Spatial Sampling and Prediction

Lina Schelin
To my family
## Contents

List of papers vi

Abstract vii

Preface ix

1 Introductory remark 1

2 Introduction 2

3 Spatial sampling 4
  3.1 Spatial probability sampling 7
  3.2 Spatial non-probability sampling 9

4 Spatial prediction 10
  4.1 Stationarity 12
  4.2 The variogram 13
    4.2.1 Empirical estimators 14
    4.2.2 Variogram models 15
    4.2.3 Estimation of model parameters 17
  4.3 More on kriging 19
  4.4 Prediction intervals 23
  4.5 Bootstrap 24
  4.6 Censoring 25

5 Summary of papers 26
  5.1 Paper I: Kriging prediction intervals based on semiparametric bootstrap 26
  5.2 Paper II: Construction of kriging prediction intervals for non-Gaussian spatial processes 26
  5.3 Paper III: Spatially balanced sampling through the pivotal method 27
  5.4 Paper IV: How to select representative samples 27
  5.5 Paper V: Spatial prediction in the presence of left-censoring 27

Papers I–V
List of papers

This thesis is based on the following papers:


Paper I is reprinted with the kind permission of Springer Science+Business Media and paper III with the kind permission of John Wiley and Sons.
Abstract

This thesis discusses two aspects of spatial statistics: sampling and prediction. In spatial statistics, we observe some phenomena in space. Space is typically of two or three dimensions, but can be of higher dimension. Questions in mind could be; What is the total amount of gold in a gold-mine? How much precipitation could we expect in a specific unobserved location? What is the total tree volume in a forest area? In spatial sampling the aim is to estimate global quantities, such as population totals, based on samples of locations (papers III and IV). In spatial prediction the aim is to estimate local quantities, such as the value at a single unobserved location, with a measure of uncertainty (papers I, II and V).

In papers III and IV, we propose sampling designs for selecting representative probability samples in presence of auxiliary variables. If the phenomena under study have clear trends in the auxiliary space, estimation of population quantities can be improved by using representative samples. Such samples also enable estimation of population quantities in subspaces and are especially needed for multi-purpose surveys, when several target variables are of interest.

In papers I and II, the objective is to construct valid prediction intervals for the value at a new location, given observed data. Prediction intervals typically rely on the kriging predictor having a Gaussian distribution. In paper I, we show that the distribution of the kriging predictor can be far from Gaussian, even asymptotically. This motivated us to propose a semiparametric method that does not require distributional assumptions. Prediction intervals are constructed from the plug-in ordinary kriging predictor. In paper V, we consider prediction in the presence of left-censoring, where observations falling below a minimum detection limit are not fully recorded. We review existing methods and propose a semi-naive method. The semi-naive method is compared to one model-based method and two naive methods, all based on variants of the kriging predictor.

Keywords: Auxiliary variables, Censoring, Inclusion probabilities, Kriging, Local pivotal method, Minimum detection limit, Prediction intervals, Representative sample, Spatial process, Spatial sampling, Semiparametric bootstrap
Preface

Efter många år är denna avhandling tillslut tillräckligt färdig för att lämnas till tryck...

Den skulle aldrig ha blivit till om det inte vore för ett antal personer. Om inte Peter A hade erbjudit mig och uppmuntrat mig att ta en amanuensstjänst hade jag kanske inte hamnat på institutionen överhuvudtaget. Ett tack även till Peter W som redan första terminen på grundutbildningen uppmärksammar mig på att detta var en karriärväg. Jag har alltid undrat över hur mycket de till synes obetydliga kommentarerna har påverkat mig. Tack till er båda för den uppmuntran ni gett mig och andra studenter.


Jag är också glad att jag har fått lära känna Anton (och hans ego), först som doktorandkollega och sen som nära vän. Vi har pratat undervisning, forskning och en massa strunt och jag har uppskattat allt av det. Han har också lärt mig att bara för att en person låter säker så behöver det inte betyda att personen har rätt. Som synes i denna avhandling, så blev det i slutändan även forskningssamarbeten.

Det finns många tidigare och nuvarande kollegor som jag är glad över att ha fått arbeta/umgås med. Jag vill särskilt nämna våra ovärderliga administratörer, min biträdande handledare Patrik, Thomas, Ingrid S, Per, och Niklas. Det är många fler som borde nämnas, men jag är rädd att jag då glömer någon.
Tack också till er som har läst och kommit med värdefulla synpunkter på kappan.


Umeå, Mars 2012

Lina Schelin
1 Introductory remark

Within most areas of research statistics is a frequently used tool. It is used to evaluate new drugs in the pharmaceutical industry, to find genes causing specific diseases in medicine, to estimate volume of trees in forestry, to evaluate the effect of different teaching activities in education research and so on. There is a large set of methods for answering statistical questions, and there are a lot of methods not yet invented. The following four steps are typically among the steps in the research process

1. Specify the research question (formulate hypothesis);
2. Collect data;
3. Analyze data;
4. Draw conclusions from analysis.

Statistics is, or should be, involved in all of the above steps, especially steps 2 - 4. This thesis is about step two and three. Sampling is the process of selecting a sample from a population, in order to make inference for the whole population. Paper III and IV belong to this research area, with a specific interest in spatial sampling. In spatial sampling, we use a specific sampling design to collect data in space. A space is typically of two or three dimensions, but can be of any dimension $d$, where $d$ is an integer. When it comes to analyzing data, we may ask: Which statistical method is most appropriate? What assumptions can be made? Are there methods available? Do we need to improve old methods or construct new ones? How well do the methods work? How large are the errors? How sensitive are the methods to violated assumptions? These are all questions that (should) arise when data are about to be analyzed. The remaining work described in this thesis concerns development and improvement of methods within spatial statistics. More specifically it considers spatial prediction. Possible fields of applications are image analysis, climate, and environmental statistics. Two general problem, in spatial statistics, concern spatial dependency and spatial prediction, where kriging is one frequently used predictor. Papers I, II and V are related to prediction by kriging.
Introduction

Let \( \{Z(x) : x \in A\} \), where \( A \subseteq \mathbb{R}^d \), be a spatial process describing some phenomenon of interest, where \( x \) is a location in a space of dimension \( d \). This phenomenon might be the temperature, the precipitation, the amount of gold in the ground, etc. In general, we assume that

\[
Z(x) = f(x, \beta) + \varepsilon(x).
\]

The part \( f(x, \beta) \) is deterministic and describes the trend while \( \varepsilon(x) \) is the stochastic part modeling the short-range fluctuations allowing for spatial correlation. The parameter vector \( \beta \) contains all parameters describing the trend. Figure 1 shows four different realizations of one specific spatial process \( Z(x) \).

