1

An Introduction to Model-based Geostatistics

Peter J. Diggle
Paulo J. Ribeiro Jr
Ole F. Christensen

October 10, 2003

The text will appear as Chapter 2 in :
Møller, J. (ed.) *Spatial statistics and computational methods*, Springer Verlag, 2003.
ATTENTION : PLEASE DO NOT DISTRIBUTE

1.1 Introduction

The term *geostatistics* identifies the part of spatial statistics which is concerned with continuous spatial variation, in the following sense. The scientific focus is to study a spatial phenomenon, $s(x)$ say, which exists throughout a continuous spatial region $A \subset \mathbb{R}^2$ and can be treated as if it were a realisation of a stochastic process $S(\cdot) = \{S(x) : x \in A\}$. In general, $S(\cdot)$ is not directly observable. Instead, the available data consist of measurements $Y_1, \ldots, Y_n$ taken at locations $x_1, \ldots, x_n$ sampled within A, and Y_i is a noisy version of $S(x_i)$. We shall assume either that the sampling design for $x_1, \ldots, x_n$ is deterministic or that it is stochastic but independent of the process $S(\cdot)$, and all analyses are carried out conditionally on $x_1, \ldots, x_n$.

The subject has its origins in problems connected with estimation of ore reserves in the mining industry (Krige 1951). Its subsequent development by Matheron and colleagues at École des Mines, Fontainebleau took place largely independently of "mainstream" spatial statistics. Standard references to this "classical" approach to geostatistics include Journel & Huijbregts (1978) and Chilés & Delfiner (1999). Parallel developments by Matérn (1960) and Whittle (1954, 1962, 1963) eventually led to the integration of classical geostatistics within spatial statistics. For example, Ripley

(1981) re-cast the common geostatistical technique known as kriging within the framework of stochastic process prediction, whilst Cressie (1993) identified geostatistics as one of the three main sub-branches of spatial statistics. Significant cross-fertilisation continued throughout the 1980's and 1990's, but there is still vigorous debate on practical issues, such as the need (or not) for different approaches to prediction and parameter estimation, and the role of explicit probability models. The term *model-based geostatistics* was coined by Diggle, Tawn & Moyeed (1998) to mean the application of explicit parametric stochastic models and formal, likelihood-based methods of inference to geostatistical problems.

Our goal in this chapter is to introduce the reader to the model-based approach, in the sense intended by Diggle et al. (1998). We first describe two motivating examples, and formulate the general modelling framework for geostatistical problems, emphasising the key role of spatial prediction within the general framework. We then investigate the widely used special case of the Gaussian model, and discuss both maximum likelihood and Bayesian methods of inference. We present the results from an illustrative case-study based on one of our two motivating examples. We then consider non-Gaussian models, with a particular focus on generalised linear spatial models. The chapter concludes with some discussion, information on software and further references.

1.2 Examples of geostatistical problems

1.2.1 Swiss rainfall data

This is a standard data-set which has been widely used for empirical comparison of different methods of spatial interpolation (further information can be found at *ftp://ftp.geog.uwo.ca/SIC97*). The scientific problem posed by the data is to construct a continuous spatial map of rainfall values from observed values at a discrete set of locations. The original data consist of rainfall measurements on 8 May 1986 from 467 locations in Switzerland. The convention adopted in earlier analyses of these data is to use 100 of the data-points, as shown in Figure 1.1, to formulate and fit models to the data, and for prediction at locations without observations, whilst reserving the remaining 367 for empirical validation of the resulting predictions. In our illustrative analysis reported in Section 1.8 we use only the first 100 data points.

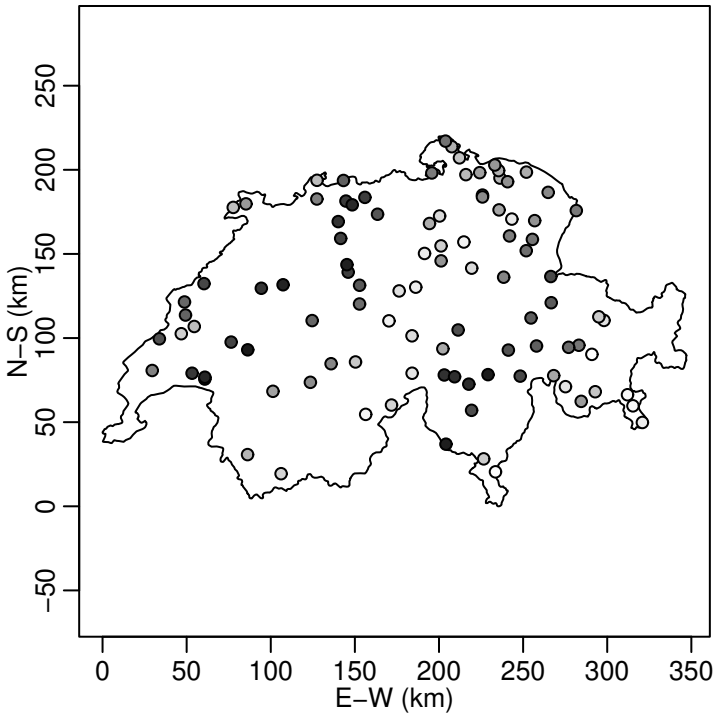


FIGURE 1.1. Swiss rainfall data at sample locations $x_1, \ldots, x_{100}$. Grey scale from white (low values) to black (high values) corresponds to the values of the observed rainfall, $y_1, \ldots, y_{100}$

1.2.2 Residual contamination of Rongelap Island

These data are from a study of residual contamination on a Pacific island, Rongelap, following the USA's nuclear weapons testing programme during the 1950's (Diggle, Harper & Simon 1997). The island was evacuated in 1985, and a large, multi-disciplinary study was subsequently undertaken to determine whether the island is now safe for re-settlement. Within this overall objective, a specific goal was to estimate the spatial variation in residual contamination over the island, with a particular interest in the maximum level of contamination. To this end, a survey was carried out and noisy measurements Y_i of radioactive caesium concentrations were obtained initially on a grid of locations x_i at 200m spacing which was later supplemented by in-fill squares at 40m spacing. Figure 1.2 shows a map of the sampling locations $x_1, \ldots, x_{157}$. The in-fill squares are particularly useful for identifying and fitting a suitable model to the data, because they give direct information about the small-scale spatial correlation structure. Design issues of this kind can have an important effect on the efficiency of any subsequent inferences. Generally speaking, placing sampling locations

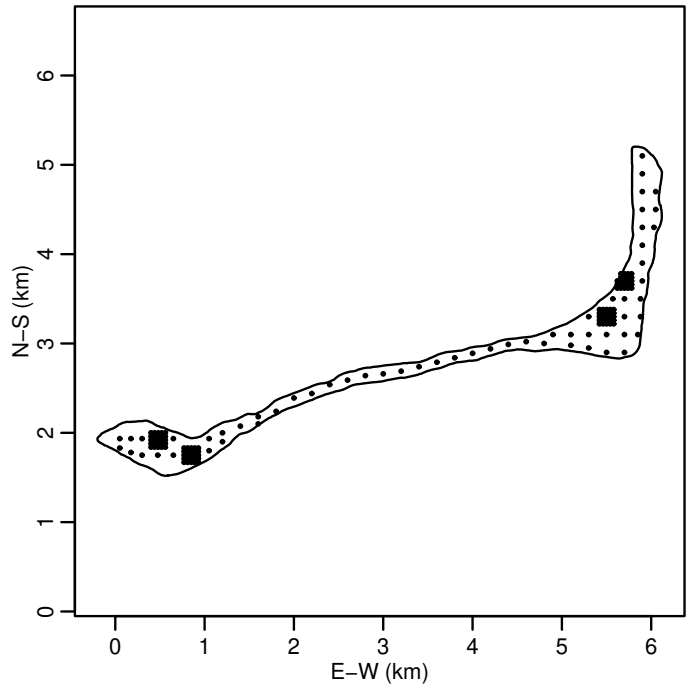


FIGURE 1.2. Sampling locations for the survey of residual contamination on Rongelap Island

in a regular grid to cover the study region would be efficient for spatial prediction if all model parameters were known, whereas deliberately including more closely spaced sub-sets of sampling locations leads to more efficient estimation of certain model parameters. In this introductory account we shall not discuss design issues further.

A full analysis of the Rongelap island data, using a spatial Poisson log-linear model and Bayesian inference, is given in Diggle et al. (1998).

1.3 The general geostatistical model

We shall adopt the following general model and notation. Firstly, the data for analysis are of the form $(x_i, y_i) : i = 1, ..., n$, where $x_1, \ldots, x_n$ are locations within a study region $A \subset \mathbb{R}^2$ and $y_1, \ldots, y_n$ are measurements associated with these locations. We call $\{x_i : i = 1, ..., n\}$ the *sampling design* and assume that y_i is a realisation of $Y_i = Y(x_i)$, where $Y(\cdot) = \{Y(x) : x \in A\}$ is the *measurement process*. We postulate the existence of an unobserved stochastic process $S(\cdot) = \{S(x) : x \in A\}$, called the *signal process*; often in practice, $Y_1, \ldots, Y_n$ are noisy versions of $S(x_1), \ldots, S(x_n)$.

Prediction is an integral part of a geostatistical analysis. We call $T = T(S(\cdot))$ the *target for prediction*. A *geostatistical model* is a specification of the joint distribution of the measurement process and the signal process, of the form $[S(\cdot), Y(\cdot)] = [Y(\cdot)|S(\cdot)][S(\cdot)]$, where $[\cdot]$ means "the distribution of." Note in particular that this model does not specify the distribution of the sampling design, which as noted earlier is assumed to be independent of both $S(\cdot)$ and $Y(\cdot)$. A *predictor* of T is any function $\hat{T} = \hat{T}(Y)$ where $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. The *minimum mean square error* predictor minimises $MSE(\hat{T}) = \mathrm{E}[(T - \hat{T})^2]$, where the expectation is taken with respect to the joint distribution of T and Y. We have the following general result.

Proposition 1. Provided that $\mathrm{Var}[T] < \infty$, the minimum mean square error predictor of T is $\hat{T} = \mathrm{E}_T[T|Y]$, with associated prediction mean square error $\mathrm{E}[(T - \hat{T})^2] = \mathrm{E}_Y \mathrm{Var}_T[T|Y]$.

It is easy to show that $\mathrm{E}[(T - \hat{T})^2] \leq \mathrm{Var}[T]$, with equality if T and Y are independent random variables.

For point prediction, it is common practice to use $\mathrm{E}[T|y]$, the minimum mean square error predictor evaluated at the observed y. Similarly, for an estimate of the achieved mean square error, we would use the value of the prediction mean square error at the observed y, also called the *prediction variance*, $\mathrm{Var}[T|y]$. However, the complete answer to a prediction problem should be expressed as a probability distribution, $[T|y]$, called the *predictive distribution*. Within the Bayesian inferential paradigm which we shall eventually adopt, the predictive distribution coincides with the *posterior distribution* of T. From this point of view, the mean and variance of this posterior distribution are just two of many possible summary statistics. In particular, the mean is not transformation invariant; if $\hat{T}$ is the best predictor for T (in a mean square sense), this does not necessarily imply that $g(\hat{T})$ is the best predictor for $g(T)$.

1.4 The Gaussian Model

In the basic form of the *Gaussian* geostatistical model, $S(\cdot)$ is a stationary Gaussian process with $\mathrm{E}[S(x)] = \mu$, $\mathrm{Var}[S(x)] = \sigma^2$ and correlation function $\rho(u) = \mathrm{Corr}[S(x), S(x')]$, where $u = \|x - x'\|$, the Euclidean distance between x and x'. Also, the conditional distribution of Y_i given $S(\cdot)$ is Gaussian with mean $S(x_i)$ and variance τ^2, and $Y_i : i = 1, ..., n$ are mutually independent, conditional on $S(\cdot)$. Figure 1.3 shows a simulation of this model in one spatial dimension.

An equivalent formulation is that

$$Y_i = S(x_i) + Z_i : i = 1, ..., n,$$

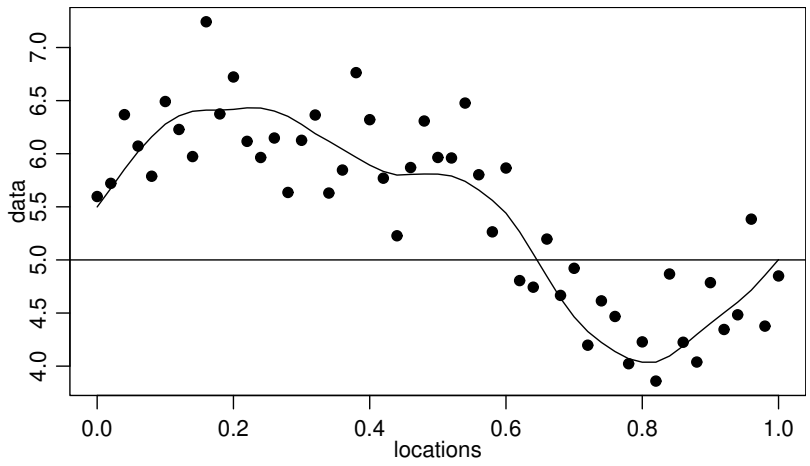


FIGURE 1.3. A simulation of the Gaussian model, illustrating the data $Y_1, \ldots, Y_n$ (dots), the signal $S(\cdot)$ (smooth curve) and the mean μ (horizontal line).

where $Z_1, \ldots, Z_n$ are mutually independent, identically distributed with $Z_i \sim N(0, \tau^2), i = 1, ..., n$. The distribution of Y is multivariate Gaussian,

$$Y \sim N(\mu \mathbf{1}, \sigma^2 R + \tau^2 I)$$

where $\mathbf{1}$ denotes an n-element vector of ones, I is the $n \times n$ identity matrix and R is the $n \times n$ matrix with $(i,j)^{th}$ element $\rho(u_{ij})$ where $u_{ij} = ||x_i - x_j||$.

The specification of the correlation function, $\rho(u)$, determines the smoothness of the resulting process $S(\cdot)$. A formal mathematical description of the smoothness of a spatial surface $S(\cdot)$ is its degree of differentiability. A process $S(\cdot)$ is *mean-square continuous* if, for all x, $\mathrm{E}[\{S(x) - S(x')\}^2] \to 0$ as $\|x - x'\| \to 0$. Similarly, $S(x)$ is *mean square differentiable* if there exists a process $S'(\cdot)$ such that, for all x,

$$\mathrm{E}\left[\left\{\frac{S(x) - S(x')}{\|x - x'\|} - S'(x)\right\}^2\right] \to 0 \text{ as } \|x - x'\| \to 0.$$

The mean-square differentiability of $S(\cdot)$ is directly linked to the differentiability of its covariance function, according to the following result, a proof of which can be found in Chapter 2.4 in Stein (1999) or Chapter 5.2 in Cramér & Leadbetter (1967).

