(\phi)$ is the $n \times n$ matrix with $(i, j)^{th}$ element $\rho(u_{ij})$ being $u_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\|$.

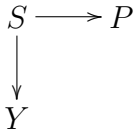
The distribution of the *preferential sampling design*, defined by the point process P, is not independent of S (see Remark 6.1). Hence, it will have to be considered in the definition of our likelihood function. Its representation is assumed to be

$$P \sim \text{Poisson}\,(\Lambda(\mathbf{x}))$$

where $\log(\Lambda(\mathbf{x})) \sim$ SGP [4], being $\Lambda(\mathbf{x}) = \exp\{\alpha + \beta S(\mathbf{x})\}$ with $|\beta|$ a parameter for the degree of preferability and α some constant value.

7.3.1 Complete data likelihood

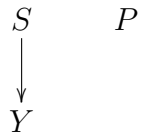
As stated in Remark 6.2, the geostatistical model for preferential sampling is specified by the joint distribution $[S, Y, P]$. The related complete data likelihood will then be given by this distribution, for which a diagram representation might be



It describes the stochastic dependencies between processes, with $Y \not\perp S$ and $P \not\perp S$. In the case of the traditional geostatistical model, note that this diagram

[4] Remember that SGP stands for stationary Gaussian process.

representation simplifies to



The presence of the Gaussian field S involves an infinite number of random variables and its likelihood function is not in closed form. S can be, however, approximated by a random discretization function, here represented by S^*, becoming $[S, Y, P] \simeq [S^*][P|S^*][Y|S^*, P]$.

If we assume that the distribution of P is not relevant for the conditional distribution of Y, as described in the diagram representation of the preferential sampling model, then $[S, Y, P] \simeq [S^*][Y|S^*][P|S^*] = [S^*, Y][P|S^*]$.

Suppose $\boldsymbol{\theta}$ is the vector of all model parameters, then the complete likelihood can be represented by

$$L_C(\boldsymbol{\theta}|\ S^*, Y, P) = [S^*|\mu, \sigma^2, \phi]\ [Y|S^*, \tau^2]\ [P|S^*, \beta] = [S^*, Y|\mu, \sigma^2, \phi, \tau^2]\ [P|S^*, \beta]$$

or, invoking an equivalent formulation, by

$$f(\mathbf{s}^*, \mathbf{y}, \mathbf{x}) \equiv f(\mathbf{s}^*) f(\mathbf{y}|\mathbf{s}^*) f(\mathbf{x}|\mathbf{s}^*) \equiv f(\mathbf{s}^*, \mathbf{y}) f(\mathbf{x}|\mathbf{s}^*). \tag{7.2}$$

Suppose now that S^* is regarded as a very fine grid with dimension $N = n_{grid} \times n_{grid}$, then

$$\mathbf{S}^* = (S(\mathbf{x}_\mathbf{1}^*), ..., S(\mathbf{x}_\mathbf{N}^*))^t \ \sim\ \text{MVN}(\mu\mathbf{1}, \boldsymbol{\Sigma}_\mathbf{S}),$$

where $\mathbf{x}_\mathbf{1}^*, ..., \mathbf{x}_\mathbf{N}^*$ are the centre points of all small squares on the fine grid, $\mathbf{1}$ denotes the N-element vector of ones and $\boldsymbol{\Sigma}_\mathbf{S} = \sigma^2 \mathbf{R}_\mathbf{S}(\phi)$, being $\mathbf{R}_\mathbf{S}(\phi)$ the $N \times N$ correlation matrix defined for all possible distances between points of the fine grid.

In this case, as in most cases, it is easier to take the logarithm of the likelihood function to proceed with parameter estimation. This happens mainly for computational reasons, as we are dealing with very high dimensional multivariate normals

and very small values for the inhomogeneous Poisson density. The log-likelihood is given by

$$l_C(\boldsymbol{\theta}|\ S^*, Y, P) = \log L_C(\boldsymbol{\theta}|\ S^*, Y, P) = \log f(\mathbf{s}^*, \mathbf{y}) + \log f(\mathbf{x}|\mathbf{s}^*) \tag{7.3}$$

a) being

$$\log f(\mathbf{s}^*, \mathbf{y}) = -0.5 \left\{ \log |\boldsymbol{\Sigma}| + \mathbf{v}^t \boldsymbol{\Sigma}^{-1} \mathbf{v} \right\}$$

where $\mathbf{v}^t = (\mathbf{s}^*\ \mathbf{y}) - \mu \mathbf{1}^t$, $\boldsymbol{\Sigma}_{\mathbf{SY}} = \sigma^2 \mathbf{R}_{\mathbf{SY}}(\phi)$ and $\mathbf{R}_{\mathbf{SY}}(\phi)$ is the $N \times n$ correlation matrix defined for all possible distances between points of the fine grid and the sample points; finally, $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{S}} & \boldsymbol{\Sigma}_{\mathbf{SY}} \\ \boldsymbol{\Sigma}_{\mathbf{SY}}{}^t & \boldsymbol{\Sigma}_{\mathbf{Y}} \end{pmatrix}$;

b) and

$$\log f(\mathbf{x}|\mathbf{s}^*) = \beta \sum_{i=1}^{n} S(\mathbf{x_i}) - n\ \log \widehat{\mu}_D. \tag{7.4}$$

Here, as described in Diggle (2003), we are writing the likelihood function of our inhomogeneous Poisson process as a product of a Poisson distribution for the n independent $\mathbf{x_i}$ locations, whose common distribution has density $\frac{\lambda(\mathbf{x})}{\mu_D}$ with $\mu_D = \int_D \lambda(\mathbf{x}) d\mathbf{x}$. Consequently, we are considering

$$f((\mathbf{x_1}, ..., \mathbf{x_n})|\mathbf{s}^*) = \frac{\prod_{i=1}^{n}\ \exp(\beta S(\mathbf{x_i}))}{(\mu_D)^n}$$

where a numerical approximation to μ_D is given by $\widehat{\mu}_D = \sum_{k=1}^{N} \lambda(\mathbf{x_k^*}) \times w_k$ with w_k indicating the area of each quadrature cell (Baddeley and Turner 2000); in a square grid, we will be assuming $\widehat{\mu}_D = \sum_{k=1}^{N} \exp(\beta S(\mathbf{x_k^*})) \times \frac{1}{N}$.

In Figure 7.7, we plot the complete log-likelihood given in (7.3) for a simulated data set. Aiming at reducing the dimensionality of this likelihood surface, we choose to plot here the likelihood for each model parameter by fixing the other parameters at their true values. Bear in mind that these are not profile likelihoods, as it will be described in Section 7.4.1, but they can be referred to as *slices* of the

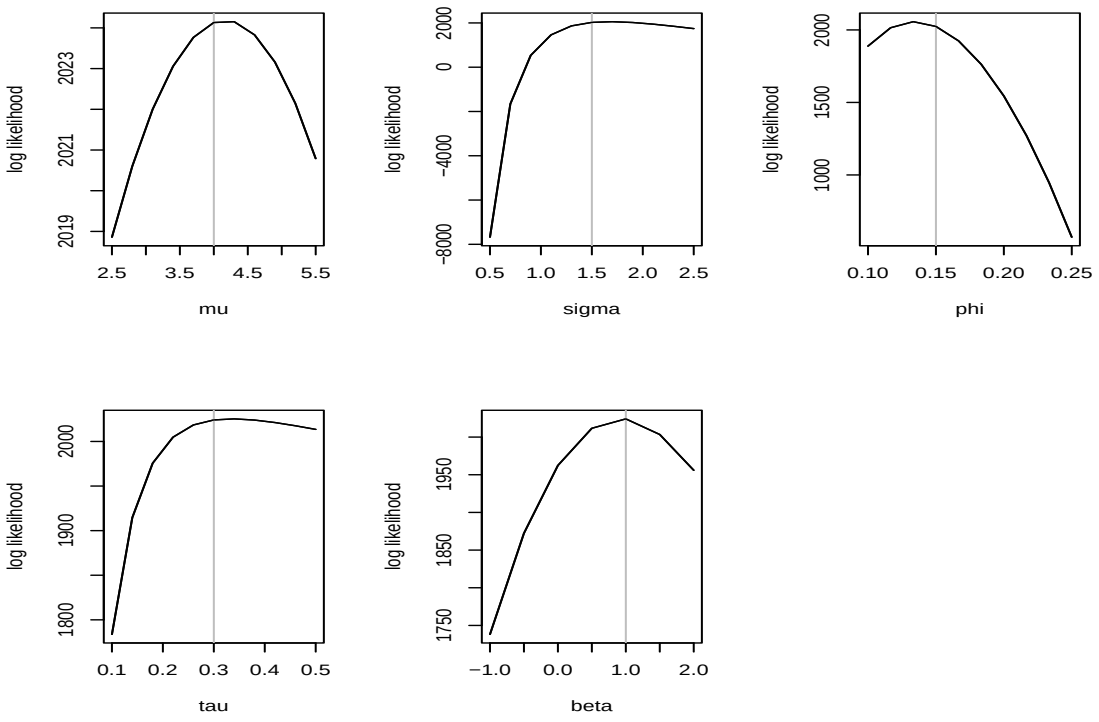


Figure 7.7: Complete log-likelihood given in (7.3) for a simulated data set (slices, not profiles). The grey line gives the true value of each model parameter.

likelihood. The true value of each parameter is identified by a grey line in the corresponding chart.

Alternatively, each plot in Figure 7.7 could be obtained by fixing the remaining parameters at some initial and rough estimate of $\boldsymbol{\theta}_0$. As we will discuss in later Sections, a good candidate for $\widehat{\boldsymbol{\theta}}_0$ can be derived by ignoring the preferability issue. An iterative procedure could then be applied to improve the final $\widehat{\boldsymbol{\theta}}$.

7.3.2 Likelihood of observed data

The marginal likelihood function of the sample data, $L(\boldsymbol{\theta}|\ Y, P)$, is derived from the density $f(\mathbf{y}, \mathbf{x}) = \int f(\mathbf{s}, \mathbf{y}, \mathbf{x}) d\mathbf{s}$. According to Equation (7.2), one has

$$L(\boldsymbol{\theta}|\ Y, P) = \mathrm{E}_{\mathbf{S}^*}\left[f(\mathbf{y}|\mathbf{s}^*) f(\mathbf{x}|\mathbf{s}^*)\right].$$

The expectation can now be approximated by Monte Carlo, becoming

$$L_1^{MC}(\boldsymbol{\theta}|\ Y, P) = \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}|\mathbf{s_j^*}) f(\mathbf{x}|\mathbf{s_j^*})$$

Consequently, the log of the marginal likelihood function is

$$l_1^{MC}(\boldsymbol{\theta}|\ Y, P) = \log\left\{\sum_{j=1}^{m} \exp\left(\log f(\mathbf{y}|\mathbf{s_j^*}) + \log f(\mathbf{x}|\mathbf{s_j^*})\right)\right\} - \log m$$

where

- $\mathbf{S_1^*}, ..., \mathbf{S_m^*} \sim \text{MVN}(\mu\mathbf{1}, \boldsymbol{\Sigma_S})$
- $\log f(\mathbf{y}|\mathbf{s_j^*}) = \log \prod_{i=1}^{n} f(y_i|S_j(\mathbf{x_i})) = \sum_{i=1}^{n} \log f(y_i|S_j(\mathbf{x_i}))$,

  because the $Y_1, ..., Y_n$ are mutually independent, conditional on $S(.)$. Note that the conditional distribution of Y_i given $S(.)$ is Gaussian with mean $S(\mathbf{x_i})$ and variance τ^2.
- $\log f(\mathbf{x}|\mathbf{s_j^*})$ is given in (7.4).

