Model-based Geostatistics

Peter J. Diggle and Paulo J. Ribeiro Jr.

May 22, 2006

Peter J. Diggle
Department of Mathematics and Statistics
Lancaster University, Lancaster, UK
LA1 4YF
p.diggle@lancaster.ac.uk

Paulo J. Ribeiro Jr
Departamento de Estatística
Universidade Federal do Paraná
C.P. 19.081
Curitiba, Paraná, Brasil
81.531-990
paulojus@est.ufpr.br

Preface

Geostatistics refers to the sub-branch of spatial statistics in which the data consist of a finite sample of measured values relating to an underlying spatially continuous phenomenon. Examples include: heights above sea-level in a topographical survey; pollution measurements from a finite network of monitoring stations; determinations of soil properties from core samples; insect counts from traps at selected locations. The subject has an interesting history. Originally, the term *geostatistics* was coined by Georges Matheron and colleagues at Fontainebleau, France, to describe their work addressing problems of spatial prediction arising in the mining industry. See, for example, Matheron (1963, 1971). The ideas of the Fontainebleau school were developed largely independently of the mainstream of spatial statistics, with a distinctive terminology and style which tended to conceal the strong connections with parallel developments in spatial statistics. These parallel developments included work by Kolmogorov (1941), Matérn (1960, reprinted as Matérn, 1986), Whittle (1954, 1962, 1963), Bartlett (1964, 1967) and others. For example, the core geostatistical method known as *simple kriging* is equivalent to minimum mean square error prediction under a linear Gaussian model with known parameter values. Papers by Watson (1971,1972) and the book by Ripley (1981) made this connection explicit. Cressie (1993) considered geostatistics to be one of three main branches of spatial statistics, the others being discrete spatial variation (covering distributions on lattices and Markov random fields) and spatial point processes. Geostatistical methods are now used in many areas of application, far beyond the mining context in which they were originally developed.

Despite this apparent integration with spatial statistics, much geostatistical practice still reflects its independent origins, and from a mainstream statistical perspective this has some undesirable consequences. In particular, explicit

stochastic models are not always declared and *ad hoc* methods of inference are often used, rather than the likelihood-based methods of inference which are central to modern statistics. The potential advantages of using likelihood-based methods of inference are two-fold: they generally lead to more efficient estimation of unknown model parameters; and they allow for the proper assessment of the uncertainty in spatial predictions, including an allowance for the effects of uncertainty in the estimation of model parameters.

Diggle, Tawn & Moyeed (1998) coined the phrase *model-based geostatistics* to describe an approach to geostatistical problems based on the application of formal statistical methods under an explicitly assumed stochastic model. This book takes the same point of view.

We aim to produce an applied statistical counterpart to Stein (1999), who gives a rigorous mathematical theory of kriging. Our intended readership includes postgraduate statistics students and scientific researchers whose work involves the analysis of geostatistical data. The necessary statistical background is summarised in an Appendix, and we give suggestions of further background reading for readers meeting this material for the first time.

Throughout the book, we illustrate the statistical methods by applying them in the analysis of real data-sets. Most of the data-sets which we use are publically available and can be obtained from the book's web-page, *http://www.maths.lancs.ac.uk/∼diggle/mbg*.

Most of the book's chapters end with a section on computation, in which we show how the R software (R Development Core Team 2005) and contributed packages **geoR** and **geoRglm** can be used to implement the geostatistical methods described in the corresponding chapters. This software is freely available from the R Project web-page (*http://www.r-project.org*).

The first two chapters of the book provide an introduction and overview. Chapters 3 and 4 then describe geostatistical models whilst chapters 5 to 8 cover associated methods of inference. The material is mostly presented for univariate problems, i.e. those for which the measured response at any location consists of a single value, but Chapter 3 includes a discussion of some multivariate extensions to geostatistical models and associated statistical methods.

The connections between classical and model-based gostatistics are closest when, in our terms, the assumed model is the linear Gaussian model. Readers who wish to confine their attention to this class of models on a first reading may skip Sections 3.11, 3.12, Chapter 4, Sections 5.5, 7.5, 7.6 and Chapter 8.

Many friends and colleagues have helped us in various ways: by improving our understanding of geostatistical theory and methods; by working with us on a range of collaborative projects; by allowing us to use their data-sets; and by offering constructive criticism of early drafts. We particularly wish to thank Ole Christensen, with whom we have enjoyed many helpful discussions. Ole is also the lead author of the **geoRglm** package.

Peter J Diggle, Paulo J Ribeiro Jr, March 2006.

Contents

1
Introduction

1.1 Motivating examples

The term *spatial statistics* is used to describe a wide range of statistical models and methods intended for the analysis of spatially referenced data. Cressie (1993) provides a general overview. Within spatial statistics, the term *geostatistics* refers to models and methods for data with the following characteristics. Firstly, values $Y_i : i = 1, \ldots, n$ are observed at a discrete set of sampling locations x_i within some spatial region A. Secondly, each observed value Y_i is either a direct measurement of, or is statistically related to, the value of an underlying continuous spatial phenomenon, $S(x)$, at the corresponding sampling location x_i. This rather abstract formulation can be translated to a variety of more tangible scientific settings, as the following examples demonstrate.

Example 1.1. *Surface elevations*

The data for this example are taken from Davis (1972). They give the measured surface elevations y_i at each of 52 locations x_i within a square, A, with side-length 6.7 units. The unit of distance is 50 feet (≈15.24 meters), whereas one unit in y represents 10 feet (≈3.05 meters) of elevation.

Figure 1.1 is a *circle plot* of the data. Each datum (x_i, y_i) is represented by a circle with centre at x_i and radius proportional to y_i. The observed elevations range between 690 and 960 units. For the plot, we have subtracted 600 from each observed elevation, to heighten the visual contrast between low and high values. Note in particular the cluster of low values near the top-centre of the plot.

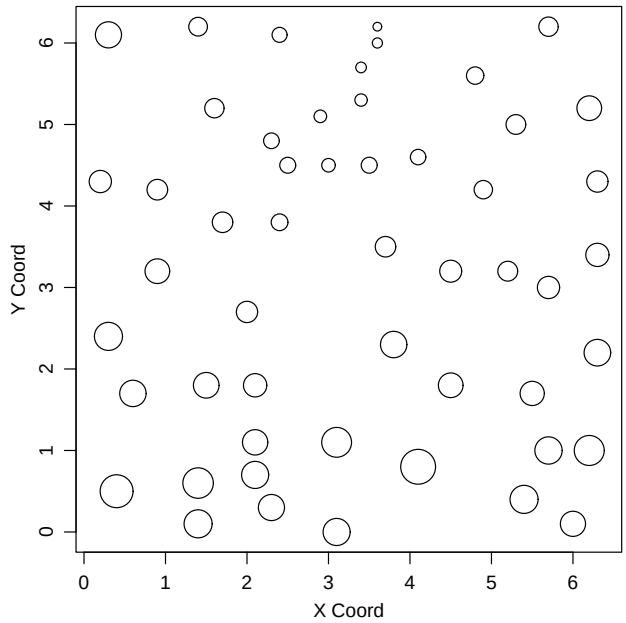

Figure 1.1. Circle plot of the surface elevation data. For the coordinates, the unit of distance is 50 feet. The observed elevations range from 690 to 960 units, where 1 unit represents 10 feet of elevation. Circles are plotted with centres at the sampling locations and radii determined by a linear transformation of the observed elevations (see Section 1.6).

The objective in analysing these data is to construct a continuous elevation map for the whole of the square region A. Let $S(x)$ denote the true elevation at an arbitrary location x. Since surface elevation can be measured with negligible error, in this example each y_i is approximately equal to $S(x_i)$. Hence, a reasonable requirement would be that the map resulting from the analysis should interpolate the data. Our notation, distinguishing between a measurement process Y and an underlying true surface S, is intended to emphasise that this is not always the case.

Example 1.2. *Residual contamination from nuclear weapons testing*

The data for this example were collected from Rongelap Island, the principal island of Rongelap Atoll in the South Pacific, which forms part of the Marshall Islands. The data were previously analysed in Diggle et al. (1998), and have the format $(x_i, y_i, t_i) : i = 1, \ldots, 157$, where x_i identifies a spatial location, y_i is a photon emission count attributable to radioactive caesium, and t_i is the time (in seconds) over which y_i was accumulated.

These data were collected as part of a more wide-ranging, multi-disciplinary investigation into the extent of residual contamination from the USA nuclear weapons testing programme, which generated heavy fall-out over the island

in the 1950's. Rongelap island has been uninhabited since 1985, when the inhabitants left on their own initiative after years of mounting concern about the possible adverse health effects of the residual contamination. Each ratio y_i/t_i gives a crude estimate of the residual contamination at the corresponding location x_i but, in contrast to Example 1.1, these estimates are subject to non-negligible statistical error. For further discussion of the practical background to these data, see Diggle, Harper & Simon (1997).

Figure 1.2 gives a circle plot of the data, using as response variable at each sampling location x_i the observed emission count per unit time, y_i/t_i. Spatial coordinates are in metres, hence the east-west extent of the island is approximately 6.5 kilometres. The sampling design consists of a primary grid covering the island at a spacing of approximately 200 metres together with four secondary 5 by 5 sub-grids at a spacing of 50 metres. The role of the secondary sub-grids is to provide information about short-range spatial effects, which have an important bearing on the detailed specification and performance of spatial prediction methods.

The clustered nature of the sampling design makes it difficult to construct a circle plot of the complete data-set which is easily interpretable on the scale of the printed page. The inset to Figure 1.2 therefore gives an enlarged circle plot for the western extremity of the island. Note that the variability in the emission counts per unit time within each sub-grid is somewhat less than the overall variability across the whole island, which is as we would expect if the underlying variation in the levels of contamination is spatially structured.

In devising a statistical model for the data, we need to distinguish between two sources of variation: spatial variation in the underlying true contamination surface, $T(x)$ say; and statistical variation in the observed photon emission counts, y_i, given the surface $T(x)$. In particular, the physics of photon emissions suggests that a Poisson distribution would provide a reasonable model for the conditional distribution of each y_i given the corresponding value $T(x_i)$. The gamma camera which records the photon emissions integrates information over a circular area whose effective diameter is substantially smaller than the smallest distance (50 metres) between any two locations x_i. It is therefore reasonable to assume that the y_i are conditionally independent given the whole of the underlying surface $T(x)$. In contrast, there is no scientific theory to justify any specific model for $T(x)$, which represents the long-term cumulative effect of variation in the initial deposition, soil properties, human activity and a variety of natural environmental processes. We return to this point in Section 1.2.

One scientific objective in analysing the Rongelap data is to obtain an estimated map of residual contamination. However, in contrast to Example 1.1, we would argue that in this example the map should not interpolate the observed ratios y_i/t_i because each such ratio is a noisy estimate of the corresponding value of $T(x_i)$. Also, because of the health implications of the pattern of contamination across the island, particular properties of the map are of specific interest, for example the location and value of the maximum of $T(x)$, or areas within which $T(x)$ exceeds a prescribed threshold.

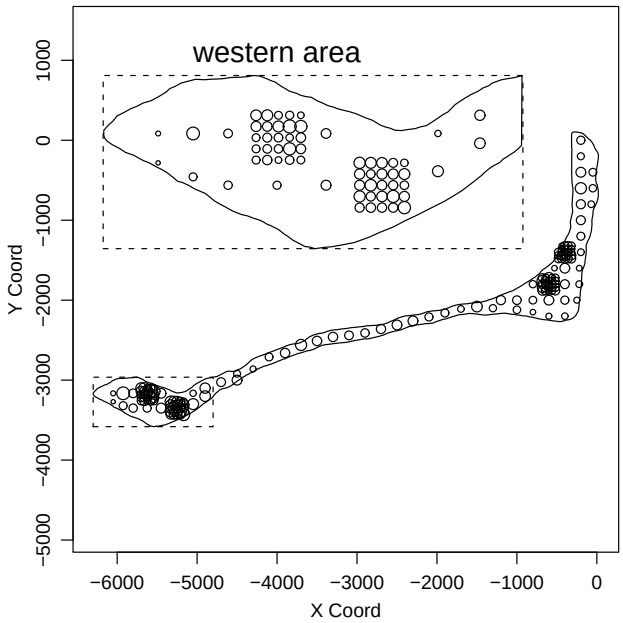

Figure 1.2. Circle plot for data from Rongelap island. Circles are plotted with centres at the sampling locations and radii proportional to observed emission counts per unit time. The unit of distance is 1 metre. The inset shows an enlargement of the western extremity of the island.

Example 1.3. *Childhood malaria in The Gambia*

These data are derived from a field-survey into the prevalence of malaria parasites in blood-samples taken from children living in village communities in The Gambia, West Africa. For practical reasons, the sampled villages were concentrated into five regions rather than being sampled uniformly across the whole country. Figure 1.3 is a map of The Gambia showing the locations of the sampled villages. The clustered nature of the sampling design is clear.

Within each village, a random sample of children was selected. For each child, a binary response was then obtained, indicating the presence or absence of malaria parasites in a blood-sample. Covariate information on each child included their age, sex, an indication of whether they regularly slept under a mosquito net and, if so, whether or not the net was treated with insecticide. Information provided for each village, in addition to its geographical location, included a measure of the green-ness of the surrounding vegetation derived from satellite data, and an indication of whether or not the village belonged to the primary health care structure of The Gambia Ministry for Health.

The data-format for this example is therefore $(x_i, y_{ij}, d_i, d_{ij})$ where the subscripts i and j identify villages, and individual children within villages, respectively, whilst d_i and d_{ij} similarly represent explanatory variables recorded at the village level, and at the individual level, as described below. Note that if only village-level explanatory variables are used in the analysis, we might choose

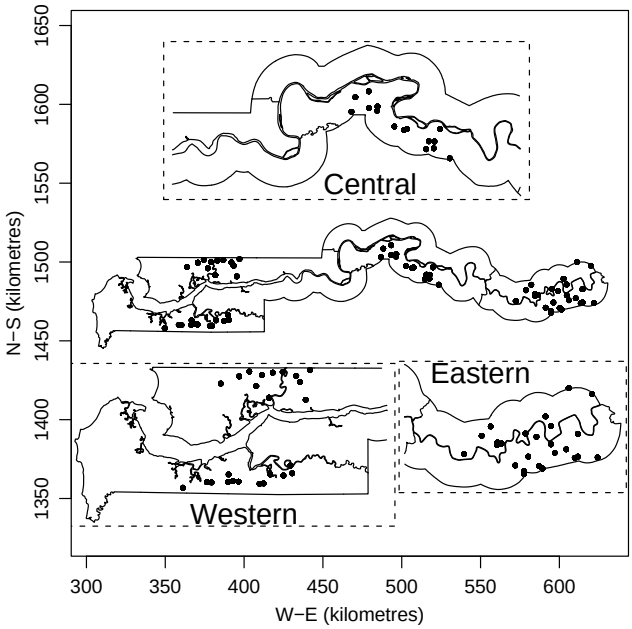

Figure 1.3. Sampling locations for the Gambia childhood malaria survey. The inset plots are enlarged maps of the western, central and eastern regions of The Gambia.

to analyse the data only at the village level, in which case the data-format could be reduced to (x_i, n_i, y_i, d_i) where n_i is the number of children sampled in the ith village, and $y_i = \sum_{j=1}^{n_i} y_{ij}$ the number who test positive

Figure 1.4 is a scatterplot of the observed prevalences, y_i/n_i, against the corresponding green-ness values, u_i. This shows a weak positive correlation.

The primary objective in analysing these data is to develop a predictive model for variation in malarial prevalence as a function of the available explanatory variables. A natural starting point is therefore to fit a logistic regression model to the binary responses y_{ij}. However, in so doing we should take account of possible unexplained variation within or between villages. In particular, unexplained spatial variation between villages may give clues about as-yet unmeasured environmental risk factors for malarial infection.