The purpose of studying such phenomena is often one of the two following:

A. Observe the phenomenon at sampled locations, and \textit{predict} at a new unobserved location.

B. Observe the phenomenon at sampled locations, and \textit{estimate} a total for the population (all possible locations).

De Gruijter et al. (2006) refers to these two aims as estimation of local (A) and global (B) quantities. This introduction is intended to give the reader an insight in these two fundamentally different approaches. The following is a direct quote from Valliant et al. (2000).

"Finite population sampling is distinguished from the rest of statistics by its focus on the actual population of which the sample is a part. In other parts of statistics, observations are typically represented as realizations of random variables, and the inference refer not to any actual population of units, but to the probability law that governs the random variables."

Finite population sampling is what de Gruijter et al. (2006) call classical survey sampling. It primarily deals with the second aim (B), and is also referred to as the design-based approach. It is compared, in de Gruijter et al. (2006), to the model-based approach which primarily focuses on the first aim (A). The design-based and model-based approaches have totally different foundations and, depending on the purpose, either one of them...
Figure 1: Four realizations of a stationary and isotropic zero mean Gaussian process with an exponential dependence structure.

might be recommended. If it is possible to randomly select sample locations, and if the goal is to estimate a parameter of the target distribution, such as the expected value, the variance or a quantile, then a design-based approach should be used. Inference in this case requires probability sampling. A model-based approach should be used if a reliable model describing the variation is available; there is spatial correlation, and the goal is to predict at one or several new locations. It requires in general a larger sample size. A model-based approach can be combined with probability sampling, but this is not necessarily the optimal sampling strategy. The optimal sampling strategy for prediction is not the same as the optimal strategy for estimation of model parameters. For a wider discussion about the model-based versus the design-based approach, we refer to Brus and de Gruijter (1997) and de Gruijter et al. (2006).
As already indicated, the difference between the design-based and the model-based approach lies in the source of variation. In a design-based approach (i.e. in classical survey sampling) it is assumed that the process only have the deterministic part $Z(x) = f(x, \beta)$, and that the domain is partitioned into $N$ population units. From the $N$ population units we sample $n < N$ units according to a sampling design. The variance for the estimator (e.g. of the total) comes from the sample locations being randomly chosen according to some design. On the contrary, in a model-based approach, the sample locations are fixed. Here the variance is due to the fact that the realization is random. The realization refers to the observed values of the random process. The main goal is prediction or estimation of model parameters. Prediction is performed given the observed data at locations $x_1, \ldots, x_n$.

The next two sections are devoted to a deeper review of spatial sampling and spatial prediction. Section 3 describes the design-based approach in more detail. It is also discussed how to choose sample locations under the model-based approach. The main focus of Section 4 is the model-based approach. An overview of deterministic methods for prediction is also given.

### 3 Spatial sampling

In classical survey sampling the population is usually finite, and some auxiliary information is assumed known for the whole population. Survey sampling is for many associated with populations consisting of individuals, households or other specified units. One example of a population, is all the trees in a forest area. Based on a selected sample the final goal is to estimate some unknown population quantity, most often the mean or the total of the variable of interest (the target variable). For the trees it can be the total volume of all trees. This unknown quantity is estimated using an estimator. The variance for this estimator is also of interest.

It is less common to associate survey sampling with continuous populations. However, continuous populations can also be considered. One way is by partition the continuous population into a countable number of population units.

Let $U$ denote the population of $N$ units. The indicator $I_i$ is used to indicate if unit $i$ is included in the sample or not. We have $I_i = 1$ if unit $i$ is
included in the sample and $I_i = 0$ otherwise. Each unit $i$ has a prescribed inclusion probability $0 < \pi_i \leq 1$. The first order inclusion probability, $\pi_i = E[I_i] = P(I_i = 1)$, is the probability that unit $i$ is included in a sample. The second order inclusion probability, $\pi_{ij} = E[I_i I_j] = P(I_i = 1, I_j = 1)$, is the probability that both units $i$ and $j$ are included in a sample. When the sample size $n$ is fixed, the inclusion probabilities must satisfy $n = \sum_{i=1}^{N} \pi_i$. The inclusion probabilities can be equal for all population units or proportional to an auxiliary variable. The values of all auxiliary variables are known prior to sampling for all population units. Finally, let $y_i$ denote the value of the target variable for unit $i$. Prior to sampling, this value is unknown. In classical sampling this value is assumed to be without measurement error. Measurement errors occurs when the observed value differ from the true value, see, e.g., Särndal et al. (2003, Ch. 16) for comments on the effect of measurement error on the estimates.

Let $S$ be the set of all possible samples of size $n$ from the population $U$. A sampling design is the set of all possible samples and a function $p(\cdot) > 0$, where $p(s), s \in S$, is the probability that sample $s$ is selected. For a formal definition, we refer to Wolter (2007). In some literature the sampling design also includes the estimation method, but this is a less common definition. Samples can be selected with or without replacement, nevertheless in what follows we only consider designs without replacement.

When a sample has been selected, the total $Y = \sum_{i=1}^{N} y_i$ can be estimated by the unbiased Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952)

$$\hat{Y}_{HT} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} I_i.$$

The variance for this estimator is

$$V(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{i,j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.$$

To calculate the variance, the second order inclusion probabilities must be known. For many sampling designs this is not the case. However, if they are known, the Sen-Yates-Grundy estimator (Yates and Grundy, 1953; Sen, 1953)

$$\hat{V}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{i,j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 I_i I_j,$$
can be used to estimate the variance without bias if all second order inclusion probabilities are positive.

The variance of a survey statistic is actually a function of both the statistic and the nature of the design. We now briefly overview some commonly used designs. A detailed overview of estimators and variance of estimators, for different sampling designs, is found in Wolter (2007).