Alternative expression

In practice to obtain an expression for the likelihood of the observed data, it was found more convenient, as more computationally efficient, to consider instead

$$f(\mathbf{y}, \mathbf{x}) = \int f(\mathbf{y}) f(\mathbf{x}|\mathbf{s}) f(\mathbf{s}|\mathbf{y}) d\mathbf{s},$$

that is an equivalent formulation also resulting from Equation (7.2). In this way, the marginal likelihood function can be written as $L(\boldsymbol{\theta}|\ Y, P) = f(\mathbf{y})\ \mathrm{E}_{\mathbf{S}^*|\mathbf{Y}}\left[f(\mathbf{x}|\mathbf{s}^*)\right]$, and the corresponding log-likelihood function becomes

$$l(\boldsymbol{\theta}|\ Y, P) = \log f(\mathbf{y}) + \log\ \mathrm{E}_{\mathbf{S}^*|\mathbf{Y}}\left[f(\mathbf{x}|\mathbf{s}^*)\right].$$

Here, we highlight that the first term is, by now, assumed to be the one used under Gaussian assumptions in classical geostatistics (see e.g. Diggle et al. 2003), and the second is the *correction term* obtained for the preferability issue.

The Monte Carlo approximation becomes

$$l_2^{MC}(\boldsymbol{\theta}|\ Y, P) = \log f(\mathbf{y}) + \log\ \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{x}|\mathbf{s_j^*}) =$$

$$= \log f(\mathbf{y}) + \left(\log\left\{\sum_{j=1}^{m}\exp\left(\log f(\mathbf{x}|\mathbf{s_j^*})\right)\right\} - \log m\right) \tag{7.5}$$

where

- $\mathbf{S_1^*}, ..., \mathbf{S_m^*} \ \sim\ \text{MVN}(\boldsymbol{\mu}_{\mathbf{S^*|Y}}, \boldsymbol{\Sigma}_{\mathbf{S^*|Y}})$

  Applying properties of the multivariate Gaussian distribution (see e.g. Anderson 1984), the joint distribution was used to derive the above conditional distribution $[\mathbf{S}^*|\mathbf{Y}]$, so that

  - $\boldsymbol{\mu}_{\mathbf{S^*|Y}} = \mu\mathbf{1} + \boldsymbol{\Sigma}_{\mathbf{SY}}{\boldsymbol{\Sigma}_{\mathbf{Y}}}^{-1}(\mathbf{y} - \mu\mathbf{1})$
  - $\boldsymbol{\Sigma}_{\mathbf{S^*|Y}} = \boldsymbol{\Sigma}_{\mathbf{S}} - \boldsymbol{\Sigma}_{\mathbf{SY}}{\boldsymbol{\Sigma}_{\mathbf{Y}}}^{-1}{\boldsymbol{\Sigma}_{\mathbf{SY}}}^{t}$.

- $\log f(\mathbf{y}) = -0.5\left\{\log|\boldsymbol{\Sigma}_{\mathbf{Y}}| + (\mathbf{y} - \mu\mathbf{1})^t{\boldsymbol{\Sigma}_{\mathbf{Y}}}^{-1}(\mathbf{y} - \mu\mathbf{1})\right\}$

- $\log f(\mathbf{x}|\mathbf{s_j^*})$ is given in (7.4).

In Figure 7.8, we compare the log-likelihood of the observed data against the complete log-likelihood for a simulated data set, whose expressions are given in (7.5) and (7.3), respectively. This example aims at illustrating how much one loses by not having $\mathbf{S}^*$, it being unavailable for real data applications, and having instead $\mathbf{S}^*|\mathbf{Y}$. It results in a larger variability for the covariance parameters, σ, ϕ and τ, suggesting a cautious Monte Carlo approximation for expression (7.5).

7.4 Estimation of model parameters

The maximization of the likelihood function, under an assumed model, provides estimates which are unbiased and efficient when applied to large samples (see e.g. Azzalini 1996). In the case of the preferential sampling model, we are considering that the estimation of parameter $\boldsymbol{\theta} = (\mu, \sigma, \phi, \tau, \beta)^t$ is required.

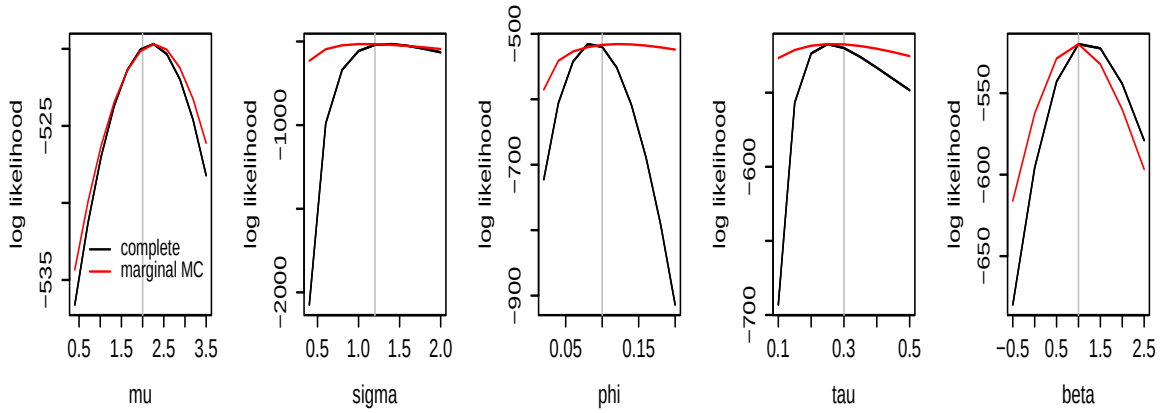


Figure 7.8: Complete and marginal log-likelihoods given in (7.3) and (7.5), respectively, for a simulated data set (slices, not profiles). The grey line gives the true value of each model parameter.

In principle, the log-likelihood surface defined by (7.5) gives us information about the parameter estimators and their corresponding variability. However, the dimensionality of this surface does not allow us a direct analysis. This is a typical difficulty of multiparameter models when applying likelihood-based methods. This problem is sometimes approached by using the profile log-likelihood, as briefly described in next Section.

7.4.1 Profile log-likelihoods

Suppose that $\boldsymbol{\theta}$ is divided into two parts $(\boldsymbol{\theta}_{\mathbf{I}}, \boldsymbol{\theta}_{\mathbf{II}})$: the part of primary interest or the parameter of interests $\boldsymbol{\theta}_{\mathbf{I}}$, and the part of secondary interest or the nuisance parameter $\boldsymbol{\theta}_{\mathbf{II}}$. Additionally, suppose $l(\boldsymbol{\theta}_{\mathbf{I}}, \boldsymbol{\theta}_{\mathbf{II}})$ to be the log-likelihood. The profile log-likelihood is defined as

$$l_p(\boldsymbol{\theta}_{\mathbf{I}}) = l(\boldsymbol{\theta}_{\mathbf{I}}, \widehat{\boldsymbol{\theta}}_{\mathbf{II}}(\boldsymbol{\theta}_{\mathbf{I}})) = \max_{\boldsymbol{\theta}_{\mathbf{II}}} \{l(\boldsymbol{\theta}_{\mathbf{I}}, \boldsymbol{\theta}_{\mathbf{II}})\}.$$

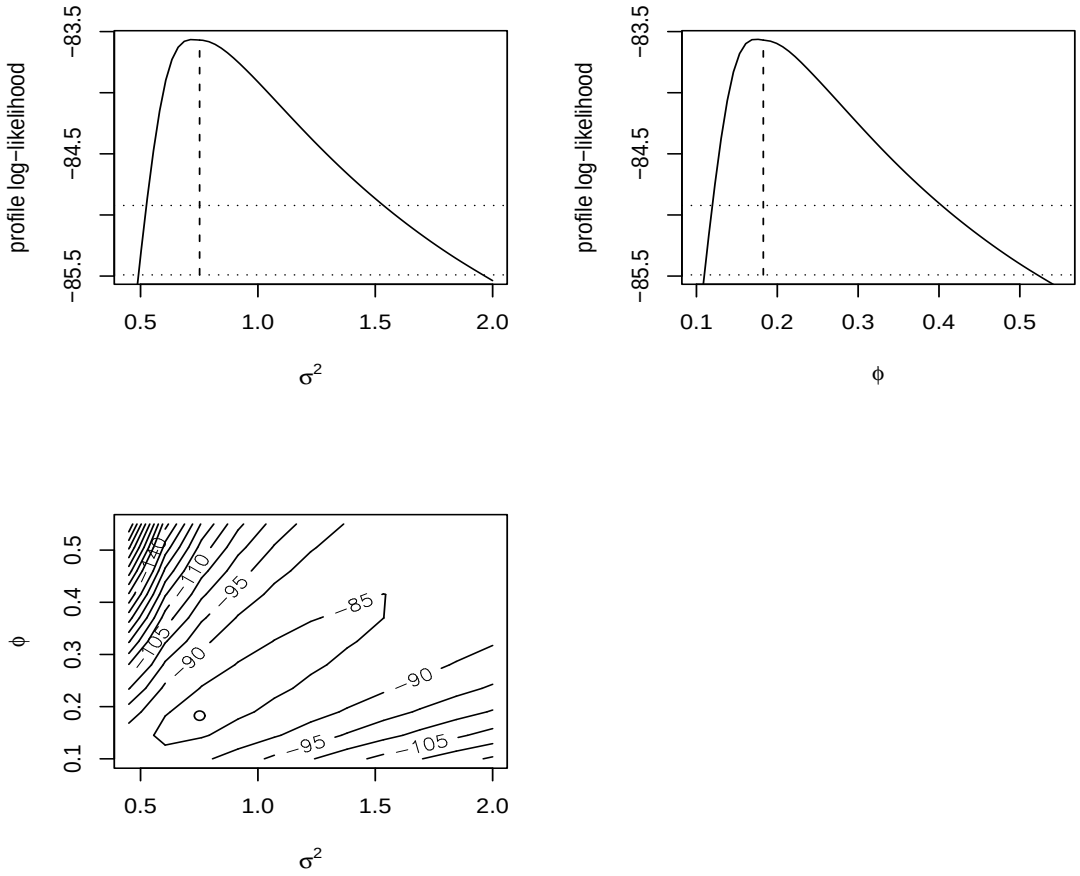


Figure 7.9: Profile log-likelihood from the *s100* simulated data set; computed for the covariance parameters σ^2 and ϕ.

The profile log-likelihood $l_p(\boldsymbol{\theta}_\mathbf{I})$ is used as if it is the likelihood from a model with only parameter $\boldsymbol{\theta}_\mathbf{I}$, becoming

$$\widehat{\boldsymbol{\theta}}_\mathbf{II} = \max_{\boldsymbol{\theta}_\mathbf{I}} \{l_p(\boldsymbol{\theta}_\mathbf{I})\}.$$

The idea is to use the profile likelihood in exactly the same way in which the *ordinary* likelihood is used when drawing inferences. In the case of point estimation, maximizing the profile log-likelihood with respect to $\boldsymbol{\theta}_\mathbf{I}$ leads to the same estimator $\widehat{\boldsymbol{\theta}}_\mathbf{I}$ as is obtained by maximizing the log-likelihood simultaneously with respect to $\boldsymbol{\theta}_\mathbf{I}$ and $\boldsymbol{\theta}_\mathbf{II}$ [5] (see e.g. Murphy and van der Vaart 2000 for further

[5] Note that this equivalence does not carry over to interval estimation or hypothesis testing.

details).

Figure 7.9 includes a simple example to illustrate this approach application, when making $\beta = 0$ and restricting $\boldsymbol{\theta} = (\sigma^2, \phi)^t$. It shows the log-likelihood surface and the corresponding profile log-likelihoods when analysing the *s100* data set, a simulated data set available in R package from Ribeiro Jr and Diggle (2001). As expected, the estimator derived for $\boldsymbol{\theta}$ from the top-graphs and the one from the bottom-graph are the same.

7.4.2 Importance sampling

This Section includes a short description of the Monte Carlo maximum likelihood method that we have considered adopting for our parameter estimation. This method has been proved to be useful for those cases where the likelihood is analytically intractable (Geyer 1994, Moller and Waagepetersen 2003 and Geyer 1999). It suggests that the likelihood function $L(\boldsymbol{\theta})$ be written as a ratio against a fixed $\boldsymbol{\theta_0}$. This type of approach for classical model-based geostatistics is discussed in Christensen (2004).