Example 1.4. *Soil data*

These data have the format $(x_i, y_{i1}, y_{i2}, d_{i1}, d_{i2})$, where x_i identifies the location of a soil sample, the two y-variables give the calcium and magnesium content whilst the two d-covariates give the elevation and sub-area code of each sample.

The soil samples were taken from the 0-20cm depth layer at each of 178 locations. Calcium and magnesium content were measured in $mmol_c/dm^3$ and the elevation in metres. The study region was divided into three sub-regions which have experienced different soil management regimes. The first, in the upper-left corner, is typically flooded during each rainy season and is no longer

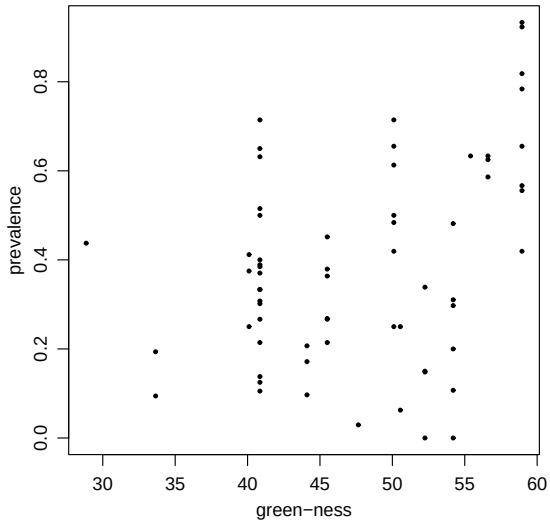

Figure 1.4. Observed prevalences against green-ness for villages in the Gambia childhood malaria survey.

used as an experimental area because of its varying elevation. The calcium and magnesium levels in this region therefore represent the pattern of natural spatial variation in background content. The second, corresponding to the lower half of the study region, and the third, in the upper-right corner, have received fertilisers in the past: the second is typically occupied by rice fields, whilst the third is frequently used as an experimental area. Also, the second sub-region was the most recent of the three to which calcium was added to neutralise the effect of aluminium in the soil, which partially explains the generally higher measured calcium values within this sub-region.

The sampling design is an incomplete regular lattice at a spacing of approximately 50 metres. The data were collected by researchers from PESAGRO and EMBRAPA-Solos, Rio de Janeiro, Brasil (Capeche 1997).

The two panels of Figure 1.5 show circle plots of the calcium (left panel) and magnesium (right panel) data separately, whilst Figure 1.6 shows a scatterplot of calcium against magnesium, ignoring the spatial dimension. This shows a moderate positive correlation between the two variables; the value of the sample correlation between the 178 values of calcium and magnesium content is $r = 0.33$.

Figure 1.7 shows the relationship between the potential covariates and the calcium content. There is a clear trend in the north-south direction, with generally higher values to the south. The relationships between calcium content and either east-west location or elevation are less clear. However, we have included on each of the three scatterplots a lowess smooth curve (Cleveland 1981) which, in the case of elevation, suggests that there may be a relationship with calcium beyond an elevation threshold. Finally, the boxplots in the bottom right panel of Figure 1.7 suggest that the means of the distributions of calcium content are

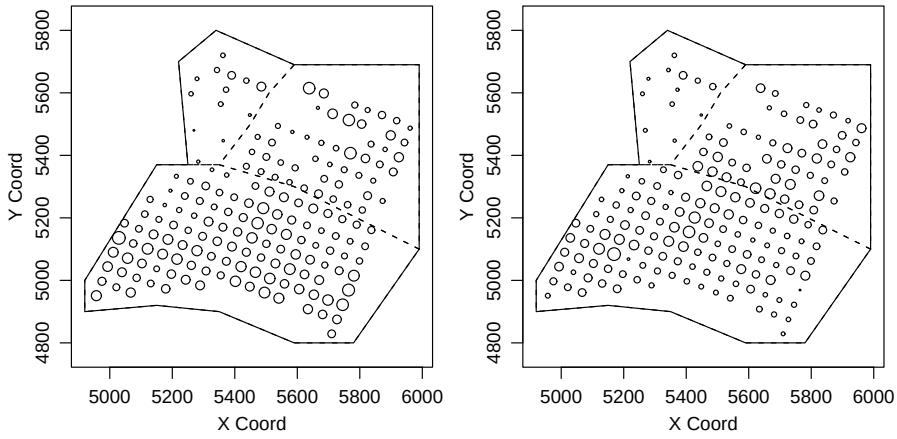

Figure 1.5. Circle plots of calcium and magnesium content with dashed lines delimiting sub-regions with different soil management practices.

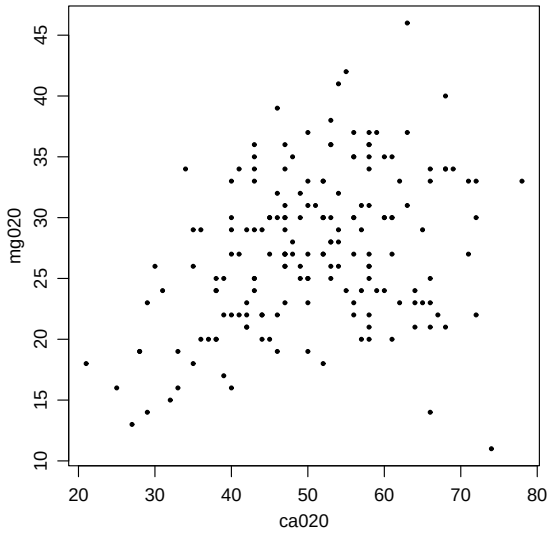

Figure 1.6. Scatterplot of calcium content against magnesium content in the 0-20cm soil layer.

different in the different sub-regions. In any formal modelling of these data, it would also be sensible to examine covariate effects after allowing for a different mean response in each of the three sub-regions, in view of their different management histories.

One objective for these data is to construct maps of the spatial variation in calcium or magnesium content. Because these characteristics are determined from small soil cores, and repeated sampling at effectively the same location would yield different measurements, the constructed maps should not necessar-

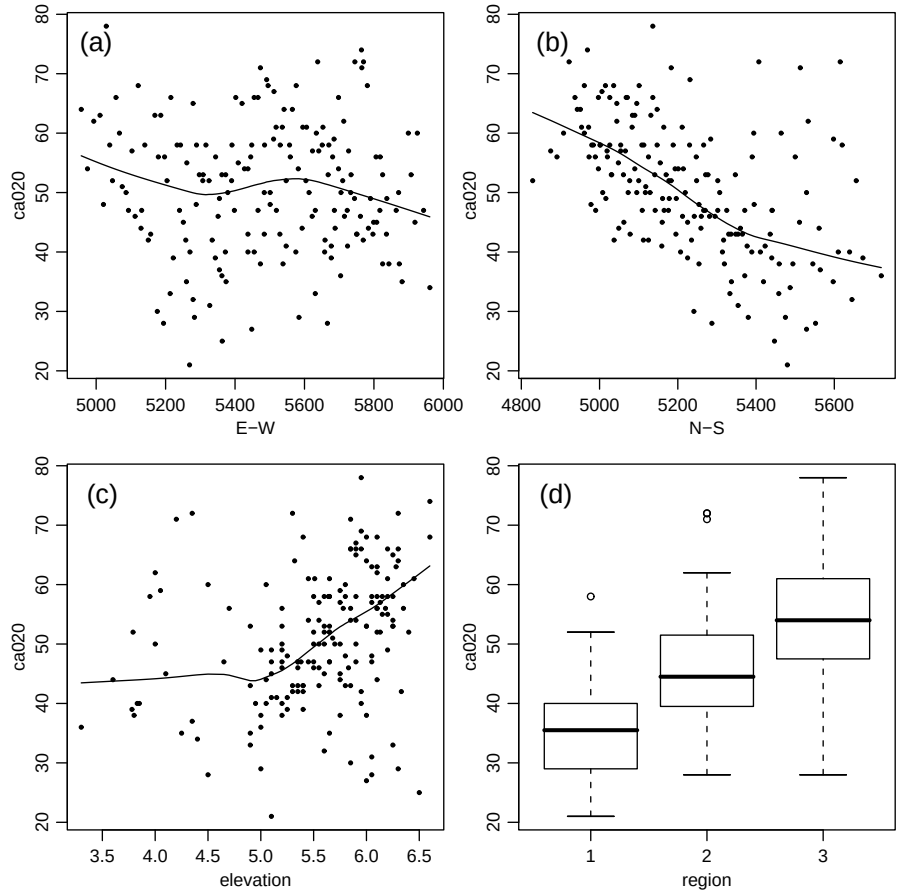

Figure 1.7. (a, b, c) Scatterplots of calcium content against: (a) $E - W$ coordinate, (b) $N - S$ coordinate, (c) elevation. Lines are lowess curves. (d) Box-plots of calcium content in each of the three sub-regions.

ily interpolate the data. Another goal is to investigate relationships between calcium or magnesium content and the two covariates. The full data-set also includes the values of the calcium and magnesium content in the 20-40cm depth layer.

We shall introduce additional examples in due course. However, these four are sufficient to motivate some basic terminology and notation, and to indicate the kinds of problems which geostatistical methods are intended to address.

1.2 Terminology and notation

The basic format for univariate *geostatistical data* is

$$(x_i, y_i) : i = 1, \ldots, n,$$

where x_i identifies a spatial location (typically in two-dimensional space, although one-dimensional and three-dimensional examples also occur) and y_i is a scalar value associated with the location x_i. We call y the *measurement variable* or *response*. A defining characteristic of geostatistics is that the measurement variable is, at least in principle, defined throughout a continuous study-region, A say. Furthermore, we shall assume that the sampling design for the locations x_i is either deterministic (for example, the x_i may form a grid over the study-region), or stochastically independent of the process which generates the measurements y_i. Each y_i is a realisation of a random variable Y_i whose distribution is dependent on the value at the location x_i of an underlying spatially continuous stochastic process $S(x)$ which is not directly observable. In particular cases, such as in our Example 1.1, we might reasonably assume that $Y_i = S(x_i)$, but in general it is important to preserve a distinction between the observable quantities Y_i and the unobservable, or latent process $S(x)$.

The basic form of a *geostatistical model* therefore incorporates at least two elements: a real-valued stochastic process $\{S(x) : x \in A\}$, which is typically considered to be a partial realisation of a stochastic process $\{S(x) : x \in \mathbb{R}^2\}$ on the whole plane; and a multivariate distribution for the random variable $Y = (Y_1, \ldots, Y_n)$ conditional on $S(\cdot)$. We call $S(x)$ the *signal* and Y_i the *response*. Often, Y_i can be thought of as a noisy version of $S(x_i)$ and the Y_i can be assumed to be conditionally independent given $S(\cdot)$.

1.2.1 Support

Examples 1.2 and 1.4 illustrate a general issue with geostatistical data, concerning the *support* of each measured response. Formally, we associate each y_i with a point location x_i. However, in many cases y_i derives from a finite area for which x_i is a convenient reference point. In Example 1.4, the support is clearly identifiable as the circular cross-section of the soil core used to obtain each sample, and x_i denotes the centre of the cross-section. In Example 1.2, definition of the support is more difficult. The gamma camera integrates positron emissions over a circular neighbourhood of each sample location x_i, but rather than a sharp cut-off at a known distance, the camera traps a smaller proportion of the actual emissions with increasing distance from the centre of the circle. This implies that the modelled signal, $S(x)$, should strictly be interpreted as a weighted integral of an underlying spatially continuous signal, $S^*(x)$ say, so that

$$S(x) = \int w(r) S^*(x - r) dr.$$

Under this formulation, $S(x)$ is still a real-valued, spatially continuous process, i.e. it is well-defined for all $x \in \mathbb{R}^2$. Its genesis as an integral does, however, have

implications for what covariance structure we can strictly assume for the process $S(\cdot)$, since any smoothness in the behaviour of the weighting function $w(\cdot)$ constrains the allowable form of covariance structure for $S(\cdot)$. In this particular example we do not need to model the effect of the weighting function explicitly, because its effective range is much smaller than the minimum distance of 50 metres between any two points in the design.

The idea that geostatistical measurements have finite, rather than infinitesimal, support is to be contrasted with problems in which measurements are derived from a partition of a spatial region into discrete spatial units $i = 1, \ldots, n$, each of which yields a measurement y_i. This is often the case, for example, in spatial epidemiology, where data on disease prevalence may be recorded as counts in administrative sub-regions, for example counties or census tracts. In that context, the modelling options are either to deal explicitly with the effects of the spatial integration of an underlying spatially continuous process $S^*(x)$ or, more pragmatically, to specify a model at the level of the discrete spatial units, i.e. a multivariate distribution for random variables $Y_i : i = 1, \ldots, n$. Models of the second kind have an extensive literature and are widely used in practice to analyse data arising as a result of spatial aggregation into discrete units. Less commonly, the actual spatial units are genuinely discrete; an example would be data on the yields of individual fruit-trees in an orchard.

Evidently, a common feature of geostatistical models and discrete spatial models is that they both specify the joint distribution of a spatially referenced, n-dimensional random variable $(Y_1, \ldots, Y_n)$. An important difference is that a geostatistical model automatically embraces any n, and any associated set of sampling locations, whereas a discrete spatial model is specific to a particular set of locations. A classic early reference to the modelling and analysis of data from discrete spatial units is Besag (1974). See also Cressie (1993, chapters 6 and 7).

1.2.2 Multivariate responses and explanatory variables

As our motivating examples llustrate, in many applications the basic (x_i, y_i)-format of geostatistical data will be extended in either or both of two ways. There may be more than one measurement variable, so defining a *multivariate response*, $y_i = \{y_{i1}, ..., y_{id}\}$, or the data may include spatial explanatory variables, $\{d_k(x) : x \in A\}$, sometimes also called *covariates*.

The distinction between the two is not always clear-cut. From a modelling point of view, the difference is that a model for a multivariate response requires the specification of a vector-valued stochastic process over the study-region A, whereas spatial explanatory variables are treated as deterministic quantities with no associated stochastic model. One consequence of this is that a spatial explanatory variable must, at least in principle, be available at any location within A if it is to be used to predict responses at unsampled locations x. An example would be the green-ness index in Example 1.3. The index is calculated on a 1 km pixel grid and can therefore be used to predict malaria prevalence without making any assumptions about its spatial variation. Even then, in our

experience the distinction between a stochastic signal $S(x)$ and a spatial explanatory variable $d(x)$ is largely a reflection of our scientific goals. Again using Example 1.3 to illustrate the point, the goal in this example is to understand how environmental factors affect malaria prevalence. Elevation is one of several factors which determine the suitability of a particular location to support breeding mosquitos, and is a candidate for inclusion as an explanatory variable in a stochastic model for prevalence. In contrast, in Example 1.1 the goal is to interpolate or smooth a spatially sparse set of measured elevations so as to obtain a spatially continuous elevation map, hence elevation is treated as a stochastic response.

In most geostatistical work, the adoption of a stochastic model for $S(x)$ reflects its unknown, unobserved quality rather than a literal belief that the underlying spatial surface of interest is generated by the laws of probability. Indeed, in many applications the role of the signal process $S(x)$ is as a surrogate for unmeasured explanatory variables which influence the response variable. In modelling $S(x)$ as a stochastic process we are using stochasticity at least in part as a metaphor for ignorance.

For this reason, when relevant explanatory variables are only available at the data-locations x_i and we wish to use their observed values for spatial prediction at an unsampled location x, a pragmatic strategy is to treat such variables as additional responses, and accordingly to formulate a multivariate model. Example 1.4 illustrates both situations: the calcium and magnesium contents form a bivariate spatial stochastic process, whereas region and, to a good approximation, elevation, are available at any location, are not of scientific interest in themselves, and can therefore be treated as explanatory variables. In this example, both components of the bivariate response are measured at each data-location. More generally, measurements on different components of a multivariate response need not necessarily be made at a common set of locations.