Simple random sampling (SRS) is probably the most well known sampling design. For SRS all possible samples have the same probability, and we randomly select one of the samples with equal probability, i.e., \( p(s) = \left( \frac{N}{n} \right)^{-1} \), for all \( s \in S \).

Another approach to select \( n \) units among \( N \) units is by systematic sampling (see, e.g., Lohr, 1999). Assume that the population units can be ordered in a list. If the inclusion probabilities are equal for all units, we first select one unit randomly among the first \( k \) in that list, where \( k \cdot n = N \). Then we choose every \( k \)th element, starting from the randomly selected unit. If the units are listed in increasing or decreasing order, using an auxiliary variable, estimation based on a systematic sample is likely to be more efficient, in comparison to SRS. On the other hand, systematic sampling can be less precise than SRS if the list has some periodic pattern. Unequal inclusion probabilities can be incorporated into the systematic sampling design. The procedure is then slightly modified.

If we are interested in estimation within subgroups of the population we can divide the population into strata and produce samples within each stratum. This is known as stratified sampling. Auxiliary variables are used to construct strata. Strata could be described as groups which have similar values on one or several auxiliary variables. Stratified sampling is also likely to be more efficient than SRS, when estimating quantities for the whole population. Basic theory for stratified sampling can be found in Lohr (1999, Ch. 4). In a spatial setting stratification can be used to produce samples that are well spread over the spatial domain. One of the drawbacks with stratified sampling is that it is not trivial how to divide into strata.

Sampling designs where the inclusion probabilities are known and unequal are referred to as \( \pi \)ps designs. The abbreviation indicates that the inclusion probability is proportional to size (of some auxiliary variable). Several \( \pi \)ps designs are described in Lundquist (2009) and Grafström (2010).
3.1 Spatial probability sampling

Up to now, all designs are presented for a general situation and not specific for populations on a spatial domain. Now, we assume that spatial coordinates are known for all population units. We also assume that there are spatial trends, in the meaning that two nearby units are more similar than units further apart. More precisely, we assume that the values of the target variable are similar for locations close in space. When this is the case, we can estimate the unknown population quantity more efficient if the sample is well spread over the population, i.e. over the spatial domain. A well spread sample is said to be spatially balanced. As a measure of spatial balance, we follow Stevens and Olsen (2004) and use the concept of Voronoi polygons. The Voronoi polygon for sample unit \( i \in s \) includes all population units closer to \( i \) than to any other sample unit \( j \). The total inclusion probability within polygon \( i \) is denoted \( v_i \). For a spatially balanced sample, we should have \( v_i \approx 1 \) for all \( i \). Comparisons, between different designs ability to produce spatially balanced samples, can be done by calculating the mean of

\[
B = \frac{1}{n} \sum_{i \in s} (v_i - 1)^2,
\]

over repeated samples \( s \in S \).

As mentioned, stratification can be used to assure that the sample is well spread over the population. Stratification is difficult, especially if unequal inclusion probabilities are used. Another design that is spreading the sample well over space is the Generalized Random-Tessellation Stratified (GRTS) design, introduced by Stevens and Olsen (2003, 2004). The design can handle unequal inclusion probabilities. However, it is only defined for two dimensions. In paper III, we suggest sampling designs that produce spatially balanced samples. The two local pivotal methods (LPM) we present there have the same foundation, but somewhat different algorithms. Figure 2 show the difference between samples produced by one of the local pivotal methods compared to simple random sampling. It clearly shows that LPM samples are much more well spread over the population.

The local pivotal methods, proposed by us in paper III, have no limitations in dimension. This is used in paper IV, where we present a strategy to produce representative samples, i.e. samples that are well spread in the auxiliary variable space. The space can be spanned by several auxiliary
variables, not only spatial coordinates. Representative samples are important in multi-purpose surveys, when several target variables are of interest, and they enable estimation in subspaces.

Let us consider an example where the concept, of representative samples, is applicable in environmental monitoring. Assume that we want to decide the locations for new monitoring stations. Beforehand we have access to spatial locations and additional information on different land and vegetation types. We want to choose the monitoring locations based on a probability design, and we want to guarantee that all land and vegetation types are represented. At the same time, we want to spread the locations in space. This can be achieved by using the local pivotal methods.

In classical survey literature the model-based approach is referred to as the prediction theory approach. The population is considered as a realization from a superpopulation model. Model assisted inference, which is the main theme of Särndal et al. (2003), should also be mentioned. In that case, models are used to improve design-based estimators.
3.2 Spatial non-probability sampling

So far, we have discussed the design-based approach where the focus is to estimate the total or some other quantity of the population. All variation of the corresponding estimate is due to the design. Before we proceed to spatial prediction, we briefly discuss how select samples with aim (A) in mind. We now step away from the design-based approach.

In the model-based approach, where the aim is to estimate parameters describing the dependence and make prediction at new locations the sampling locations do not have to be random. In practice it is common to use a grid or transect, which is also called lattice designs. The term design is not completely unambiguous, since the definition in the design-based approach and the model-based approach differ. Nevertheless, in what follows, the meaning should be clear from the context. Lattice designs are widely used since they are efficient for prediction and simple to implement. In environmental monitoring, it is also very common to use so called opportunistic designs. In this case, observations are made at already existing monitoring locations (Diggle and Ribeiro, 2007).

Different designs are optimal for different purposes. The two purposes in mind are parameter estimation (dependence) and prediction. There have been different attempts to find optimal sampling schemes based on different optimality criterions. The main conclusions from Lark (2002), Kerry and Oliver (2007), Marchant and Lark (2007) and Diggle and Ribeiro (2007, Ch. 8) are the following. In order to estimate the dependence parameters the distance between observations must be shorter than the range of dependence. The optimal strategy changes with the spatial dependence. A scattered cluster is optimal for small distance parameters and/or short range dependencies, while a regular pattern is optimal for strong spatial dependencies (Lark, 2002). For spatial prediction the recommendation is a well spread sample (over the domain) complemented with a few additional clusters.

Returning to the design-based approach, Särndal et al. (2003, Ch. 14.4) give some examples of non-probability sampling methods, i.e. expert choice sampling and judgment sampling. A design that combines probabilistic and non-probabilistic techniques is quota sampling. There also exist non-probability designs that are closely related to the probability sampling designs. However, they do not completely satisfy the definition, since some samples do not have a positive probability of being selected. Two such
designed are cut-off sampling and balanced sampling, see Deville and Tillé (2004) for a description of the cube method.