Considering our estimation problem as a *missing data problem*, the likelihood can be approximated by making importance sampling as follows. First, suppose the case where τ^2 and β are fixed, and $\boldsymbol{\theta} = (\mu, \sigma^2, \phi)^t$. We can suppress the dependence on τ^2 and β and the likelihood function can then be written as

$$L(\boldsymbol{\theta}) = \int f(\mathbf{y}|\mathbf{s}^*) f(\mathbf{s}^*|\boldsymbol{\theta}) f(\mathbf{x}|\mathbf{s}^*) d\mathbf{s}^* =$$

$$= \int \frac{f(\mathbf{y}|\mathbf{s}^*) f(\mathbf{s}^*|\boldsymbol{\theta}) f(\mathbf{x}|\mathbf{s}^*)}{f(\mathbf{y}|\mathbf{s}^*) f(\mathbf{x}|\mathbf{s}^*) f(\mathbf{s}^*|\boldsymbol{\theta} = \boldsymbol{\theta}_0)} f(\mathbf{y}|\mathbf{s}^*) f(\mathbf{x}|\mathbf{s}^*) f(\mathbf{s}^*|\boldsymbol{\theta} = \boldsymbol{\theta}_0) \propto$$

$$\propto \int \frac{f(\mathbf{s}^*|\boldsymbol{\theta})}{f(\mathbf{s}^*|\boldsymbol{\theta} = \boldsymbol{\theta}_0)} \; f(\mathbf{s}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} = \boldsymbol{\theta}_0) \; d\mathbf{s}^* = \mathrm{E}_{\boldsymbol{\theta}_0} \left[\frac{f(\mathbf{s}^*|\boldsymbol{\theta})}{f(\mathbf{s}^*|\boldsymbol{\theta} = \boldsymbol{\theta}_0)} \;\middle|\; \mathbf{y}, \mathbf{x} \right] \qquad (7.6)$$

where the conditional density $f(\mathbf{s}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} = \boldsymbol{\theta}_0) \propto f(\mathbf{y}|\mathbf{s}^*) f(\mathbf{x}|\mathbf{s}^*) f(\mathbf{s}^*|\boldsymbol{\theta} = \boldsymbol{\theta}_0)$, and $\mathrm{E}_{\boldsymbol{\theta}_0}[. \,|\mathbf{y}, \mathbf{x}]$ denotes expectation with respect to $f(. \,|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} = \boldsymbol{\theta}_0)$.

Maximum likelihood estimates can now be calculated by maximizing next Monte Carlo approximation to (7.6)

$$L_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m}\frac{f(\mathbf{s_j^*}|\boldsymbol{\theta})}{f(\mathbf{s_j^*}|\boldsymbol{\theta}=\boldsymbol{\theta}_0)}$$

where $\mathbf{s_1^*}, ..., \mathbf{s_m^*}$ are sampled from the conditional multivariate normal distribution $f(\mathbf{s}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} = \boldsymbol{\theta}_0)$.

In case of $\boldsymbol{\theta} = (\mu, \sigma^2, \phi, \tau^2, \beta)^t$, the equivalent expression for the Monte Carlo approximation would be

$$L_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m}\frac{f(\mathbf{y}|\mathbf{s}_j^*;\tau)f(\mathbf{x}|\mathbf{s}_j^*;\beta)f(\mathbf{s}_j^*|\mu,\sigma,\phi)}{f(\mathbf{y}|\mathbf{s}_j^*;\tau_0)f(\mathbf{x}|\mathbf{s}_j^*;\beta_0)f(\mathbf{s}_j^*|\mu_0,\sigma_0,\phi_0)}$$

All simulations are performed taking into account the starting value $\boldsymbol{\theta_0}$, so they share $\boldsymbol{\theta_0}$. Therefore, if $\boldsymbol{\theta_0}$ is significantly different from $\boldsymbol{\theta}$, one may get too much variability and inferior estimates. Additionally, as stated in Christensen (2004), the maximization of the approximation $L_m(.)$ can be sensitive to $\boldsymbol{\theta_0}$, because likelihood functions of this type are relatively flat and they can be multimodal. These disadvantages require a careful investigation of $\boldsymbol{\theta_0}$ by considering a variety of starting values.

7.4.3 Direct Monte Carlo approximation

We have decided to proceed with parameter estimation through a direct Monte Carlo approximation, by adopting a numerical general-purpose optimization function together with the algorithm described in this Section. The selected optimization method is the one suggested by Byrd, Lu, Nocedal and Zhu (1995), using a limited-memory modification of the BFGS quasi-Newton method. According to which, each parameter subject to estimation can be given a lower and upper bound. The initial value $\boldsymbol{\theta_0} = (\mu_0, \sigma_0, \phi_0, \tau_0, \beta_0)^t$ must satisfy the constraints.

The marginal log-likelihood $l_2^{MC}(\boldsymbol{\theta})$ of the sample data, given in (7.5), depends on the Monte Carlo approximation for the expectation $\mathrm{E}_{\mathbf{S}^*|\mathbf{Y}}[.]$. The chosen

approach to this approximation becomes an important issue. In principle, this approximation can be as accurate as we wish by

- increasing the number of replicas m;
- increasing[6] the number of grid points $N = n_{grid} \times n_{grid}$;
- reducing the numerical noise, resultant from the Monte Carlo simulations.

Obviously, previous items should be balanced against the resulting computational cost. Bear in mind that, when applying our suggested parameter estimation approach in the case of simulated sample data sets, the available complete log-likelihood serves as a benchmark against which to refine our solution.

The proposed algorithm, aiming to acquire $\boldsymbol{\theta_1}$, can be summarized as follows.

1. Define $\boldsymbol{\theta_0}$:
   - Ignore the preferability issue and obtain μ_0, σ_0^2, τ_0^2 and ϕ_0 from the reduced-likelihood function, as suggested in Diggle et al. (2003).
   - Obtain β_0 by 1. deriving a NP intensity estimator $\widehat{\lambda}(\mathbf{x})$;
     2. choosing β such that $\log \widehat{\lambda}(\mathbf{x}) \simeq const + \beta\, Y(\mathbf{x})$.
2. Define $\nu^2 = \tau^2/\sigma^2$ as the relative nugget, $\boldsymbol{\psi} = (\phi, \nu^2)^t$ and

$$\mathbf{V}(\boldsymbol{\psi}) = (\mathbf{R_Y}(\phi) + \nu^2 \mathbf{I_n})^{-1}.$$

   Then, $l_2^{MC}(\boldsymbol{\theta})$ is modified to consider:

   - $\log f(\mathbf{y}) = -0.5\,\{2n \log \sigma - \log |\mathbf{V}(\boldsymbol{\psi})| + \sigma^{-2}(\mathbf{y} - \mu\mathbf{1})^t \mathbf{V}(\boldsymbol{\psi})(\mathbf{y} - \mu\mathbf{1})\}$
   - $\boldsymbol{\mu}_{\mathbf{S}^*|\mathbf{Y}} = \mu\mathbf{1} + \mathbf{R_{SY}}(\phi)\mathbf{V}(\boldsymbol{\psi})(\mathbf{y} - \mu\mathbf{1})$
   - $\boldsymbol{\Sigma}_{\mathbf{S}^*|\mathbf{Y}} = \sigma^2\{\mathbf{R_S}(\phi) - \mathbf{R_{SY}}(\phi)\mathbf{V}(\boldsymbol{\psi})\mathbf{R_{SY}}(\phi)^t\} = \sigma^2 \mathbf{D}(\boldsymbol{\psi})\mathbf{D}^t(\boldsymbol{\psi})$
     where matrix $\mathbf{D}(\boldsymbol{\psi})$ is the corresponding Cholesky factorization

[6] A larger n_{grid}, *degree of discretization*, mainly helps to reduce the variability of $\widehat{\sigma}^2$ and $\widehat{\tau}^2$.

Hereafter, this modified function will be referred to as the *reparametrized* marginal log-likelihood or $l_3^{MC}(\mu, \sigma^2, \beta, \boldsymbol{\psi})$.

3. Given $\boldsymbol{\psi_0} = (\phi_0, \nu_0)^t$, apply a numerical procedure as the one represented in Figure 7.10 to find $\boldsymbol{\psi_1} = (\phi_1, \nu_1)^t$, those values which maximize function $l_3^{MC}(\mu = \mu_0, \sigma^2 = \sigma_0^2, \beta = \beta_0, \boldsymbol{\psi})$.

4. Derive $\mathbf{V}(\boldsymbol{\psi_1})$ and $\mathbf{D}(\boldsymbol{\psi_1})$. Finally, the remaining parameters from $\boldsymbol{\theta_1}$ can be obtained as the MLE's of the likelihood function $l_3^{MC}(\mu, \sigma^2, \beta, \boldsymbol{\psi} = \boldsymbol{\psi}_1)$.

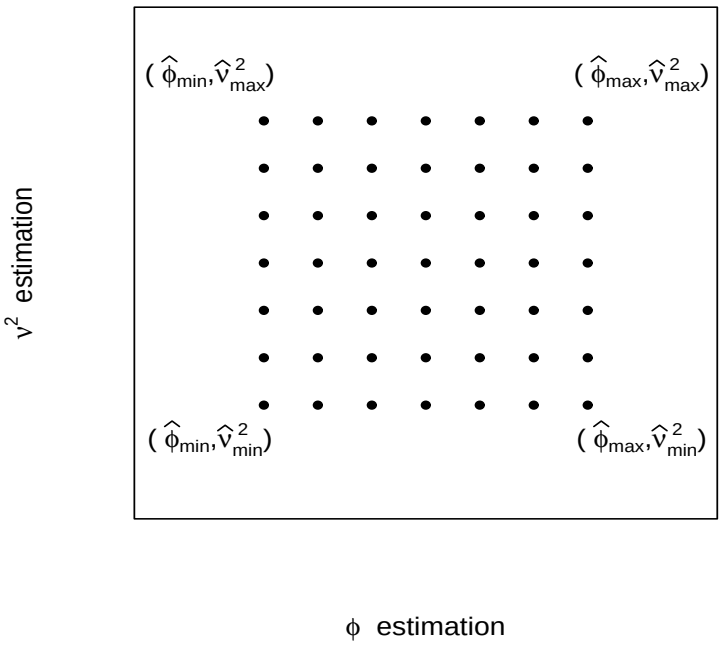


Figure 7.10: Numerical approach to the estimation of ϕ and ν^2. A grid of values around $\boldsymbol{\psi_0} = (\phi_0, \nu_0)^t$ are tested into the reparametrized marginal log-likelihood. The resulting MLE's define $\boldsymbol{\psi_1} = (\phi_1, \nu_1)^t$.

This algorithm might be used iteratively to improve estimate $\boldsymbol{\theta_1}$ and so construct new estimates of $\boldsymbol{\theta}$. Once more, the computational cost must be balanced against the resulting improvements.

Note that the reparametrization $\nu^2 = \tau^2/\sigma^2$ is already adopted in Diggle et al. (2003). Furthermore, Christensen (2004) argues that the main computational burden in evaluating the marginal likelihood function is the inversion of the matrix $\mathbf{R_Y}(\phi) + \nu^2\mathbf{I_n}$, so a numerical procedure with few evaluations of this matrix to be preferred.

Reducing Monte Carlo error

We now explain how we approach the Monte Carlo approximation for the expectation $\mathrm{E}_{\mathbf{S}^*|\mathbf{Y}}[.]$. Bear in mind that we need m realizations of $\mathrm{MVN}(\boldsymbol{\mu}_{\mathbf{S}^*|\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{S}^*|\mathbf{Y}})$, in such a way that $\mathbf{S_1^*}, ..., \mathbf{S_m^*}$ are identically distributed and, then,

$$\mathrm{E}_{\mathbf{S}^*|\mathbf{Y}}[\mathrm{f}(\mathbf{x}|\mathbf{s}^*)] \simeq \frac{1}{\mathrm{m}} \sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{f}(\mathbf{x}|\mathbf{s_j^*}). \tag{7.7}$$

A traditional method for generating the multivariate normal, with an arbitrary covariance matrix, is based on the Cholesky decomposition. In our case, we have

$$\boldsymbol{\Sigma}_{\mathbf{S}^*|\mathbf{Y}} = \sigma^2\{\mathbf{R_S}(\phi) - \mathbf{R_{SY}}(\phi)\mathbf{V}(\boldsymbol{\psi})\mathbf{R_{SY}}(\phi)^t\} = \sigma^2\mathbf{D}(\boldsymbol{\psi})\mathbf{D}^t(\boldsymbol{\psi}).$$

So, we can define

$$\mathbf{S_j^*} = \boldsymbol{\mu}_{\mathbf{S}^*|\mathbf{Y}} + \sigma\mathbf{D}(\boldsymbol{\psi})\mathbf{Z_j^*}$$

where $\mathbf{Z_j^*} \sim \mathrm{MVN}(0, \mathbf{I_N})$ and $\mathbf{I_N}$ is the $N \times N$ identity matrix.