Note that the locations x_i potentially play a dual role in geostatistical analysis. Firstly, spatial location is material to the model for the signal process $S(x)$ in that the stochastic dependence between $S(x)$ and $S(x')$ is typically modelled as a function of the locations in question, x and x'. Secondly, each location defines the values of a pair of explanatory variables corresponding to the two spatial coordinates. The convention in geostatistics is to use the term *trend surface* to mean a spatially varying expectation of the response variable which is specified as a function of the coordinates of the x_i, whereas the term *external trend* refers to a spatially varying expectation specified as a function of other explanatory variables $d(x)$. For example, the elevation data as presented in Example 1.1 do not include any explanatory variables which could be used in an external trend model, but as we shall see in Chapter 2 a low-order polynomial trend surface can explain a substantial proportion of the observed spatial variation in the data.

1.2.3 Sampling design

The locations x_i at which measurements are made are collectively called the *sampling design* for the data. A design is *non-uniform* if the sampling intensity varies systematically over the study-region, in the sense that before the actual actual sampling points are chosen, some parts of the study-region are deliberately sampled more intensively than others. This is as distinct from the sampling intensity varying by chance; for example, if sample points are located as an independent random sample from a uniform distribution over the study-region, it may (indeed, will) happen that some parts of the study-region are more intensively sampled than others but we would still describe this as a uniform design because of its method of construction.

A design is *non-preferential* if it is deterministic, or if it is stochastically independent of $S(\cdot)$. Conventional geostatistical methods assume, if only implicitly, that the sampling design is non-preferential, in which case we can legitimately analyse the data conditional on the design. Provided that sampling is non-preferential, the choice of design does not impact on the assumed model for the data, but does affect the precision of inferences which can be made from the data. Furthermore, different designs are efficient for different kinds of inference. For example, closely spaced pairs of sample locations are very useful for estimating model parameters, but would be wasteful for spatial prediction using a known model.

1.3 Scientific objectives

In most applications, the scientific objectives of a geostatistical analysis are broadly of two kinds: estimation and prediction.

Estimation refers to inference about the parameters of a stochastic model for the data. These may include parameters of direct scientific interest, for example those defining a regression relationship between a response and an explanatory variable, and parameters of indirect interest, for example those defining the covariance structure of a model for $S(x)$.

Prediction refers to inference about the realisation of the unobserved signal process $S(x)$. In applications, specific prediction objectives might include prediction of the realised value of $S(x)$ at an arbitrary location x within a region of interest, A, typically presented as a map of the predicted values of $S(x)$, or prediction of some property of the complete realisation of $S(x)$ which is of particular relevance to the problem in hand. For example, in the mining applications for which geostatistical methods were originally developed, the average value of $S(x)$ over an area potentially to be mined would be of direct economic interest, whereas in the Rongelap Island example an identification of those parts of the island where $S(x)$ exceeds some critical value would be more useful than the average as an indicator of whether the island is fit for re-habitation. Geostatistical models and methods are particularly suited to scientific problems whose objectives include prediction, in the sense defined here.

A third kind of inferential problem, namely *hypothesis testing*, can also arise in geostatistical problems, although often only in a secondary sense, for example in deciding whether or not to include a particular explanatory variable in a regression model. For the most part, in this book we will tacitly assume that testing is secondary in importance to estimation and prediction.

1.4 Generalised linear geostatistical models

Classical generalised linear models, introduced by Nelder & Wedderburn (1972), provide a unifying framework for the analysis of many superficially different kinds of independently replicated data. Several different ways to extend the generalised linear model class to dependent data have been proposed, amongst which perhaps the most widely used are *marginal models* (Liang & Zeger 1986) and *mixed models* (Breslow & Clayton 1993). What we shall call a *generalised linear geostatistical model* is a generalised linear mixed model of a form specifically oriented to geostatistical data.

The first ingredient in this class of models is a stationary Gaussian process $S(x)$. A stochastic process $S(x)$ is *Gaussian* if the joint distribution of $S(x_1), \ldots, S(x_n)$ is multivariate Normal for any integer n and set of locations x_i. The process is *stationary* if the expectation of $S(x)$ is the same for all x, the variance of $S(x)$ is the same for all x and the correlation between $S(x)$ and $S(x')$ depends only on $u = ||x - x'||$, the Euclidean distance between x and x'. We shall use the class of stationary Gaussian processes as a flexible, empirical model for an irregularly fluctuating, real-valued spatial surface. Typically, the nature of this surface, which we call the *signal*, is of scientific interest but the surface itself cannot be measured directly. The range of applicability of the model can be extended by the use of mathematical transformations. For example, in the suggested model for the Rongelap island photon emission data, the Gaussian process $S(x)$ is the logarithm of the underlying contamination surface $T(x)$. We discuss the Gaussian model, including non-stationary versions, in more detail in Chapter 3.

The second ingredient in the generalised linear geostatistical model is a statistical description of the data-generating mechanism conditional on the signal. This part of the model follows a classical generalized linear model as described by McCullagh & Nelder (1989), with $S(x)$ as an offset in the linear predictor. Explicitly, conditional on $S(\cdot)$ the responses $Y_i : i = 1, \ldots, n$ at locations $x_i : i = 1, \ldots, n$ are mutually independent random variables whose conditional expectations, $\mu_i = \mathrm{E}[Y_i|S(\cdot)]$, are determined as

$$h(\mu_i) = S(x_i) + \sum_{k=1}^{p} \beta_k d_k(x_i), \tag{1.1}$$

where $h(\cdot)$ is a known function, called the *link function*, the $d_k(\cdot)$ are observed *spatial explanatory variables* and the β_k are unknown *spatial regression parameters*. The terms on the right-hand side of (1.1) are collectively called the *linear*

predictor of the model. The conditional distribution of each Y_i given $S(\cdot)$ is called the *error distribution*.

For each of our introductory examples, there is a natural candidate model within the generalized linear family.

For Example 1.1, in which the response is real-valued, we might adopt a *linear Gaussian* model, in which the link function $h(\cdot)$ is the identity and the error distribution is Gaussian with variance τ^2. Hence, the true surface elevation at a location x is given by $S(x)$ and, conditional on the realisation of $S(x)$ at all locations the measured elevations y_i are mutually independent, Normally distributed with conditional means $S(x_i)$ and common conditional variance τ^2. A possible extension of this model would be to include spatial explanatory variables to account for a possible non-stationarity of $S(\cdot)$. For example, the circle plot of the data (Figure 1.1) suggests that elevations tend to decrease as we move from south to north. We might therefore consider including the north-south coordinate of the location as an explanatory variable, $d_1(\cdot)$ say, so defining a non-constant plane over the area. The conditional mean of each y_i given $S(x)$ would then be modelled as $d_1(x_i)\beta + S(x_i)$.

For Example 1.2, in which the response is a photon emission count, the underlying physics motivates the Poisson distribution as a suitable error distribution whilst the log-linear formulation suggested earlier is an empirical device which constrains the expected count to be non-negative, as required. The photon emission counts Y_i can then be modelled as conditionally independent Poisson-distributed random variables, given an underlying surface $T(\cdot)$ of true levels of contamination. Also, the expectation of Y_i is directly proportional both to the value of $T(x_i)$ and to the time, t_i, over which the observed count is accumulated. Hence, the conditional distribution of Y_i should be Poisson with mean $t_i T(x_i)$. In the absence of additional scientific information a pragmatic model for $T(x)$, recognising that it necessarily takes non-negative values, might be that $\log T(x) = S(x)$ is a Gaussian stochastic process with mean μ, variance σ^2 and correlation function $\rho(x, x') = \mathrm{Corr}\{S(x), S(x')\}$. Like any statistical model, this is an idealisation. A possible refinement to the Poisson assumption for the emission counts conditional on the signal $S(x)$ would be to recognise that each y_i is a so-called nett count, calculated by subtracting from the raw count an estimate of that part of the count which is attributable to broad-band background radiation. With regard to the model for $S(x)$, the assumed constant mean could be replaced by a spatially varying mean if there were evidence of systematic variation in contamination across the island.

For Example 1.3, the sampling mechanism leads naturally to a binomial error distribution at the village-level or, at the child-level, a Bernoulli distribution with the conditional mean μ_{ij} representing the probability of a positive response from the jth child sampled within the ith village. A logit-linear model, $h(\mu_{ij}) = \log\{\mu_{ij}/(1-\mu_{ij})\}$, constrains the μ_{ij} to lie between 0 and 1 as required, and is one of several standard choices. Others include the probit link, $h(\mu) = \Phi^{-1}(\mu)$ where $\Phi(\cdot)$ denotes the standard Normal distribution function, or the complementary-log-log, $h(\mu) = \log\{-\log(\mu)\}$. In practice, the logit and probit links are hard to distinguish, both corresponding to a symmetric

S-shaped curve for μ as a function of the linear predictor with the point of symmetry at $\mu = 0.5$, whereas the complementary-log-log has a qualitatively different, asymmetric form.

Example 1.4 features a bivariate response, and therefore falls outside the scope of the (univariate) generalized linear geostatistical model as described here. However, a separate linear Gaussian model could be used for each of the two responses, possibly after appropriate transformation, and dependence between the two response variables could then be introduced by extending the unobserved Gaussian process $S(x)$ to a bivariate Gaussian process, $S(x) = \{S_1(x), S_2(x)\}$. This example also includes explanatory variables as shown in Figure 1.7. These could be added to the model as indicated in equation (1.1), using the identity link function.

1.5 What is in this book?

This books aims to describe and explain statistical methods for analysing geostatistical data. The approach taken is model-based, by which we mean that the statistical methods are derived by applying general principles of statistical inference based on an explicitly declared stochastic model of the data-generating mechanism.

In principle, we place no further restriction on the kind of stochastic model to be specified. Our view is that a model for each particular application should ideally be constructed by collaboration between statistician and subject-matter scientist with the aim that the model should incorporate relevant contextual knowledge whilst simultaneously avoiding unnecessary over-elaboration and providing an acceptable fit to the observed data. In practice, a very useful and flexible model-class is the generalized linear geostatistical model, which we described briefly in Section 1.4. Chapters 3 and 4 develop linear and generalized linear geostatistical models in more detail. We also include in Chapter 4 some cautionary examples of spatial modelling problems for which the generalized linear model is inadequate.

We shall develop both classical and Bayesian approaches to parameter estimation. The important common feature of the two approaches is that they are based on the likelihood function. However, we also describe simpler, more ad hoc approaches and indicate why they are sometimes useful.

For problems involving prediction, we shall argue that a Bayesian approach is natural and convenient because it provides a ready means of allowing uncertainty in model parameters to be reflected in the widths of our prediction intervals.

Within the Bayesian paradigm, there is no formal distinction between an unobserved spatial stochastic process $S(x)$ and an unknown parameter θ. Both are modelled as random variables. Nevertheless, although we use Bayesian methods extensively, we think that maintaining the distinction between *prediction* of $S(x)$ and *estimation* of θ is important in practice. As noted in Section 1.3 above, prediction is concerned with learning about the particular realisation of

the stochastic process $S(x)$ which is assumed to have generated the observed data y_i, whereas estimation is concerned with *properties* of the process $S(\cdot)$ which apply to all realisations. Section 2.4 discusses some of the inferential implications of this distinction in the context of a specific, albeit hypothetical, example.

1.5.1 Organisation of the book

Chapters 3 and 4 of the book discuss geostatistical models, whilst Chapters 5 to 8 discuss associated methods for the analysis of geostatistical data. Embedded within these chapters is a model-based counterpart to classical, linear geostatistics, in which we assume that the linear Gaussian model is applicable, perhaps after transformation of the response variable. We do not necessarily believe that the Gaussian is a correct model, only that it provides a reasonable approximation. Operationally, its significance is that it gives a theoretical justification for using linear prediction methods, which under the Gaussian assumption have the property that they minimise mean squared prediction errors. In Chapter 8 we give a model-based perspective on design issues for geostatistical studies.

Our aim has been to give a thorough description of core topics in model-based geostatistics. However, in several places we have included shorter descriptions of some additional topics, together with suggestions for further reading. These additional topics are ones for which model-based geostatistical methods are, at the time of writing, incompletely developed. They include constructions for multivariate Gaussian models, preferential sampling and point process models.

Throughout the book, we intersperse methodological discussion with illustrative examples using real or simulated data. Some of the data-sets which we use are not freely available. Those which are can be downloaded from the book's web-page, *http://www.maths.lancs.ac.uk/∼diggle/mbg*.

Most chapters, including this one, end with a section on "Computation". In each such section we give examples of R code to implement the geostatistical methods described in the corresponding chapters, and illustrate some of the optional input parameters for various functions within the contributed R packages **geoR** and **geoRglm**. These illustrations are intended to be less formal in style than the help-pages which form part of the package documentation. The web-sites, *http://www.est.ufpr.br/geoR* and *http://www.est.ufpr.br/geoRglm*, also include illustrative sessions using these two packages. Material from the computation sections is also available from the book's web-page.

The "Computation" sections assume that the reader is familiar with using R for elementary statistics and graphics. For readers who are not so familiar, a good introductory textbook is Dalgaard (2002), whilst general information about the R project can be found in documentation available in the R-Project web page, *http://www.r-project.org*. These sections are also optional, in the sense that they introduce no new statistical ideas, and the remainder of the book can be read without reference to this material.

1.5.2 Statistical pre-requisites

We assume that the reader has a general knowledge of the standard tools for exploratory data-analysis, regression modelling and statistical inference. With regard to regression modelling, we use both linear and generalised linear models. One of many good introductions to linear models is Draper & Smith (1981). The standard reference to generalised linear models is McCullagh & Nelder (1989). We make extensive use of likelihood-based methods, for both non-Bayesian and Bayesian inference. Appendix A gives a short summary of the key ideas. A good treatment of likelihood-based methods in general is Pawitan (2001), whilst O'Hagan (1994) specifically discusses the Bayesian method.

Readers will also need some knowledge of elementary probability and stochastic process theory. Introductory books at a suitable level include Ross (1976) for elementary probability and Cox & Miller (1965) for stochastic processes.

We shall also use a variety of computer-intensive methods, both for simulating realisations of stochastic processes and more generally in Monte Carlo methods of inference, including Markov chain Monte Carlo. A good general introduction to simulation methods is Ripley (1987). Tanner (1996) presents a range of computational algorithms for likelihood-based and Bayesian inference. Gelman, Carlin, Stern & Rubin (2003) focus on Bayesian methods for a range of statistical models. Gilks, Richardson & Spiegelhalter (1996) discuss both theoretical and practical aspects of Markov chain Monte Carlo.

1.6 Computation

The examples in this section, and in later chapters, use the freely available software R and the contributed R packages **geoR** and **geoRglm**. Readers should consult the R project web page, *http://www.r-project.org*, for further information on the software and instructions on its installation.

In the listing of the R code for the examples, the `>` sign is the R prompt and the remainder of the line denotes the R command entered by the user in response to the prompt. R commands are shown in *`slanted verbatim font like this`*. When a single command is spread over two or more lines, the second and subsequent lines of input are prompted by a `+` sign, rather than the `>` sign. The R system is based on subroutines called *functions*, which in turn can take *arguments* which control their behaviour. Function names are followed by parentheses, in the format `function()`, whereas arguments are written within the parentheses. Any lines without the `>` prompt represent outputs from a function which, by default, are passed back to the screen. They are shown in `verbatim font like this`.

1.6.1 Elevation data

In our first example, we give the commands needed to load the **geoR** package, and to produce the circle plot of the elevation data, as shown in Figure 1.1.