Proposition 2. Let $S(\cdot)$ be a stationary Gaussian process with correlation function $\rho(u) : u \in \mathbb{R}$. Then, $S(\cdot)$ is mean-square continuous if and only if $\rho(u)$ is continuous at $u = 0$; $S(\cdot)$ is k times mean-square differentiable if and only if $\rho(u)$ is at least $2k$ times differentiable at $u = 0$.

In general, continuity and/or differentiability in mean square do not imply the corresponding properties for realisations. However, within the Gaussian

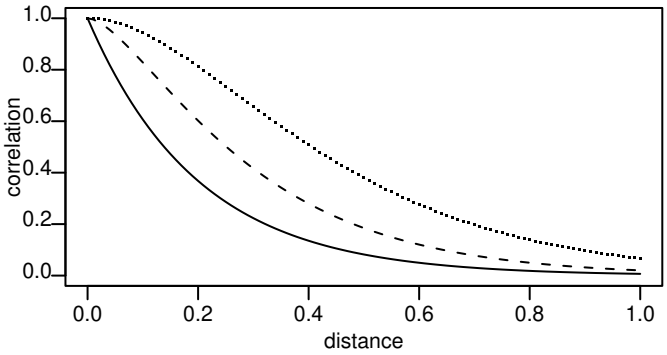


FIGURE 1.4. The Matérn correlation function with $\phi = 0.2$ and $\kappa = 1$ (solid line), $\kappa = 1.5$ (dashed line) and $\kappa = 2$ (dotted line).

framework continuity or differentiability of realisations can be achieved by imposing slightly more strict smoothness conditions on the correlation function. For details, see Chapter 9 in Cramér & Leadbetter (1967), Adler (1981) and Kent (1989).

Amongst the various families of correlation function which have been proposed, the *Matérn* family is particularly attractive. Its algebraic form is given by

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

where $\kappa > 0$ and $\phi > 0$ are parameters, and $K_{\kappa}(\cdot)$ denotes a Bessel function of order κ. Special cases include the *exponential* correlation function, $\rho(u) = \exp(-u/\phi)$, when $\kappa = 0.5$, and the *squared exponential* or *Gaussian* correlation function, $\rho(u) = \exp(-(u/\tilde{\phi})^2)$, when $\phi = \tilde{\phi}/(2\sqrt{\kappa+1})$ and $\kappa \to \infty$. What makes the family particularly attractive is that the corresponding process $S(\cdot)$ is mean-square $\lceil \kappa - 1$ times differentiable where $\lceil \kappa$ denotes the largest integer less or equal to κ. Hence κ, which can be difficult to estimate from noisy data, can be chosen to reflect scientific knowledge about the smoothness of the underlying process which $S(\cdot)$ is intended to represent. Figure 1.4 shows examples of the Matérn correlation function for $\kappa = 1, 1.5$ and 2.

Other families include the *powered exponential*,

$$\rho(u) = \exp\{-(u/\phi)^{\kappa}\},$$

defined for $\phi > 0$ and $0 < \kappa \leq 2$. This is less flexible than it first appears, because the corresponding process $S(\cdot)$ is mean-square continuous (but non-differentiable) if $\kappa < 2$, but mean-square infinitely differentiable if $\kappa = 2$, in which case the correlation matrix R may be very ill-conditioned. Figure 1.5 shows three examples of the powered exponential correlation function.

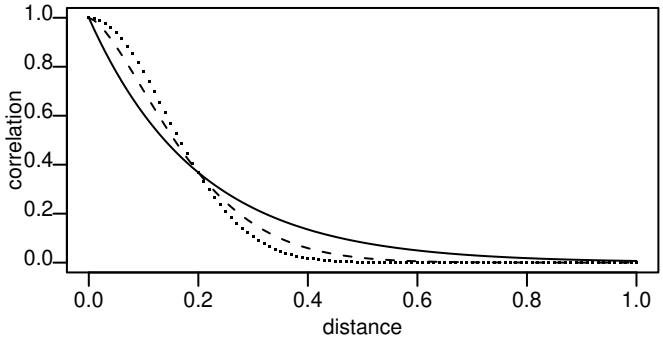


FIGURE 1.5. The powered exponential correlation function with $\phi = 0.2$ and $\kappa = 1$ (solid line), $\kappa = 1.5$ (dashed line) and $\kappa = 2$ (dotted line).

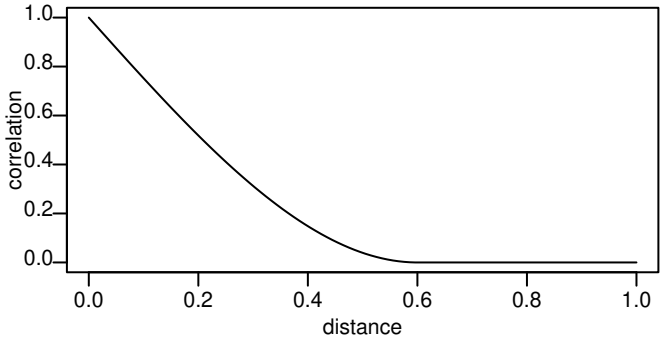


FIGURE 1.6. The spherical correlation function with $\phi = 0.6$.

In classical geostatistics, the *spherical* family is widely used. This has

$$\rho(u;\phi) = \begin{cases} 1 - \frac{3}{2}(u/\phi) + \frac{1}{2}(u/\phi)^3 & : \quad 0 \leq u \leq \phi \\ 0 & : \quad u > \phi \end{cases}$$

where $\phi > 0$ is a single parameter. One qualitative difference between this and the earlier families is that it has a finite range, i.e. $\rho(u) = 0$ for sufficiently large u. With only a single parameter it lacks the flexibility of the Matérn class. Also, the function is only once differentiable at $u = \phi$ which can cause difficulties with maximum likelihood estimation (Warnes & Ripley 1987, Mardia & Watkins 1989). Figure 1.6 shows an example of the spherical correlation function with correlation parameter $\phi = 0.6$.

Note that all the correlation functions presented here have the property that $\rho(u;\phi) = \rho_0(u/\phi)$; i.e. ϕ is a scale parameter with units of distance.

It is instructive to compare realisations of Gaussian processes with different

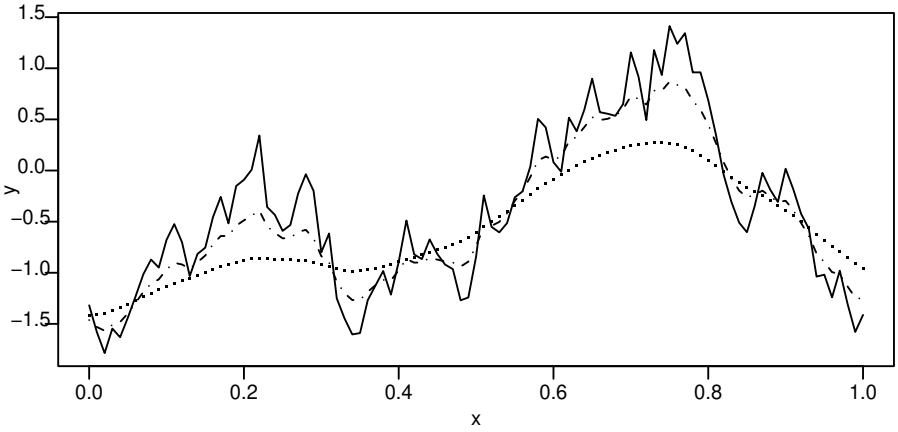


FIGURE 1.7. Simulations of Gaussian processes with Matérn correlation functions, using $\phi = 0.2$ and $\kappa = 0.5$ (solid line), $\kappa = 1$ (dashed line) or $\kappa = 2$ (dotted line).

correlation functions. For example, Figure 1.7 shows realisations of three different processes within the Matérn class, all generated from the same random number stream; the differences in smoothness as κ varies are very clear.

1.4.1 Prediction Under The Gaussian Model

Assume initially that the target for prediction is $T = S(x_0)$, the value of the signal process at a particular location x_0, where x_0 is not necessarily included within the sampling design. Under the Gaussian model, $[T, Y]$ is multivariate Gaussian. Therefore, $\hat{T} = \mathrm{E}[T|Y]$, the prediction variance $\mathrm{Var}[T|Y]$ and the predictive distribution $[T|Y]$ can be easily derived from the following standard result.

Proposition 3. Let $X = (X_1, X_2)$ be multivariate Gaussian, with mean vector $\mu = (\mu_1, \mu_2)$ and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

i.e. $X \sim N(\mu, \Sigma)$. Then, the conditional distribution of X_1 given $X_2 = x_2$ is also multivariate Gaussian, $X_1|X_2 = x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$, where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

and

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

For the geostatistical model $[T, Y]$ is multivariate Gaussian with mean vector $\mu\mathbf{1}$ and variance matrix

$$\begin{bmatrix} \sigma^2 & \sigma^2\mathbf{r}^{\mathrm{T}} \\ \sigma^2\mathbf{r} & \tau^2 I + \sigma^2 R \end{bmatrix}$$

where $\mathbf{r}$ is a vector with elements $r_i = \rho(||x_0 - x_i||) : i = 1, ..., n$. Hence, using Proposition 3 with $X_1 = T$ and $X_2 = Y$, we find that the minimum mean square error predictor for $T = S(x_0)$ is

$$\hat{T} = \mu + \sigma^2\mathbf{r}^{\mathrm{T}}(\tau^2 I + \sigma^2 R)^{-1}(y - \mu\mathbf{1}) \tag{1.1}$$

with prediction variance

$$\mathrm{Var}[T|y] = \sigma^2 - \sigma^2\mathbf{r}^{\mathrm{T}}(\tau^2 I + \sigma^2 R)^{-1}\sigma^2\mathbf{r}. \tag{1.2}$$

Note that in the Gaussian model, for fixed values of the parameters, the conditional variance does not depend on y but only on the spatial configuration of the data and prediction location(s) defining R and r. In conventional geostatistical terminology, construction of the surface $\hat{S}(\cdot)$, where for each location x_0, $\hat{T} = \hat{S}(x_0)$ is given by (1.1), is called *simple kriging*. This name was given by G. Matheron as a reference to D.G. Krige, who pioneered the use of statistical methods in the South African mining industry (Krige 1951).

The minimum mean square error predictor for $S(x_0)$ can be written explicitly as a linear function of the data y

$$\begin{aligned} \hat{T} = \hat{S}(x_0) &= \mu + \sum_{i=1}^{n} w_i(x_0)(y_i - \mu) \\ &= \{1 - \sum_{i=1}^{n} w_i(x_0)\}\mu + \sum_{i=1}^{n} w_i(x_0)y_i. \end{aligned}$$

Thus, the predictor $\hat{S}(x_0)$ compromises between its unconditional mean μ and the observed data y, the nature of the compromise depending on the target location x_0, the data-locations $x_1, \ldots, x_n$ and the values of the model parameters. We call $w_1(x_0), \ldots, w_n(x_0)$ the *prediction weights*. In general, the weight $w_i(x_0)$ tends to be large when x_i is close to x_0, $i = 1, \ldots, n$, and conversely, but this depends on the precise interplay between the sampling design and the assumed covariance structure of the data; in particular, even when the assumed correlation function is decreasing in distance, there is no guarantee that the weights will decrease with distance. Nor are they guaranteed to be positive, although in most practical situations large negative weights are rare.

One way to gain insight into the behaviour of the simple kriging predictor, $\hat{S}(\cdot)$, is to compute it for particular configurations of data under a range

of assumed covariance structures. Note in particular the following general features of $\hat{S}(\cdot)$. Firstly, the surface $\hat{S}(\cdot)$ interpolates the data (meaning that $\hat{S}(x_i) = y_i$ for all x_i in the sampling design) if and only if $\tau^2 = 0$, since in this case $Y(x_i) = S(x_i)$ for $i = 1, ..., n$. When $\tau^2 > 0$, $\hat{S}(\cdot)$ tends to smooth out extreme fluctuations in the data. Secondly, for the correlation models considered here, $\hat{S}(\cdot)$ inherits the analytic smoothness at the origin of the assumed correlation function of $S(\cdot)$. So, for example, within the Matérn class, $\kappa \leq 0.5$ leads to a continuous but non-differentiable surface $\hat{S}(\cdot)$ whereas $\kappa > 0.5$ produces a smoother, differentiable surface. Finally, for typical correlation models in which $\rho(u) \to 0$ as $u \to \infty$, $\hat{S}(x_0) \approx \mu$ for a location x_0 sufficiently remote from all x_i in the sampling design, whereas when x_0 is close to one or more x_i, the corresponding $\hat{S}(x_0)$ will be more strongly influenced by the y_i's at these adjacent sampling locations.

Figure 1.8 illustrates some of these points, in the case of a small, one-dimensional data-set. The lines in the upper panel are the point predictions $\hat{S}(x), x \in [0; 1]$ obtained using the data indicated by the circles. The data y are assumed to follow the model $Y_i = S(x_i) + Z_i$ where $S(\cdot)$ has mean $\mu = 0$, signal variance $\sigma^2 = 1$ and a Matérn correlation function with $\phi = 0.2$ and $\kappa = 2$, and Z_i are mutually independent with zero mean and variance τ^2. Holding the data fixed, Figure 1.8 shows the predictions which result when we assume each of $\tau^2 = 0$, 0.25 and 0.5. We observe that at data locations, when $\tau^2 = 0$ the predicted values coincide with the data. The higher the value of τ^2 the more the predictions approach the overall mean. The lower panel shows the corresponding prediction variances with tick-marks indicating the data locations.