We highlight here that $\mathbf{Z_1^*}, ..., \mathbf{Z_m^*}$ are not necessarily independent. A variance reduction technique in Monte Carlo simulation, based on "antithetic pairs" and duplicating the number of replicas, can be applied. In this way:

1. For $j = 1, ..., m$, we generate "antithetic pairs" within $\boldsymbol{\theta}$

   $$\mathbf{Z_{2j-1}^*} = \mathbf{Z_j} \qquad \text{and} \qquad \mathbf{Z_{2j}^*} = -\mathbf{Z_j},$$

   where $\mathbf{Z_j} \sim \mathrm{MVN}(0, \mathbf{I_N})$.

2. Hence, we fix previous $\mathbf{Z_1}, ..., \mathbf{Z_m}$ for different values of $\boldsymbol{\theta}$.

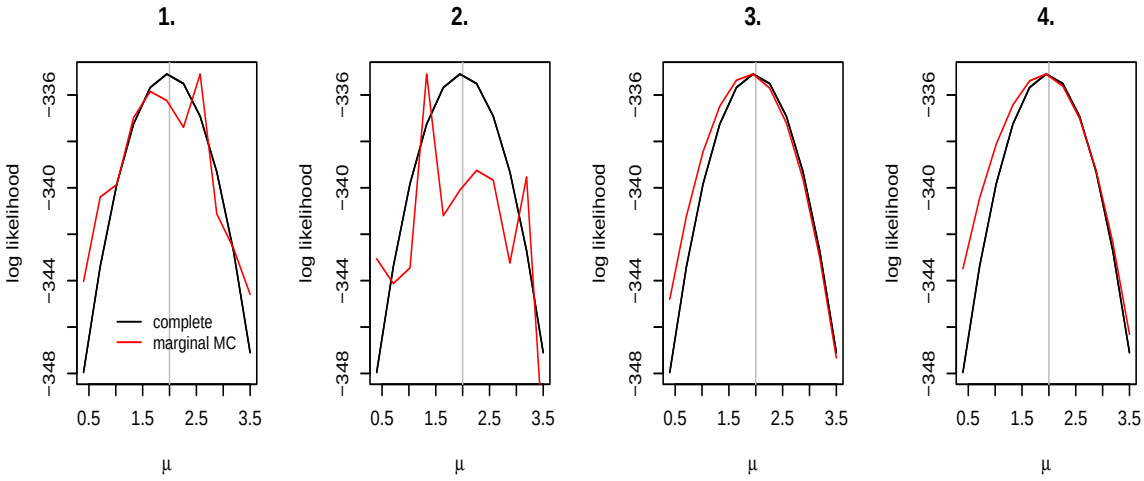


Figure 7.11: Reducing simulation variability for μ estimation. In plots 1.-2., we consider independent $\mathbf{Z_1^*}, ..., \mathbf{Z_m^*}$ and not fixed for all $\boldsymbol{\theta}$. In plots 3.-4., we consider "antithetic pairs" within $\mathbf{Z_1^*}, ..., \mathbf{Z_{2m}^*}$ and fixed for all $\boldsymbol{\theta}$.

The results obtained using this technique are illustrated in Figure 7.11. For the same sample data set $(\mathbf{x}, \mathbf{y})$, the replicas $\mathbf{S_j^*}$ were generated considering the two alternatives:

- Independent $\mathbf{Z_1^*}, ..., \mathbf{Z_m^*}$ and not fixed for all $\boldsymbol{\theta}$ (results for two distinct group of replicas were included in plots 1. and 2.).

- "Antithetic pairs" $\mathbf{Z_1}, ..., \mathbf{Z_m}$ and fixed for all $\boldsymbol{\theta}$ (results for two distinct group of replicas were included in plots 3. and 4.). Note that, here, one has the double of replicas, so expression (7.7) must be changed accordingly.

Figure 7.11 shows a smoother marginal likelihood function, when fixed "antithetic pairs" are adopted. The variability of an estimator may be considered to contain two main components, the *simulation variability* and the *data variability*[7]. The technique just proposed aims at reducing the importance of the first component and at rendering the second component dominant.

[7]The second component corresponds to the classical MLE variance.

7.5 Simulation study

To study how well the estimation procedure presented in the previous Section works, we performed the following simulation study. To simulate the spatial data, we considered the theoretical model

$$Y(\mathbf{x}) = S(\mathbf{x}) + \mathrm{N}(0, \tau^2)$$

where $S(.)$ is a stationary Gaussian process with $\mathrm{E}[S(.)] = \mu$, $\mathrm{Var}[S(.)] = \sigma^2$ and spatial correlation function $\rho(.;\phi)$ from the Matérn family given in (2.2) with a smoothness parameter $\kappa = 1$. Moreover, $\log(\Lambda(\mathbf{x})) = \alpha + \beta S(\mathbf{x})$.

Parameters μ, σ, ϕ, τ and β are the target of estimation. The maximization of $l_3^{MC}(.)$, the reparametrized marginal log-likelihood, yields the maximum likelihood estimates of the model parameters.

To simulate the sample data sets, we tried a variety of values for the model parameters. Namely, we analysed the impact of range parameter ϕ, of the variance ratio τ/σ and, as expected, the impact of the preferability degree β. For the latter, we wished to compare the results of the MLE's with a null and a non-null degree of preferability.

To perform the maximization of $l_3^{MC}(.)$, we tried a variety of values for the *tuning quantities* of the algorithm defined in Section 7.4.3. Namely, we analysed the impact of the number of grid points $N = n_{grid} \times n_{grid}$ on the simulation of $\mathbf{S}^*|\mathbf{Y}$ and the number of replicas m on the approximation of the expectation $\mathrm{E}_{\mathbf{S}^*|\mathbf{Y}}[.]$.

We also explored a modified version of the algorithm defined for the direct Monte Carlo approximation, to reduce the estimation of all five parameters to a single step. This version, aimed at maximizing the $l_2^{MC}(.)$ given in (7.5) instead of the reparametrized $l_3^{MC}(.)$, becomes prohibitive in terms of computational cost when applied together with the Cholesky factorization. Aware of the advantages of faster solutions for the conditional simulation of Gaussian random fields, we compared two different simulating methods: spectral turning bands and circulant embedding (see Schlather 2001 for further details). Table 7.3 includes the time

N.Grid Points	20		30		40		50	
N.Replicas	20	40	20	40	20	40	20	40
Spectral/Circ.Emb.	2.6	4.2	3.2	4.7	3.0	4.3	3.2	4.9

Table 7.3: Time ratio between the spectral turning bands method and the circulant embedding method used for simulating the conditional Gaussian random fields.

ratio between these two methods, illustrating the superiority of the latter method. The relative superiority increases when one uses more replicas to approximate the expectation $\mathrm{E}_{\mathbf{S}^*|\mathbf{Y}}[.]$.

The circulant embedding method has proved to be faster than the spectral turning bands method, because it exploits the speed and efficiency of the fast Fourier transform (Chan and Wood 1997). However, the algorithm described in Section 7.4.3, with the two-steps maximization of function $l_3^{MC}(.)$, presents the least time-consuming results (about 25% less time than the maximization when applying the circulant embedding method). This adds weight to the importance of adopting a numerical procedure, such as that described in Figure 7.10, demanding few evaluations of matrices $\mathbf{V}(\phi, \nu^2)$ and $\mathbf{D}(\phi, \nu^2)$.

Table 7.4 exemplifies the results of our simulation study. The idea is to compare the MLEs assuming the traditional Gaussian model in classical geostatistics with the MLEs obtained under the proposed preferential sampling model. This assessment is implemented for spatial data in both the preferentially and non-preferentially sampled cases.

We performed 100 simulations of sample data sets with $\mu = 4$, $\sigma = 1.4$, $\phi = 0.2$, $\tau = 0.3$ and $\beta = 0$, and 100 more simulations for $\beta = 2$. For the maximization of our Monte Carlo log-likelihood function we considered a total of grid points $N = 900$ and a total number of replicas $m = 40$.

The main conclusions from Table 7.4 may be summarized as follows. As expected, if spatial data were not preferentially sampled, the traditional and the PS

	Preferential sampling ?			
	No		Yes	
	PS model	Traditional	PS model	Traditional
$\widehat{\mu}$	$(3.880, 4.134)$	$(3.886, 4.141)$	$(3.749, 4.038)$	$(5.090, 5.372)$
$\widehat{\sigma}$	$(1.198, 1.326)$	$(1.211, 1.337)$	$(0.911, 1.046)$	$(0.807, 0.888)$
$\widehat{\phi}$	$(0.165, 0.190)$	$(0.167, 0.192)$	$(0.137, 0.163)$	$(0.112, 0.130)$
$\widehat{\tau}$	$(0.296, 0.322)$	$(0.295, 0.321)$	$(0.296, 0.311)$	$(0.305, 0.318)$
$\widehat{\beta}$	$(-0.015, 0.011)$	—	$(1.752, 1.833)$	—

Table 7.4: MLE's confidence intervals obtained from a total of 100 independent samples. The true values of model parameters are $\mu = 4$, $\sigma = 1.4$, $\phi = 0.2$ and $\tau = 0.3$. In the case of a non-null degree of preferability, $\beta = 2$.

models present very similar MLE results. However, the *correction term* proposed for the preferability issue in (7.5) proves to be important, when the preferability degree is non-null. Note that the MLE's confidence intervals derived using the PS model do not include the true values of the parameters in the case of σ, ϕ and β, nevertheless they are always closer to those true values than the MLEs of the traditional model.

An accurate estimation of parameters σ and ϕ seems to be more difficult than of other parameters, but observe that this difficulty occurs also when the spatial data are not preferentially sampled. Furthermore, parameter β seems to be under-estimated under preferential sampling; the corresponding MLE allows us, however, to decide if we are under this type of sampling.

We performed a second iteration of the direct MC approximation algorithm, using as starting values the output from the first iteration (presented in Table 7.4). As expected, there was only a slight improvement in the accuracy of the MLEs, and the computational cost almost doubled. To decide whether the algorithm should be used iteratively, the specific requirements of one's application should be considered.

7.5.1 Issues of parameter identifiability

The estimation procedure in multi-parameter models may lead to issues of parameter identifiability. The simultaneous estimation of several parameters may significantly affect the global results. For example, the accuracy of the MLE of β in the foregoing simulation study has diminished significantly when compared with the case presented in Figure 7.8, which shows the slice of the likelihood for parameter β. Additionally it should be borne in mind that, in estimation problems of this type, the likelihood functions may be relatively flat.

We have accordingly decided to apply some sensitivity analysis to the preferential sampling data. We believe that a small amount of extra non preferentially sampled data can assist in solving the problem of parameter identifiability. In Figure 7.12, we give an example of an *hybrid* sampling design with 16% extra data collected on a grid.