The example assumes that the data are stored in a standard three-column text-file `elevation.dat` located in the R working directory. The first two columns on each line give the (x, y)-coordinates of a location whilst the third column gives the corresponding value of the measured elevation. The version of the data which can be downloaded from the book web page is already formatted in this way.

```
> require(geoR)
> elevation <- read.geodata("elevation.dat")
> points(elevation, cex.min = 1, cex.max = 4)
```

The first command above uses the built-in R function `require()` to load the **geoR** package. The second command reads the data and converts them to an object of the class `geodata` using `read.table()` and `as.geodata()` internally. The last command invokes a method for `points()` which is provided by the package **geoR**. In this way, the generic R function `points()` is able to use the **geoR** function `points.geodata()` to produce the required plot of the data. The example includes optional settings for arguments which control the sizes of the plotted circles. By default, the diameters of the plotted circles are defined by a linear transformation of the measured elevations onto a scale ranging between `cex.min` and `cex.max` times the default plotting character size.

The output returned when typing *`args(points.geodata)`* will show other arguments which can be used to modify the resulting plot. For example,

```
> points(elevation, cex.min = 2, cex.max = 2, col = "gray")
```

will plot the locations as filled circles with gray shades proportional to the measured elevation values, whereas

```
> points(elevation, cex.min = 2, cex.max = 2, pt.div = "quint")
```

will result in points filled with different colours according to the quintiles of the empirical distribution of measured elevations.

Because the elevation data are also included in the **geoR** package, they can be loaded from within R, once the package itself has been loaded, by using the `data()` function, and explanatory documentation accessed using the `help()` function, as follows.

```
> data(elevation)
> help(elevation)
```

There are several data-sets included in the package **geoR** which can be loaded with `data()`. Typing the command *`data(package="geoR")`* will show a list of the available data sets with respective names and a short description. For each of them there is a help file explaining the data contents and format.

Another, and often more convenient, way of running a sequence of R commands is to use `source()`. To do so, we first type the required sequence of commands, without the `>` at the beginning of each line, into a text-file, say `elevation.R` although any other legal file-name could be used. We then invoke the whole sequence by responding to the R prompt with the single command

```
> source("elevation.R")
```

This option, or an equivalent mode of operation based on toggling between an editor and an R command window, is usually more efficient than typing R commands directly in response to the > prompt.

The next example shows the output generated by applying the `summary()` function to the elevation data. The output includes the number of data points, the minimum and maximum values of the x and y-coordinates and of the distances between pairs of points, together with summary statistics for the measured elevations.

```
> summary(elevation)

Number of data points: 52

Coordinates summary
      x   y
min 0.2 0.0
max 6.3 6.2

Distance summary
     min      max
0.200000 8.275869

Data summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  690.0   787.5   830.0   827.1   873.0   960.0
```

Another function which is useful for initial exploration of a set of data is the method `plot.geodata()`, which is invoked by default when a `geodata` object is supplied as an argument to the built-in `plot()` function. Its effect is to produce a 2×2 display showing the point locations, the measured values at each location against each of the coordinates, and a histogram of the measured values. This plot for the elevation data is shown in Figure 1.8, which is produced by the command

```
> plot(elevation, lowess = T)
```

The optional argument `lowess=T` adds a smooth line to the scatterplots of the measured values against each of the spatial coordinates. The top-right panel of Figure 1.8 has been rotated by 90 degrees from the conventional orientation, i.e. the measured values correspond to the horizontal rather than the vertical axis, so that the spatial coordinate axes have the same interpretation throughout. These plots aim to investigate the behaviour of the data along the coordinates, which can be helpful in deciding whether a trend surface should be included in the model for the data. By default, the plot of the data locations shown in the top-left panel of Figure 1.8 uses circles, triangles, vertical and diagonal crosses to correspond to the quartiles of the empirical distribution of measured values. On a computer screen, these points would also appear in different colours: blue,

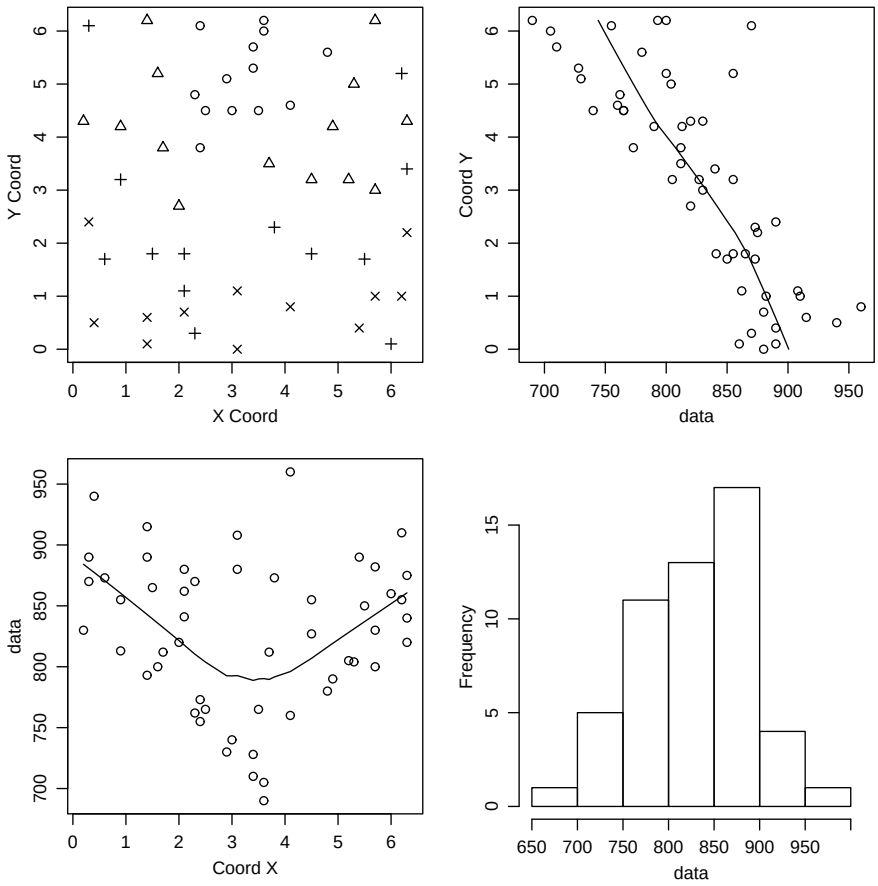

Figure 1.8. Point locations (top-left), data values against coordinates (top-right and bottom-left) and histogram (bottom-right) of the measured elevations.

green, yellow and red, respectively. The use of four distinct colours is the default for this function.

1.6.2 More on the `geodata` *object*

The functions `read.geodata()` and `as.geodata()` store a geostatistical dataset in a particular format called a `geodata` object. A `geodata` object is a list which has two obligatory components: a matrix with the two-dimensional coordinates (`coords`) of the sampling design and a vector giving the corresponding measured value at each of the locations in the design (`data`). Four additional, optional components are: a matrix with coordinates defining the boundary of the polygonal study area (`borders`); a vector or data-frame with covariates (`covariate`); an offset variable (`units.m`); and a vector indexing the number of the realisation of the process if more than one is available (`realisation`), as for instance for data collected at different time points. These additional

components, if present, are then used automatically by some of the **geoR** functions.

The example below shows the components of some of the data-sets which are included in the **geoR** package as `geodata` objects.

```
> names(elevation)

$coords
[1] "x" "y"

$data
[1] "data"

> data(parana)
> names(parana)

$coords
[1] "east"  "north"

$data
[1] "data"

$other
[1] "borders"    "loci.paper"

> data(ca20)
> names(ca20)

$coords
[1] "east"  "north"

$data
[1] "data"

$covariate
[1] "altitude" "area"

$other
[1] "borders" "reg1"    "reg2"    "reg3"

> names(unclass(ca20))

[1] "coords"    "data"      "covariate" "borders"   "reg1"
[6] "reg2"      "reg3"
```

The slightly different results returned from the calls `names(ca20)` and `names(unclass(ca20))` illustrate that some special *methods* have been provided to modify the way that standard R functions handle `geodata` objects; in this case the standard command `names(ca20)` recognises that `ca20` is a `geodata` object, and invokes the non-standard method `names.geodata()` whereas the command `unclass(ca20)` gives the standard result of the `names` function by removing the class `geodata` from the object `ca20`.

Other, perhaps more useful methods to facilitate data manipulation are also implemented such as `as.data.frame.geodata()` which converts a `geodata` object to a data-frame and `subset.geodata()` which facilitates extracting subsets of `geodata` objects. Below we illustrate the usage of `subset.geodata()` on the `ca20` data-set selecting data only within sub-area 3 in the first command and selecting only data greater than 70 in the second.

```
> ca20.3 <- subset(ca20, area == 3)
> ca20.g70 <- subset(ca20, data > 70)
```

1.6.3 Rongelap data

Our next example produces a circle plot for the Rongelap data, together with an enlarged inset of the western part of the island. The `rongelap` data-set is included with the **geoRglm** package.

```
> require(geoRglm)
> data(rongelap)
```

The response to the command `names(rongelap)` reveals that the rongelap `geodata` object has four components: `coords` contains the spatial coordinates; `data` contains the photon emission counts y_i attributable to radioactive caesium; `units.m` is an off-set variable which gives the values of t_i, the time (in seconds) over which y_i was accumulated; `borders` contains the coordinates of a digitisation of the island's coastline. The function `summary()` recognises and summarises all four components.

```
> names(rongelap)

$coords
NULL

$data
[1] "data"

$units.m
[1] "units.m"

$other
[1] "borders"

> summary(rongelap)

Number of data points: 157

Coordinates summary
    Coord.X Coord.Y
min   -6050   -3430
max     -50       0
```

```
Distance summary
     min      max
  40.000 6701.895

Borders summary
           [,1]       [,2]
min -6299.31201 -3582.2500
max    20.37916   103.5414

Data summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     75    1975    2639    3011    3437   21390

Offset variable summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  200.0   300.0   300.0   401.9   400.0  1800.0
```

We can use `points()` to visualise the data on a map of the study area as shown in Figure 1.2. For the enlargement of the western part of the island, we have used `subarea()` to select a subset of the original data-set whose spatial coordinates lie within a specified sub-area. The function `subarea()` accepts arguments `xlim` and/or `ylim` defining a rectangular sub-area. If these arguments are not provided the user is prompted to click on two points which then define the opposite corners of the required rectangular area. To produce the Figure, we use the following sequence of commands.

```
> points(rongelap)
> rongwest <- subarea(rongelap, xlim = c(-6300, -4800))
> rongwest.z <- zoom.coords(rongwest, xzoom = 3.5, xoff = 2000,
+     yoff = 3000)
> points(rongwest.z, add = T)
> rect.coords(rongwest$sub, lty = 2, quiet = T)
> rect.coords(rongwest.z$sub, lty = 2, quiet = T)
> text(-4000, 1100, "western area", cex = 1.5)
```

The object `rongwest` is a `geodata` object which is generated by `subarea()`. It has the same components as the original `geodata` object but is restricted to the area whose x-coordinates are in the range -6300 to -4800; because the `ylim` argument was not used, the y-coordinate range is unrestricted.

Note that, by default, if the element `units.m` is present in the data object, as for this case, the size of the circle plotted at each location is determined by the corresponding emission count per unit time, rather than by the emission count itself. Setting `data=rongelap$data` the effect of the argument is that the raw data on emission count would be plotted. If preferred, the argument `pt.div="equal"` could be used to specify that all the points should have the same size. The coastline is included in the plot by default because the element `borders` is present in the `geodata` object. If this is unwanted the argument

`borders` can be set to `NULL`. Alternatively, another object with the polygon defining the region bondaries can be passed using this argument.

1.6.4 The Gambia malaria data

The Gambia malaria data shown in Example 1.3 are available as a `data-frame` in the **geoR** package. The commands below load the data and display the first three lines of the resulting data-frame, with variable names printed at the head of each column of data.

```
> data(gambia)
> gambia[1:3, ]
```

```
            x       y pos  age netuse treated green phc
1850 349631.3 1458055   1 1783      0       0 40.85   1
1851 349631.3 1458055   0  404      1       0 40.85   1
1852 349631.3 1458055   0  452      1       0 40.85   1
```

Each line corresponds to one child. The columns are the coordinates of the village where the child lives (x and y), whether or not the child tested positive for malaria (*pos*), their age in days (*age*), usage of bed-net(*netuse*), whether the bed-net is treated with insecticide (*treated*), the vegetation index measured at the village location (*green*) and the presence or absence of a health centre in the village (*phc*).

To display the data as show in Figure 1.3 we use the `gambia.map()` function which is also included in **geoR**.

```
> gambia.map()
```

1.6.5 The soil data

The soil data shown in Example 1.4 are included in **geoR** and can be loaded with the commands `data(ca20)` and `data(camg)`. The former loads only the calcium data, stored as a `geodata` object, whereas the latter loads a data-frame which includes both the calcium and the magnesium data. In order to produce the right-hand panel in Figure 1.5 we use the sequence of commands below.

```
> data(camg)
> mg20 <- as.geodata(camg, data.col = 6)
> points(mg20, cex.min = 0.2, cex.max = 1.5, pch = 21)
> data(ca20)
> polygon(ca20$reg1, lty = 2)
> polygon(ca20$reg2, lty = 2)
> polygon(ca20$reg3, lty = 2)
```

The first command loads the combined data using `data()`, the second creates a `geodata` object for plotting the magnesium data. Borders of the region and sub-regions included in the plot use extra information provided in the calcium data object `ca20`, which is included in the **geoR** package.

We now inspect the ca20 object in more detail using the summary() function. Remember that *help(ca20)* gives the documentation for this data-set.

```
> summary(ca20)

Number of data points: 178

Coordinates summary
    east north
min 4957  4829
max 5961  5720

Distance summary
       min        max
  43.01163 1138.11774

Borders summary
    east north
min 4920  4800
max 5990  5800

Data summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   43.00   50.50   50.68   58.00   78.00

Covariates summary
    altitude     area
 Min.   :3.300   1: 14
 1st Qu.:5.200   2: 48
 Median :5.650   3:116
 Mean   :5.524
 3rd Qu.:6.000
 Max.   :6.600

Other elements in the geodata object
[1] "reg1" "reg2" "reg3"
```

The output above shows that the data contain 178 locations, with E-W coordinates ranging from 4957 to 5961 and N-S coordinates ranging from 4829 to 5720. The minimum distance between any two locations is about 43 units and the maximum 1138. The object also has a borders component which is a two-column matrix with rows corresponding to a set of coordinates defining the polygonal boundary of the study-area. The function also shows summary statistics for the response variable and for the covariates. For the covariate area the summary indicates that 14, 48 and 116 locations lie within the sub-areas 1, 2 and 3, respectively.

1.7 Exercises

1.1. Produce a plot of the Rongelap data in which a continuous colour-scale or grey-scale is used to indicate the value of the emission count per unit time at each location, and the two sub-areas with the 5 by 5 sub-grids at 50 metre spacing are shown as insets.

1.2. Construct a polygonal approximation to the boundary of The Gambia. Construct plots of the malaria data which show the spatial variation in the values of the observed prevalence in each village, and of the green-ness covariate.

1.3. Consider the elevation data as a simple regression problem with elevation as the response and north-south location as the explanatory variable. Fit the standard linear regression model using ordinary least squares. Examine the residuals from the linear model, with a view to deciding whether any more sophisticated treatment of the spatial variation in elevation might be necessary.

1.4. Find a geostatistical data-set which interests you.

(a) What scientific questions are the data intended to address? Do these concern estimation, prediction or testing?
(b) Identify the *study-region*, the *design*, the *response* and the *covariates*, if any.
(c) What is the *support* of each response?
(d) What is the underlying *signal*?
(e) If you wished to predict the signal throughout the study-region, would you choose to interpolate the response data?

1.5. Load the Paraná data set using the command `data(parana)` and inspect its documentation using `help(parana)`. For these data, consider the same questions as were raised in Exercise 1.4.

2
An overview of model-based geostatistics

The aim of this chapter is to provide a short overview of model-based geostatistics, using the elevation data of Example 1.1 to motivate the various stages in the analysis. Although this example is very limited from a scientific point of view, its simplicity makes it well-suited to the task in hand. Note, however, that Handcock & Stein (1993) show how to construct a useful explanatory variable for these data using a map of streams which run through the study-region.