4 Spatial prediction

In this section, we discuss how to predict a spatial process \( \{Z(x) : x \in A\} \) at one or several unobserved locations, where \( x \in A \subseteq \mathbb{R}^d \) denotes a spatial location. We assume that the process is continuous, and that it is possible to observe the process at every single location. Examples of a phenomenon that could be described as spatial processes are for instance temperature, precipitation or quicksilver concentration in the ground. It is most common to consider spatial processes located in two or three dimensions \( (d = 2 \text{ or } 3) \), but the methodology is not restricted to those dimensions. The value of the process at a given location is a random variable; it has a variation specified by a probability distribution. The value at location \( x \) can be described by \( Z(x) = f(x, \beta) + \varepsilon(x) \). The model has a deterministic part \( f(x, \beta) \) and spatially correlated noise \( \varepsilon(x) \). The error term, \( \varepsilon(x) \), is assumed to have mean zero.

We aim to predict the values of one single realization of the process at new locations given that the process is observed at the locations \( x_1, ..., x_n \). Let \( Z(x_p) \) denote the value of the interesting variable at a new location \( x_p \), that has not been observed.

Given the observed data, there are many techniques for spatial prediction, and several of them are closely related. Many of them rely on the assumption that values tend to be similar for locations close to each other. The majority of the prediction methods can be seen as weighted averages of the observed data. We make a distinction between deterministic and non-deterministic prediction techniques. The deterministic techniques do not take the error term into account.

One deterministic technique is based on Voronoi polygons, introduced in Section 3. The region of interest is divided into \( n \) Voronoi polygons based on the \( n \) observed locations. The Voronoi polygon \( i \) contains all locations closer to the observed location \( x_i \) than to any other observed location. For all locations belonging to the Voronoi polygon \( i \) the spatial process is
predicted to be equal to the observed value $z(x_i)$. This method is also called Thiessens polygons or Dirichlet tessellation.

Inverse distance weighting interpolation is a technique (Shepard, 1968), where the predicted value $\hat{Z}(x_p)$ at a new location $x_p$ is a weighted average of the observed values, i.e.

$$\hat{Z}(x_p) = \sum_{i=1}^{n} \lambda_i \cdot z(x_i).$$

The weight $\lambda_i$ controls how much the value at location $x_i$ affects the prediction at location $x_p$. Usually the weight $\lambda_i$ is inverse proportional to the distance between $x_i$ and $x_p$, resulting in smaller weights for units further away from $x_p$.

Regression techniques can also be used for prediction purposes. Global polynomial regression models the process by a polynomial function in the spatial coordinates. The regression coefficients are estimated by least square methods, and the obtained regression equation is used for prediction. The error terms in the regression models are typically assumed to be independent. If there is dependence left in the residuals, the prediction can be improved by taking the spatial dependence into account.

Splines is another technique used for prediction. There are several different spline techniques, but they are all based on describing the spatial process by piecewise polynomials in the coordinates. The smoothness of the predictor can be controlled by incorporating additional conditions (Ruppert et al., 2003, Ch. 13). In the spatial prediction context the thin plate spline function is common.

Now we proceed to the non-deterministic prediction methods. They are commonly used in geostatistics. Geostatistics is one branch of spatial statistics that originates from the mining industry. Methods to characterize and predict spatial phenomena were developed side by side with other subareas of spatial statistics. In present time, geostatistics is used in many different fields, of which soil science is one specific example.

One way to predict the process, $Z(x)$, at a new location $x_p$ based on the observed data $z = (z(x_1), z(x_2), ..., z(x_n))^T$ is, similarly to the deterministic methods, to construct a weighted average of the data. We use the best linear unbiased predictor given the observed data, known as the kriging predictor. This predictor minimizes the mean squared prediction error,
when the dependence structure is assumed known. If the spatial process is Gaussian the kriging predictor is best among all predictors (Cressie, 1993). The dependence can be described by the correlation function, the covariance function or by the variogram. We return to this in Section 4.2.

Kriging is for many equated with geostatistical interpolation or with optimal prediction in space. The term interpolation can be somewhat misleading, since kriging includes both spatial interpolation and extrapolation. The theory of kriging is dated back to the 1960’s and it is named after the South African mining engineer D. G. Krige. The technique was formalized by G. Matheron, and the term *kriging* appears (in English) in Matheron (1963). It should be mentioned that contemporaneous researchers developed similar (optimal spatial interpolation) methods within other research areas. The origins of kriging is described in Cressie (1990).

It is common to make assumptions about the spatial process, to simplify estimation of unknown model parameters. The process \( Z(x) \), or the residual process \( \varepsilon(x) \), is typically assumed to be *stationary*. The concept of stationarity is described in more detail in Section 4.1. Another assumption that make estimation simpler is that of *isotropy*. A process is said to be isotropic if the dependence (i.e. the variogram) is a function of the distance only, and independent of direction. Otherwise, we talk about *anisotropy*. In paper I, II and V, we mainly work with stationary isotropic processes.

### 4.1 Stationarity

Stationarity is not a necessary, but a very common assumption on spatial processes that facilitates parameter estimation. There are different types of stationarity. If the joint distribution of \( Z(x) \) and \( Z(x+h) \) is the same for all \( x \) and \( h \), then the process is said to be *strongly stationary*. Sometimes this is referred to as strictly stationary or simply stationary. If the stationarity assumption is relaxed only to require that \( E[Z(x)] = \mu \) is independent of location \( x \) and that \( C(h) = \text{Cov}(Z(x+h), Z(x)) \), only depend on \( h \), then \( Z(x) \) is a *second order stationary* process. For Gaussian processes, second order stationarity is equivalent to strong stationarity. Finally, \( Z(x) \) is called an *intrinsically stationary* random process if \( E[Z(x+h) - Z(x)] = 0 \) and \( \text{Var}(Z(x+h) - Z(x)) \) is a function of \( h \) solely. Note that second order stationarity always implies intrinsic stationarity, but the opposite need not be true (see, e.g., Myers, 1989).
4.2 The variogram