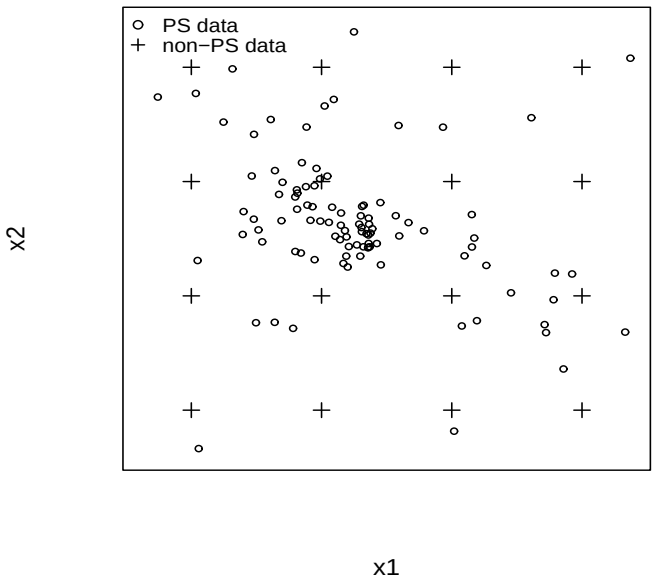


Figure 7.12: Example of *hybrid* sampling design: 16% extra points sampled in a grid (not preferentially sampled).

	Extra non-PS data ?		
	No	No	Yes
	Traditional	PS model	PS model
$\widehat{\mu}$	$(5.090, 5.372)$	$(3.749, 4.038)$	$(3.915, 4.193)$
$\widehat{\sigma}$	$(0.807, 0.888)$	$(0.911, 1.046)$	$(1.045, 1.159)$
$\widehat{\phi}$	$(0.112, 0.130)$	$(0.137, 0.163)$	$(0.153, 0.176)$
$\widehat{\tau}$	$(0.305, 0.318)$	$(0.296, 0.311)$	$(0.299, 0.312)$
$\widehat{\beta}$	—	$(1.752, 1.833)$	$(1.712, 1.784)$

Table 7.5: MLE's confidence intervals obtained from a total of 100 independent samples. The true values of model parameters are $\mu = 4$, $\sigma = 1.4$, $\phi = 0.2$, $\tau = 0.3$ and $\beta = 2$. In the case of extra non PS data, 16% extra points were sampled.

Bear in mind that the marginal log-likelihood must now be re-written as

$$l(\mu, \sigma^2, \beta, \boldsymbol{\psi} \mid Y_1, Y_2, P) = \log f(\mathbf{y_1}, \mathbf{y_2}) + \mathrm{E}_{S^*|Y_1,Y_2}[f(\mathbf{x}|\mathbf{s}^*)]$$

where $\mathbf{y_1}$ represents the original data and $\mathbf{y_2}$ represents the extra data. Note that the second term involves replicas $\mathbf{s_1^*}, ..., \mathbf{s_m^*}$ from $[\mathbf{S}^*|\mathbf{Y_1}, \mathbf{Y_2}]$, which should represent the unobserved field $S(.)$ better than those replicas from $[\mathbf{S}^*|\mathbf{Y_1}]$.

The new MLE values obtained from a total of 100 independent samples are presented in Table 7.5. The right column shows the resulting confidence intervals from the hybrid sampling, confirming that introducing some extra non preferentially sampled data improves the estimation of most parameters.

7.5.2 Spatial prediction for the PS model

Before closing the current Chapter, we would like to make some comments about how we could proceed with spatial prediction. Bear in mind that the interest lies usually in predicting the value of $S(\mathbf{x})$ at an arbitrary location $\mathbf{x}$ within a region of interest D, resulting into a map of the entire surface $S(\mathbf{x})$. Alternatively, the prediction target might be some property of the complete realization of $S(\mathbf{x})$ for all

$\mathbf{x}$ in D which is pertinent to the application in hand. The property of interest can be expressed as a functional of the complete surface $S(.)$, denoted by $T = T(S(.))$. Examples are the estimation of the average value of $S(\mathbf{x})$ over D, or its maximum value, or even the probability that $S(\mathbf{x})$ is above a certain threshold value.

Both for parameter inference and for prediction, conditional simulation of the unobserved gaussian field given the measurement data is needed. So that the prediction problem can be reduced to studying the conditional distribution of $S(.)$ given the observed data $\mathbf{y}$. Let us write $\mathbf{S} = (S(\mathbf{x_1}), ..., S(\mathbf{x_n}))^t$ for the unobserved values of the underlying process at the sampling locations $\mathbf{x_1}, ..., \mathbf{x_n}$ and $\mathbf{S}^*$ for the values of $S(.)$ at all other locations of interest, typically a fine grid of locations covering the study area. The prediction of (functionals of) $\mathbf{S}^*$, where $\mathbf{S}^* = (S(\mathbf{x_{n+1}}), ..., S(\mathbf{x_{n+q}}))^t$ and $q \geq 1$, requires the predictive distribution, which can be written as

$$[\mathbf{S}^*|\mathbf{y}] \equiv \int f(\mathbf{s}^*|\mathbf{s}, \boldsymbol{\theta}) f(\mathbf{s}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{s} d\boldsymbol{\theta}$$

or, equivalently, the predictive density of $\mathbf{S}^*$ can be given by

$$f(\mathbf{s}^*|\mathbf{y}) = \mathrm{E}[f(\mathbf{s}^*|\mathbf{s})|\mathbf{y}]$$

where $f(\mathbf{s}^*|\mathbf{s})$ is the conditional density of $\mathbf{S}^*$ given $\mathbf{S}$, and dependence on parameters is suppressed. As referred in Christensen (2004), prediction at unobserved locations is then naturally decomposed into two phases: (1) conditional simulation of $[S|\mathbf{y}]$, and (2) prediction of some functional of $\mathbf{S}^*$ based on simulations from $[S|\mathbf{y}]$. The approach of plug-in prediction, where $\widehat{\boldsymbol{\theta}}_{MLE}$ is plugged into the predictive distribution as being the truth, would be useful for phase one.

7.6 Closing remarks and future work

Typically in geostatistics, one wishes to make inferences about a real-valued stochastic process $S(\mathbf{x})$ on a continuous space $\mathbf{x}$, being $Y(\mathbf{x})$ the corresponding measurement process (probably a noisy version of $S(\mathbf{x})$). The points of $\mathbf{x}$ at which one

observes data can be determined either by the scientist himself (example of the deterministic sampling design) or by a stochastic point process P. In any case the distribution of the point process, which we write as $[P]$, is regarded as irrelevant. Thus the natural way to formulate a model for the resulting data is using the marginal distribution $[Y]$.

Under preferential sampling, the points of $\mathbf{x}$ are always given by a stochastic point process and, more precisely, they are determined by a point process P which is stochastically dependent on the spatial variable of interest S, so locations are realizations of $[P|S]$. This is the case of the pollution data, for which this dependency can be justified by favoring the siting of moss near suspected sources of pollution. Thus the natural way to formulate a model for the resulting data can not ignore the distribution of the point process, and it should take into account the marginal distribution $[Y, P]$.

In Chapter 6, we have investigated the consequences of assuming $[P|S] = [P]$. Unsurprisingly, inferences which ignore this stochastic dependence can be inadequate, and the simple approach of ignoring preferential sampling can be unacceptable. We have shown that the classic approaches for declustering of the preferred locations help to reduce the problem, but they continue to be inadequate. In Section 7.5, the new approach based on strong parametric assumptions has been evaluated through a simulation study. The results strongly suggest that correct inferences can be recovered by treating data as a realization of a marked point process.

Recall from Section 7.3.2 that two equivalent formulations were given as the marginal distribution of the sample data:

1. $[Y, P] = \mathrm{E}_S\{[P|S][Y|S]\}$;

2. $[Y, P] = [Y]\ \mathrm{E}_{S|Y}\{[P|S]\}$.

The second formulation was pointed out as an intuitive model, because the cor-

responding log-likelihood function can be written as

$$l(\theta|Y,P) = \log\ [Y] + \log\ \mathrm{E}_{S|Y}\{[P|S]\}$$

where the first term represents the conventional multivariate log-normal and the second represents a *correction term* for the preferability issue. We believe, however, that an important issue should be taken into account when adopting the foregoing model: under preferential sampling, the above multivariate normal is not exactly the conventional one. We believe that the resulting parameters estimates of $[Y]$ are being subject to a similar phenomenon of the so-called *length-bias*. For example, in survival analysis, the *length-bias* phenomenon arises because persons who do not live long after the diagnosis tend to be excluded from studies of existing disease.

As a goal for future research, we plan to investigate different approaches to the correction of bias in $[Y]$. Bear in mind that the concept of biased observations arise when a sampling procedure chooses an observation with probability that depends on the value of the observation. This is exactly the case of the preferential sampling, which favors some observations and neglects others. One possible approach may be to assume μ and σ as functions of the preferability degree β; or to assume the following linear relationship

$$Y = \alpha_0 + \alpha_1\ S + N(0,\tau^2),$$

where previously $\alpha_0 = 0$ and $\alpha_1 = 1$. These options will allow us to study sensitivity to *small* departures from the model, which is a useful step to solve problems of model uncertainty as discussed in Copas and Eguchi (2005). These authors also address problems of bias resulting from data with incomplete observations, which, at least in principle, may be corrected if we have further data available; note that this was the idea behind the hybrid sampling described in Section 7.5, which results indeed suggest some bias reduction.

In addition to the attempts to correct for bias, the true distribution of Y can be investigated. In the univariate case, non-parametric estimates of the length-biased

probability density function have been presented, and relationships between the *exact* and the *nearby* distributions have been established (see e.g. Oluyede 2003). It may be challenging to look for a similar parametric solution for the multivariate case, which could allow us to obtain the true model for the above formulation $[Y, P] = [Y]\ \mathrm{E}_{S|Y}\{[P|S]\}$.

Furthermore, as future work, we want to investigate the alternative formulation $[Y, P] = \mathrm{E}_S\{[P|S][Y|S]\}$, which does not explicitly involve the classic multivariate normal. Our preliminary analysis of the corresponding log-likelihood function has indicated a need for a careful numerical approximation. We may consider the adoption of the Monte Carlo maximum likelihood method *importance sampling* for the estimation of our model parameters. Note that in both formulations, we have an approximation to an expectation, $\mathrm{E}_{S|Y}$ and E_S, respectively; such requires the simulation of replicas from $[S|Y]$ and $[S]$, respectively; thus, in both cases, the effect of the *length-bias* should be analysed and, if needed, we should suggest a corrector.

Finally, it would also be interesting to take into account research topics, such as: the analytical expression of the theoretical mark variogram γ_M mentioned in Section 6.3; and the application of the discussed methods and spatial prediction to other datasets, apart from moss data.

Appendix A

Extended abstract (spanish)

En la actualidad la estadística espacial juega un papel relevante, gracias a la influencia del desarrollo tecnológico en el tratamiento de datos espaciales. Más allá de la existencia de las fuentes tradicionales de datos espaciales, tales como mapas o material obtenido en censos y fotos aéreas, surgen además nuevas fuentes con excelente fiabilidad, destacando los datos obtenidos por satélite. Estas nuevas tecnologías están asociadas a la disponibilidad de una mayor capacidad computacional y software específico, tal como software de procesamiento de imágenes y sistemas de información geográfica (Diblasi 2002).

Los modelos espaciales se desarrollan a partir de datos recogidos en distintas localizaciones espaciales, de manera que midan la relación entre las observaciones alcanzadas en diversas posiciones. Estos modelos deben representar la noción intuitiva de existencia de correlación entre datos situados en zonas próximas, que se reduce con el aumento de la distancia; asimismo, deben reflejar la presencia de errores espacialmente correlacionados. En este sentido, el análisis de correlación espacial permite observar la forma en que las variables, como cargas contaminantes obtenidas en distintos puntos del espacio, se encuentran relacionadas. Un hecho particularmente útil, desde el punto de vista práctico, reside en que estas relaciones hacen posible la predicción de valores en localizaciones donde no han sido efectuadas mediciones.

Se puede obtener una muestra de datos a partir de observaciones realizadas en una, dos o tres dimensiones. Por ejemplo, las mediciones efectuadas a lo largo de un curso fluvial son unidimensionales. Las medidas de pluviosidad y otras variables metereológicas se refieren a posiciones concretas que, conjuntamente, forman un campo aleatorio bidimensional. Por último, las mediciones de concentración mineral en el suelo se tratan en ocasiones como problemas tridimensionales, debido a que se toman a distintas profundidades.