2.1 Design

Statistical design is concerned with deciding what data to collect in order to address a question, or questions, of scientific interest. In this chapter, we shall assume that the scientific objective is to produce a map of surface elevation within a square study region whose side-length is 6.7 units, or 335 feet ($\approx$ 102 meters); we presume that this study-region has been chosen for good reason, either because it is of interest in its own right, or because it is representative of some wider spatial region.

In this simple setting, there are essentially only two design questions: at how many locations should we measure the elevation? and where should we place these locations within the study-region?

In practice, the answer to the first question is usually dictated by limits on the investigator's time and/or any additional cost in converting each field sample into a measured value. For example, some kinds of measurements involve expensive off-site laboratory assays whereas others, such as surface elevation, can be measured directly in the field. For whatever reason, the answer in this example is 52.

For the second question, two obvious candidate designs are a *completely random* design or a *completely regular* design. In the former, the locations x_i form an independent random sample from the uniform distribution over the study area, i.e. a homogeneous planar Poisson process (Diggle, 2003, chapter 1). In the latter, the x_i form a regular lattice pattern over the study-region. Classical sampling theory (Cochran 1977) tends to emphasise the virtue of some form of random sampling to ensure unbiased estimation of underlying population characteristics, whereas spatial sampling theory (Matérn 1960) shows that under typical modelling assumptions spatial properties are more efficiently estimated by a regular design. A compromise, which the originators of the surface elevation data appear to have adopted, is to use a design which is more regular than the completely random design but not as regular as a lattice.

Lattice designs are widely used in applications. The convenience of lattice designs for field-work is obvious, and provided there is no danger that the spacing of the lattice will match an underlying periodicity in the spatial phenomenon being studied, lattice designs are generally efficient for spatial prediction (Matérn 1960). In practice, the rigidity and simplicity of a lattice design also provide some protection against sub-conscious bias in the placing of the x_i. Note in this context that, strictly, a regular lattice design should mean a lattice whose origin is located at random, to guard against any subjective bias. The soil data of Example 1.4 provide an example of a regular lattice design.

Even more common in some areas of application is the *opportunistic design*, whereby geostatistical data are collected and analysed using an existing network of locations x_i which may have been established for quite different purposes. Designs of this kind often arise in connection with environmental monitoring. In this context, individual recording stations may be set up to monitor pollution levels from particular industrial sources or in environmentally sensitive locations, without any thought initially that the resulting data might be combined in a single, spatial analysis. This immediately raises the possibility that the design may be preferential, in the sense discussed in Section 1.2.3. Whether they arise by intent or by accident, preferential designs run the risk that a standard geostatistical analysis may produce misleading inferences about the underlying continuous spatial variation.

2.2 Model formulation

We now consider model formulation – unusually before, rather than after, exploratory data analysis. In practice, clean separation of these two stages is rare. However, in our experience it is useful to give some consideration to the kind of model which, in principle, will address the questions of interest before refining the model through the usual iterative process of data analysis followed by reformulation of the model as appropriate.

For the surface elevation data, the scientific question is a simple one – how can we use the measured elevations to construct our best guess (or, in more formal language, to predict) the underlying elevation surface throughout the study-

region? Hence, our model needs to include a real-valued, spatially continuous stochastic process, $S(x)$ say, to represent the surface elevation as a function of location, x. Depending on the nature of the terrain, we may want $S(x)$ to be continuous, differentiable or many-times differentiable. Depending on the nature of the measuring device, or the skill of its operator, we may also want to allow for some discrepancy between the true surface elevation $S(x_i)$ and the measured value Y_i at the design location x_i. The simplest statistical model which meets these requirements is a stationary Gaussian model, which we define below. Later, we will discuss some of the many possible extensions of this model which increase its flexibility.

We denote a set of geostatistical data in its simplest form, i.e. in the absence of any explanatory variables, by $(x_i, y_i) : i = 1, \ldots, n$ where the x_i are spatial locations and y_i is the measured value associated with the location x_i. The assumptions underlying the stationary Gaussian model are:

1. $\{S(x) : x \in \mathbb{R}^2\}$ is a Gaussian process with mean μ, variance $\sigma^2 = \mathrm{Var}\{S(x)\}$ and correlation function $\rho(u) = \mathrm{Corr}\{S(x), S(x')\}$, where $u = ||x - x'||$ and $||\cdot||$ denotes distance;
2. conditional on $\{S(x) : x \in \mathbb{R}^2\}$, the y_i are realisations of mutually independent random variables Y_i, Normally distributed with conditional means $\mathrm{E}[Y_i|S(\cdot)] = S(x_i)$ and conditional variances τ^2.

The model can be defined equivalently as

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

where $\{S(x) : x \in \mathbb{R}^2\}$ is defined by assumption 1 above and the Z_i are mutually independent $\mathrm{N}(0, \tau^2)$ random variables. We favour the superficially more complicated conditional formulation for the joint distribution of the Y_i given the signal, because it identifies the model explicitly as a special case of the generalized linear geostatistical model which we introduced in Section 1.4.

In order to define a legitimate model, the correlation function $\rho(u)$ must be positive-definite. This condition imposes non-obvious constraints so as to ensure that, for any integer m, set of locations x_i and real constants a_i, the linear combination $\sum_{i=1}^{m} a_i S(x_i)$ will have non-negative variance. In practice, this is usually ensured by working within one of several standard classes of parametric model for $\rho(u)$. We return to this question in Chapter 3. For the moment, we note only that a flexible, two-parameter class of correlation functions due to Matérn (1960) takes the form

$$\rho(u; \phi, \kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}K_{\kappa}(u/\phi) \tag{2.1}$$

where $K_{\kappa}(\cdot)$ denotes the modified Bessel function of the second kind, of order κ. The parameter $\phi > 0$ determines the rate at which the correlation decays to zero with increasing u. The parameter $\kappa > 0$ is called the *order* of the Matérn model, and determines the differentiability of the stochastic process $S(x)$, in a sense which we shall make precise in Chapter 3.

Our notation for $\rho(u)$ presumes that $u \geq 0$. However, the correlation function of any stationary process must by symmetric in u, hence $\rho(-u) = \rho(u)$.

The stochastic variation in a physical quantity is not always well described by a Normal distribution. One of the simplest ways to extend the Gaussian model is to assume that the model holds after applying a transformation to the original data. For positive-valued response variables, a useful class of transformations is the Box-Cox family (Box & Cox 1964):

$$Y^* = \begin{cases} (Y^\lambda - 1)/\lambda & : \quad \lambda \neq 0 \\ \log Y & : \quad \lambda = 0 \end{cases} \tag{2.2}$$

Another simple extension to the basic model is to allow a spatially varying mean, for example by replacing the constant μ by a linear regression model for the conditional expectation of Y_i given $S(x_i)$, so defining a spatially varying mean $\mu(x)$.

A third possibility is to allow $S(x)$ to have non-stationary covariance structure. Arguably, most spatial phenomena exhibit some form of non-stationarity, and the stationary Gaussian model should be seen only as a convenient approximation to be judged on its usefulness rather than on its strict scientific provenance.

2.3 Exploratory data analysis

Exploratory data analysis is an integral part of modern statistical practice, and geostatistics is no exception. In the geostatistical setting, exploratory analysis is naturally oriented towards the preliminary investigation of spatial aspects of the data which are relevant to checking whether the assumptions made by any provisional model are approximately satisfied. However, non-spatial aspects can and should also be investigated.

2.3.1 Non-spatial exploratory analysis

For the elevation data in Example 1.1 the 52 data values range from 690 to 960, with mean 827.1, median 830 and standard deviation 62. A histogram of the 52 elevation values (Figure 2.1) indicates only mild asymmetry, and does not suggest any obvious outliers. This adds some support to the use of a Gaussian model as an approximation for these data. Also, because geostatistical data are, at best, a correlated sample from a common underlying distribution, the shape of their histogram will be less stable than that of an independent random sample of the same size, and this limits the value of the histogram as a diagnostic for non-Normality.

In general, an important part of exploratory analysis is to examine the relationship between the response and available covariates, as illustrated for the soil data in Figure 1.7. For the current example, the only available covariates to consider are the spatial coordinates themselves.

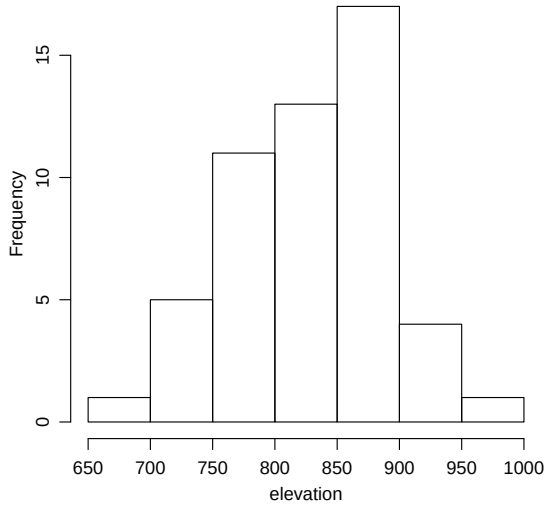

Figure 2.1. Histogram of the surface elevation data.

2.3.2 Spatial exploratory analysis

The first stage in spatial exploratory data analysis is simply to plot the response data in relation to their locations, for example using a circle plot as shown for the surface elevation data in Figure 1.1. Careful inspection of this plot can reveal spatial outliers, i.e. responses which appear grossly discordant with their spatial neighbours, or spatial trends which might suggest the need to include a trend surface model for a spatially varying mean, or perhaps qualitatively different behaviour in different sub-regions.

In our case, the most obvious feature of Figure 1.1 is the preponderance of large response values towards the southern end of the study region. This suggests that a trend surface term in the model might be appropriate. In some applications, the particular context of the data might suggest that there is something special about the north-south direction – for example, for applications on a large geographical scale, we might expect certain variables relating to the physical environment to show a dependence on latitude. Otherwise, our view would be that if a trend surface is to be included in the model at all, then both of the spatial coordinates should contribute to it because the orientation of the study region is essentially arbitrary.

Scatterplots of the response variable against each of the spatial coordinates can sometimes reveal spatial trends more clearly. Figure 2.2 show the surface elevations plotted against each of the coordinates, with lowess smooths (Cleveland, 1979, 1981) added to help visualisation. These plots confirm the north-south trend whilst additionally suggesting a less pronounced, non-monotone east-west trend, with higher responses concentrated towards the eastern and western edges of the study-region.

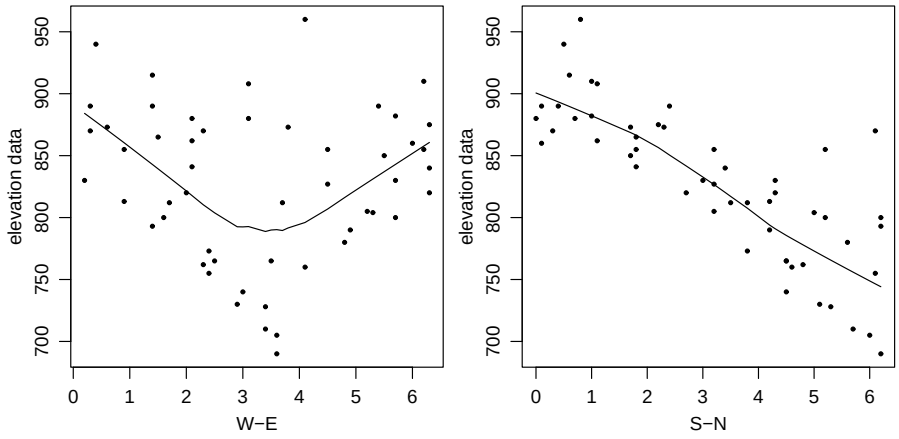

Figure 2.2. Elevation data against the coordinates.

When interpreting plots of this kind it can be difficult, especially when analysing small data-sets, to distinguish between a spatially varying mean response and correlated spatial variation about a constant mean. Strictly speaking, without independent replication the distinction between a deterministic function $\mu(x)$ and the realisation of a stochastic process $S(x)$ is arbitrary. Operationally, we make the distinction by confining ourselves to "simple" functions $\mu(x)$, for example low-order polynomial trend surfaces, using the correlation structure of $S(x)$ to account for more subtle patterns of spatial variation in the response. In Chapter 5 we shall use formal, likelihood-based methods to guide our choice of model for both mean and covariance structure. Less formally, we interpret spatial effects which vary on a scale comparable to or greater than the dimensions of the study-region as variation in $\mu(x)$ and smaller-scale effects as variation in $S(x)$. This is in part a pragmatic strategy, since covariance functions which do not decay essentially to zero at distances shorter than the dimensions of the study region will be poorly identified, and in practice indistinguishable from spatial trends. Ideally, the model for the trend should also have a natural physical interpretation; for example, in an investigation of the dispersal of pollutants around a known source, it would be natural to model $\mu(x)$ as a function of the distance, and possibly the orientation, of x relative to the source.

To emphasise this point, the three panels of Figure 2.3 compare the original Figure 1.1 with circle plots of residuals after fitting linear and quadratic trend surface models by ordinary least squares. If we assume a constant spatial mean for the surface elevations themselves, then the left-hand panel of Figure 2.3 indicates that the elevations must be very strongly spatially correlated, to the extent that the correlation persists at distances beyond the scale of the study region. As noted above, fitting a model of this kind to the data would result in poor identification of parameters describing the correlation structure. If, in contrast, we use a linear trend surface to describe a spatially varying mean, then the central panel of Figure 2.3 still suggests spatial correlation because

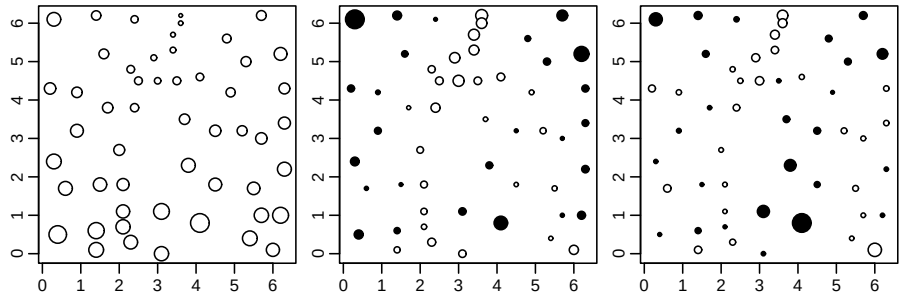

Figure 2.3. Circle plot of the surface elevation data. The left-hand panel shows the original data. The center and right-hand panels show the residuals from first-order (linear) and second-order (quadratic) polynomial trend surfaces, respectively, using empty and filled circles to represent negative and positive residuals and circle radii proportional to the absolute values of the residuals.

positive and negative residuals tend to occur together, but the scale of the spatial correlation is smaller. The right-hand panel of 2.3 has a qualitatively similar appearance to the centre panel, but the range of the residuals has been reduced, because some additional variation is taken up by the quadratic terms in the fitted trend surface. The range of the residuals is from -61.1 to $+110.7$ in the centre panel, and from -63.3 to $+97.8$ in the right-hand panel.

Notwithstanding the above discussion, visual assessment of spatial correlation from a circle plot is difficult. For a sharper assessment, a useful exploratory tool is the *empirical variogram*. We discuss theoretical and empirical variograms in more detail in Chapters 3 and 5, respectively. Here, we give only a brief description.