The dependence structure of a spatial process is often described by the covariance or correlation function, but in geostatistics it is more common to consider the variogram. The fact that the variogram is defined for intrinsic stationary processes is one reason for its usage in geostatistics. Another reason is that parameter estimation is less biased (Cressie, 1993, p. 70). If the process is (at least) intrinsically stationary, the variogram is defined as

$$2\gamma(h) = \text{Var}[Z(x) - Z(x + h)].$$

If the variogram is the same in all directions it is called isotropic, otherwise, we have an anisotropic variogram. The semivariogram is usually defined as $\gamma(h)$, but the literature is not consistent. For second order stationary processes the semivariogram is related to the covariance function by

$$\gamma(h) = C(0) - C(h).$$

This is exemplified in Figure 3. For intrinsically stationary processes the covariance function is not always defined and hence the relation is not always valid.
4.2.1 Empirical estimators

An empirical or experimental variogram estimator is computed from the observed data. It can be constructed in different ways, but it is generally based on squared differences between values at pairs of locations. The most frequently mentioned estimator is the method of moment estimator, proposed by Matheron (see, e.g., Journel and Huijbregts, 1978)

\[ 2\hat{\gamma}(h) = \frac{1}{|N_h|} \sum_{i \in N_h} (Z(x_i) - Z(x_j))^2, \]

where \(|N_h|\) is the number of distinct pairs in \(N_h = \{(x_i, x_j) : x_i - x_j = h, i, j = 1, \ldots, n\}\). This estimator is also referred to as the classical estimator. The empirical variogram can only be estimated for lags found in the observed data. The estimator is sensitive to outliers, and more robust estimators have been proposed by, e.g., Cressie and Hawkins (1980) and Genton (1998). In Lark (2000a) a class of robust estimators is compared with the overall conclusion that they do not necessary solve the problems with outliers.

If the observations are irregularly spaced, adjustments of the method of moments estimator are needed. Instead of considering pair of observations separated by a specific lag we have to group them into lag classes. Journel and Huijbregts (1978, p. 194) recommend 30-50 pairs per lag as one practical rule. They also suggest that only distances less than half the size of the area under study are used. An empirical estimator for the irregularly spaced situation is

\[ 2\hat{\gamma}(h_k) = \text{ave}\{(Z(x_i) - Z(x_j))^2 : (x_i, x_j) \in N_h, h \in T(h_k)\} \]

where \(k = 1, \ldots, K\), \(T(h_k)\) is some specified region around lag \(h_k\), and \(\text{ave}\{\cdot\}\) denotes a possible weighted average (cf. Cressie, 1993, p. 70). How smooth or noisy this estimator is depends upon the lag classes (Lark, 2000b). One argument against using empirical estimators is that the estimate at one specific lag is correlated with the estimate at another lag, since the same (correlated) observations are part of both estimates.

The empirical variogram estimators cannot be used for all lags \(h\). This is needed to enable prediction at arbitrary locations. Another problem is that the variogram must be conditionally negative semidefinite (to ensure positive variance). Hence, the empirical estimator cannot be used directly
for kriging. To overcome this problem, in general a model is assumed. Only functions that are conditionally negative semidefinite are valid as variogram models. In a completely non-parametric framework, we would like to avoid model assumptions. However, in general a variogram model is assumed.

### 4.2.2 Variogram models

Before we consider some of the most common variogram models, for isotropic processes, we briefly go through related concepts. The limit of a bounded semivariogram model, \( \gamma(h) \), \( h = \|h\| \), is referred to as the *sill*. The *range* is defined as the lag where the sill is reached. If the sill is reached only asymptotically the *practical range* is the lag where 95% of the sill is reached. The commonly used models have one parameter that controls the range. The relation between the actual parameter value and the practical range depends on the model. By definition, the variogram is always zero at lag 0. If the variogram approaches a positive value, and not zero, when the lag approaches 0, we talk about a *nugget effect*. This causes a discontinuity at the origin. The nugget effect can occur due to measurement errors.

There are many different variogram models, of which four are discussed in this thesis. Many of the models are similar in the sense that they mainly differ in the behavior close to the origin. When choosing among the available models, we could use prior knowledge or do visual inspection of the empirical variogram.

The simplest isotropic variogram model, without nugget effect, is the exponential model,

\[
2\gamma(h) = 2\sigma^2(1 - \exp(-h/r)),
\]

(2)

where \( \sigma^2 \geq 0 \) and \( r > 0 \). The exponential model has a linear decay as \( h \to 0 \). For the exponential model, the practical range is about \( 3r \). Figure 4 show how different parameter values for the exponential model changes the realization of a Gaussian process. The same set of independent standard normal random numbers is used to generate all six figures. In the left panel the variance is held constant and the range parameter is varied, and in the right panel the range is constant while the variance is varied.

Another isotropic variogram model is the spherical model

\[
2\gamma(h) = \begin{cases} 
2\sigma^2 \left( \frac{3h}{2r} - \frac{h^3}{2r^3} \right) & 0 \leq h \leq r \\
2\sigma^2 & h > r
\end{cases},
\]
Figure 4: Realizations of Gaussian spatial process with an exponential dependence structure for various parameter combinations.
where $\sigma^2 \geq 0$ and $r > 0$. Alike the exponential model, the spherical model has a linear decay when $h \to 0$. According to Webster and Oliver (2007, p. 88) this is the most frequently used model in geostatistics in one, two or three dimensions.

The third model discussed is the Gaussian variogram model,

$$2\gamma(h) = 2\sigma^2 (1 - \exp(-h^2 / r^2)), \tag{3}$$

where $\sigma^2 \geq 0$ and $r > 0$. This model has a parabolic behavior at the origin, and a practical range of about $1.73r$. Chiles and Delfiner (1999) do not recommend the use of this model (especially without nugget effect), mainly due to numerical instability.

In recent years the Matérn model, named after the Swedish statistician Bertil Matérn, has received much attention. Proponents such as Stein (1999) emphasize its flexibility. The general model is

$$2\gamma(h) = 2\sigma^2 \left(1 - \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{h}{r}\right)^\nu K_\nu \left(\frac{h}{r}\right)\right),$$

where $\sigma^2 \geq 0$ is the variance, $r > 0$ is the range parameter, $\nu > 0$ is a smoothness parameter, and $K_\nu$ is a modified Bessel function of the second kind of order $\nu$. The smoothness parameter controls how many times the function is differentiable. There exist other parameterizations of the model which does not change the model, but can affect the estimation of it. The exponential variogram model (2) is a special case of the Matérn model, with $\nu = 0.5$. When $\nu \to \infty$, we have the Gaussian model (3), and when $\nu = 1$, we have Whittle’s elementary correlation model (Webster and Oliver, 2007, p. 92). In Figure 5 we exemplify the four presented variogram models.