Un proceso espacial es un proceso estocástico que puede representarse como un conjunto de variables aleatorias (o vectores) $Z(\mathbf{x})$, con $\mathbf{x} \in D \subset I\!R^d$, donde D es un espacio euclídeo d-dimensional con $d = 1, 2, 3$; se suele denotar como $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$. Si se consideran las localizaciones espaciales $\mathbf{x_1}, ..., \mathbf{x_n}$, entonces $Z(\mathbf{x_1}), ..., Z(\mathbf{x_n})$ identifica los datos observados en estas localizaciones. Estas observaciones pueden ser obtenidas a partir de una o más variables discretas o continuas.

De acuerdo con Cressie (1993), la naturaleza de la región de observación D permite diferenciar tres tipos de procesos espaciales, a saber, los procesos reticulados, los procesos puntuales y los procesos continuos. Estos últimos constituyen el objetivo de la geoestadística (véase Matheron 1963) y en ellos, a diferencia de lo que ocurre con los procesos reticulados, dados dos puntos espaciales asociados a observaciones existentes, siempre es posible encontrar otro punto donde la variable aleatoria puede ser igualmente observada.

El trabajo desarrollado en esta memoria se enmarca en el ámbito de la geoestadística y además hace uso de la teoría de procesos puntuales, ya que se supone que las localizaciones espaciales han sido generadas por un determinado mecanismo estocástico. En lo que respecta al tema de "evaluación de la dependencia espacial" (primera parte del título de este trabajo, "*assessing spatial dependency*"), se consideran dos cuestiones diferenciadas, como son la estimación de la estructura de dependencia espacial y el consiguiente procedimiento de predicción espacial. En

el contexto de la geoestadística se asume con frecuencia que las localizaciones de muestreo se encuentran igualmente distribuidas a lo largo de la región observada; también se supone que los procesos puntuales para las localizaciones de los datos no dependen de estos mismos. En consecuencia y haciendo referencia al tema "muestreo no estándar" (segunda parte del título de esta memoria, "*non-standard sampling*"), se aborda el incumplimiento de uno o ambos de los supuestos anteriores.

El contenido de este trabajo se ha dividido en distintos artículos; uno de ellos ha sido aceptado para su publicación, Menezes et al. (2005a), otro se encuentra en proceso de revisión, Menezes et al. (2005b), y los dos restantes están en fase de preparación. La estructura de la memoria que aquí se presenta se describe a continuación.

En el Capítulo 2 se introduce la notación y conceptos básicos relativos a la modelización de datos geoestadísticos. Se examinan ideas, ampliamente tratadas en este contexto, sobre la conveniencia de imponer hipótesis como la isotropía o la estacionariedad del proceso subyacente. asimismo, se hace una breve revisión de los métodos usados para el tratamiento de datos con tendencia espacial o para abordar problemas más complejos referidos a los procesos espaciales no estacionarios.

Además, se pone de manifiesto la utilidad del variograma como herramienta para medir la dependencia espacial existente entre los datos. Teniendo en cuenta que el variograma de un proceso espacial intrínseco e isotrópico $Z(\mathbf{x})$ verifica que:

$$\gamma(\|\mathbf{x_i} - \mathbf{x_j}\|) = \frac{1}{2}\mathrm{Var}[Z(\mathbf{x_i}) - Z(\mathbf{x_j})] = \frac{1}{2}\mathrm{E}[(Z(\mathbf{x_i}) - Z(\mathbf{x_j}))^2],$$

es posible estimar el variograma a partir de los datos muestrales $\{(\mathbf{x_i}, Z(\mathbf{x_i})),\ i = 1, ..., n\}$ reemplazando la esperanza teórica anterior por la correspondiente media muestral. Se pueden incluir pesos o ponderaciones sobre los valores promediados para suavizar la estimación del variograma (de la forma descrita en el Capítulo 3). En cualquier caso, la esperanza anterior marca la pauta para determinar cómo se podría estimar el variograma de un proceso estacionario a partir de los datos

observados.

El Capítulo 3 se enfoca hacia la estimación de la dependencia espacial bajo muestreo estándar. La mayoría de los estimadores empíricos no pueden utilizarse directamente para la predicción espacial, ya que no cumplen la propiedad de ser condicionalmente definido negativos y, por tanto, podrían dar lugar a estimaciones negativas del error cuadrático medio de predicción, como se menciona en Cressie (1993).

Típicamente, la resolución del problema anterior se divide en tres etapas: *estimación empírica del variograma*, *selección del modelo válido* y *ajuste del modelo*. Para llevar a cabo estas tareas se han propuesto distintos procedimientos. En este sentido, el principal objetivo de este capítulo fue el de identificar estos métodos y compararlos en un estudio numérico que abarcase una amplia variedad de situaciones de dependencia espacial. Las comparaciones se basaron principalmente en los valores estimados de los errores cuadráticos de los estimadores válidos resultantes. Además, en este trabajo se ha propuesto un método empírico, fácilmente implementable, para comparar las principales características del variograma estimado.

En general, los resultados que aquí se presentan muestran que un estimador válido del variograma, obtenido a partir de un procedimiento no paramétrico, es una buena alternativa frente a los mecanismos paramétricos tradicionales. Los métodos no paramétricos tienen la ventaja adicional de que evitan los problemas asociados a una mala especificación del modelo paramétrico, que puede ocurrir en muchas ocasiones. Estas ventajas son todavía más evidentes si bajo los datos muestrales subyace una dependencia espacial atípica, como la producida por un modelo oscilatorio. No obstante, supone pagar un coste computacional extra frente al coste asociado al uso de un procedimiento paramétrico sencillo, como el que ajusta un modelo válido a alguno de los estimadores empíricos a través de los criterios de mínimos cuadrados ponderados.

El método paramétrico basado en la máxima verosimilitud restringida como criterio de ajuste sólo puede competir con los otros procedimientos cuando se dispone de pocos datos y, simultáneamente, cuando los datos observados no siguen ningún tipo de estructura oscilatoria.

Los métodos geoestadísticos introducidos en el Capítulo 3 asumen implícitamente que el diseño de muestreo para las localizaciones $\mathbf{x_i}$, $i = 1, ..., n$ es determinista o es estocástico pero independiente del proceso, y todos los análisis se llevan a cabo condicionalmente sobre $\mathbf{x_i}$ (Diggle et al. 2003). Se supone entonces que los puntos de muestreo han sido elegidos independientemente de los valores de la variable espacial. Sin embrago, puede haber dependencias debidas al método de muestreo utilizado, tales como la selección preferente de áreas específicas que se consideran críticas (por ejemplo, la búsqueda de valores máximos).

Schlather et al. (2004) propone métodos para detectar la dependencia entre marcas y localizaciones para procesos puntuales marcados. Como se describe en Mateu and Ribeiro Jr (1999), el campo aleatorio (que hemos estado estudiando) y los procesos puntuales marcados son dos tipos de procesos espaciales tales que:

- El primero se define en cada punto de la región observada y el propio investigador puede determinar las posiciones muestrales (ejemplo de un diseño muestral determinista).

- Para el segundo, las localizaciones siempre se generan a partir de un proceso estocástico puntual y normalmente se espera que existan interacciones entre las localizaciones y las marcas. De otro modo, se tendría el conocido *modelo de campo aleatorio* (un proceso puntual marcado constituye una clase de campo aleatorio).

Si los datos son consistentes con un modelo de campo aleatorio, el diseño puntual y las marcas pueden analizarse separadamente utilizando técnicas estándar para procesos puntuales (e.g. Ripley 1981 y Diggle 2003) y para datos geoes-

tadísticos (e.g. diversas referencias mencionadas en el Capítulo 3). Por lo tanto, este análisis se simplifica enormemente.

El estudio de características de segundo orden, como el variograma, de un proceso espacial debería tener presente si los datos provienen de un campo aleatorio o de un verdadero proceso puntual marcado. Ejemplo de referencias relacionadas con este tema pueden ser Walder and Stoyan (1996), Mateu and Ribeiro Jr (1999) y Schlather (2002).

Schlather et al. (2004) hace referencia a dos situaciones en las que las localizaciones y los procesos subyacentes podrían ser dependientes, con el consiguiente incumplimiento de esta importante hipótesis geoestadística. En primer lugar, cuando la dependencia es una propiedad intrínseca de los propios datos; por ejemplo las posiciones relativas de los árboles influyen en sus tamaños debido a su competencia por la luz y los nutrientes. Este es el caso de un verdadero proceso puntual marcado. De forma alternativa, esta dependencia se podría justificar mediante un conocimiento científico a priori de la variable espacial de interés; por ejemplo del nivel de contaminación local esperado en la contaminación atmosférica. Esto puede dar lugar a la obtención de muestras en áreas con valores atípicos.

Nuestro trabajo se refiere al problema resultante de la segunda situación, que pensamos que es de mayor importancia en geoestadística ya que ocurre con frecuencia en las medidas de campo reales y, a menudo, o bien es ignorado o bien es resuelto con técnicas generales como las de desagrupación (e.g. Goovaerts 1997 y Isaaks and Srivastava 1989).

El capítulo 4 está motivado por el ejemplo de aplicación sobre los datos de radioactividad de la isla de Rongelap, donde se ha utilizado un procedimiento de recogida de datos en dos etapas, que conlleva a la presencia de datos agrupados. Por ello, comenzamos por restringir nuestra atención a las muestras multietápicas, lo cual permite valorar la presencia de dependencia multietápica o también llamada *dependencia secuencial*, donde la elección de los puntos de muestreo se lleva a cabo

a partir de medidas previas.

Se proponen algunos métodos de exploración de datos que tratan de detectar la obtención multietápica sesgada de datos espaciales. Estos métodos se definen como extensión del test de significación de Montecarlo descrito en Schlather et al. (2004). Más aún, se investigan modelos correctores que permitan minimizar el impacto en la estimación del variograma debido a la adopción de diseños de muestreo no estándar que actualmente se están utilizando. A saber, con respecto al tema de la agrupación, se propone una compensación para las áreas no pobladas mediante una corrección sobre las áreas de alta densidad (ya que podrían no ser suficientemente representativas del conjunto de datos, desde el punto de vista espacial) y, simultáneamente, aprovechando los beneficios de un estimador de tipo núcleo.

A continuación se valora el efecto de los métodos resultantes sobre los datos de la isla de Rongelap, concluyendo que sobre este conjunto de datos subyacen características, tales como una reducida varianza espacial y que las localizaciones forman una línea recta debido a la disposición de la isla, que requieren una cuidadosa estimación. En consecuencia, esto hace aconsejable la utilización de métodos correctores como los propuestos.

El Capítulo 5 está dedicado al estudio de un nuevo estimador de tipo núcleo del variograma que proponemos para datos agrupados, probándose además que es asintóticamente insesgado y consistente. Además, se proponen valores óptimos para sus *parámetros de suavizado* desconocidos, dos cantidades que, apropiadamente elegidas, afectan al comportamiento del estimador.

Teniendo en cuenta que $\{Z(\mathbf{x}) : \mathbf{x} \in D \subset I\!R^d\}$ representa un proceso aleatorio intrínseco e isotrópico y denotando por $Z(\mathbf{x_1}), ..., Z(\mathbf{x_n})$ los valores del proceso observados en las localizaciones $\mathbf{x_1}, ..., \mathbf{x_n}$, el estimador del variograma sugerido se define como sigue:

$$\widehat{\gamma}(u) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right)[Z(\mathbf{x_i}) - Z(\mathbf{x_j})]^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right)}, \quad u \geq 0,$$

donde $n_i = \sum_k I_{\{\|\mathbf{x_i}-\mathbf{x_k}\|\leq\delta\}}$ y $n_j = \sum_k I_{\{\|\mathbf{x_j}-\mathbf{x_k}\|\leq\delta\}}$; h y δ representan los parámetros ventana y de radio del entorno, respectivamente.

Los métodos de desagrupación son bastante intuitivos y en la literatura sobre estadística espacial se reconoce su necesidad para estimar tendencias espaciales promedio representativas para datos agrupados (véase e.g. Goovaerts (1997) y Isaaks and Srivastava (1989), o Dubois and Saisana (2002) para una comparación clásica de los métodos de desagrupación). En contraste con lo anterior, normalmente no se considera en este contexto la necesidad de una estimación fiable de las estructuras espaciales de segundo orden. La presencia de datos muestrales agrupados no es, sin embargo, despreciable en absoluto como se muestra en el Capítulo 4; véase por ejemplo la figura 4.2. o la tabla 4.1., que ilustran el mal comportamiento de los estimadores tradicionales del variograma bajo densidades muestrales desiguales.