For a set of geostatistical data $(x_i, y_i) : i = 1, \ldots, n$, the *empirical variogram ordinates* are the quantities $v_{ij} = \frac{1}{2}(y_i - y_j)^2$. For obvious reasons, some authors refer to these as the *semi-variogram ordinates*. If the y_i have spatially constant mean and variance, then v_{ij} has expectation $\sigma^2\{1 - \rho(x_i, x_j)\}$ where σ^2 is the variance and $\rho(x_i, x_j)$ denotes the correlation between y_i and y_j. If the y_i are generated by a stationary spatial process, then $\rho(\cdot)$ depends only on the distance between x_i and x_j and typically approaches zero at large distances, hence the expectation of the v_{ij} approaches a constant value, σ^2, as the distance u_{ij} between x_i and x_j increases. If the y_i are uncorrelated, then all of the v_{ij} have expectation σ^2. These properties motivate the definition of the *empirical variogram* as a plot of v_{ij} against the corresponding distance u_{ij}. A more easily interpretable plot is obtained by averaging the v_{ij} within distance bands.

The left-hand panel of Figure 2.4 shows a variogram for the original surface elevations, whilst the right-hand panel shows variograms for residuals from the linear and quadratic trend-surface models, indicated by solid and dashed lines, respectively. In the left-hand panel, the variogram increases throughout the plotted range, indicating that *if* these data were generated by a stationary stochastic process, then the range of its spatial correlation must extend beyond

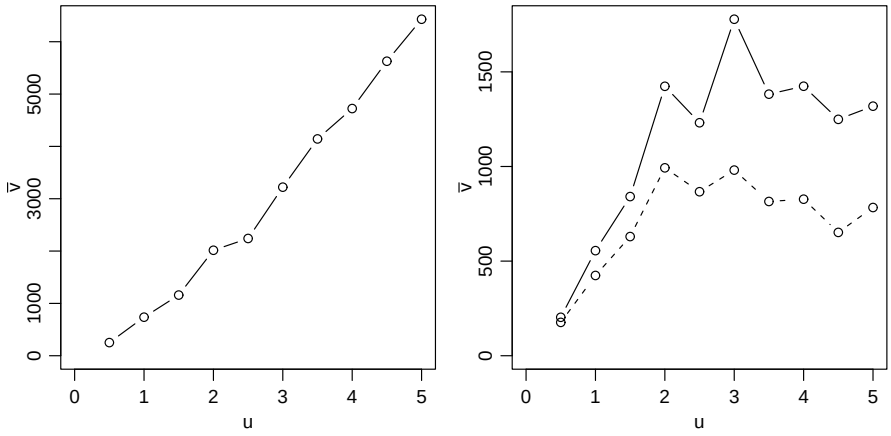

Figure 2.4. Empirical variograms for the original data (left-panel) and for residuals (right panel) from a linear (solid lines) or quadratic (dashed lines) trend surface. In all three cases, empirical variogram ordinates have been averaged in bins of unit width.

the scale of the study-region. Pragmatically, including a spatially varying mean is a better modelling strategy. The solid line on right hand panel shows behaviour more typical of a stationary, spatially correlated process, i.e. an initial increase levelling off as the correlation decays to zero at larger distances. Finally, the shape of the variogram in the dashed line on the right-hand panel is similar to the solid one but its range is smaller by a factor of about 0.6. The range of values in the ordinates of the empirical variogram is approximately equal to the variance of the residuals, hence the reduction in range again indicates how the introduction of progressively more elaborate models for the mean accounts for correspondingly more of the empirical variation in the original data. Note also that in all panels of Figure 2.4 the empirical variogram approaches zero at small distances. This indicates that surface elevation is being measured with negligible error, relative to either the spatial variation in the surface elevation itself (left-hand panel), or the residual spatial variation about the linear or quadratic trend surface (right-hand panel). This interpretation follows because the expectation of v_{ij} corresponding to two independent measurements, y_i and y_j, at the same location is simply the variance of the measurement error.

We emphasise that, for reasons explained in Chapter 5, we prefer to use the empirical variogram only as an exploratory tool, rather than as the basis for formal inference. With this proviso, Figure 2.4 gives a strong indication that a stationary model is unsuitable for these data, whereas the choice between the linear and quadratic trend-surface models is less clear-cut.

When an empirical variogram appears to show little or no spatial correlation, it can be useful to assess more formally whether the data are compatible with an underlying model of the form $y_i = \mu(x_i) + z_i$ where the z_i are uncorrelated residuals about a spatially varying mean $\mu(x)$. A simple way to do this is to compute residuals about a fitted mean $\hat{\mu}(x)$ and to compare the residual empirical variogram with the envelope of empirical variograms com-

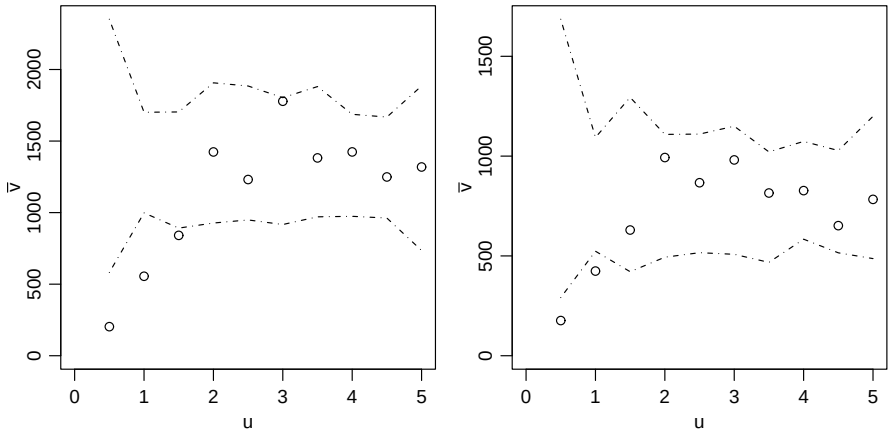

Figure 2.5. Monte Carlo envelopes for the variogram of ordinary least squares residuals of the surface elevation data after fitting linear (left-hand panel) or quadratic (right-hand panel) trend surface models.

puted from random permutations of the residuals, holding the corresponding locations fixed. The left-hand panel of Figure 2.5 shows a variogram envelope obtained from 99 independent random permutations of the residuals from a linear trend surface fitted to the surface elevations by ordinary least squares. This shows that the increasing trend in the empirical variogram is statistically significant, confirming the presence of positive spatial correlation. The same technique applied to the residuals from the quadratic trend surface produces the diagram shown as the right-hand panel of Figure 2.5. This again indicates significant spatial correlation, although the result is less clear-cut than before, as the empirical variogram ordinates at distances 0.5 and 1.0 fall much closer to the lower simulation envelope than they do in the left-hand panel.

2.4 The distinction between parameter estimation and spatial prediction

Before continuing with our illustrative analysis of the surface elevation data, we digress to expand on the distinction between estimation and prediction.

Suppose that $S(x)$ represents the level of air pollution at the location x, that we have observed (without error, in this hypothetical example) the values $S_i = S(x_i)$ at a set of locations $x_i : i = 1, \ldots, n$ forming a regular lattice over a spatial region of interest, A, and that we wish to learn about the average level of pollution over the region A. An intuitively reasonable estimate is the sample mean,

$$\bar{S} = n^{-1} \sum_{i=1}^{n} S_i. \tag{2.3}$$

What precision should we attach to this estimate?

Suppose that $S(x)$ has aconstant expectation, $\theta = \mathrm{E}[S(x)]$ for any location x in A. One possible interpretation of $\bar{S}$ is as an estimate of θ, in which case an appropriate measure of precision is the mean square error, $\mathrm{E}[(\bar{S} - \theta)^2]$. This is just the variance of $\bar{S}$, which we can calculate as

$$n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(S_i, S_j). \tag{2.4}$$

For a typical geostatistical model, the correlation between any two S_i and S_j will be either zero or positive, and (2.4) will therefore be larger than the naive expression for the variance of a sample mean, σ^2/n where $\sigma^2 = \mathrm{Var}\{S(x)\}$.

If we regard $\bar{S}$ as a *predictor* of the *spatial average*,

$$S_A = |A|^{-1} \int_A S(x)dx,$$

where $|A|$ is the area of A, then the mean square prediction error is $\mathrm{E}[(\bar{S} - S_A)^2]$. Noting that S_A is a random variable, we write this as

$$\begin{aligned} \mathrm{E}[(\bar{S} - S_A)^2] &= n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(S_i, S_j) \\ &+ |A|^{-2} \int_A \int_A \mathrm{Cov}\{S(x), S(x')\} dx dx' \\ &- 2(n|A|)^{-1} \sum_{i=1}^{n} \int_A \mathrm{Cov}\{S(x), S(x_i)\} dx. \end{aligned} \tag{2.5}$$

In particular, the combined effect of the second and third terms on the right hand side of (2.5) can easily be to make the mean square prediction error smaller than the naive variance formula. For example, if we increase the sample size n by progressively decreasing the spacing of the lattice points x_i, (2.5) approaches zero, whereas (2.4) does not.

2.5 Parameter estimation

For the stationary Gaussian model, the parameters to be estimated are the mean μ and any additional parameters which define the covariance structure of the data. Typically, these include the signal variance σ^2, the conditional or measurement error variance τ^2 and one or more correlation function parameters ϕ.

In geostatistical practice, these parameters can be estimated in a number of different ways which we shall discuss in detail in Chapter 5. Our preference here is to use the method of maximum likelihood within the declared Gaussian model.

For the elevation data, if we assume a stationary Gaussian model with a Matérn correlation function and a fixed value $\kappa = 1.5$, the maximum likelihood

estimates of the remaining parameters are $\hat{\mu} = 848.3$, $\hat{\sigma}^2 = 3510.1$, $\hat{\tau}^2 = 48.2$ and $\hat{\phi} = 1.2$.

However, our exploratory analysis suggested a model with a non-constant mean. Here, we assume a linear trend surface,

$$\mu(x) = \beta_0 + \beta_1 d_1 + \beta_2 d_2$$

where d_1 and d_2 are the north-south and east-west coordinates. In this case the parameter estimates are $\hat{\beta}_0 = 912.5$, $\hat{\beta}_1 = -5$, $\hat{\beta}_2 = -16.5$, $\hat{\sigma}^2 = 1693.1$, $\hat{\tau}^2 = 34.9$ and $\hat{\phi} = 0.8$. Note that because the trend surface accounts for some of the spatial variation, the estimate of σ^2 is considerably smaller than for the stationary model, and similarly for the parameter ϕ which corresponds to the range of the spatial correlation. As anticipated, for either model the estimate of τ^2 is much smaller than the estimate of σ^2. The ratio of $\hat{\tau}^2$ to $\hat{\sigma}^2$ is 0.014 for the stationary model, and 0.021 for the linear trend surface model.

2.6 Spatial prediction

For prediction of the underlying, spatially continuous elevation surface we shall here illustrate perhaps the simplest of all geostatistical methods: *simple kriging*. In our terms, simple kriging is minimum mean square error prediction under the stationary Gaussian model, but ignoring parameter uncertainty, i.e. estimates of all model parameters are plugged into the prediction equations as if they were the true parameter values. As discussed earlier, we do not claim that this is a good model for the surface elevation data.

The minimum mean square error predictor, $\hat{S}(x)$ say, of $S(x)$ at an arbitrary location x is the function of the data, $y = (y_1, \ldots, y_n)$, which minimises the quantity $\mathrm{E}[\{\hat{S}(x) - S(x)\}^2]$. A standard result, which we discuss in Chapter 6, is that $\hat{S}(x) = \mathrm{E}[S(x)|y]$. For the stationary Gaussian process, this conditional expectation is a linear function of the y_i, namely

$$\hat{S}(x) = \mu + \sum_{i=1}^{n} w_i(x)(y_i - \mu) \tag{2.6}$$

where the $w_i(x)$ are explicit functions of the covariance parameters σ^2, τ^2 and ϕ.

The top-left panel of Figure 2.6 gives the result of applying (2.6) to the surface elevation data, using as values for the model parameters the maximum likelihood estimates reported in Section 2.5, whilst the bottom-left panel shows the corresponding prediction standard errors, $SE(x) = \sqrt{\mathrm{Var}\{S(x)|y\}}$. The predictions follow the general trend of the observed elevations whilst smoothing out local irregularities. The prediction variances are generally small at locations close to the sampling locations, because $\hat{\tau}^2$ is relatively small; had we used the value $\tau^2 = 0$ the prediction standard error would have been exactly zero at each sampling location and the predicted surface $\hat{S}(x)$ would have interpolated the observed responses y_i.

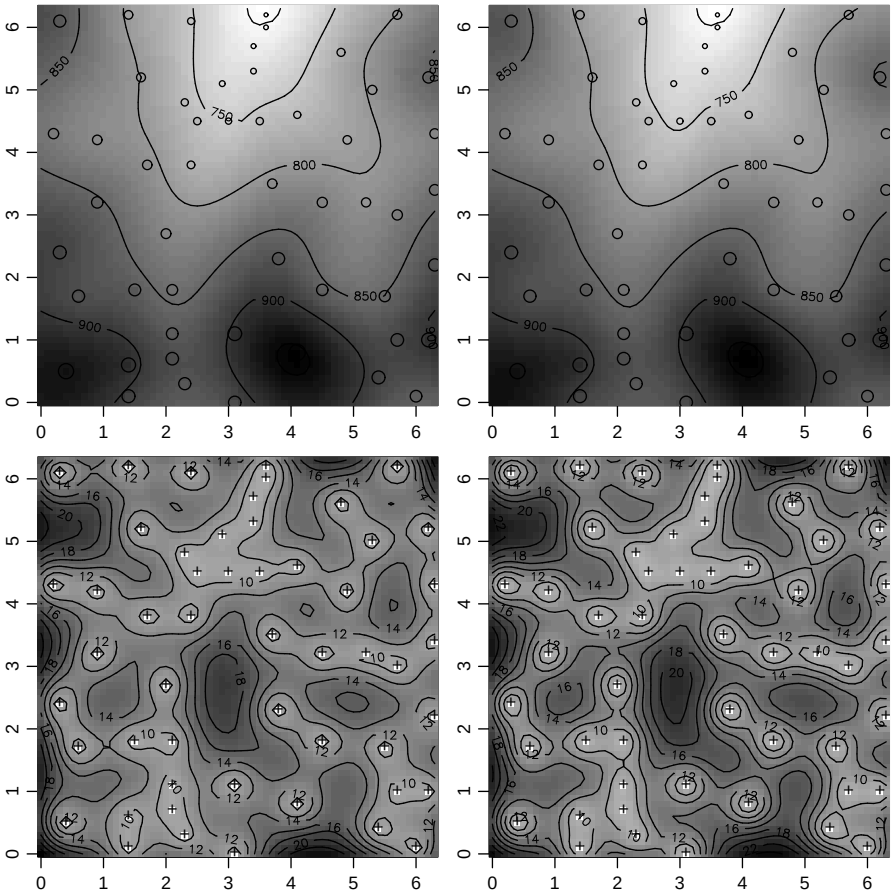

Figure 2.6. Simple kriging predictions for the surface elevation data. The top-left panel shows the simple kriging predictor as a grey-scale image and contour plot; sampling locations are plotted as circles with radii proportional to observed elevations. The bottom-left panel shows the prediction standard deviations; sampling locations are plotted as small crosses. The top-right and bottom-right panels give the same information, but based on the model with a linear trend-surface.

It is straightforward to adapt the simple kriging formula (2.6) to incorporate a spatially varying mean. We simply replace the constant μ on the right-hand-side of (2.6) by a spatial trend, $\mu(x)$. If we do this, using the linear trend surface model and its associated maximum likelihood parameter estimates we obtain the results summarised in the top-right and bottom-right panels of Figure 2.6. The plots corresponding to the two different models are directly comparable because they use a common grey-scale within each pair. Note in particular that in this simple example, the dubious assumption of stationarity has not prevented the simple kriging methodology from producing a predicted surface which captures qualitatively the apparent spatial trend in the data, and which is almost identical to the predictions obtained using the more reasonable linear trend

surface model. The two models produce somewhat different prediction standard errors; these range between 0 and 25.5 for the stationary model, between 0 and 24.4 for the model with the linear trend surface and between 0 and 22.9 for the model with the quadratic trend surface. The differences amongst the three models are rather small. They are influenced by several different aspects of the data and model, including the data-configuration and the estimated values of the model parameters. In other applications, the choice of model may have a stronger impact on the predictive inferences we make from the data, even when this choice does not materially affect the point predictions of the underlying surface $S(x)$. Note also that the plug-in standard errors quoted here do not account for parameter uncertainty.