In many applications, the inferential focus is not on $S(x_0)$ at a specific location x_0, but on some other property which can be expressed as a functional of the complete surface $S(\cdot)$, for example an areal average or maximum value. Firstly, let T be any *linear* functional of $S(\cdot)$,

$$T = \int_A w(x)S(x)dx$$

for some prescribed weighting function $w(x)$. Under the Gaussian model, $[T, Y]$ is multivariate Gaussian, hence $[T|y]$ is univariate Gaussian and the conditional mean and variance are

$$\mathrm{E}[T|y] = \int_A w(x)\mathrm{E}[S(x)|y]dx$$

and

$$\mathrm{Var}[T|y] = \int_A \int_A w(x)w(x')\mathrm{Cov}[S(x), S(x') \mid y]dxdx'.$$

Note in particular that

$$\hat{T} = \int_A w(x)\hat{S}(x)dx.$$

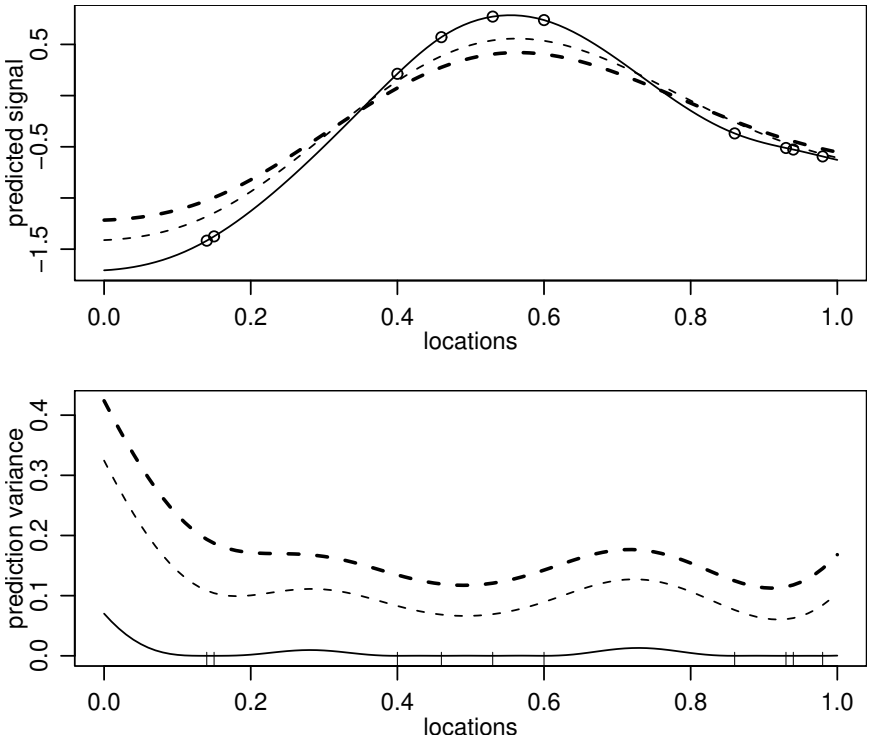


FIGURE 1.8. Point predictions and the data indicated by circles (upper panel) and prediction variances (lower panel) from 10 randomly spaced sampling locations indicated by the tick-marks in lower panel, assuming a Matérn correlation function with $\phi = 0.2$ and $\kappa = 2$, $\sigma^2 = 1$ and τ^2: 0 (solid line), 0.25 (thin dashed line) and 0.5 (thick dashed line).

In other words, given a predicted surface $\hat{S}(\cdot)$, it is reasonable simply to calculate any linear property of this surface and to use the result as the predictor for the corresponding linear property of the true surface $S(\cdot)$. However, this is not the case for prediction of non-linear properties. Note in particular that in practice the point predictor $\hat{S}(\cdot)$ tends to under-estimate peaks and over-estimate troughs in the true surface $S(\cdot)$. Hence, for example, the maximum of $\hat{S}(\cdot)$ would be a poor predictor for the maximum of $S(\cdot)$.

1.4.2 Extending the Gaussian model

The Gaussian model discussed so far is, of course, not appropriate for all applications. In later sections, we will discuss a range of non-Gaussian models. Here, we discuss briefly how some of the assumptions may be relaxed whilst remaining within the Gaussian framework.

Firstly, we need to be able to deal with a non-constant mean value surface $\mu(x)$. Technically the simplest case is when $\mu(x)$ is specified by a linear model, $\mu(x) = \sum_{j=1}^{p} \beta_j f_j(x)$, where $f_1(x), \ldots, f_p(x)$ are observed functions of location, x. A special case, known as polynomial trend surface modelling, arises when $f_1(x), \ldots, f_p(x)$ are powers of the spatial coordinates $x_{(1)}$ and $x_{(2)}$. In our opinion, linear or possibly quadratic trend surfaces are occasionally useful as pragmatic descriptions of spatial variation in an overall level of the responses $Y_1, \ldots, Y_n$, but more complicated polynomial trend surfaces are seldom useful, since they often lead to unrealistic extrapolations beyond the convex hull of the sampling design. Another possibility is to define the $f_j(x)'s$ above as functions of observed covariates. Note that this requires covariate measurements also to be available at prediction locations. The procedure of obtaining predictions using a polynomial trend of the coordinates is often called *universal kriging* in the geostatistics literature, while the case when other covariates are used is called *kriging with a trend model* (Goovaerts 1997). Non-linear models for $\mu(x)$ will often be more realistic on physical grounds. However, fitting non-linear models is technically less straightforward than in the linear case and needs to be approached with caution.

Secondly, in some applications we may find empirical evidence of directional effects in the covariance structure. The simplest way to deal with this is by introducing a *geometric anisotropy* into the assumed covariance structure. Physically, this corresponds to a rotation and stretching of the original spatial coordinates. Algebraically, it adds to the model two more parameters: the *anisotropy angle* ψ_A and the *anisotropy ratio* $\psi_R > 1$. These define a transformation of the space of locations $x = (x_{(1)}, x_{(2)})$ according to

$$(x'_{(1)}, x'_{(2)}) = (x_{(1)}, x_{(2)}) \begin{pmatrix} \cos(\psi_A) & -\sin(\psi_A) \\ \sin(\psi_A) & \cos(\psi_A) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \psi_R^{-1} \end{pmatrix}$$

and the correlation between two locations is modelled as a function of distance in this transformed space.

A third possible extension is to assume an additional component for the variance, the so-called micro-scale variation, hence in the stationary case with no covariates the model is extended to

$$Y_i = S(x_i) + S_0(x_i) + Z_i : i = 1, ..., n$$

where $S(\cdot)$ and Z_i are as before but additionally $S_0(\cdot)$ is a stationary Gaussian process with rapidly decaying spatial correlation. If we formally assume that $S_0(\cdot)$ is uncorrelated spatial Gaussian white noise, then the terms $S_0(x_i)$ and Z_i are indistinguishable. In practice, they will also be indistinguishable if the correlation of $S_0(\cdot)$ decays within a distance smaller than the smallest distance between any two sampling locations. In mining applications the micro-scale component is assumed to be caused by the existence

of small nuggets of enriched ore and is approximated by a white noise process. Hence, in practice the term "nugget effect" applied to the independent error term Z_i is interpreted, according to context, as measurement error, micro-scale variation or a non-identifiable combination of the two.

Stationarity itself is a convenient working assumption, which can be relaxed in various ways. A functional relationship between mean and variance can sometimes be resolved by a transformation of the data. When the responses $Y_1, \ldots, Y_n$ are continuous but the Gaussian model is clearly inappropriate, some additional flexibility is obtained by introducing an extra parameter λ defining a Box-Cox transformation of the response. The resulting model assumes that the data, denoted $y = (y_1, ..., y_n)$, can be transformed by

$$\tilde{y}_i = h_\lambda(y_i) = \begin{cases} (y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log y_i & \text{if } \lambda = 0, \end{cases} \tag{1.3}$$

such that $(\tilde{y}_1, \ldots, \tilde{y}_n)$ is a realisation from a Gaussian model. De Oliveira, Kedem & Short (1997) propose formal Bayesian methods of inference within this model class, one consequence of which is that their predictions are averages over a range of models corresponding to different values of λ. An alternative approach is to estimate λ, but then hold λ fixed when performing prediction (Christensen, Diggle & Ribeiro Jr 2001). This avoids the difficulty of placing a physical interpretation on a predictive distribution which is averaged over different scales of measurement.

Intrinsic variation, a weaker hypothesis than stationarity, states that the process has stationary increments. This represents a spatial analogue of the random walk model for time series, and is widely used as a default model for discrete spatial variation, see Chapter 3 and (Besag, York & Mollié 1991).

Finally, *spatial deformation* methods (Sampson & Guttorp 1992) seek to achieve stationarity by a non-linear transformation of the geographical space, $x = (x_{(1)}, x_{(2)})$.

It is important to remember that the increased flexibility of less restrictive modelling assumptions is bought at a price. In particular, over-complex models fitted to sparse data can easily lead to poor identifiability of model parameters, and to poorer predictive performance than simpler models.

1.5 Parametric estimation of covariance structure

1.5.1 Variogram analysis

In classical geostatistics, the standard summary of the second-moment structure of a spatial stochastic process is its variogram. The *variogram*

of a stochastic process $Y(\cdot)$ is the function

$$V(x, x') = \frac{1}{2}\text{Var}\{Y(x) - Y(x')\}.$$

For the linear Gaussian model, with $u = ||x - x'||$,

$$V(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}.$$

The basic structural covariance parameters of the linear Gaussian model are the *nugget variance*, τ^2, the *total sill*, $\tau^2 + \sigma^2 = \text{Var}\{Y(x)\}$, and the *range*, ϕ, such $\rho(u) = \rho_0(u/\phi)$. Thus, any reasonable version of the linear Gaussian model will involve at least three covariance parameters. However, we would need abundant data (or contextual knowledge) to justify estimating more than three parameters. Note in particular that the Matérn family uses a fourth parameter to determine the differentiability of $S(\cdot)$. Our view is that it is sensible to choose κ from amongst a small set of values to reflect contextual knowledge about the smoothness of $S(\cdot)$, rather than formally to estimate it from sparse data.

The *variogram cloud* of a set of geostatistical data is a scatterplot of the points (u_{ij}, v_{ij}), derived from the quantities

$$\begin{aligned} u_{ij} &= ||x_i - x_j|| \\ v_{ij} &= (y_i - y_j)^2/2. \end{aligned}$$

The left-hand panel of Figure 1.9 shows an example of a variogram cloud, calculated from the Swiss rainfall data. Its diffuse appearance is entirely typical. Note in particular that under the linear Gaussian model, $v_{ij} \sim V(u_{ij})\chi^2_1$ and different v_{ij}'s are correlated. The variogram cloud is therefore unstable, both pointwise and in its overall shape.

When the underlying process has a spatially varying mean $\mu(x)$ the variogram cloud as defined above is not a sensible summary. Instead, we replace the data y_i in the expression for v_{ij} by residuals $r_i = y_i - \hat{\mu}(x_i)$, where $\hat{\mu}(\cdot)$ is an estimate of the underlying mean value surface, typically an ordinary least squares estimate within an assumed linear model.

A more stable variant of the variogram cloud is the *empirical variogram* $\bar{V}(\cdot)$, as illustrated on the right-hand panel of Figure 1.9. For a separation distance u, $\bar{V}(\cdot)$ is obtained by averaging those v_{ij}'s for which $|u - u_{ij}| < h/2$, where h is a chosen bin width. The averaging addresses the first objection to the variogram cloud, namely its pointwise instability, but the difficulties caused by the inherent correlation amongst different variogram ordinates remain. Note also that the empirical variogram is necessarily sensitive to mis-specification of the mean value surface $\mu(x)$. Specifically, failure to adjust for long-range variation in the mean response will induce spurious evidence of long-range correlation in $Y(\cdot)$.

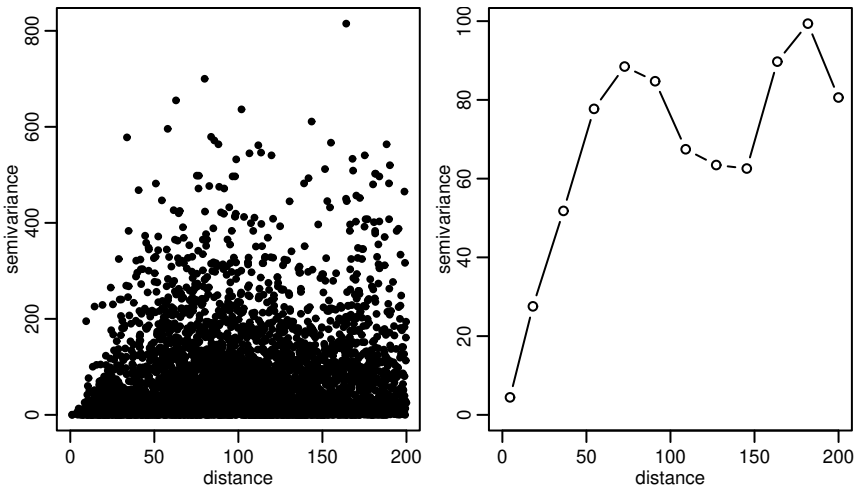


FIGURE 1.9. The variogram cloud (left panel) and binned variogram (right panel) for the Swiss rainfall data

Fitting a parametric covariance function to an empirical variogram provides one possible way to estimate covariance parameters. Frequently in practice this is done "by eye", without a formal criterion. Alternatively, ordinary or weighted least squares methods for curve fitting are sometimes used. These methods estimate the covariance parameters θ by minimising

$$S(\theta) = \sum_k w_k[\bar{V}(u_k) - V(u_k;\theta)]^2$$

where $w_k = 1$ for ordinary least squares, whereas for weighted least squares w_k is the number of pairs of measurements which contribute to $\bar{V}(u_k)$. The resulting fits are often visually convincing, but this begs the question of whether matching theoretical and empirical variograms is optimal in any sense. In fact, empirical variograms calculated from typical sizes of data-set are somewhat unstable. To illustrate this, Figure 1.10 compares the empirical variograms from three independent simulations of the same model with the true underlying variogram, where the correlation function is exponential, the parameters $\sigma^2 = 1$, $\phi = 0.25$, $\tau^2 = 0$, and 100 locations randomly distributed in a unit square. The inherently high autocorrelations amongst $\hat{V}(u)$ for successive values of u impart a misleading smoothness into the empirical variograms, suggesting greater precision than is in fact the case.

Parameter estimation via the variogram is a deeply rooted part of classical geostatistical methodology but its popularity is, in our view, misplaced. It does have a role to play in exploratory analysis, at model formulation stage and as a graphical diagnostic. For formal inference, we prefer likelihood-based methods. These have the compelling (to us) advantage that they are

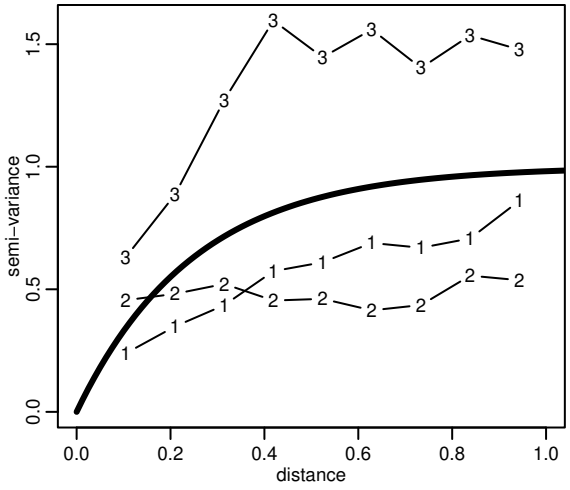


FIGURE 1.10. Empirical variograms from three independent realisations of the same stationary Gaussian process.

optimal under the stated assumptions, although they are computationally expensive for large data-sets, and a legitimate concern is that they may lack robustness. The likelihood function also plays a central role in Bayesian inference, in which estimation and prediction are naturally combined. We discuss this in greater detail in Section 1.7.