Algunas de las principales razones para la agrupación de datos muestrales son:

- Factores externos, como la selección de localizaciones condicionadas a zonas geográficas o demográficas específicas.

- La necesidad de obtener una mejor caracterización de la variabilidad a corto alcance, que requiere un muestreo más denso pero algunas veces demasiado costoso para llevarlo a cabo sobre toda la región de observación.

- La adopción de un muestreo más denso en áreas que se consideran críticas. Por ejemplo, la búsqueda de valores máximos basados en algún conocimiento a priori.

El trabajo reciente de Kovitz and Christakos (2004) se refiere al tema de la agrupación y de la estimación de la estructura de segundo orden. Estos autores sugieren una versión modificada del estimador de Matheron que también incorpora pesos desagrupadores, pero basados en zonas de proximidad. Cada zona de cualquier conjunto de datos se define mediante el área del polígono de Voronoi que

contiene todos los puntos más próximos a ese punto del interior de los datos que a cualquier otro conjunto de datos. El comportamiento de este estimador modificado del variograma se analiza en un estudio numérico.

En nuestro caso, se prueba que el estimador del variograma verifica buenas propiedades teóricas. En Menezes et al. (2004), se presenta una versión simplificada de estos resultados teóricos preliminares. Como este estimador requiere la selección del parámetro ventana h y del radio δ, recomendamos que el primero sea tratado utilizando el error cuadrático medio y que el segundo se obtenga del análisis de la estimación de la densidad derivada sobre la región de observación. Las principales contribuciones de este trabajo se describen de forma más detallada en Menezes et al. (2005*b*).

El cuerpo principal del Capítulo 5 se organiza de la forma siguiente. Primero se introduce notación adicional y se resumen las principales hipótesis consideradas en el estudio asintótico. A continuación se trata el tema de selección del radio del entorno. Luego se establecen las principales propiedades del estimador no paramétrico propuesto y se desarrollan las demostraciones correspondientes. Los resultados obtenidos para el sesgo y la varianza se utilizan para la selección del parámetro de suavizado. Finalizamos con algunos estudios numéricos y detalles de implementación sobre el estimador propuesto.

Se ha comentado anteriormente que, en geoestadística, tanto en el contexto de la predicción como en el de inferencia, se supone habitualmente que la selección de las localizaciones no depende de los valores de la variable espacial. En el Capítulo 6, se introduce una definición formal directamente relacionada con el incumplimiento de la mencionada suposición de independencia y no restringida a la recogida de datos multietápica. Supongamos que, por la naturaleza del procedimiento de muestreo, incluyendo el que persigue la búsqueda de valores máximos, existe una relación estocástica subyacente entre los datos y las localizaciones, entonces este procedimiento se corresponde con el *muestreo preferencial*.

Siguiendo la notación de Diggle et al. (2003), consideraremos que los datos para el análisis son de la forma $(\mathbf{x_i}, y_i) : i = 1, ..., n$, donde $\mathbf{x_1}, ..., \mathbf{x_n}$ son localizaciones dentro de la región de observación $D \subset I\!R^2$ e $y_1, ..., y_n$ son medidas asociadas a estas localizaciones. El conjunto $\{\mathbf{x_i} : i = 1, ..., n\}$ es el *diseño muestral* e y_i representa una realización de $Y_i = Y(\mathbf{x_i})$, donde $\{Y(\mathbf{x}) : \mathbf{x} \in D\}$ es el *proceso de medida*. También supondremos la existencia de un proceso no observado $\{S(\mathbf{x}) : \mathbf{x} \in D\}$, que generalmente es nuestro objetivo de predicción. A menudo, Y_i se puede considerar como una versión de ruido del proceso subyacente $S(\mathbf{x_i})$ o del valor en la localización $\mathbf{x_i}$ del proceso $S(.)$.

La figura 6.1., en el panel izquierdo, ilustra el ejemplo de un campo real S y, en el panel derecho, el correspondiente conjunto de datos $(\mathbf{x_i}, y_i)$.

Observación: *Bajo* **muestreo preferencial**, *el proceso de diseño muestral se supone que es estocásticamente dependiente del proceso $S(.)$. En consecuencia, el correspondiente modelo geoestadístico (especificado por la distribución conjunta del proceso involucrado) debe tener en cuenta la distribución condicional del diseño muestral.*

Considerando la presencia de muestreo preferencial en el contexto Gaussiano, proponemos un mecanismo basado en el modelo para la predicción espacial. Este nuevo modelo paramétrico se enmarca en una clase flexible de procesos de Cox log-Gaussianos. El Capítulo 6 se dedica principalmente al análisis de las consecuencias resultantes de ignorar el muestreo preferencial y de aplicar los métodos geoestadísticos clásicos. En el Capítulo 7, se planteará la inferencia a partir de la función de verosimilitud para la estimación de los parámetros del modelo.

Queremos ahora enfatizar la diferencia entre muestreo preferencial y agrupado. Como ya se ha mencionado, la agrupación de las localizaciones puede deberse a la existencia de zonas geográficas o demográficas específicas, o también podría utilizarse para poder explicar mejor la variabilidad a corto alcance. Estos son buenos

ejemplos para mostrar que el muestreo agrupado puede que no implique necesariamente el muestreo preferencial. Por el contrario, la implicación en el otro sentido suele ocurrir, ya que las localizaciones muestrales preferidas suelen estar en las áreas concentradas. Por ejemplo, algún conocimiento científico a priori sobre $S(.)$, tal como el grado de mineral esperado en una explotación minera, puede ocasionar la concentración de muestras en áreas con valores atípicamente altos.

El Capítulo final 7 está dividido en dos partes principales. La primera describe un ejemplo de aplicación a datos reales, que permite ilustrar la aplicación de varios métodos de estadística espacial que se mencionan a lo largo de este trabajo. A saber, la utilidad del variograma no paramétrico robusto a la presencia de agrupaciones, estudiado en el Capítulo 5, junto con la técnica de interpolación clásica kriging.

Se analiza la intensidad de la contaminación atmosférica basada en los datos de Galicia (noroeste España). Se dispone de dos conjuntos distintos de datos, ya que fueron recogidos en dos años diferentes: 1995 y 2000. En el último año, dado que se obtuvo una subvención mayor para el proyecto, se consideró una rejilla con mayor número de puntos para el diseño muestral. En lo que se refiere al año 1995, parece haber claros indicios de que el muestreo fuese preferencial, ya que los datos fueron obtenidos en su mayoría en las proximidades de las fuentes de contaminación con objeto de reducir costes. Los resultados confirman la importancia de aplicar un análisis paramétrico bajo muestreo preferencial.

En la segunda parte del Capítulo 7, se lleva a cabo el análisis de un mecanismo basado en el modelo para el muestreo preferencial. Se propone un modelo intuitivo, que incorpora un término corrector del criterio de preferencia.

Se describe un algoritmo para la aproximación de Montecarlo directa a la función de verosimilitud. Luego se aplica la inferencia basada en la verosimilitud para estimar los parámetros del modelo propuesto. Se lleva a cabo un estudio numérico para mostrar los beneficios de este modelo frente al tradicional. Nuestro

análisis sugiere que este modelo intuitivo debería considerar la presencia de algún sesgo. De este modo, finalizamos el Capítulo 7, incluyendo una breve discusión de los objetivos de futuras investigaciones relacionados con la evaluación de un modelo *insesgado* para el muestreo preferencial.

Appendix B

Acronyms

BLUE	best linear unbiased estimator
CSR	complete spatial randomness
GLS	generalized least squares
GLSE	generalized least squares by Genton (1998*b*)
ISE	integrated square error
LS	least squares
MINQU	minimum norm quadratic unbiased
MISE	mean integrated square error
MIVQU	minimum variance quadratic unbiased
ML	maximum likelihood
MLE	maximum likelihood estimate
MSE	mean squared error
MSSE	mean standardized squared error
NP	non parametric
OLS	ordinary least squares
P	parametric
PE	prediction error
PPB	parts per billion
PPM	parts per million

PS	preferential sampling
PV	prediction variance
REML	restricted maximum likelihood
SGP	stationary Gaussian process
WLS	weighted least squares

Bibliography

Aboal, J., Real, C., Fernández, J. and Carballeira, A. (2005), 'Mapping the results of extensive surveys: The case of atmospheric biomonitoring and terrestrial mosses', *To appear in The science of the total environment* .

Anderson, T. (1984), *An introduction to multivariate statistical analysis*, 2nd edn, Wiley.

Armstrong, M. and Delfiner, P. (1980), 'Towards a more robust variogram: A case study on coal', *Centre of Geostatistique, Fontainebleau, France* **Internal Note N-671**.

Azzalini, A. (1996), *Statistical inference: based on the likelihood*, Chapman & Hall.

Baddeley, A., Moller, J. and Waagepetersen, R. (2000), 'Non- and semi-parametric estimation of interaction in inhomogeneous point patterns', *Statistica Neerlandica* **54(3)**, 329–350.

Baddeley, A. and Turner, R. (2000), 'Practical maximum pseudolikelihood for spatial point patterns', *Australian and NZ Journal of Statistics* **42(3)**, 283–322.

Barnard, G. (1963), 'Discussion of professor bartlett's paper', *Journal of the Royal Statistical Society, Series B* **25**, 294.

Besag, J. and Diggle, P. (1977), 'Simple monte carlo tests for spatial pattern', *Journal of the Royal Statistical Society, Series C* **26**, 327–333.

Bochner, S. (1955), *Harmonic analysis and the theory of probability*, University of California Press, Los Angeles.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995), 'A limited memory algorithm for bound constrained optimization', *SIAM J.Scientific Computing* **16**, 1190–1208.

Chan, G. and Wood, A. (1997), 'An algorithm for simulating stationary gaussian random fields', *Journal of the Royal Statistical Society, Series C* **46**, 171–181.

Cherry, S. (1997), 'Non-parametric estimation of the sill in geostatistics', *Environmetrics* **8**, 13–27.

Cherry, S., Banfield, J. and Quimby, W. (1996), 'An evaluation of a non-parametric method of estimating semivariograms of isotropic spatial processes', *Journal of Applied Statistics* **17**, 563–586.

Chilès, J. and Delfiner, P. (1999), *Geostatistics: Modelling spatial uncertainty*, Wiley, New York, Chichester.

Christakos, G. (1984), 'On the problem of permissible covariance and variogram models', *Water Resources Research* **20**, 251–265.

Christensen, O. (2004), 'Monte carlo maximum likelihood in model-based geostatistics', *Journal of Computational and Graphical Statistics* **13(3)**, 702–718.

Cliff, A. and Ord, J. (1981), *Spatial processes: models and applications*, Pion, London.

Copas, J. and Eguchi, S. (2005), 'Local model uncertainty and incomplete-data bias', *Journal of the Royal Statistical Society, Series B* **67(4)**, 459–513.

Cox, D. (1972), The statistical analysis of dependencies in point processes, *in* 'Stochastic Point Processes', P.A.W.Lewis, New York, pp. 55–66.

Cressie, N. (1985), 'Fitting variogram models by weighted least squares', *Journal of the International Association for Mathematical Geology* **17(5)**, 563–586.

Cressie, N. (1993), *Statistics for spatial data*, John Wiley and Sons Inc., New York.

Cressie, N. and Hawkins, D. (1980), 'Robust estimation of the variogram', *Journal of the International Association for Mathematical Geology* **12(2)**, 115–125.

Cuevas, A., Febrero, M. and Fraiman, R. (2001), 'Cluster analysis: a further approach based on density estimation', *Computational Statistics & Data Analysis* **36**, 441–459.

Curriero, F., Hohn, M., Liebhold, A. and Lele, S. (2002), 'A statistical evaluation of non-ergodic variogram estimators', *Environmental and Ecological Statistics* **9(1)**, 89–110.