### 4.2.3 Estimation of model parameters

Geostatistics is partitioned into classical geostatistics, where the model parameters can be estimated by least squared methods (e.g. Cressie, 1993; Lahiri et al., 2002), and model-based geostatistics, where maximum likelihood methods are used (Stein, 1999). When using least squared methods, we fit a parametric variogram model to the empirical variogram estimator. In the papers included in this thesis, we use a weighted-least-square method. Let $\theta$ denote the vector of unknown parameters. Parameter estimates are
Figure 5: The four different variogram models exemplified in the text. The first three models have about the same practical range.

obtained by minimizing

\[
Q(\theta) = \sum_{k=1}^{K} \frac{(n - h_k)}{(\gamma(h_k, \theta))^2} \left( \hat{\gamma}(h_k) - \gamma(h_k, \theta) \right)^2,
\]

where \( \gamma(h_k, \theta) \) is the assumed variogram model with unknown parameters \( \theta \) and \( \hat{\gamma}(h_k) \) is defined by (1), see Cressie (1993, p. 99). When using this method, we neglect the correlation between the estimates at different lags in the empirical estimator \( \hat{\gamma}(h_k) \).
According to Webster and Oliver (2007, p. 122) around 150 observations are ideal in the two dimensions if the variogram is assumed to be isotropic. It should not be less than 100 observations. For the anisotropic case at least 250 observations are recommended. However, the reliability of the empirical variogram is not only affected by the sample size, it is also affected by the distribution of the data and the sample locations.

If we can motivate a completely distributional approach the variogram model can be estimated by maximum likelihood methods and we have model-based geostatistics. More specifically, it is usually assumed that the observations are a realization from a Gaussian process. If the observations seem to be a realization of some skewed process the data are transformed to become approximately Gaussian. Lark (2000b) compared the method of moment (using a weighted least squared method) and maximum likelihood methods. Both methods have their strength and weaknesses. For the maximum likelihood approach typically the log-likelihood is studied. For some models the global minima is hard to find. Another drawback of this method is its computational complexity. On the other hand, when few observations are available, it might be the only alternative. Further, it does not require any specification of lag classes, which should be seen as a strength. The main conclusion from Lark (2000b) is that maximum likelihood methods are to be recommended when the spatial structure is week. If on the other hand there are strong spatial dependencies, the method-of-moment is more suitable. Both methods seem to be affected in case of skewed data. Stein (1999) is a proponent of likelihood based methods as well as of the Matérn model. Maximum likelihood and least square methods for estimation of the Matérn variogram model are also discussed in Minasny and McBratney (2005). One of their conclusions is that the parameter $\nu$ in the Matérn model can be difficult to estimate correctly.

### 4.3 More on kriging

There are different variants of kriging, depending on the model assumptions for the process. In simple kriging, we assume that

$$Z(x) = \mu + \varepsilon(x),$$

where $\mu$ is known. The predictor $\hat{Z}_p(\theta)$ at a unknown location $x_p$ is

$$\hat{Z}_p(\theta) = \lambda(\theta)^T z,$$  \hspace{1cm} (4)
where $\mathbf{z} = (z(x_1), z(x_2), ..., z(x_n))^T$ are the observed data and $\boldsymbol{\theta}$ denotes the vector of variogram parameters. The weights, $\mathbf{\lambda}(\boldsymbol{\theta}) = (\lambda_1(\boldsymbol{\theta}), ..., \lambda_n(\boldsymbol{\theta}))^T$, are obtained by minimizing the mean squared prediction error

$$
\sigma^2(\boldsymbol{\theta}) = E[(\hat{Z}_p(\boldsymbol{\theta}) - Z(x_p))^2],
$$

subject to no constraints on the weights. More commonly used is ordinary kriging, where the trend is assumed constant, but unknown. The included papers on spatial prediction all deal with ordinary kriging. The weights in (4) are now obtained by minimizing the mean squared prediction error (5) subject to the constrain that the weights must sum to 1, to guarantee unbiasedness. It turns out that the ordinary kriging weights and the ordinary kriging variance are functions of the dependence structure solely (Cressie, 1993, p. 122)

$$
\mathbf{\lambda}(\boldsymbol{\theta})^T = \left( \gamma + \mathbf{1} \frac{1 - \mathbf{1}^T \Gamma^{-1} \gamma}{\mathbf{1}^T \Gamma^{-1} \mathbf{1}} \right)^T \Gamma^{-1},
$$

where $\mathbf{1} = (1, ..., 1)^T$ is of length $n$, $\gamma = (\gamma(x_1 - x_p), ..., \gamma(x_n - x_p))^T$, and where $\Gamma$ is the semivariogram matrix with entries $\Gamma_{ij} = \gamma(x_i - x_j)$, $i, j = 1, ..., n$. The ordinary kriging variance is

$$
\sigma^2(\boldsymbol{\theta}) = \gamma^T \Gamma^{-1} \gamma - \frac{(\mathbf{1}^T \Gamma^{-1} \gamma - 1)^2}{\mathbf{1}^T \Gamma^{-1} \mathbf{1}}.
$$

Up to this point, we have assumed that the dependence structure is known. In reality, it is not known and thus has to be estimated. The predictor is obtained by plugging in the estimated parameters in the ordinary kriging predictor; resulting in a plug-in kriging predictor. By using the plug-in estimators, we do not take the parameter uncertainty into account.

The variogram estimation is crucial since it solely determines the kriging weights and hence the predictions. Minor misspecifications might not affect the predictions, but they can affect the kriging variance (Chiles and Delfiner, 1999, p. 175). The behavior near the origin is most important. The parameter $\nu$ in the Matérn model enables various local behaviors, which is one of its advantages.
One natural and widely used extension of ordinary kriging is to allow the trend \( f(x, \beta) \) to vary over the spatial domain, such that

\[
Z(x) = \sum_{j=0}^{k} f_j(x) \beta_j + \epsilon(x),
\]

where \( \beta = (\beta_0, ..., \beta_k)^T \) are unknown parameters and \( f_0(x), ..., f_k(x) \) are known functions. This is referred to as universal kriging.