1.5.2 Maximum likelihood estimation

Under the Gaussian model

$$Y \sim N(F\beta, \sigma^2 R + \tau^2 I)$$

where F is the $n \times p$ matrix of covariates, β is the vector of parameters, and R depends on (ϕ, κ). The log-likelihood function is

$$\begin{aligned} l(\beta, \tau^2, \sigma^2, \phi, \kappa) \quad \propto \quad & -0.5\{\log|(\sigma^2 R + \tau^2 I)| \\ & +(y - F\beta)^{\mathrm{T}}(\sigma^2 R + \tau^2 I)^{-1}(y - F\beta)\}, \end{aligned} \tag{1.4}$$

maximisation of which yields the maximum likelihood estimates of the model parameters.

Computational details are as follows. Firstly, we reparameterise to $\nu^2 = \tau^2/\sigma^2$ and denote $V = (R + \nu^2 I)$. Given V, the log-likelihood function is maximised for

$$\hat{\beta}(V) = (F^{\mathrm{T}} V^{-1} F)^{-1} F^{\mathrm{T}} V^{-1} y$$

and

$$\hat{\sigma}^2(V) = n^{-1}(y - F\hat{\beta})^{\mathrm{T}} V^{-1}(y - F\hat{\beta}).$$

Hence, substituting $(\hat{\beta}(V), \hat{\sigma}^2(V))$ into the log-likelihood function, we obtain the reduced log-likelihood

$$l(\nu^2, \phi, \kappa) \propto -0.5\{n \log |\hat{\sigma}^2(V)| + \log |V|\}.$$

This must then be optimised numerically, followed by back-substitution to obtain $\hat{\sigma}^2$ and $\hat{\beta}$. In practice, for the Matérn correlation function we suggest choosing κ from the discrete set $\{0.5, 1, 1.5, 2, 2.5, ..., K/2\}$ for some small integer K.

If geometric anisotropy parametrised by (ψ_A, ψ_R) is included in the model, the same procedure is used, except that the additional parameters need to be incorporated into the matrix R, thereby adding two dimensions to the numerical maximisation of the likelihood.

For the transformed Gaussian model defined by (1.3), the associated log-likelihood is

$$\begin{aligned} \ell(\beta, \sigma^2, \phi, \nu^2, \kappa, \lambda) &= (\lambda - 1) \sum_{i=1}^{n} \log y_i - 0.5 \log |\sigma^2 V| \\ &\quad -0.5(h_\lambda(y) - F\beta)^{\mathrm{T}} \{\sigma^2 V\}^{-1} (h_\lambda(y) - F\beta)\}. \end{aligned}$$

Here we use the procedure above, but adding optimisation with respect to λ in the numerical maximisation.

A popular variant of maximum likelihood estimation is *restricted maximum likelihood estimation* (REML). Under the assumed model for $\mathrm{E}[Y] = F\beta$, we can transform the data linearly to $Y^* = AY$ such that the distribution of Y^* does not depend on β. Then, the REML principle is to estimate $\theta = (\nu^2, \sigma^2, \phi, \kappa)$ by maximum likelihood applied to the transformed data Y^*. We can always find a suitable matrix A without knowing the true values of β or θ, for example a projection to ordinary least squares residuals,

$$A = I - F(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}.$$

The REML estimators for θ is computed by maximising

$$\begin{aligned} l^*(\theta) &\propto -0.5\{\log |\sigma^2 V| - \log |F^{\mathrm{T}} \{\sigma^2 V\}^{-1} F| \\ &\quad +(y - F\tilde{\beta})^T \{\sigma^2 V\}^{-1} (y - F\tilde{\beta}))\}, \end{aligned}$$

where $\tilde{\beta} = \hat{\beta}(V)$. Note the extra determinant term by comparison with the ordinary log-likelihood given by (1.4).

REML was introduced in the context of variance components estimation in designed experiments (Patterson & Thompson 1971) and some early references in the geostatistical context are Kitanidis (1983) and Zimmerman (1989). In general, it leads to less biased estimators of variance parameters in small samples (for example, the elementary unbiased sample variance

is a REML estimator). Note that $l^*(\theta)$ depends on F, and therefore on a correct specification of the model for $\mu(x)$. For designed experiments, the specification of the mean $\mu(x)$ is usually not problematic. However, in the spatial setting the specification of the mean $\mu(x)$ is often a pragmatic choice. Although REML is widely recommended for geostatistical models, our experience has been that it is more sensitive than ML to misspecification of the model for $\mu(x)$.

Another generic likelihood-based idea which is useful in the geostatistical setting is that of profile likelihoods. In principle, variability of parameter estimators can be investigated by inspection of the log-likelihood surface. However, the typical dimension of this surface does not allow direct inspection. Suppose, in general, that we have a model with parameters (α, ψ) and denote its likelihood by $L(\alpha, \psi)$. To inspect the likelihood for α, we replace the nuisance parameters ψ by their ML estimators $\hat{\psi}(\alpha)$, for each value of α. This gives the *profile likelihood* for α,

$$L_p(\alpha) = L(\alpha, \hat{\psi}(\alpha)) = \max_{\psi}(L(\alpha, \psi)).$$

The profile log-likelihood can be used to calculate approximate confidence intervals for individual parameters, exactly as in the case of the ordinary log-likelihood for a single parameter model.

1.6 Plug-in prediction

We use this term to mean the simple approach to prediction whereby estimates of unknown model parameters are plugged into the prediction equations as if they were the truth. This tends to be optimistic in the sense that it leads to an under-estimation of prediction uncertainty by ignoring variability between parameter estimates and their true, unknown values. Nevertheless, it is widely used, corresponds to standard geostatistical methods collectively known as kriging, and is defensible in situations where varying model parameters over reasonable ranges produces only small changes in the sizes of the associated prediction variances.

1.6.1 The Gaussian model

For the Gaussian model we have seen that the minimum MSE predictor for $T = S(x_0)$ is

$$\hat{T} = \mu + \sigma^2 \mathbf{r}^{\mathrm{T}}(\tau^2 I + \sigma^2 R)^{-1}(y - \mu \mathbf{1})$$

with prediction variance

$$\mathrm{Var}[T|y] = \sigma^2 - \sigma^2 \mathbf{r}^{\mathrm{T}}(\tau^2 I + \sigma^2 R)^{-1}\sigma^2 \mathbf{r}.$$

A *plug-in* prediction consists of replacing the true parameters in the prediction equations above by their estimates. As noted earlier, simple kriging is prediction where estimates of the mean and covariance parameters are plugged-in. Another approach often used in practice is *ordinary kriging*, which only requires covariance parameters to be plugged-in (Journel & Huijbregts 1978). Ordinary kriging uses a linear predictor which minimises the mean square prediction error under an unbiasedness constraint which implies that the prediction weights must sum to one. This filters out the mean parameter from the expression for the predictor.

1.6.2 The transformed Gaussian model

For the Box-Cox transformed Gaussian model, assume $Y(x_0)$ is the target for prediction, and denote $T_\lambda = h_\lambda(Y(x_0))$. The minimum mean square error predictor $\hat{T}_\lambda$ and the corresponding prediction variance $\mathrm{Var}[T_\lambda \mid y]$ are found as above using simple kriging. Back-transforming to the original scale is done using formulas for moments. For $\lambda = 0$ we use properties of the exponential of a normal distribution and get

$$\hat{T} = \exp(\hat{T}_0 + 0.5\mathrm{Var}[T_0 \mid y])$$

with prediction variance

$$\mathrm{Var}[T|y] = \exp(2\hat{T}_0 + \mathrm{Var}[T_0 \mid y])(\exp(\mathrm{Var}[T_0 \mid y]) - 1).$$

For $\lambda > 0$ we can approximate $\hat{T}$ and $\mathrm{Var}[T|y]$ by a sum of moments for the normal distribution. For $\lambda = 0.5$ we get

$$\hat{T} \approx (0.5\hat{T}_{0.5} + 1)^2 + 0.25\mathrm{Var}[T_{0.5} \mid y]$$

with prediction variance

$$\mathrm{Var}[T|y] \approx (0.5\hat{T}_{0.5}+1)^4+1.5(0.5\hat{T}_{0.5}+1)^2\mathrm{Var}[T_{0.5} \mid y]+3(\mathrm{Var}[T_{0.5} \mid y])^2/16.$$

Alternatively, back-transformation to the original scale can be done by simulation as discussed in the next sub-section.

1.6.3 Non-linear targets

In our experience, the plug-in approach and the Bayesian approach presented in the next section usually give similar point predictions when predicting $T = S(x_0)$, but often the prediction variances differ and the two approaches can produce very different results when predicting non-linear targets.

Consider prediction of the non-linear target $T = T(S^*)$ where S^* are values of $S(\cdot)$ at some locations of interest (for example, a fine grid over the entire area). A general way to calculate the predictor $\hat{T}$ is by simulation. The procedure consists of the following three steps.

- Calculate $\mathrm{E}[S^*|y]$ and $\mathrm{Var}[S^*|y]$ using simple kriging.
- Simulate $s^*(1), \ldots, s^*(m)$ from $[S^*|y]$ (multivariate Gaussian).
- Approximate the minimum mean square error predictor

$$\mathrm{E}[T(S^*)|y] \approx \frac{1}{m}\sum_{j=1}^{m} T(s^*(j)).$$

For the transformed Gaussian model we use a procedure similar to above, we just need to back-transform the simulations by $h_\lambda^{-1}(\cdot)$ before taking averages.

1.7 Bayesian inference for the linear Gaussian model

Bayesian inference treats parameters in the model as random variables, and therefore makes no formal distinction between parameter estimation problems and prediction problems. This provides a natural means of allowing for parameter uncertainty in predictive inference.

1.7.1 Fixed correlation parameters

To derive Bayesian inference results for the linear Gaussian model, we first consider the situation in which we fix $\tau^2 = 0$, all other parameters in the correlation function have known values, and we allow for uncertainty only in the parameters β and σ^2. In this case the predictive distributions can be derived analytically.

For fixed ϕ, the conjugate prior family for (β, σ^2) is the Gaussian-Scaled-Inverse-χ^2. This specifies priors for β and σ^2 with respective distributions

$$[\beta|\sigma^2, \phi] \sim N\left(m_b, \sigma^2 V_b\right) \quad \text{and} \quad [\sigma^2|\phi] \sim \chi^2_{ScI}\left(n_\sigma, S^2_\sigma\right),$$

where a $\chi^2_{ScI}(n_\sigma, S^2_\sigma)$ distribution has density of the form

$$\pi(z) \propto z^{-(n_\sigma/2+1)} \exp(-n_\sigma S^2_\sigma/(2z)), \quad z > 0.$$

As a convenient shorthand, we write this as

$$[\beta, \sigma^2|\phi] \sim N\chi^2_{ScI}\left(m_b, V_b, n_\sigma, S^2_\sigma\right), \tag{1.5}$$

Using Bayes' Theorem, the prior above is combined with the likelihood given by (1.4) and the resulting posterior distribution of the parameters is:

$$[\beta, \sigma^2|y, \phi] \sim N\chi^2_{ScI}\left(\tilde{\beta}, V_{\tilde{\beta}}, n_\sigma + n, S^2\right), \tag{1.6}$$

where $V_{\tilde{\beta}} = (V_b^{-1} + F^{\mathrm{T}}R^{-1}F)^{-1}$, $\tilde{\beta} = V_{\tilde{\beta}}(V_b^{-1}m_b + F^{\mathrm{T}}R^{-1}y)$ and

$$S^2 = \frac{n_\sigma S^2_\sigma + m_b^{\mathrm{T}}V_b^{-1}m_b + y^{\mathrm{T}}R^{-1}y - \tilde{\beta}^{\mathrm{T}}V_{\tilde{\beta}}^{-1}\tilde{\beta}}{n_\sigma + n}. \tag{1.7}$$

The predictive distribution of the signal at an arbitrary set of locations, say $S^* = (S(x_{n+1}), \ldots, S(x_{n+q}))$, is obtained by integration,

$$p(s^*|y, \phi) \;=\; \int\int p(s^*|y, \beta, \sigma^2, \phi)\, p(\beta, \sigma^2|y, \phi)\, d\beta d\sigma^2,$$

where $[s^*|y, \beta, \sigma^2, \phi]$ is multivariate Gaussian with mean and variance given by (1.1) and (1.2) respectively. The integral above yields a q-dimensional multivariate-t distribution defined by:

$$\begin{aligned} [S^*|y, \phi] &\sim t_{n_\sigma+n}\left(\mu^*, S^2\Sigma^*\right), \\ \mathrm{E}[S^*|y, \phi] &= \mu^*, \\ \mathrm{Var}[S^*|y, \phi] &= \frac{n_\sigma + n}{n_\sigma + n - 2}\, S^2\Sigma^*, \end{aligned} \tag{1.8}$$

where S^2 is given by (1.7) and μ^* and Σ^* are

$$\begin{aligned} \mu^* &= (F_0 - r^{\mathrm{T}}R^{-1}F)V_{\tilde{\beta}}V_b^{-1}m_b \\ &\quad + \left[r^{\mathrm{T}}R^{-1} + (F_0 - r^{\mathrm{T}}R^{-1}F)V_{\tilde{\beta}}F^{\mathrm{T}}R^{-1}\right] y, \\ \Sigma^* &= R_0 - r^{\mathrm{T}}R^{-1}r + (F_0 - r^{\mathrm{T}}R^{-1}F)(V_b^{-1} + V_{\hat{\beta}}^{-1})^{-1}(F_0 - r^{\mathrm{T}}R^{-1}F)^{\mathrm{T}}. \end{aligned}$$

The three components in the formula for the prediction variance Σ^* can be interpreted as the variability a priori, the reduction due to the conditioning on the data, and the increase due to uncertainty in the value of β, respectively.