Delfiner, P. (1976), Linear estimation of nonstationary spatial phenomena, *in* M. Guarascio, M. David and C. Huijbregts, eds, 'Advanced Geostatistics in the Mining Industry', Dordrecht: Reidel, pp. 49–68.

Díaz, M. and Ayala, G. (1999), Measuring the spatial homogeneity in corneal endotheliums by means of a randomization test, *in* A. Kuba and A. Todd-Pokropek, eds, 'Information Processing in Medical Imaging', Vol. 1613 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 418–423.

Diblasi, A. (2002). Analisis exploratoria de observaciones espaciales con S-Plus.

Diggle, P. (2003), *Statistical analysis of spatial point patterns*, Arnold, London.

Diggle, P., Harper, L. and Simon, S. (1997), Geostatistical analysis of residual contamination from nuclear weapons testing, *in* V.Barnett and K.Feridun-Turkman, eds, 'Statistics for the Environment 3: Pollution Assessment and Control', John Wiley & Sons Ltd, Chichester, pp. 89–107.

Diggle, P. and Ribeiro Jr, P. (2002), 'Bayesian inference in gaussian model based geostatistics', *Geographical and Environmental Modelling* **6(2)**, 129–146.

Diggle, P., Ribeiro Jr, P. and Christensen, O. (2003), An introduction to model-based geostatistics, *in* J.Moller, ed., 'Spatial Statistics and Computational Methods', Vol. 173 of *Lecture Notes in Statistics*, Springer, New York, pp. 43–86.

Diggle, P., Tawn, J. and Moyeed, R. (1998), 'Model-based geostatistics (with discussion)', *Journal of the Royal Statistical Society, Series C* **47(3)**, 299–350.

Dubois, G. and Saisana, M. (2002), 'Optimizing spatial declustering weights - comparison of methods', *2002 Annual Conference of the International Association for Mathematical Geology* pp. 479–484.

Dutter, R. (1996), On robust estimation of variograms in geostatistics, *in* H. Rieder, ed., 'Robust Statistics, Data Analysis and Computer Intensive Methods', Springer-Verlag, New York, pp. 153–171.

Fernández, J., Rey, A. and Carballeira, A. (2000), 'An extended study of heavy metal deposition in galicia (nw spain) based on moss analysis', *The science of the total environment* **254**, 31–44.

Foxall, R. and Baddeley, A. (2002), 'Nonparametric measures of association between a spatial point process and a random set, with geological applications', *Journal of Applied Statistics* **51**, 165–182.

Fuentes, M. (2002), 'Spectral methods for nonstationary spatial processes', *Biometrika* **89**, 197–210.

Garcia-Soidán, P., Febrero-Bande, M. and Gonzalez-Manteiga, W. (2004), 'Nonparametric kernel estimation of an isotropic variogram', *Journal of Statistical Planning and Inference* **121**, 65–92.

Garcia-Soidán, P., Gonzalez-Manteiga, W. and Febrero-Bande, M. (2003), 'Local linear regression estimation of the semivariogram', *Statistics & Probability Letters* **64**, 169–179.

Gentle, J. (2002), *Elements of computational statistics*, Springer Verlag.

Genton, M. (1998*a*), 'Highly robust variogram estimation', *Journal of the International Association for Mathematical Geology* **30(2)**, 213–221.

Genton, M. (1998*b*), 'Variogram fitting by generalized least squares using an explicit formula for the covariance structure', *Journal of the International Association for Mathematical Geology* **30(4)**, 323–345.

Genton, M. and Gorsich, D. (2002), 'Nonparametric variogram and covariogram estimation with fourrier-bessel matrices', *Computational Statistics & Data Analysis* **41**, 47–57.

Geyer, C. (1994), 'On the convergence of monte carlo maximum likelihood calculations', *Journal of the Royal Statistical Society, Series B* **56**, 261–274.

Geyer, C. (1999), Likelihood inference for spatial point processes, *in* 'Stochastic Geometry, Likelihood and Computation', O.E.Barndorff-Nielsen, W.S.Kendall and M.N.M.van Lieshout, Boca Raton, FL: Chapman and Hall/CRC, pp. 79–140.

Gneiting, T., Sasvári, Z. and Schlather, M. (2001), 'Analogies and correspondences between variograms and covariance functions', *Advances in Applied Probability* **33(3)**, 617–630.

Goovaerts, P. (1997), *Geostatistics for natural resources evaluation*, Oxford University Press, New York.

Gorsich, D. and Genton, M. (2000), 'Variogram model selection via nonparametric derivate estimation', *Journal of the International Association for Mathematical Geology* **32(3)**, 249–270.

Gribov, A., Krivoruchko, K. and Ver Hoef, J. (2000), 'Modified weighted least squares semivariogram and covariance model fitting algorithm', *Stochastic Modeling and Geostatistics. AAPG Computer Applications in Geology* **2**.

Gunst, R. and Hartfield, M. (1997), Robust semivariogram estimation in the presence of influential spatial data values, *in* T. Gregoire and al., eds, 'Modeling Longitudinal and Spatially Correlated Data', Springer-Verlag, pp. 265–274.

Hall, P. (1991), *The bootstrap and edgeworth expansion*, Springer Verlag, New York.

Hall, P., Fisher, I. and Hoffmann, B. (1994), 'On the parametric etimation of covariance funtions.', *Annals of Statistics* **22(4)**, 2115–2134.

Higdon, D., Swall, J. and Kern, J. (1999), Non-stationary spatial modelling, *in* 'Bayesian Statistics', Vol. 6, Oxford University Press, pp. 761–768.

Hope, A. (1968), 'A simplified monte carlo significance test procedure', *Journal of the Royal Statistical Society, Series B* **30**, 582–598.

Isaaks, E. and Srivastava, R. (1989), *An introduction to applied geostatistics*, Oxford University Press.

Journel, A. and Huijbregts, C. (1978), *Mining geostatistics*, Academic Press, London.

Kottegoda, N. and Rosso, R. (1997), *Statistics, probability, and reliability for civil and environmental engineers*, McGraw Hill.

Kovitz, J. and Christakos, G. (2004), 'Spatial statistics of clustered data', *Stochastic Environmental Research* **18**, 147–166.

Kyung-Joon, C. and Shucany, W. (1998), 'Nonparametric kernel regression estimation near endpoints', *Journal of Statistical Planning and Inference* **66**, 289–304.

Lewis, P. and Shedler, G. (1979), 'Simulation of non-homogeneous poisson processes by thinning', *Naval Research Logistics Quarterly* **26**, 403–413.

Maglione, D. and Diblasi, A. (2001), 'Choosing a valid model for the variogram of an isotropic spatial process', *2001 Annual Conference of the International Association for Mathematical Geology* .

Manly, B. (1991), *Randomization and Monte Carlo methods in biology*, Chapman and Hall, London.

Mateu, J. and Ribeiro Jr, P. (1999), Geostatistical data versus point process data: analysis of second-order characteristics, *in* J. Goméz-Hernandéz and R. Froidevaux, eds, 'GeoENV-II - Geostatistics for environmental applications', Vol. 10 of *Quantitative Geology and Geostatistics*, Kluwer Academic, pp. 213–224.

Matheron, G. (1962), *Traite de geostatistique appliquee, Tome* I, Memories du Bureau de Recherches Geologiques et Minieres, vol.14, Ediciones Bureau de Recherches Geologiques et Minieres, Paris.

Matheron, G. (1963), 'Principles of geostatistics', *Economic Geology* **58**, 1246–1266.

Matheron, G. (1971), *The theory of regionalized variables and its applications*, Cahiers du Centre Morphologie Mathematique,n.5, Fontainebleau,France.

Menezes, R. (2002), 'Analysis of existing variogram estimators', *Tesina del Departamento de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela* .

Menezes, R., Garcia-Soidán, P. and Febrero-Bande, M. (2004), 'Properties of a kernel variogram estimator for clustered data', *Proceedings of the XXVIII Congreso Nacional de Estadística e Investigación Operativa, Cádiz* .

Menezes, R., Garcia-Soidán, P. and Febrero-Bande, M. (2005*a*), 'A comparison of approaches for valid variogram achievement', *To appear in Computational Statistics* .

Menezes, R., Garcia-Soidán, P. and Febrero-Bande, M. (2005*b*), 'A kernel variogram estimator for clustered data', *Under revision* .

Menezes, R. and Tawn, J. (2003), 'Análise da presença de clusters na estimação de um semivariograma', *Estatística com Acaso e Necessidade– Proceedings of the XI Annual Congress of the Portuguese Society of Statistics, Faro* pp. 451–457.

Moller, J., Syversveen, A. and Waagepetersen, R. (1998), 'Log gaussian cox processes', *Scandinavian Journal Statistics* **25**, 451–482.

Moller, J. and Waagepetersen, R. (2003), An introduction to simulation-based inference for spatial point processes, *in* J.Moller, ed., 'Spatial Statistics and Computational Methods', Vol. 173 of *Lecture Notes in Statistics*, Springer-Verlag, New York, pp. 143–198.

Murphy, S. and van der Vaart, A. (2000), 'On profile likelihood (with discussion)', *Journal of the American Statistical Association* **95**, 449–485.

Oluyede, B. (2003), 'Inequalities and bounds for kernel length-biased density estimation', *Applied Mathematics and Computation* **135(2-3)**, 541–551.

Pintore, A. and Holmes, C. (2004), 'Non-stationary covariance functions via spatially adaptive spectra', *Tentatively accepted in Journal of the American Statistical Association* .

Ploner, A. and Dutter, R. (2000), 'New directions in geostatistics', *Journal of Statistical Planning and Inference* **91**, 499–509.

Reilly, C. and Gelman, A. (2004), 'Weighted classical variogram estimation for data with clustering', *Under revision at Technometrics* .

Ribeiro Jr, P. (2002), 'Model based geostatistics, applications and computational implementation'. unpublished Ph.D. dissertation, Lancaster University, Department of Mathematics and Statistics.

Ribeiro Jr, P. and Diggle, P. (2001), 'geoR: A package for geostatistical analysis', *R News* **1(2)**, ISSN 1609–3631.

Ripley, B. (1981), *Spatial statistics*, Wiley, New York.

Rowlingson, B. and Diggle, P. (1993), 'Splancs: spatial point pattern analysis code in s-plus', *Computers in Geosciences* **19**, 627–655.

Sampson, P. and Guttorp, P. (1992), 'Bayesian inference in gaussian model based geostatistics', *Journal of the American Statistical Association* **87**, 108–119.

Schlather, M. (2001), 'Simulation of stationary and isotropic random fields', *R News 1* **2**, 18–20.

Schlather, M. (2002), 'Characterization of point processes with gaussian marks independent of locations', *Mathematische Nachrichten* **240**, 204–214.

Schlather, M., Ribeiro Jr, P. and Diggle, P. (2004), 'Detecting dependence between marks and locations of marked point processes', *Journal of the Royal Statistical Society, Series B* **66(1)**, 79–93.

Shapiro, A. and Botha, J. (1991), 'Variogram fitting with a general class of conditionally nonnegative definite functions', *Computational Statistics & Data Analysis* **11**, 87–96.

Silverman, B. (1986), *Density estimation for statistics and data analysis*, Chapman & Hall, London.

Stein, M. (1999), *Interpolation of spatial data - some theory for kriging*, Springer, New York.

Stein, M. (2005), 'Non-stationary spatial covariance functions', *Technical Report No.21* . The University of Chicago. Center for Integrating Statistical and Environmental Science.

Wackernagel, H. (1998), *Multivariate geostatistics : an introduction with applications*, 2nd edn, Springer Verlag, Berlin.

Walder, O. and Stoyan, D. (1996), 'On variograms in point processes statistics', *Biometrical Journal* **38(8)**, 895–905.

Wong, M. and Lane, T. (1983), 'A kth nearest neighbour clustering procedure', *Journal of the Royal Statistical Society, Series B* **45**, 362–368.

Yu, K. and Mateu, J. (2002), Nonparametric nearest-neighbour variogram estimation, *in* J. Mateu and F. Montes, eds, 'Spatial statistics through applications', WIT Press, Southampton, Boston.

Zimmerman, D. and Zimmerman, M. (1991), 'A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors', *Technometrics* **33(1)**, 77–91.