2.7 Definitions of distance

A fundamental stage in any geostatistical analysis is to define the metric for calculating the distance between any two locations. By default, we use the standard planar Euclidean distance, i.e. the "straight-line distance" between two locations in $\mathbb{R}^2$. Non-Euclidean metrics may be more appropriate for some applications. For example, Rathbun (1998) discusses the measurement of distance between points in an estuarine environment where, arguably, two locations which are close in the Euclidean metric but separated by dry land should not be considered as near neighbours. It is not difficult to think of other settings where natural barriers to communication might lead the investigator to question whether it is reasonable to model spatial correlation in terms of straight-line distance.

Even when straight-line distance is an appropriate metric, if the study-region is geographically extensive, distances computed between points on the earth's surface should strictly be great-circle distances, rather than straight-line distances on a map projection. Using (θ, ϕ) to denote a location in degrees of longitude and latitude, and treating the earth as a sphere of radius $r = 6378$ kilometres, the great-circle distance between two locations is

$$r \cos^{-1}\{\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\theta_1 - \theta_2)\}.$$

Section 3.2 of Waller & Gotway (2004) gives a nice discussion of this issue from a statistical perspective. Banerjee (2005) examines the effect of distance computations on geostatistical analysis and concludes that the choice of metric may influence the resulting inferences, both for parameter estimation and for prediction. Note in particular that degrees of latitude and longitude represent approximately equal distances only close to the equator.

Distances calculations are especially relevant to modelling spatial correlation, hence parameters which define the correlation structure are particularly sensitive to the choice of metric. Furthermore, the Euclidean metric plays an integral part in determining valid classes of correlation functions using Bochner's theorem (Stein 1999). Our **geoR** software implementation only calculates planar Euclidean distances.

2.8 Computation

The non-spatial exploratory analysis of the surface elevation data reported in this chapter uses only built-in R functions as follows.

```
> with(elevation, hist(data, main = "", xlab = "elevation"))
> with(elevation, plot(coords[, 1], data, xlab = "W-E",
+     ylab = "elevation data", pch = 20, cex = 0.7))
> lines(lowess(elevation$data ~ elevation$coords[, 1]))
> with(elevation, plot(coords[, 2], data, xlab = "S-N",
+     ylab = "elevation data", pch = 20, cex = 0.7))
> lines(with(elevation, lowess(data ~ coords[, 2])))
```

To produce circle plots of the residual data we use the **geoR** function `points.geodata()`, which is invoked automatically when a `geodata` object is passed as an argument to the built-in function `points()`, as indicated below. The argument `trend` defines a linear model on the covariates from which the residuals are extracted for plotting. The values `"1st"` and `"2nd"` passed to the argument `trend` are aliases to indicate first and second degree polynomials on the coordinates. More details and other options to specify the trend are discussed later in this Section and in the documentation for `trend.spatial()`. Setting `abs=T` instructs the function to draw the circles with radii proportional to the absolute values of the residuals.

```
> points(elevation, cex.max = 2.5)
> points(elevation, trend = "1st", pt.div = 2, abs = T,
+     cex.max = 2.5)
> points(elevation, trend = "2nd", pt.div = 2, abs = T,
+     cex.max = 2.5)
```

To calculate and plot the empirical variograms shown in Figure 2.4 for the original data and for the residuals, we use `variog()`. The argument `uvec` defines the classes of distance used when computing the empirical variogram, whilst `plot()` recognises that its argument is a variogram object, and automatically invokes `plot.variogram()`. The argument `trend` is used to indicate that the variogram should be calculated from the residuals about a fitted trend surface.

```
> plot(variog(elevation, uvec = seq(0, 5, by = 0.5)),
+     type = "b")
> res1.v <- variog(elevation, trend = "1st", uvec = seq(0,
+     5, by = 0.5))
> plot(res1.v, type = "b")
> res2.v <- variog(elevation, trend = "2nd", uvec = seq(0,
+     5, by = 0.5))
> lines(res2.v, type = "b", lty = 2)
```

To obtain the residual variogram and simulation envelopes under random permutation of the residuals, as shown in Figure 2.5, we proceed as in the following example. By default, the function uses 99 simulations, but this can be changed using the optional argument `nsim`.

```
> set.seed(231)
> mc1 <- variog.mc.env(elevation, obj = res1.v)
> plot(res1.v, env = mc1, xlab = "u")
> mc2 <- variog.mc.env(elevation, obj = res2.v)
> plot(res2.v, env = mc2, xlab = "u")
```

To obtain maximum likelihood estimates of the Gaussian model, with or without a trend term, we use the **geoR** function `likfit()`. Because this function uses a numerical maximisation procedure, the user needs to provide initial values for the covariance parameters, using the argument `ini`. In this example we use the default value 0 for the parameter τ^2, in which case `ini` specifies initial values for the parameters σ^2 and ϕ. Initial values are not required for the mean parameters.

```
> ml0 <- likfit(elevation, ini = c(3000, 2), cov.model = "matern",
+     kappa = 1.5)
> ml0

likfit: estimated model parameters:
      beta      tausq    sigmasq        phi
" 848.317" "  48.157" "3510.096" "   1.198"

likfit: maximised log-likelihood = -242.1
```

```
> ml1 <- likfit(elevation, trend = "1st", ini = c(1300,
+     2), cov.model = "matern", kappa = 1.5)
> ml1

likfit: estimated model parameters:
      beta0       beta1       beta2       tausq     sigmasq
" 912.4865" "  -4.9904" " -16.4640" "  34.8953" "1693.1329"
        phi
"   0.8061"

likfit: maximised log-likelihood = -240.1
```

To carry out the spatial interpolation using simple kriging we first define, and store in the object `locs`, a grid of locations at which predictions of the values of the underlying surface are required. The function `krige.control()` then defines the model to be used for the interpolation, which is carried out by `krige.conv()`. In the example below, we first obtain predictions for the stationary model, and then for the model with a linear trend on the coordinates. If required, the user can restrict the trend surface model, for example by specifying a linear trend is the north-south direction. However, as a general rule we prefer our inferences to be invariant to the particular choice of coordinate axes, and would therefore fit both linear trend parameters or, more generally, full polynomial trend surfaces.

```
> locs <- pred_grid(c(0, 6.3), c(0, 6.3), by = 0.1)
> KC <- krige.control(type = "sk", obj.mod = ml0)
```

```
> sk <- krige.conv(elevation, krige = KC, loc = locs)
> KCt <- krige.control(type = "sk", obj.mod = ml1, trend.d = "1st",
+     trend.l = "1st")
> skt <- krige.conv(elevation, krige = KCt, loc = locs)
```

Finally, we use a selection of built-in graphical functions to produce the maps shown in Figure 2.6, using optional arguments to the graphical functions to ensure that pairs of corresponding plots use the same grey-scale.

```
> pred.lim <- range(c(sk$pred, skt$pred))
> sd.lim <- range(sqrt(c(sk$kr, skt$kr)))
> image(sk, col = gray(seq(1, 0, l = 51)), zlim = pred.lim)
> contour(sk, add = T, nlev = 6)
> points(elevation, add = TRUE, cex.max = 2)
> image(skt, col = gray(seq(1, 0, l = 51)), zlim = pred.lim)
> contour(skt, add = T, nlev = 6)
> points(elevation, add = TRUE, cex.max = 2)
> image(sk, value = sqrt(sk$krige.var), col = gray(seq(1,
+     0, l = 51)), zlim = sd.lim)
> contour(sk, value = sqrt(sk$krige.var), levels = seq(10,
+     27, by = 2), add = T)
> points(elevation$coords, pch = "+")
> image(skt, value = sqrt(skt$krige.var), col = gray(seq(1,
+     0, l = 51)), zlim = sd.lim)
> contour(skt, value = sqrt(skt$krige.var), levels = seq(10,
+     27, by = 2), add = T)
> points(elevation$coords, pch = "+")
```

In **geoR**, covariates which define a linear model for the mean response can be specified by passing additional arguments to plotting or model-fitting functions. In the examples above, we used `trend="1st"` or `trend="2nd"` to specify a linear or quadratic trend surface. However, these are simply short-hand aliases to formulae which define the corresponding linear models, and are provided for users' convenience. For example, the *model formula* `trend=~coords[,1] + coords[,2]` would produce the same result as `trend="1st"`. The `trend` argument will also accept a matrix representing the design matrix of a general linear model, or the output of the trend definition function, `trend.spatial()`. For example, the call below to `plot()` can be used in order to inspect the data after taking out the linear effect of the north-south coordinate. By setting the argument `trend=~coords[,2]` the function fits a standard linear model on this covariate and uses the residuals to produce the plots shown in Figure 2.7, rather than plotting the original response data. Similarly, we could fit a quadratic function on the x-coordinate by setting `trend=~coords[,2] + poly(coords[,1], degree=2)`. We invite the reader to experiment with different options for the argument `trend` and `trend.spatial()`. The procedure of taking out the effect of a covariate is sometimes called *trend removal*.

```
> plot(elevation, low = TRUE, trend = ~coords[, 2], qt.col = 1)
```

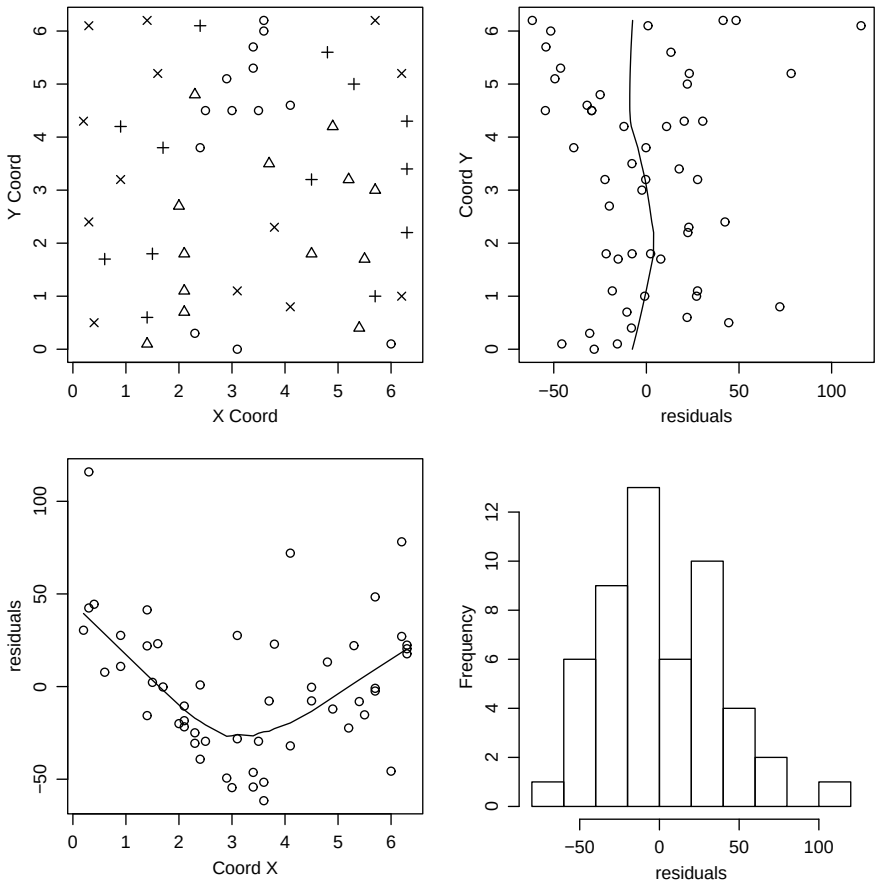

Figure 2.7. Output of `plot.geodata()` when setting the argument `trend=~coords[,2]`.

The trend argument can also be used to take account of covariates other than functions of the coordinates. For example, the data set `ca20` included in **geoR** stores the calcium content from soil samples, as discussed in Example 1.4, together with associated covariate information. Recall that in this example the study region is divided in three sub-regions with different histories of soil management. The covariate `area` included in the data-set indicates for each datum the sub-region in which it was collected. Figure 2.8 shows the exploratory plot for the residuals after removing a separate mean for calcium content in each sub-region. This diagram was produced using the following code.

```
> data(ca20)
> plot(ca20, trend = ~area, qt.col = 1)
```

The plotting functions in `geoR` also accept an optional argument `lambda` which specifies the numerical value for the parameter of the Box-Cox family

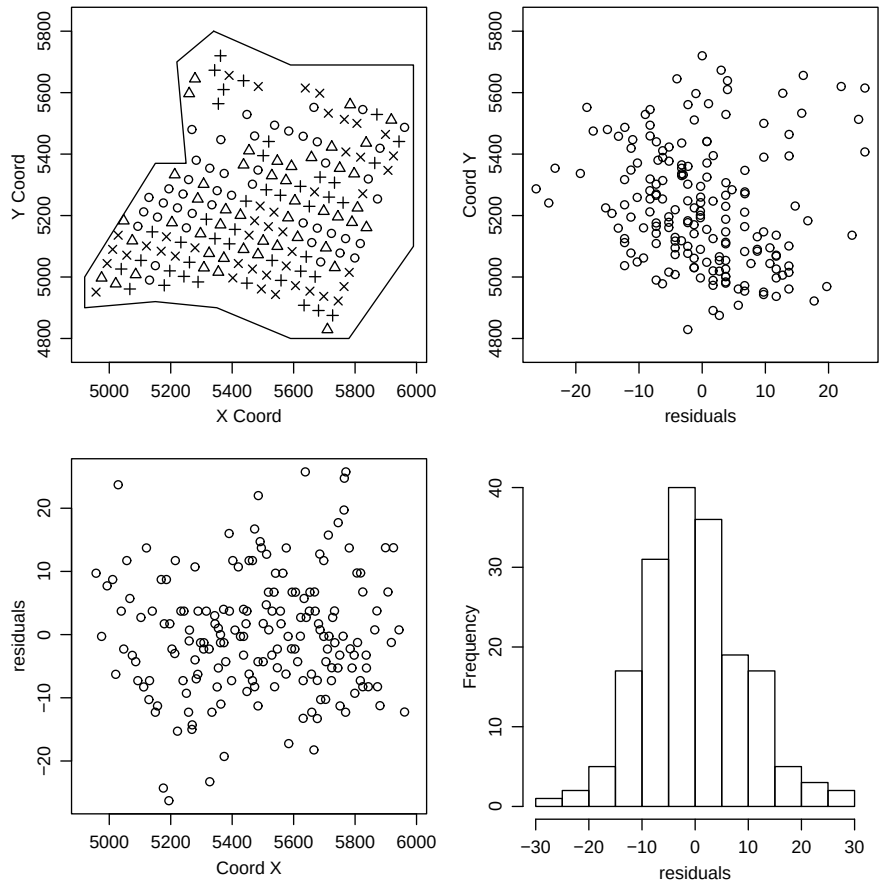

Figure 2.8. Exploratory plot for the `ca20` data-set obtained when setting `trend=~area`.

of transformations, with default `lambda=1` corresponding to no transformation. For example, the command

```
> plot(ca20, lambda = 0)
```

sets the Box-Cox transformation parameter to $\lambda = 0$, which will then produce plots using the logarithm of the original response variable.

2.9 Exercises

2.1. Investigate the R packages **splancs** or **spatstat**, both of which provide functions for the analysis of spatial point pattern data. Use either of these packages to confirm (or not, as the case may be) that the design used for the surface elevation data is more regular than a completely random design.

2.2. Consider the following two models for a set of responses, $Y_i : i = 1, \ldots, n$ associated with a sequence of positions $x_i : i = 1, \ldots, n$ along a one-dimensional spatial axis x.

(a) $Y_i = \alpha + \beta x_i + Z_i$, where α and β are parameters and the Z_i are mutually independent with mean zero and variance σ^2_Z.