It may be difficult to elicit informative priors in practice, and flat or non-informative improper priors might therefore be adopted. A non-informative prior often used in Bayesian analysis of linear models is $\pi(\beta, \sigma^2) \propto 1/\sigma^2$

(see for example, O'Hagan (1994)). Formal substitution of $V_b^{-1} = 0$ and $n_\sigma = 0$ into the formulas above for the posterior and predictive distributions gives the equivalent formulas for the non-informative prior, except that the degrees of freedom in the χ^2 posterior distribution and the multivariate-t predictive distribution are $n - p$ where p is the dimension of β, rather than n.

For the transformed Gaussian model, when $\lambda > 0$, we can back-transform predictions to the original scale using formulas for moments of the t-distribution, similar to the approach in Section 1.6. Note, however, that the exponential of a t-distribution does not have finite moments, hence when $\lambda = 0$ the minimum mean square error predictor does not exist. Prediction of non-linear targets is done using a procedure similar to the one in Section 1.6.3.

1.7.2 Uncertainty in the correlation parameters

More realistically, we now allow for uncertainty in all of the model parameters. We first consider the case of a model without measurement error, i.e. $\tau^2 = 0$ and a single correlation parameter ϕ. We adopt a prior $\pi(\beta, \sigma^2, \phi) = \pi(\beta, \sigma^2|\phi)\ \pi(\phi)$, the product of (1.5) and a proper density for ϕ. In principle a continuous prior $\pi(\phi)$ would be assigned. However, in practice we always use a discrete prior, obtained by discretising the distribution of ϕ in equal width intervals. The posterior distribution for the parameters is then given by

$$p(\beta, \sigma^2, \phi|y) \quad = \quad p(\beta, \sigma^2|y, \phi)\ p(\phi|y)$$

with $[\beta, \sigma^2|y, \phi]$ given by (1.6) and

$$p\,(\phi|y) \quad \propto \quad \pi(\phi)\ |V_{\tilde{\beta}}|^{\frac{1}{2}}\ |R|^{-\frac{1}{2}}\ (S^2)^{-\frac{n+n_\sigma}{2}}, \tag{1.9}$$

where $V_{\tilde{\beta}}$ and S^2 are given by (1.6) and (1.7) respectively. For the case where the prior is $\pi(\beta, \sigma^2, \phi) \propto \pi(\phi)/\sigma^2$, the equation above holds with $n_\sigma = -p$. Berger, De Oliveira & Sansó (2001) use a special case of this as a non-informative prior for the parameters of a spatial Gaussian process

To simulate samples from this posterior, we proceed as follows. We apply (1.9) to compute posterior probabilities $p(\phi|y)$ noting that in practice the support set will be discrete. We then simulate a value of ϕ from $[\phi|y]$, attach the sampled value to $[\beta, \sigma^2|y, \phi]$ and obtain a simulation from this distribution. By repeating the simulation as many times as required, we obtain a sample of triplets (β, σ^2, ϕ) from the joint posterior distribution of the model parameters.

The predictive distribution for the value, $S_0 = S(x_0)$ say, of the signal process at an arbitrary location x_0 is given by

$$\begin{aligned} p(s_0|y) &= \int\int\int p(s_0, \beta, \sigma^2, \phi|y)\, d\beta\, d\sigma^2\, d\phi \\ &= \int\int\int p\big(s_0, \beta, \sigma^2|y, \phi\big)\ d\beta\, d\sigma^2\, p(\phi|y)\ d\phi \\ &= \int p(s_0|y, \phi)\ p(\phi|y)\, d\phi. \end{aligned}$$

The discrete prior for ϕ allows analytic calculation of the moments of this predictive distribution. For each value of ϕ we compute the moments of the multivariate-t distribution given by (1.8) and calculate their weighted sum with weights given by the probabilities $p(\phi|y)$.

To sample from this predictive distribution, we proceed as follows. We compute the posterior probabilities $p(\phi|y)$ on the discrete support set for $[\phi]$, and simulate values of ϕ from $[\phi|y]$. Attaching a sampled value of ϕ to $[S_0|y, \phi]$ and simulating from this distribution we obtain a realisation from the predictive distribution.

Finally, when $\tau^2 > 0$, in practice we use a discrete joint prior $[\phi, \nu^2]$, where $\nu^2 = \tau^2/\sigma^2$. This adds to the computational load, but introduces no new principles. Similarly, if we wish to incorporate additional parameters in the covariance structure of $S(\cdot)$, we would again use a discretisation method to render the computations feasible.

In principle, the prior distributions for the parameters should reflect scientific prior knowledge. In practice, we will often be using the Bayesian framework pragmatically, under a vague prior specification. However, a word of caution is necessary here, as we have found that even apparently vague prior specifications can materially affect the corresponding posteriors. It seems to be a general feature of geostatistical problems that the models are poorly identified, in the sense that widely different combinations of parameter values lead to very similar fits. This may not matter if parameter estimates themselves, as opposed to the prediction target T, are not of direct interest. Also, the Bayesian paradigm at least brings this difficulty into the open, whereas plugging in more or less arbitrary point estimates merely hides the problem.

1.8 A Case Study: the Swiss rainfall data

In this case study, we follow convention by using only the first 100 of the data-locations in the Swiss rainfall data for model formulation. We consider a transformed Gaussian model, with a Matérn correlation structure.

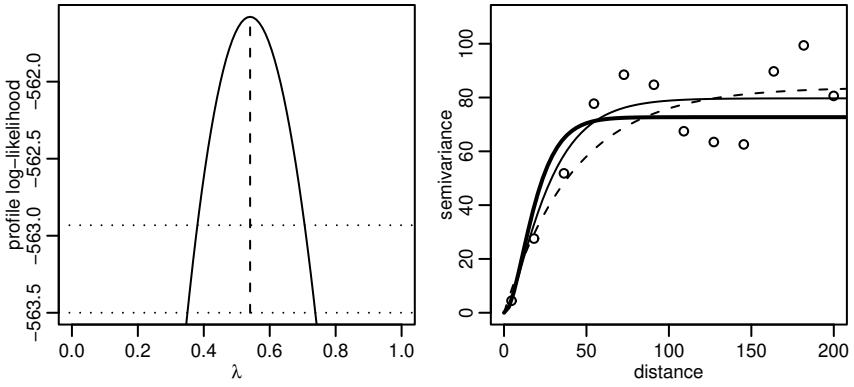


FIGURE 1.11. Left panel: profile likelihood for transformation parameter λ for the model with Matérn correlation function with $\kappa = 1$. Right panel: Estimated variograms for transformed ($\lambda = 0.5$) data (open circles), compared with the theoretical Matérn model with parameters equal to the maximum likelihood estimates. The three fits correspond to $\kappa = 0.5$ (dashed line), $\kappa = 1$ (thick solid line), $\kappa = 2$ (thin solid line).

κ	$\hat{\lambda}$	$\log \hat{L}$
0.5	0.496	-564.857
1	0.540	-561.579
2	0.561	-563.115

TABLE 1.1. Maximum likelihood estimates $\hat{\lambda}$ and the corresponding values of the log-likelihood function $\log \hat{L}$ for the Swiss rainfall data, assuming different values of the Matérn shape parameter κ.

Table 1.1 shows the maximum likelihood estimates of the Box-Cox transformation parameter λ, holding the Matérn shape parameter κ fixed at each of the three values $\kappa = 0.5, 1, 2$. The consistent message is that $\lambda = 0.5$, or a square root transformation, is a reasonable choice. The profile log-likelihood for λ shown in the left-hand panel of Figure 1.11 indicates that neither the log-transformation $\lambda = 0$, nor an untransformed Gaussian assumption ($\lambda = 1$) is tenable for these data. The right-hand panel of Figure 1.11 shows the empirical and fitted variograms, for each of $\kappa = 0.5, 1, 2$. Visually, there is little to choose amongst the three fits.

Table 1.2 shows maximum likelihood estimates for the model with $\lambda = 0.5$. The overall conclusion is that $\kappa = 1$ gives a better fit than $\kappa = 0.5$ and $\kappa = 2$. Furthermore, in each case $\hat{\tau}^2 = 0$. Figure 1.12 shows the profile log-likelihoods of the two covariance parameters σ^2, ϕ holding κ, λ and τ^2 fixed at these values. Note in particular the wide, and asymmetric, confidence intervals for the signal variance σ^2 and the range parameter ϕ. These serve to warn against over-interpretation of the corresponding point estimates.

κ	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	$\log \hat{L}$
0.5	21.205	83.865	42.388	0	-564.858
1.0	22.426	79.694	17.583	0	-561.664
2.0	23.099	72.698	8.358	0	-563.292

TABLE 1.2. Maximum likelihood estimates $\hat{\beta}$, $\hat{\phi}$, $\hat{\sigma}$, $\hat{\tau}^2$ and the corresponding value of the likelihood function $\log \hat{L}$ for the Swiss rainfall data, assuming different values of the Matérn parameter κ, and transformation parameter $\lambda = 0.5$.

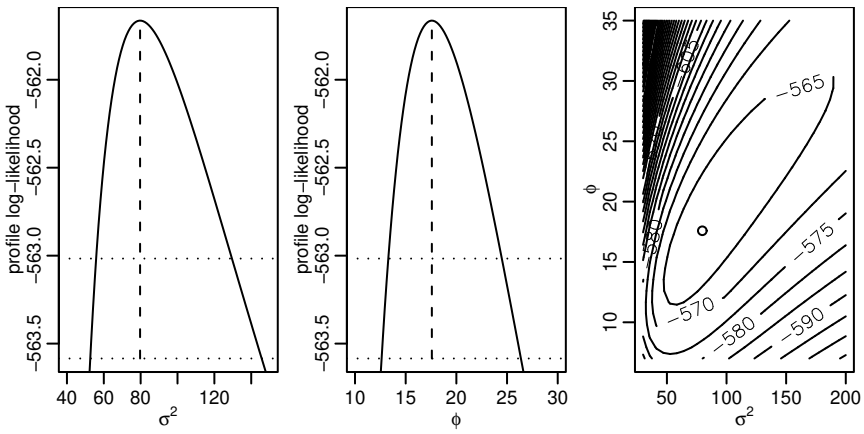


FIGURE 1.12. Profile likelihood for covariance parameters in the Matérn model fitted to the Swiss rainfall data with $\kappa = 1$ and $\lambda = 0.5$. Left panel σ^2, middle panel ϕ, right panel the 2-D profile likelihood.

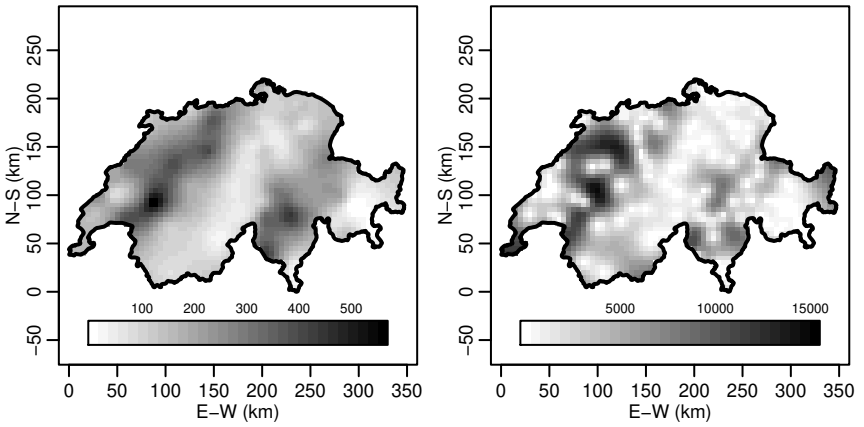


FIGURE 1.13. Maps of predictions (left panel) and prediction variances (right panel) for the Swiss rainfall data.

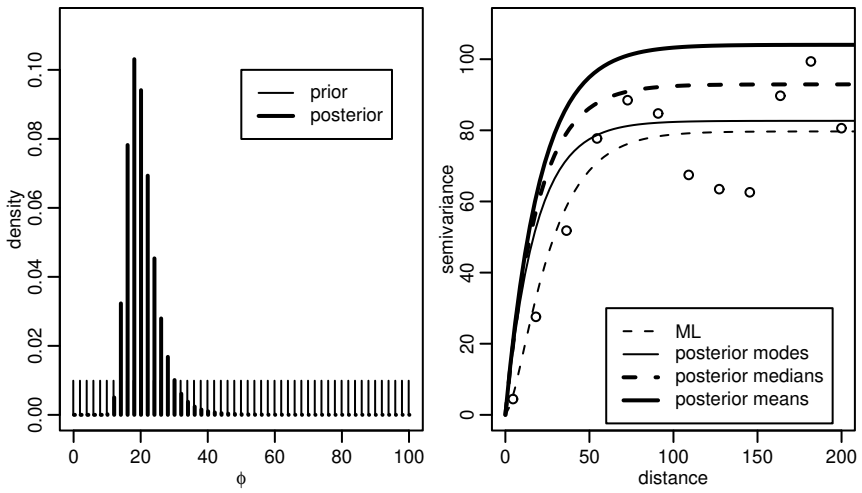


FIGURE 1.14. Left panel: uniform prior and corresponding posterior distribution. Right panel: variograms based on summaries of the posterior and on the ML estimator.

Figure 1.13 maps the point predictions of rainfall values and associated prediction variances from a plug-in prediction using the transformed Gaussian model with $\lambda = 0.5$ and $\kappa = 1$. The grid spacing for prediction corresponds to a distance of 5 km between adjacent prediction locations. The values of the prediction variances shows a positive association with predicted values, as a consequence of the transformation adopted; recall that in the untransformed Gaussian model, the prediction variance depends only on the model parameters and the study design, and not directly on the measured values.

We now turn to the Bayesian analysis, adopting the prior $\pi(\beta, \sigma^2|\phi) \propto 1/\sigma^2$ and a discrete uniform prior for ϕ with 101 points equally spaced in the interval $[0; 100]$. The posterior distribution for ϕ is then obtained by computing (1.9) for each discrete value, and standardising such that the probabilities add to one. The left-hand panel of Figure 1.14 shows the uniform prior adopted and the posterior distribution obtained for this data-set. The right-hand panel of Figure 1.14 displays variograms based on different summaries of the posterior $[\sigma^2, \phi|y]$ and on the ML estimates $(\hat{\sigma}^2, \hat{\phi})$. The differences between the Bayesian estimates reflect the asymmetry in the posterior distributions of ϕ and σ^2. Note that in all three cases the Bayesian estimate of σ^2 is greater than the ML estimate.

Values of the parameters ϕ and σ^2 sampled from the posterior are displayed by the histograms in the left and centre panels of Figure 1.15. The right-hand panel of Figure 1.15 shows that there is a strong correlation in the posterior, despite the fact that priors for these two parameters are independent. This echoes the shape of the two-dimensional profile likelihood

shown earlier in Figure 1.12. Similar results were obtained for other choices of prior.