Another important kriging method to be mentioned is trans-Gaussian kriging, of which lognormal kriging is a special case. Parameter estimation and kriging are performed on transformed data. Data are transformed to become Gaussian using some link function; \( \log(\cdot) \) for lognormal kriging. The final predictions are backtransformed to the original scale by the same link function. Some adjustments are usually done because the backtransformed predictor is not unbiased. Trans-Gaussian kriging is studied in paper II.

Other kriging methods are for example; cokriging (which include explanatory variables in the model), indicator kriging (the process is converted to indicator variables), disjunctive kriging (based on functions of the observed values). For a general overview of kriging methods, we refer to Journel and Huijbregts (1978), Cressie (1993), and Webster and Oliver (2007).

We often want to predict at several locations, to construct a complete map based on the observed locations. Figure 6 illustrates the effect of the sampling locations on the predictions. The top row shows one realization of a Gaussian process with exponential variogram, generated at a dense grid (2500 grid points in a 10×10 square). A sample of about 200 observations is used to predict the whole map. The sample locations are chosen with simple random sampling, the (second) local pivotal method and on a lattice. In the bottom row of the figure, we plot the residuals; the difference between the predicted values and the true values. For simple random sampling, we clearly see that the residuals are larger in areas where we do not have any sample locations. With the two other methods, we do not have the same problem, since the observations are more evenly spaced.

In kriging, we first estimate parameters and next make predictions. As previously mentioned, the parameter uncertainty is not always taken into account when presenting the predicted values and the corresponding variance. One way of handling this is by using Bayesian methods, e.g. Diggle.
Figure 6: A realization of a Gaussian process with an exponential dependence structure generated at 2500 grid points (top row). The difference between the ordinary kriging predictor and the true values, based on a sample of about 200 locations (bottom row). Samples are selected using simple random sampling (SRS), (second) local pivotal method (LPM), and on a lattice (GRID).

and Ribeiro (2007, Ch. 7). In the Bayesian setting both the process and the parameters are seen as random variables and thus parameter uncertainty is taken into account. Prior belief/knowledge specified by the user in terms of a prior distribution of $\theta$ is combined with observed data into a posterior distribution used to find the final predictive distribution. The posterior distribution of $\theta$ is the conditional distribution of $\theta$ given observed data. Based on this, one can find the conditional distribution of the process based on the data. Bayesian methods require a prior distribution for $\theta$, which is not always trivial to choose. Bayesian methods are computationally demanding since they often include numerical evaluation of integrals.
4.4 Prediction intervals

If the spatial process is Gaussian, the predictor $\hat{Z}_p(\theta)$ is also Gaussian, and a prediction interval for $Z(x_p)$ with nominal coverage $1 - \alpha$ can be constructed as

$$[\hat{Z}_p(\theta) - z_{1-\alpha/2}\sigma(\theta), \hat{Z}_p(\theta) + z_{1-\alpha/2}\sigma(\theta)],$$

where $\hat{Z}_p(\theta)$ is the ordinary kriging predictor, $\sigma(\theta)$ is the corresponding kriging variance and $z_\alpha$ is the $\alpha$ quantile of the standard normal distribution. Both the predictor and the kriging variance are functions of unknown parameters that have to be estimated. A prediction interval is instead constructed using estimated parameters resulting in a so-called plug-in prediction interval,

$$[\hat{Z}_p(\hat{\theta}) - z_{1-\alpha/2}\hat{\sigma}(\hat{\theta}), \hat{Z}_p(\hat{\theta}) + z_{1-\alpha/2}\hat{\sigma}(\hat{\theta})].$$

This plug-in prediction interval will not end up with the nominal coverage probability.

Moreover, the Gaussian assumption need not to be fulfilled. This also affects the coverage probability. Attempts have been made to adjust the kriging variance and/or the corresponding prediction interval, so that parameter uncertainty is incorporated, see, e.g., Wang and Wall (2003).

In paper I and II, we propose and evaluate an approach to construct valid prediction intervals for the ordinary kriging predictor based on semiparametric bootstrap. This method does not require a Gaussian assumption. Let $F$ denote the distribution of $(\hat{Z}_p(\hat{\theta}) - Z(x_p))$. If $F$ is known, a nominal $1 - \alpha$ prediction interval for $Z(x_p)$ is given by

$$[\hat{Z}_p(\hat{\theta}) - q_{1-\alpha/2}, \hat{Z}_p(\hat{\theta}) - q_{\alpha/2}],$$

where $q_\alpha$ is the $\alpha$ quantile of $F$. A semiparametric bootstrap method is used to estimate the unknown distribution of $F$. The prediction interval for $Z(x_p)$ based on the semiparametric bootstrap approach is thus

$$[\hat{Z}_p(\hat{\theta}) - q^*_{1-\alpha/2}, \hat{Z}_p(\hat{\theta}) - q^*_{\alpha/2}],$$

where $q^*_\alpha$ is the $\alpha$ quantile of the empirical distribution used to estimated $F$. A brief overview of bootstrap is now given.
4.5 Bootstrap

Bootstrap is a resampling method introduced by Efron (1979) to estimate the accuracy of estimators. Assume that $z_1, z_2, \ldots, z_n$ are independent and identically distributed observations from the distribution $F(z, \xi)$, where $\xi$ is some unknown parameter. Let $T = T(n)$ denote the estimator of $\xi$. Our interest lies in the distribution of $T$. If we want to make inference for $\xi$, for example a confidence interval, we need the distribution of $T$.

If we can generate new samples from $F(z, \xi)$, we can use those to estimate the distribution of $T$. The problem is of course that $F$ is unknown. The idea of bootstrap is instead to approximate $F$ with $\hat{F}$. $\hat{F}$ can be estimated with parametric or nonparametric methods. Parametric methods involve distributional assumptions and maximum likelihood methods while nonparametric methods uses the empirical distribution of the sample data. The empirical distribution of the data is $\hat{F}(z) = \sum_{i=1}^{n} I_{z_i \leq z} / n$ where $I_{z_i \leq z} = 1$ if $z_i \leq z$ and zero otherwise. The distribution function will hence look like a step function with jumps at $z_1, \ldots, z_n$. New samples of size $n$ can be generated from $\hat{F}$. From each sample, we can estimate $\xi$ by $t^*$, resulting in $t^*_1, \ldots, t^*_B$, where $B$ is the number of bootstrap samples. Then, $t^*_1, \ldots, t^*_B$ is used to estimate the distribution of $T$.