To predict the values of rainfall in a grid of points over Switzerland we can compute moments analytically, as described in Section 1.7.1 and Section 1.7.2. Figure 1.16 shows a comparison between "plug-in" and Bayesian point predictions (left panel) and their standard errors (right panel). The strong concentration of points along the diagonal in the left-hand panel of Figure 1.16 shows that, for this particular example, the Bayesian point predictions do not differ too much from the "plug-in" predictions. However, as indicated in the right-hand panel of Figure 1.16 there are differences in the estimated uncertainty associated with the predictions, with the plug-in variances tending to slightly under-estimate the variance of the predictive distribution, especially where the prediction variance itself is relatively large.

Inferences about non-linear functionals can be performed by sampling from the predictive distribution and processing the sampled values according to the functional of interest. This generates a sample from the posterior distribution of the target for prediction. As an example, consider inference for the target $T_{max} = max\{Y(x) : x \in A\}$, the maximum rainfall over the whole of Switzerland. In practice we redefined T_{max} to be the maximum over the 5 km spaced grid. Taking 2000 simulations from the predictive distribution and computing the maximum for each simulation we find values in the interval $[531, 1114]$ with a mean of 667.4 and standard deviation of 73.9. Simulations from the "plug-in" predictive distribution generated with the same seed for the random number generator showed a mean of 655.8 and standard deviation of 67.4. So for this prediction target the Bayesian prediction is larger than the plug-in prediction. Also, the Bayesian predic-

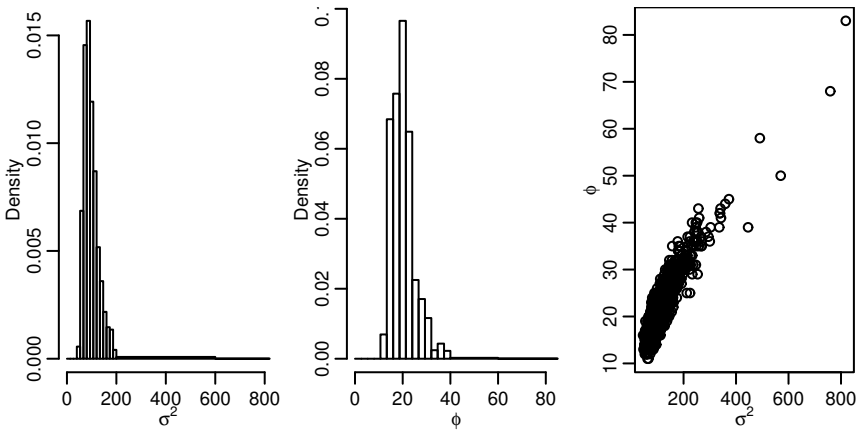


FIGURE 1.15. Histograms for samples from the posteriors for the parameters σ^2 (left) and ϕ (middle), and corresponding scatterplot.

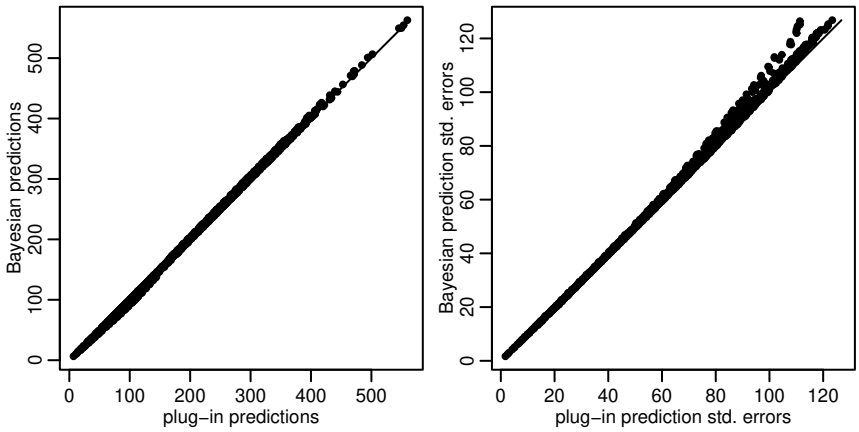


FIGURE 1.16. Comparing "plug-in" and Bayesian predicted values (left) in a 5 km spaced grid over the area, and associated standard errors (right).

tion standard error is larger than the plug-in prediction standard error, which is often seen in practice, but is not always the case.

From our experience with a variety of real and simulated data-sets, we consider this particular data-set to be an exceptionally well behaved one. The profile likelihoods are sharp and not too wide. No extra residual variation was found after fitting the spatial part of the model. The results were insensitive to different choices of prior for ϕ. However, in our experience this situation is somewhat atypical. Rather, noisy data are common and inferences tend to have greater associated uncertainty than in this example. In these situations, the discrepancy between Bayesian and plug-in methods becomes more pronounced.

In the Bayesian analysis reported here we have used vague priors. Ideally, more informative priors relevant to the problem at hand should be considered, although elicitation of such priors is often a difficult task.

1.9 Generalised linear spatial models

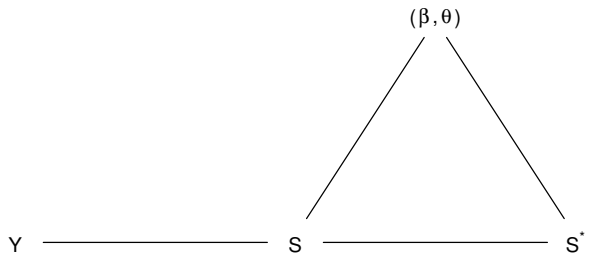
The classical generalised linear model (GLM) is defined for a set of mutually independent responses $Y_1, ..., Y_n$. The expectations $\mu_i = \mathrm{E}[Y_i]$ are specified by a *linear predictor* $h(\mu_i) = \sum_{j=1}^{k} f_{ij}\beta_j$, in which $h(\cdot)$ is a known function, called the *link* function (McCullagh & Nelder 1989). An important extension of this basic class of models is the *generalised linear mixed model* or GLMM (Breslow & Clayton 1993), in which $Y_1, \ldots, Y_n$ are mutually independent conditional on the realised values of a set of latent random variables $U_1, \ldots, U_n$, and the conditional expectations are given

by $h(\mu_i) = U_i + \sum_{j=1}^{k} f_{ij}\beta_j$. A *generalised linear spatial model* (GLSM) is a GLMM in which the $U_1, \ldots, U_n$ are derived from a spatial process $S(\cdot)$. This leads to the following model-specification.

Let $S(\cdot) = \{S(x) : x \in A\}$ be a Gaussian stochastic process with $\mathrm{E}[S(x)] = \sum_{j=1}^{p} f_j(x)\beta_j$, $\mathrm{Var}[S(x)] = \sigma^2$ and $\rho(u) = \mathrm{Corr}[S(x), S(x')]$ where $u = \|x - x'\|$. Assume that measurements $Y_1, \ldots, Y_n$ are conditionally independent given $S(\cdot)$, with conditional expectations μ_i and $h(\mu_i) = S(x_i), i = 1, \ldots, n$, for a known link function $h(\cdot)$. In this model the signal process is $\{h^{-1}(S(x)) : x \in A\}$.

As in the case of the classical GLM, the GLSM embraces the linear Gaussian model as a special case, whilst providing a natural extension to deal with response variables for which a standard distribution other than the Gaussian more accurately describes the sampling mechanism involved. In what follows, we focus on the Poisson-log-linear model for count data and the logistic model for binomial data.

We denote the regression parameters by β and covariance parameters in the model by θ. We write $Y = (Y_1, ..., Y_n)^{\mathrm{T}}$ for the observed responses at locations $x_1, \ldots, x_n$ in the sampling design, $S = (S(x_1), ..., S(x_n))^{\mathrm{T}}$ for the unobserved values of the underlying process at $x_1, \ldots, x_n$, and S^* for the values of $S(\cdot)$ at all other locations of interest, typically a fine grid of locations covering the study region. The conditional independence structure of the GLSM is then indicated by the following graph.



The likelihood for a model of this kind is in general not expressible in closed form, but only as a high-dimensional integral

$$L(\beta, \theta) = \int \prod_{i=1}^{n} g(y_i; h^{-1}(s_i)) p(s; \beta, \theta) ds_1, \ldots, s_n, \qquad (1.10)$$

where $g(y; \mu)$ denotes the density of the error distribution parameterised by

the mean μ, and $p(s;\beta,\theta)$ is the multivariate Gaussian density for the vector S. The integral above is also the normalising constant in the conditional distribution of $[S|y,\beta,\theta]$,

$$p(s \mid y,\beta,\theta) \propto \prod_{i=1}^{n} g(y_i; h^{-1}(s_i))p(s;\beta,\theta). \tag{1.11}$$

In practice, the high dimensionality of the integral prevents direct calculation of the predictive distribution $[S^* \mid y,\beta,\theta]$.

Standard methods of approximating the integral (1.10) and hence evaluating (1.11) are of unknown accuracy in the geostatistical setting, but Markov chain Monte Carlo methods (see Chapter 1) provide a possible solution.

1.9.1 Prediction in a GLSM

Assume first that the parameters in the model are known. From the figure with the graphical model above we see that prediction of $T = T(S^*)$ can be separated into three steps.

- Simulate $s(1),\ldots,s(m)$ from $[S|y]$ (using MCMC).
- Simulate $s^*(j)$ from $[S^*|s(j)]$, $j = 1,\ldots,m$ (multivariate Gaussian).
- Approximate the minimum mean square error predictor

$$\mathrm{E}[T(S^*)|y] \approx \frac{1}{m}\sum_{j=1}^{m} T(s^*(j)).$$

Whenever possible, it is desirable to replace Monte Carlo sampling by direct evaluation. For example, if it is possible to calculate $\mathrm{E}[T(S^*)|s(j)]$, $j = 1,\ldots,m$ directly, we would use the approximation

$$\mathrm{E}[T(S^*)|y] \approx \frac{1}{m}\sum_{j=1}^{m} \mathrm{E}[T(S^*)|s(j)],$$

thereby reducing the Monte Carlo error due to simulation.

To simulate from $[S \mid y]$ we use the truncated Langevin-Hastings algorithm as in Christensen, Møller & Waagepetersen (2001). This algorithm uses gradient information in the proposal distribution and has been found to work well in practice by comparison with a random walk Metropolis algorithm. First we make a reparametrisation defining $S = F^{\mathrm{T}}\beta + \Omega^{1/2}\Gamma$ where $\Omega^{1/2}$ is a square root of $\Omega = \mathrm{Var}[S]$, say a Cholesky factorisation, and a priori $\Gamma \sim N(0,I)$. Using an MCMC-algorithm to obtain a sample $\gamma_1,\ldots,\gamma_n$

from $[\Gamma \mid y]$, we multiply by $\Omega^{1/2}$ and obtain a sample $s(1), \ldots, s(m)$ from $[S|y]$.

The MCMC-algorithm used is a Metropolis-Hastings algorithm where all components of Γ are updated simultaneously. The proposal distribution is a multivariate Gaussian distribution with mean $m(\gamma) = \gamma + (\delta/2)\nabla(\gamma)$ where $\nabla(\gamma) = \frac{\partial}{\partial\gamma} \log f(\gamma \mid y)$, and variance δI_n. For a GLSM with canonical link function h, the gradient $\nabla(\gamma)$ has the following simple form:

$$\nabla(\gamma) = \frac{\partial}{\partial\gamma} \log f(\gamma \mid y) = -\gamma + (\Omega^{1/2})^{\mathrm{T}}\{y - h^{-1}(s)\}, \tag{1.12}$$

where $s = F^{\mathrm{T}}\beta + \Omega^{1/2}\gamma$ and h^{-1} is applied coordinatewise. If we modify the gradient $\nabla(\gamma)$ (by truncating, say) such that the term $\{y - h^{-1}(s)\}$ is bounded, the algorithm can be shown to be geometrically ergodic, and a Central Limit Theorem therefore exits. The Central Limit Theorem with asymptotic variance estimated by Geyer's monotone sequence estimate (Geyer 1992), can be used to assess the Monte Carlo error of the calculated prediction. This algorithm is not specific to the canonical case since the formula in (1.12) can be generalised to accommodate models with a non-canonical link function.

In practice one has to choose the proposal variance δ. We tune the algorithm by running a few test runs and choosing δ such that approximately 60% of the proposals are accepted. To avoid storing a large number of high-dimensional simulations $s(1), \ldots, s(m)$ we also thin the sample such that, say, only every 100th simulation is stored.

1.9.2 Bayesian inference for a GLSM

First we consider Bayesian inference for a GLSM, using the Gaussian-Scaled-Inverse-χ^2 prior for (β, σ^2) defined in (1.5), holding ϕ fixed. The marginal density of S, obtained by integrating over β and σ^2, becomes an n-dimensional multivariate-t density, $t_{n_\sigma}(m_b, S_\sigma^2(R + FV_bF^{\mathrm{T}}))$. Therefore the posterior density of S is

$$p(s \mid y) \propto \prod_{i=1}^{n} g(y_i; h^{-1}(s_i))p(s) \tag{1.13}$$

where $p(s)$ is the marginal density of S.

In order to obtain a sample $s(1), \ldots, s(m)$ from this distribution we use a Langevin-Hastings algorithm, the reparametrisation $S = F^{\mathrm{T}}m_b + S_\sigma(R + FV_bF^{\mathrm{T}})\Omega^{1/2}\Gamma$, where $\Omega = S_\sigma^2(R + FV_bF^{\mathrm{T}})$, and a priori $\Gamma \sim t_{n+n_\sigma}(0, I_n)$. The gradient $\nabla(\gamma)$ which determines the mean of the proposal distribution

has the following form when h is the canonical link function,

$$\nabla(\gamma) = \frac{\partial}{\partial\gamma}\log f(\gamma \mid y) = -\gamma(n+n_\sigma)/(n_\sigma + \|\gamma\|^2) + (\Omega^{1/2})^{\mathrm{T}}\{y - h^{-1}(s)\}. \tag{1.14}$$

By using a conjugate prior for (β, σ^2) we find that $[\beta, \sigma^2 \mid s(j)]$, $j = 1, \ldots, m$ are Gaussian-Scaled-Inverse-χ^2 distributions with means and variances given by (1.6). From this we can calculate the mean and the variance of the posterior $[\beta, \sigma^2 \mid y]$, and also simulate from it.

Procedures similar to the ones given in Section 1.9.1 can be used for prediction. The only difference is that from (1.8), we see that $[S^*|s(j)]$, $j = 1, \ldots, m$, are now multivariate-t distributed rather than multivariate Gaussian.

Concerning the use of flat or non-informative priors for β and σ^2 in a GLSM, a word of caution is needed. The prior $1/\sigma^2$ for σ^2, recommended as a non-informative prior for the Bayesian linear Gaussian model in Section 1.7, results in an improper posterior distribution for a GLMM (see Natarajan & Kass (2000)), and should therefore be avoided. Since a linear Gaussian model with a fixed positive measurement error $\tau_0^2 > 0$ can be considered as a special case of a GLSM, this is also true for such a model. There seems to be no consensus concerning reference priors for GLMM's.

We now allow for uncertainty also in ϕ, and adopting as our prior $\pi(\beta, \sigma^2, \phi) = \pi_{\{N\chi^2_{ScI}\}}(\beta, \sigma^2)\pi(\phi)$, where $\pi(\phi)$ is any proper prior.

When using an MCMC-algorithm updating ϕ, we need to calculate $(R(\phi) + FV_bF^{\mathrm{T}})^{1/2}$ for each new ϕ value, which is the most time-consuming part of the algorithm. To avoid this significant increase in computation time, we adopt a discrete prior for ϕ on a set of values covering the range of interest, and precompute and store $(R(\phi) + FV_bF^{\mathrm{T}})^{1/2}$ for each value of ϕ.

To simulate from $[S, \phi|y]$, after integrating out β and σ^2, we use a hybrid Metropolis-Hastings algorithm where S and ϕ are updated sequentially. The update of S is of the same type as used earlier, with ϕ equal to the present value in the MCMC iteration. To update ϕ we use a random walk Metropolis update where the proposal distribution is a Gaussian distribution rounded to the nearest ϕ value in the discrete set for the prior. The output of this algorithm is a sample $(s(1), \phi(1)), \ldots, (s(m), \phi(m))$ from the distribution $[S, \phi \mid y]$.

The predictive distribution for S^* is given by

$$p(s^*|y) = \int\int p(s^* \mid s, \phi)p(s, \phi \mid y)dsd\phi$$

To simulate from this predictive distribution, we simulate $s^*(j)$ from $[S^* \mid s(j), \phi(j)]$, which is multivariate-t, $j = 1, \ldots, m$.

We may also want to introduce a nugget term into the specification of the model, replacing $S(x_i)$ by $S(x_i)+U_i$ where the U_i are mutually independent Gaussian variates with mean zero and variance τ^2 . Here, in contrast to the Gaussian case, we can make a formal distinction between the U_i as a representation of micro-scale variation and the error distribution induced by the sampling mechanism, for example Poisson for count data. In some contexts, the U_i may have a more specific interpretation. For example, if a binary response were obtained from each of a number of sampling units at each of a number of locations, a binomial error distribution would be a natural choice, and the U_i and $S(x_i)$ would then represent, respectively, non-spatial and spatial sources of extra-binomial variation. The inferential procedure is essentially unchanged, except that we now use a discrete joint prior for (ϕ, τ^2).

1.9.3 A spatial model for count data

A GLSM for modelling spatial count data is the Poisson-log-linear spatial model, in which $[Y_i \mid S(x_i)]$ follows a Poisson distribution with mean $t_i \exp(S(x_i))$, $i = 1, \ldots, n$. The term t_i may, for example, represent a time-interval over which the corresponding count Y_i is accumulated, as in Diggle et al. (1998), or an area within which the number of events Y_i is counted, as in Christensen & Waagepetersen (2002).

We assume initially that parameters are known, and that we are interested in predicting the intensity $\lambda(x_0) = \exp(S(x_0))$ at a location x_0. Given a sample $s(1), \ldots, s(m)$ from $[S|y]$, obtained using the MCMC-algorithm in Section 1.9.1, $[S(x_0)|s(j)], j = 1, ..., m$ follow multivariate Gaussian distributions. Since the moments of the exponential of a multivariate Gaussian distribution are obtainable in closed form, the following procedure can be used for predicting $\lambda(x_0) = \exp(S(x_0))$.

- Calculate $\mathrm{E}[S(x_0)|s(j)]$ and $\mathrm{Var}[S(x_0)|s(j)]$, $j = 1, \ldots, m$, using kriging.
- Calculate, for each of $j = 1, \ldots, m$,

$$\mathrm{E}[\lambda(x_0)|s(j)] = \exp(\mathrm{E}[S(x_0)|s(j)] + 0.5\mathrm{Var}[S(x_0)|s(j)])$$

- Approximate

$$\mathrm{E}[\lambda(x_0)|y] \approx \frac{1}{m} \sum_{j=0}^{m} \mathrm{E}[\lambda(x_0)|s(j)]$$

Note that $\mathrm{E}[\exp(\alpha S)|y]$ is finite for any $\alpha \in \mathbb{R}^n$, $\mathrm{E}[S(x_0)|S]$ is a linear function of S, and $\mathrm{Var}[S(x_0)|S]$ does not depend on S. Therefore $\mathrm{E}[\lambda(x_0)|y] < \infty$, and the quantity we want to approximate using MCMC

exists. As we shall see below, if we use only simulation-based methods we may unwittingly produce estimates of quantities that do not exist.

An algorithm similar to one above could, in principle, be used when we want to incorporate prior information into the predictions by using the conjugate Gaussian-Scaled-Inverse-χ^2 prior for (β, σ^2), the difference being that $[S(x_0) \mid s(j)]$, $j = 1, \ldots, m$ are now multivariate-t distributions. However, because the mean of the exponential of a multivariate-t distribution is not finite, the procedure fails. In fact, the minimum mean square error predictor does not exist in this case. Had we used a different MCMC-algorithm, sampling β and σ^2 instead of integrating them out, or had we decided to generate a sample $\exp(s_0(1)), \ldots, \exp(s_0(m))$ instead of using the formula for $\mathrm{E}[\exp(S(x_0))|s(j)]$, $j = 1, \ldots, m$, this problem might have been missed. This method would, of course, have generated a valid sample from the required predictive distribution. If we do want to quote a point prediction in a situation of this kind, we might for example use the predictive median rather than the mean.

1.9.4 Spatial model for binomial data

A GLSM for binomial data is as follows. The data are arranged as triples, (x_i, y_i, n_i), where y_i is a count of the number of successes out of n_i Bernoulli trials associated with the location x_i. Conditional on an unobserved Gaussian process $S(\cdot)$, we model the y_i as realisations of mutually independent binomial random variables with numbers of trials n_i and success probabilities $p_i = p(x_i)$, where

$$\log\{p(x)/(1 - p(x))\} = S(x). \tag{1.15}$$

As before, the process $S(\cdot)$ has spatially varying mean $\mu(x) = \sum f_j(x)\beta_j$, variance σ^2 and correlation parameter ϕ.

To illustrate the prediction problem in this context, suppose that the target for prediction is $T = p(x_0)$. Because no closed form expressions can be found for the mean and variance of $[T \mid S]$ we need to simulate from this distribution. Assuming a Gaussian-Scaled-Inverse-χ^2 prior for (β, σ^2), and a proper prior for ϕ, we proceed as follows:

- simulate $((s(1), \phi(1)), \ldots, (s(m), \phi(m))$ from $[S, \phi|y]$, using MCMC;
- calculate $\mathrm{E}[S(x_0)|s(j), \phi(j)]$ and $\mathrm{Var}[S(x_0)|s(j), \phi(j)]$ for each of $j = 1, \ldots, m$;
- simulate values $s_0(j), j = 1, ..., m$ from multivariate-t distributions with common degrees of freedom $n + n_\sigma$, means $\mathrm{E}[S(x_0)|s(j)]$ and variances $\mathrm{Var}[S(x_0)|s(j)]$;

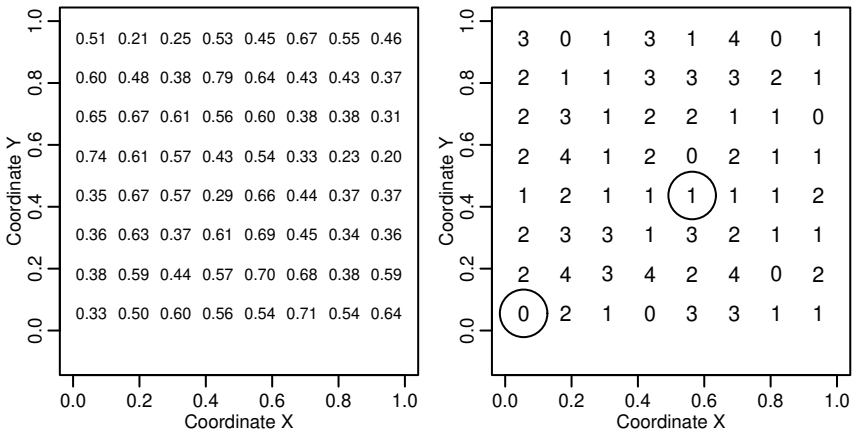


FIGURE 1.17. Map of simulated binomial data. Left: the values of underlying process $S(\cdot)$. Right: binomial data with size 4 and probability parameter equal to the inverse logit-function; circles indicates locations for which MCMC traces will be shown.

- approximate

$$\mathrm{E}[T|y] \approx \frac{1}{m}\sum_{j=0}^{m} \exp(s_0(j))/(1+\exp(s_0(j))).$$

Heagerty & Lele (1998) and De Oliveira (2000) use an apparently different model for spatial binary data which they call the clipped Gaussian field. In this model, the measurement process is $\{Y(x) = 1_{\{S(x)>0\}} : x \in A\}$, where $S(\cdot)$ is a Gaussian process. Assuming that the process $S(\cdot)$ has a positive nugget τ^2, we can write this model as $Y(x) = 1_{\{\tilde{S}(x)+U(x)>0\}}$, where $\tilde{S}(\cdot)$ is another Gaussian process and $U(\cdot)$ is a Gaussian white noise process with mean 0 and variance 1. The conditional distribution $[Y(x) \mid \tilde{S}(x)]$ is binomial of size 1 and probability $P(Y(x) = 1 \mid \tilde{S}(x)) = \Phi(\tilde{S}(x))$. The model is therefore identical to the one described above, except that the logit link in (1.15) is replaced by the probit link.

1.9.5 Example

To illustrate the inferential procedure in a GLSM we consider the simulated data-set shown in Figure 1.17 which consists of binomial data at 64 locations. The left-hand panel shows the values of the underlying Gaussian random variables $S(x_1), \ldots, S(x_n)$ simulated from a model with exponential correlation function and parameter values $(\beta, \sigma^2, \phi) = (0, 0.5, 0.2)$. The right-hand plot shows the corresponding binomial variables which are simulated from independent binomial distributions of size 4 and probability

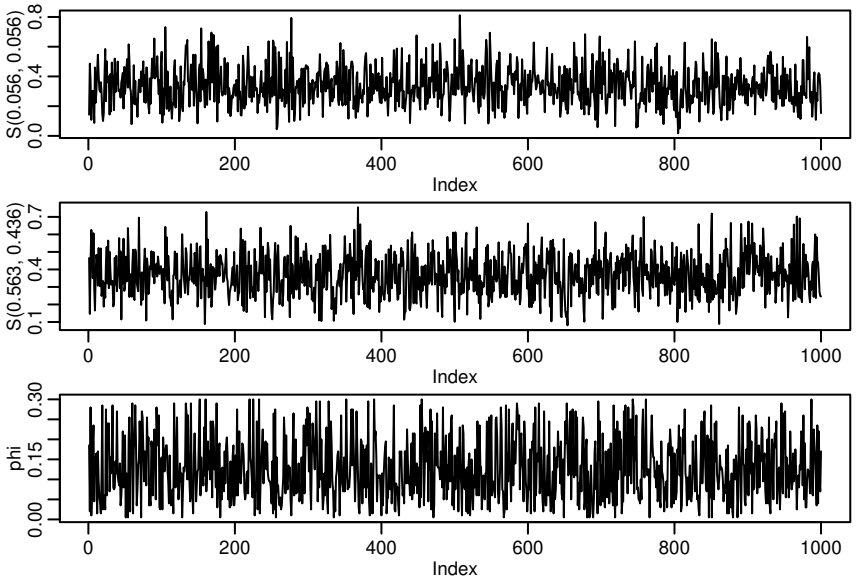


FIGURE 1.18. Time series with the MCMC output for the two locations indicated by circles in Figure 1.17 and for the parameter ϕ.

parameter $p(x_i) = \exp(S(x_i))/(1 + \exp(S(x_i)))$ $i = 1, \ldots, n$, the inverse logit transform of $S(\cdot)$. Note that inference is based on the observed binomial values in the right hand plot, with $S(x_1), \ldots, S(x_n)$ in the left hand plot considered as unobserved.

We perform a Bayesian analysis with exponential correlation function and the following priors: a Gaussian $N(0, \sigma^2)$ prior for $[\beta \mid \sigma^2]$, a $\chi^2_{ScI}(5, 0.5)$ prior for σ^2, and a discrete exponential prior for ϕ, $\pi(\phi) = \exp(-\phi/0.2)$, with 60 discretisation points in the interval $[0.005, 0.3]$.

For inference we run the MCMC-algorithm described in Section 1.9.2, discarding the first $10,000$ iterations then retaining every 100th of $100,000$ iterations to obtain a sample of size 1000. Figure 1.18 shows the output for the two S coordinates circled in Figure 1.17, and for the parameter ϕ. The estimated autocorrelations are in each case less than 0.1 for all positive lags, and the thinned sample has very low autocorrelation.

For prediction of the probabilities over the area, we consider 1600 locations in a regular square grid and use the procedure described in Section 1.9.4. The left-hand panel of Figure 1.19 shows the predictions at the 1600 locations, whilst the right-hand panel shows the associated prediction variances.

Comparing the left-hand panels of Figure 1.17 and Figure 1.19 we see that the predicted surface has less extreme values than does the true surface $S(\cdot)$, as is to be expected when predicting from noisy data. The prediction variances on the right-hand plot in Figure 1.19 show a weak dependence

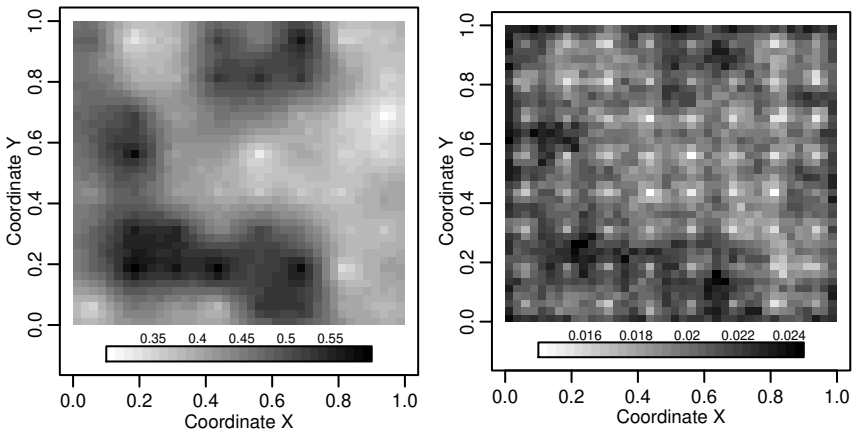


FIGURE 1.19. Left: predicted values at the grid points. Right: prediction variances.

on the means, with a preponderance of large values of the prediction variance in areas where the predicted means are close to 0.5. The effect of the sampling design is also clear, with small prediction variances at locations close to grid-points.