Confidence intervals for $\xi$ can be constructed in various ways based on the bootstrap distribution. For example, a percentile interval with nominal coverage $(1 - \alpha)$ is defined as

$$\left[ t^{\star \alpha/2}, t^{\star 1-\alpha/2} \right],$$

where $t^{\star \alpha}$ is the empirical quantile of the bootstrap replicates (see, e.g., Shao and Tu, 1995).

Under some assumptions, consistency results are available for independent data. The fact that asymptotic results are available does not give desired results for small samples. For extensive summaries of bootstrap methods for both independent and dependent data, see, eg., Shao and Tu (1995), Efron and Tibshirani (1993), Davison and Hinkley (1997) or Lahiri (2003).

In papers I and II, we propose and evaluate a semi-parametric bootstrap approach (Solow, 1985) to construct prediction intervals when data are spatially dependent. The main idea behind the method is to make the data
uncorrelated before taking bootstrap samples. To each uncorrelated bootstrap sample the dependence is "added back" again. We use the ordinary kriging predictor to construct prediction intervals based on percentiles of the bootstrap distribution.

4.6 Censoring

In geostatistics, the value at a location is sometimes only partially known. Censored data are very common in survival analysis, but can occur in other areas as well. In survival analysis the individual’s lifetimes are assumed to be independent. Proposed and commonly used parametric (likelihood based) and non-parametric methods are based on this assumption. In the spatial setting, the assumption about independence is no longer valid. Within spatial statistics the methodology for analyzing censored data is not as developed. Some results can be found in Stein (1992), Militino and Ugarte (1999), De Oliveira (2005), and Rathbun (2006).

If values below a minimum detection limit are unobservable, we have left-censored data. Left-censored data can be analyzed with naive methods, where censored data are replaced with the minimum detection limit or half of it. Prediction is performed based on the imputed data, as if the data were uncensored. Many of the more advanced methods are model-based methods that uses maximum likelihood methods. They typically rely on a Gaussian assumption. In paper V, we compare one of the model-based methods (Rathbun, 2006) to the naive methods as well as to one semi-naive method. The methods are compared on processes with different characteristics such as different dependence structures and different marginal distributions. Practitioners often prefer to use as simple (user friendly) methods and hence we study how well the naive method perform in comparison to the model-based method, both under correct and incorrect distributional assumptions.
5 Summary of papers

5.1 Paper I: Kriging prediction intervals based on semiparametric bootstrap

This is a joint work with Sara Sjöstedt-de Luna. We consider prediction intervals for the ordinary kriging predictor. Typically the construction of prediction intervals relies on Gaussian assumptions. We show that the distribution of kriging predictors for non-Gaussian processes may be far from Gaussian, even asymptotically. This motivated us to search for alternative ways to construct prediction intervals. We propose to use a semiparametric bootstrap method that does not rely on distributional assumptions about the data generating process. A simulation study for Gaussian as well as lognormal processes shows that the semiparametric bootstrap method works well. For the lognormal process, we see significant improvement in coverage probability compared to traditional methods relying on Gaussian assumptions.

5.2 Paper II: Construction of kriging prediction intervals for non-Gaussian spatial processes

This paper is a continuation of paper I. It is in collaboration with Sara Sjöstedt-de Luna. Three methods to construct prediction intervals for the value of a stationary process at a new location are compared. All three methods are based on plug-in ordinary kriging predictors. We give special attention to non-Gaussian processes, where construction of intervals is less straightforward. Methods based on asymptotic normality, Gaussian transformations, and semiparametric bootstrap are compared on both simulated and real data. Our study suggests that the semiparametric method (that does not rely on distributional assumptions) is robust, and is to be recommended for non-Gaussian processes. For practitioners the semiparametric method is an attractive alternative since the method can be used without making distributional assumptions.
5.3 Paper III: Spatially balanced sampling through the pivotal method

This is a joint work with Niklas L.P. Lundstöm and Anton Grafström. We present two simple sampling designs that produce spatially balanced samples. The designs can handle both equal and unequal inclusion probabilities. We show that the designs produce samples with a high degree of spatial balance. For populations with spatial trends, examples indicates that estimation of population totals are improved using these designs. The designs can be used for higher dimensions.

5.4 Paper IV: How to select representative samples

This is a joint work with Anton Grafström. It is a follow-up on paper III, and we assume that we have several auxiliary variables available, prior to sampling. The set of auxiliary variables can include spatial coordinates. In this paper, we introduce a new way of randomly selecting representative samples given auxiliary variables. Any number of auxiliary variables, both non-categorical and categorical, can be handled. We define a representative sample as a sample well spread in the auxiliary variable space. Representative samples are particularly useful for multi-purpose surveys, when several target variables are of interest. It also enables estimation of parameters in subspaces, and improved estimation of target variable distributions. We propose a simple distance measure that, together with two previously proposed sampling designs, enables us to produce representative samples. We evaluate the procedure on two real data sets. Both sampling designs are such that they automatically stratify on well separated clusters in the auxiliary variable space. The designs will hence work well for categorical auxiliary variables. We also gain in efficiency when estimating quantities for the target variables.

5.5 Paper V: Spatial prediction in the presence of left-censoring

This paper is joint work with Sara Sjöstedt-de Luna. In this paper, we discuss how the value at unobserved locations should be predicted in presence of left-censoring. We use the kriging predictor, that is fundamental in geostatistics, and commonly used for all model-based spatial prediction.
Censored spatial data can be analyzed with sophisticated methods, that require full parametric models, or with very naive methods based on simple imputations at censored locations. We compare these two approaches, and one semi-naive approach, on processes with different characteristics. The goal with the simulation study is to find out if advanced methods can be motivated in all situations. The general answer is no.
References


Myers, D., 1989. To be or not to be... stationary? That is the question. Math. Geol. 21, 347–362.


31