1.10 Discussion

In this short introduction to the subject of model-based geostatistics, our aim has been to set out the basic methodology for dealing with geostatistical problems from the perspective of mainstream parametric statistical modelling. Under Gaussian assumptions, the resulting prediction methodology has a very close relationship to classical geostatistical kriging methods, but the treatment of unknown parameters departs markedly from the classical geostatistical approach. The classical approach uses curve-fitting methods to match empirical and theoretical variograms, whereas we have emphasised the use of the likelihood function for parameter estimation, whether from a Bayesian or non-Bayesian point of view. The Bayesian approach has the attractive property that uncertainty in the values of model parameters is recognised in the construction of prediction intervals, leading to a more honest assessment of prediction error. This should not blind us to the uncomfortable fact that even the simplest geostatistical models may be poorly identifiable from the available data, and in these situations the choice of prior may have an unpleasantly strong influence on the resulting inferences. However, our suspicion is that more ad hoc methods based on simple plug-in methods conceal, rather than solve, this difficulty.

One general point that cannot be over-emphasised is that the model and inferential method used on a particular set of data should be informed by the scientific objective of the data analysis. The full machinery of geostatistical modelling is indicated when prediction at unobserved spatial locations is a central objective. For problems of this kind, some or all of the model parameters, for example those defining the assumed spatial covariance structure of the data, are means to an end rather than quantities to be interpreted in their own right, and it may not matter that these parameters are poorly identified. For problems in which the scientific focus is on parameter estimation, for example in investigating the regression effects of explanatory variables, a simpler approach such as the method of generalised estimating equations may be all that is required (Gotway & Stroup 1997). Note, however that this changes the interpretation of the regression parameter as affecting marginal, rather than conditional expectations.

We acknowledge that our use of the Bayesian inferential paradigm is pragmatic. We find it difficult to come up with convincing arguments for the choice of priors in a geostatistical model, but we do want a prediction methodology which acknowledges all of the major sources of uncertainty in our predictions.

We have omitted altogether a number of topics due to restrictions on space. Within the linear Gaussian setting, extensions of the methodology to multivariate and/or space-time data are straightforward in principle, although the enriched data-structure leads to a proliferation of modelling choices. Some examples are included in the suggestions given in Section 1.12 for further reading. Also, for large space-time data-sets, apparently obvious approaches may be computationally infeasible.

Outside the generalised linear model setting, the number of potential models is practically limitless and the ability to fit by Monte Carlo methods almost arbitrarily complex models is a two-edged sword. On one hand, it is right in principle that models should be informed by scientific knowledge, rather than chosen from an artificially restricted class of analytically or numerically tractable models. Against this, it is all too easy to devise a model whose complexity far outstrips the capacity of the data to provide reliable validation of its underlying assumptions.

A fundamental problem which is ignored in most geostatistical work is the possibility of stochastic interaction between the signal or measurement process and the sampling design. For example, in mineral exploration samples will be taken from locations which are thought likely to yield commercially viable grades of ore. The formal framework for handling problems of this kind is a *marked point process*, a joint probability model for a stochastic *point process* X, and an associated set of random variables, or *marks*, Y. As always, different factorisations of the joint distribution are available, and whilst these factorisations are mathematically equivalent, in practice they

lead to different modelling assumptions. The simplest structural assumption is that X and Y are independent processes, hence $[X, Y] = [X][Y]$. This is sometimes called the *random field model*, and is often assumed implicitly in geostatistical work. In the dependent case, one possible factorisation is $[X, Y] = [X|Y][Y]$. This is a natural way to describe what is sometimes called *preferential sampling*, in which sampling locations are determined by partial knowledge of the underlying mark process; an example would be the deliberate siting of air pollution monitors in badly polluted areas. The opposite factorisation, $[X, Y] = [X][Y|X]$, may be more appropriate when the mark process is only defined at the sampling locations; for example, the heights of trees in a naturally regenerated forest. The full inferential implications of ignoring violations of the random field model have not been widely studied.

1.11 Software

All the analyses reported in this chapter have been carried out using the packages `geoR` and `geoRglm`, both of which are add-on's to the freely available and open-source statistical system `R` (Ihaka & Gentleman 1996). The official web site of the `R`-project is at *www.r-project.org*. Both packages are available in the contributed section of `CRAN` (Comprehensive R Archive Network).

The package `geoR` (Ribeiro Jr & Diggle 2001) implements basic geostatistical tools and the methods for Gaussian linear models described here. Its official web site is *www.maths.lancs.ac.uk/~ribeiro/geoR*, where instructions for downloading and installation can be found, together with a tutorial on the package usage.

The package `geoRglm` is an extension of `geoR` which implements the generalised linear spatial model described in Section 1.9. It is available at *www.lancaster.ac.uk/~christen/geoRglm*, together with an introduction to the package.

Other computational resources for analysis of spatial data using `R` are reviewed in issues 2 and 3 of *R-NEWS*, available at *cran.r-project.org/doc/Rnews*. An extensive collection of geostatistics materials can be found in the AI-GEOSTATS web site at *www.ai-geostats.org*.

1.12 Further reading

Chilés & Delfiner (1999) is a standard reference for classical geostatistical methods. Cressie (1993) describes geostatistics as one of three main

branches of spatial statistics. Stein (1999) gives a rigorous account of the mathematical theory underlying linear kriging. Ribeiro Jr & Diggle (1999) give a more detailed presentation of Bayesian inference for the linear Gaussian model. Other references to Bayesian inference for geostatistical models include Kitanidis (1986), Le & Zidek (1992), Handcock & Stein (1993), De Oliveira et al. (1997), Diggle & Ribeiro Jr (2003), De Oliveira & Ecker (2002) and Berger et al. (2001). Omre & Halvorsen (1989) describe the link between Bayesian prediction and simple or ordinary kriging.

Christensen, Møller & Waagepetersen (2001) give further details and properties of the Langevin-Hastings algorithm used in Section 1.9.1. Bayesian inference for a GLSM is described in Diggle et al. (1998) and in Christensen & Waagepetersen (2002) where algorithms are used that update both the random effect S and all the parameters (including β and σ^2). These algorithms are more general than the one presented in this chapter, since they do not require conjugate priors (or limiting cases of conjugate priors). However, from our experience, in practice a consequence of the extra generality is that the algorithms need to be more carefully tuned to specific applications in order to achieve good mixing properties. Zhang (2002) analyses spatial binomial data, and develops a Monte Carlo EM gradient algorithm for maximum likelihood estimation in a GLSM.

Multivariate spatial prediction is presented in Chapter 5 in Chilés & Delfiner (1999); see also Brown, Le & Zidek (1994), Le, Sun & Zidek (1997) and Zidek, Sun & Le (2000).

Examples of space-time modelling include Handcock & Wallis (1994), Wikle & Cressie (1999), Brown, Kårensen, Roberts & Tonellato (2000), Brix & Diggle (2001) and Brown, Diggle, Lord & Young (2001).

Wälder & Stoyan (1996), Wälder & Stoyan (1998) and Schlather (2001) discuss the connection between the classical variogram and the more general second-order properties of marked point processes. Marked point processes are also discussed in Chapter 4.

Acknowledgments: We thank the UK Engineering and Physical Sciences Research Council, the European Union and CAPES (Brasil) for financial support. We also wish to thank Rasmus Waagepetersen and Peter Guttorp for their comments on the manuscript.

1.13 References

Adler, R. J. (1981). *The geometry of random fields*, Wiley, New York.

Berger, J. O., De Oliveira, V. & Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data, *Journal of the American Statistical*

Association **96**: 1361–1374.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43**: 1–59.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**: 9–25.

Brix, A. & Diggle, P. J. (2001). Spatio-temporal prediction for log-Gaussian Cox processes, *Journal of the Royal Statistical Society, Series* B **63**: 823–841.

Brown, P. E., Diggle, P. J., Lord, M. E. & Young, P. C. (2001). Space-time calibration of radar rainfall data, *Applied Statistics* **50**: 221–241.

Brown, P. E., Kårensen, K. F., Roberts, G. O. & Tonellato, S. (2000). Blur-generated non-separable space-time models, *Journal of the Royal Statistical Society, Series* B **62**: 847–860.

Brown, P. J., Le, N. D. & Zidek, J. V. (1994). Multivariate spatial interpolation and exposure to air polutants, *Canadian Journal of Statistics* **22**: 489–509.

Chilés, J. P. & Delfiner, P. (1999). *Geostatistics; modeling spatial uncertainty*, Wiley, New York.

Christensen, O. F., Diggle, P. J. & Ribeiro Jr, P. J. (2001). Analysing positive-valued spatial data: the transformed Gaussian model, *in* P. Monestiez, D. Allard & R. Froidevaux (eds), *GeoENV III - Geostatistics for Environmental Applications*, Vol. 11 of *Quantitative Geology and Geostatistics*, Kluwer, Dordrecht, pp. 287–298.

Christensen, O. F., Møller, J. & Waagepetersen, R. (2001). Geometric ergodicity of Metropolis Hastings algorithms for conditional simulation in generalised linear mixed models, *Methodology and Computing in Applied Probability* **3**: 309–327.

Christensen, O. & Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models, *Biometrics* **58**: 280–286.

Cramér, H. & Leadbetter, M. R. (1967). *Stationary and related processes*, Wiley, New York.

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

De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields, *Computational Statistics and Data Analysis* **34**: 299–314.

De Oliveira, V. & Ecker, M. D. (2002). Bayesian hot spot detection in the presence of a spatial trend: application to total nitrogen concentration in the Chesapeake Bay, *Environmetrics* **13**: 85–101.

De Oliveira, V., Kedem, B. & Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields, *Journal of the American Statistical Association* **92**: 1422–1433.

Diggle, P. J., Harper, L. & Simon, S. (1997). Geostatistical analysis of residual contamination from nuclear weapons testing, *in* V. Barnett & F. Turkman (eds), *Statistics for the environment 3: pollution assesment and control*, Wiley, Chichester, pp. 89–107.

Diggle, P. J. & Ribeiro Jr, P. J. (2003). Bayesian inference in Gaussian model based geostatistics, *Geographical and Environmental Modelling* . (to appear).

Diggle, P. J., Tawn, J. A. & Moyeed, R. A. (1998). Model based geostatistics (with discussion), *Applied Statistics* **47**: 299–350.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion), *Statistical Science* **7**: 473–511.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York.

Gotway, C. A. & Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction, *Journal of Agricultural, Biological and Environmental Statistics* **2**: 157–178.

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

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

Heagerty, P. J. & Lele, S. R. (1998). A composite likelihood approach to binary spatial data, *Journal of the American Statistical Association* **93**: 1099–1111.

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computatioanl and Graphical Statistics* **5**: 299–314.

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

Kent, J. T. (1989). Continuity properties of random fields, *Annals of probability* **17**: 1432–1440.

Kitanidis, P. K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrological applications., *Water Resources Research* **22**: 499–507.

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

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand, *Journal of the Chemical, Metallurgical and Mining Society of South Africa* **52**: 119–139.

Le, N. D., Sun, W. & Zidek, J. V. (1997). Bayesian multivariate spatial interpolation with data missing by design, *Journal of the Royal Statistical Society, Series* B **59**: 501–510.

Le, N. D. & Zidek, J. V. (1992). Interpolation with uncertain covariances: a Bayesian alternative to kriging, *Journal of Multivariate Analysis* **43**: 351–374.

Mardia, K. V. & Watkins, A. J. (1989). On multimodality of the likelihood in the spatial linear model, *Biometrika* **76**: 289–296.

Matérn, B. (1960). Spatial variation, *Technical report*, Meddelanden fran Statens Skogsforsningsinstitut, Stockholm.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, second edn, Chapman and Hall, London.

Natarajan, R. & Kass, R. E. (2000). Bayesian methods for generalized linear mixed models, *Journal of the American Statistical Association* **95**: 227–237.

O'Hagan, A. (1994). *Bayesian Inference*, Vol. 2b of *Kendall's advanced theory of statistics*, Edward Arnold.

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

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

Ribeiro Jr, P. J. & Diggle, P. J. (1999). Bayesian inference in Gaussian model-based geostatistics, *Tech. Report ST-99-09*, Lancaster University.

Ribeiro Jr, P. J. & Diggle, P. J. (2001). geoR: a package for geostatistical analysis, *R News* **1/2**: 15–18. Available from: http://www.r-project.org/doc/Rnews.

Ripley, B. D. (1981). *Spatial statistics*, Chapman and Hall, New York.

Sampson, P. D. & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure, *Journal of the American Statistical Association* **87**: 108–119.

Schlather, M. (2001). On the second-order characteristics of marked point processes, *Bernoulli* **7**: 99–117.

Stein, M. L. (1999). *Interpolation of Spatial Data: some theory for kriging*, Springer Verlag, New York.

Wälder, O. & Stoyan, D. (1996). On variograms in point process statistics, *Biometrical Journal* **38**: 895–905.

Wälder, O. & Stoyan, D. (1998). On variograms in point process statistics: Erratum, *Biometrical Journal* **40**: 109.

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

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

Whittle, P. (1962). Topographic correlation, power-law covariance functions, and diffusion, *Biometrika* **49**: 305–314.

Whittle, P. (1963). Stochastic processes in several dimensions, *Bulletin of the International Statistical Institute* **40**: 974–994.

Wikle, C. K. & Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering, *Biometrika* **86**: 815–829.

Zhang, H. (2002). On estimation and prediction for spatial generalised linear mixed models, *Biometrics* **58**: 129–136.

Zidek, J. V., Sun, W. & Le, N. D. (2000). Designing and integrating composite networks for monitoring multivariate Gaussian pollution field, *Applied Statistics* **49**: 63–79.

Zimmerman, D. L. (1989). Computationally efficient restricted maximum likelihood estimation of generalized covariance functions., *Mathematical Geology* **21**: 655–672.