(b) $Y_i = A + Bx_i + Z_i$ where the Z_i are as in (a) but A and B are now random variables, independent of each other and of the Z_i, each with mean zero and respective variances σ^2_A and σ^2_B.

For each of these models, find the mean and variance of Y_i and the covariance between Y_i and Y_j for any $j \neq i$. Given a single realisation of either model, would it be possible to distinguish between them?

2.3. Suppose that $Y = (Y_1, \ldots, Y_n)$ follows a multivariate Normal distribution with $\mathrm{E}[Y_i] = \mu$ and $\mathrm{Var}\{Y_i\} = \sigma^2$ and that the covariance matrix of Y can be expressed as $V = \sigma^2 R(\phi)$. Write down the log-likelihood function for $\theta = (\mu, \sigma^2, \phi)$ based on a single realisation of Y and obtain explicit expressions for the maximum likelihood estimators of μ and σ^2 when ϕ is known. Discuss how you would use these expressions to find maximum likelihood estimators numerically when ϕ is unknown.

2.4. Load the `ca20` data-set with *`data(ca20)`*. Check the data-set documentation with *`help(ca20)`*. Perform an exploratory analysis of these data. Would you include a trend term in the model? Would you recommend a data transformation? Is there evidence of spatial correlation?

2.5. Load the Paraná data with *`data(parana)`* and repeat Exercise 2.4.

References

Azzalini, A. (1996). *Statistical Inference: Based on the Likelihood*, Chapman and Hall, London.

Baddeley, A. & Vedel Jensen, E. B. (2005). *Stereology for Statisticians*, Chapman and Hall/CRC, Boca Raton.

Banerjee, S. (2005). On geodetic distance computations in spatial modeling, *Biometrics* **61**: 617–625.

Banerjee, S., Wall, M. M. & Carlin, B. P. (2003). Frailty modelling for spatially correlated survival data, with application to infant mortality in minnesota, *Biostatistics* pp. 123–142.

Barry, J., Crowder, M. & Diggle, P. J. (1997). Parametric estimation using the variogram, *Technical Report ST-97-06*, Dept. Maths and Stats, Lancaster University, Lancaster, UK.

Bartlett, M. S. (1955). *Stochastic Processes*, Cambridge University Press.

Bartlett, M. S. (1964). A note on spatial pattern, *Biometrics* pp. 891–892.

Bartlett, M. S. (1967). Inference and stochastic process, *Journal of the Royal Statistical Society, Series A* pp. 457–477.

Bellhouse, D. R. (1977). Some optimal designs for sampling in two dimensions, *Biometrika* **64**: 605–611.

Ben-jamma, F., Marino, M. & Loaiciga, H. (1995). Sampling design for contaminant distribution in lake sediments, *Journal of Water Resources Planning and Managmeent* **121**: 71–79.

Benes, V., Bodlak, K., Møller, J. & Waagepetersen, R. P. (2001). Bayesian analysis of log gaussian cox process models for disease mapping, *Technical Report Research Report R-02-2001*, Department of Mathematical Sciences, Aalborg University.

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. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society, Series B* **36**: 192–225.

Besag, J. & Mondal, D. (2005). First-order intrinsic autoregressions and the de Wijs process, *Biometrika* **92**: 909–920.

Boussinesq, M., Gardon, J., Kamgno, J., Pion, S. D., Gardon-Wendel, N. & Chippaux, J. P. (2001). Relationships between the prevalence and intensity of *loa loa* infection in the Central province of Cameroon, *Annals of Tropical Medicine and Parasitology* **95**: 495–507.

Bowman, A. W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*, Oxford University Press.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B* **26**: 211–252.

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.

Brix, A. & Møller, J. (2001). Space-time multitype log Gaussian Cox processes with a view to modelling weed data, *Scandinavian Journal of Statistics* **28**: 471–488.

Capeche, C. L. e. (1997). Caracterização pedológica da fazenda angra - pesagro/rio - estação experimental de campos (rj), *Informação, globalização, uso do solo*, Vol. 26, Congresso Brasileiro de Ciência do Solo, Embrapa/SBCS, Rio de Janeiro.

Chilès, J.-P. & Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*, Wiley, New York.

Christensen, O. (2001). *Methodology and applications in non-linear model based geostatistics*, PhD thesis, Aalborg University, Denmark.

Christensen, O. F. (2004). Monte Carlo maximum likelihood in model-based geostatistics, *Journal of Computational and Graphical Statistics* **13**: 702–718.

Christensen, O. F., Møller, J. & Waagepetersen, R. P. (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. F. & Ribeiro Jr., P. J. (2002). geoRglm: a package for generalised linear spatial models, *R-NEWS* pp. 26–28.
*http://cran.R-project.org/doc/Rnews

Christensen, O. F., Roberts, G. O. & Skøld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models, *Jornal of Computational and Graphical Statistics* **15**: 1–17.

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

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**: 829–836.

Cleveland, W. S. (1981). Lowess: A program for smoothing scatterplots by robust locally weighted regression, *The American Statistician* **35**: 54.

Cochran, W. G. (1977). *Sampling Techniques*, second edn, Wiley, New York.

Cox, D. R. (1955). Some statistical methods related with series of events (with discussion), *Journal of the Royal Statistical Society, Series B* **17**: 129–157.

Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**: 187–220.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Cox, D. R. & Miller, H. D. (1965). *The Theory of Stochastic Processes*, Methuen, London.

Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*, Chapman and Hall, London.

Cressie, N. (1985). Fitting variogram models by weighted least squares, *Mathematical Geology* **17**: 563–586.

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

Cressie, N. & Hawkins, D. M. (1980). Robust estimation of the variogram, *Mathematical Geology* **12**: 115–125.

Cressie, N. & Wikle, C. K. (1998). The variance-based cross-variogram: you can add apples and oranges, *Mathematical Geology* **30**: 789–799.

Dalgaard, P. (2002). *Introductory Statistics with R*, Springer. ISBN 0-387-95475-9.

Davis, J. C. (1972). *Statistics and Data Analysis in Geology*, second edn, Wiley, New York.

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

De Wijs, H. J. (1951a). Statistics of ore distribution. Part I. Frequency distribution of assay values, *Journal of the Royal Netherlands Geological and Mining Society* **13**: 365–375.

De Wijs, H. J. (1951b). Statistics of ore distribution. Part II. Theory of binomial distributions applied to sampling and engineering problems, *Journal of the Royal Netherlands Geological and Mining Society* **15**: 12–24.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*, second edn, Edward Arnold, London.

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 assessment and control*, Wiley, Chichester, pp. 89–107.

Diggle, P. J., Heagerty, P., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*, second edn, Oxford University Press, Oxford.

Diggle, P. J. & Lophaven, S. (2006). Bayesian geostatistical design, *Scandinavian Journal of Statistics* **33**: 55–64.

Diggle, P. J., Moyeed, R. A., Rowlingson, B. & Thomson, M. (2002). Childhood malaria in the Gambia: a case-study in model-based geostatistics, *Applied Statistics* **51**: 493–506.

Diggle, P. J., Ribeiro Jr, P. J. & Christensen, O. F. (2003). An introduction to model-based geostatistics, *in* J. Møller (ed.), *Spatial Statistics and Computational Methods*, Springer, pp. 43–86.

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

Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J., Boussinesq, M. & Molyneux, D. H. (2006). Spatial modeling and prediction of *loa loa*

risk: decision making under uncertainty, *International Journal of Epidemiology (submitted)* .

Diggle, P., Rowlingson, B. & Su, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance, *Environmetrics* **16**: 423–34.

Draper, N. & Smith, H. (1981). *Applied Regression Analysis*, second edn, Wiley, New York.

Dubois, G. (1998). Spatial interpolation comparison 97: foreword and introduction, *Journal of Geographic Information and Decision Analysis* **2**: 1–10.

Duchon, J. (1977). Splines minimising rotation-invariant semi-norms in Sobolev spaces, *in* W. Schempp & K. Zeller (eds), *Constructive Theory of Functions of Several Variables*, Springer, pp. 85–100.

Fedorov, V. V. (1989). Kriging and other estimators of spatial field characteristics, *Atmospheric Environment* **23**: 175–184.

Gelfand, A. E., Schmidt, A. M., Banerjee, S. & Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization (with discussion), *Test* **13**: 263–312.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003). *Bayesian Data Analysis*, second edn, Chapman and Hall, London.

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

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations, *Journal of the Royal Statistical Society, Series B* **56**: 261–274.

Geyer, C. J. & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion), *Journal of the Royal Statistical Society, Series B* **54**: 657–699.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Gneiting, T. (1997). *Symmetric Positive Definite Functions with Applications in Spatial Statistics*, PhD thesis, University of Bayreuth.

Gneiting, T., Sasvári, Z. & Schlather, M. (2001). Analogues and correspondences between variograms and covariance functions, *Advances in Applied Probability* **33**.

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.

Greig-Smith, P. (1952). The use of random and contiguous quadrats in the study of the structure of plant communities, *Annals of Botany* **16**: 293–316.

Guttorp, P., Meiring, W. & Sampson, P. D. (1994). A space-time analysis of of ground-level ozone data, *Environmetrics* **5**: 241–254.

Guttorp, P. & Sampson, P. D. (1994). Methods for estimating heterogeneous spatial covariance functions with environmental applications, *in* G. P. Patil & C. R. Rao (eds), *Handbook of Statistics X11: Environmental Statistics*, Elsevier/North Holland, New York, pp. 663–690.

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

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

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika* **61**: 383–385.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**: 97–109.

Henderson, R., Shimakura, S. E. & Gorst, D. (2002). Modelling spatial variation in leukaemia survival data, *Journal of the American Statistical Association* **97**: 965–972.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic ocean (with discussion), *Environmental and Ecological Statistics* **5**: 173–190.

Higdon, D. (2002). Space and space-time modelling using process convolutions, *in* C. Anderson, V. Barnett, P. C. Chatwin & A. H. El-Shaarawi (eds), *Quantitative Methods for Current Environmental Issues*, Wiley, Chichester, pp. 37–56.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer.

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

Kammann, E. E. & Wand, M. P. (2003). Geoadditive models, *Applied Statistics* **52**: 1–18.

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

Kitanidis, P. K. (1978). Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resources Research* **22**: 499–507.

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

Knorr-Held, L. & Best, N. (2001). A shared component model for detecting joint and selective clustering of two diseases., *Journal of the Royal Statistical Society, Series A* **164**: 73–85.

Kolmogorov, A. N. (1941). Interploation und extrapolation von stationären zufäligen folgen, *Izv. Akad. Nauk SSSR* **5**: 3–14.

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.

Lark, R. M. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood, *Geoderma* **105**: 49–80.

Laslett, G. M. (1994). Kriging and splines: an empirical comparison of their predictive performance in some applications, *Journal of the American Statistical Association* **89**: 391–409.

Lee, Y. & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society, Series B* **58**: 619–678.

Lee, Y. & Nelder, J. A. (2001). Modelling and analying correlated non-normal data., *Statistical Modelling* **1**: 3–16.

Li, Y. & Ryan, L. (2002). Modelling spatial survival data using semiparametric frailty models, *Biometrics* **58**: 287–297.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**: 13–22.

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*, Statens Skogsforsningsinstitut, Stockholm.

Matérn, B. (1986). *Spatial variation*, second edn, Springer, Berlin.

Matheron, G. (1963). Principles of geostatistics, *Economic geology* **58**: 1246–1266.

Matheron, G. (1971a). Random set theory and its application to stereology., *Journal of Microscopy* **95**: 15–23.

Matheron, G. (1971b). The theory of regionalized variables and its applications, *Technical Report 5*, Cahiers du Centre de Morphologie Mathematique.

Matheron, G. (1973). The intrinsic random functions and their applications, *Advances in Applied Probability* **5**: 508–541.

McBratney, A. B. & Webster, R. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalised variables. II. Program and examples., *Computers and Geosciences* **7**: 335–365.

McBratney, A. B. & Webster, R. (1986). Choosing functions for semi-variograms of soil properties and fitting them to sample estimates, *Journal of Soil Science* **37**: 617–639.

McBratney, A., Webster, R. & Burgess, T. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalised variables. I. Theory and methods., *Computers and Geosciences* **7**: 331–334.

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

Menezes, R. (2005). Assessing spatial dependency under non-standard sampling. Unpublished PhD Thesis.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machine, *Journal of Chemical Physics* **21**: 1087–91.

Møller, J., Syversveen, A. R. & Waagepetersen, R. P. (1998). Log-Gaussian Cox processes, *Scandinavian Journal of Statistics* **25**: 451–482.

Møller, J. & Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*, Chapman and Hall/CRC.

Muller, W. (1999). Least squares fitting from the variogram cloud, *Statistics and Probability Letters* **43**: 93–98.

Muller, W. G. & Zimmerman, D. L. (1999). Optimal designs for variogram estimation, *Environmetrics* **10**: 23–27.

Natarajan, R. & Kass, R. E. (2000). Bayesian methods for generalized linear mixed models, *Journal of the American Statistical Association* **95**: 222–37.

Naus, J. I. (1965). Clustering of random points in two dimensions, *Biometrika* **52**: 263–267.

Neal, P. & Roberts, G. O. (2006). Optimal scaling for partially updating MCMC algorithms, *Annals of Applied Probability* .

Nelder, J. A. & Wedderburn, R. M. (1972). Generalized linear models., *Journal of the Royal Statistical Society, Series A* **135**: 370–84.

O'Hagan, A. (1994). *Bayesian Inference*, Vol. 2b of *Kendall's Advanced Theory of Statistics*, Edward Arnold.

Omre, H. (1987). Bayesian kriging - merging observations and qualified guesses in kriging, *Mathematical Geology* **19**: 25–38.

Omre, H., Halvorsen, B. & Berteig, V. (1989). A Bayesian approach to kriging, *in* M. Armstrong (ed.), *Geostatistics*, Vol. I, pp. 109–126.

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.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.

Perrin, O. & Meiring, W. (1999). Identifiability for non-stationary spatial structure, *Journal of Applied Probability* **36**: 1244–1250.

R Development Core Team (2005). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Rathbun, S. L. (1996). Estimation of poisson intensity using partially observed concomitant variables, *Biometrics* **52**: 226–242.

Rathbun, S. L. (1998). Spatial modelling in irregularly shaped regions: kriging estuaries, *Environmetrics* **9**: 109–129.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion), *Journal of the Royal Statistical Society, Series B* **39**: 172–192.

Ripley, B. D. (1981). *Spatial Statistics*, Wiley, New York.

Ripley, B. D. (1987). *Stochastic Simulation*, Wiley, New York.

Ross, S. (1976). *A First Course in Probability*, Macmillan, New York.

Royle, J. A. & Nychka, D. (1988). An algorithm for the construction of spatial coverage designs with implementation in splus, *Computers and Geosciences* **24**: 479–88.

Rue, H. & Held, L. (2005). *Gaussan Markov Random Fields: Theory and Applications*, Chapman and Hall, London.

Rue, H. & Tjelmeland, H. (2002). Fitting Gaussian random fields to Gaussian fields, *Scandinavian Journal of Statistics* **29**: 31–50.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

Russo, D. (1984). Design of an optimal sampling network for estimating the variogram, *Soil Science Society of America Journal* **52**: 708–716.

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

Sarndal, C. E. (1978). Design-based and model-based inference in survey sampling (with discussion), *Scandinavian Journal of Statistics* **5**: 27–52.

Schlather, M. (1999). Introduction to positive definite functions and to unconditional simulation of random fields, *Technical Report ST-99-10*, Dept. Maths and Stats, Lancaster University, Lancaster, UK.

Schlather, M., Ribeiro Jr, P. J. & Diggle, P. J. (2004). Detecting dependence between marks and locations of marked point processes, *Journal of the Royal Statistical Society, Series B* **66**: 79–93.

Schmidt, A. M. & Gelfand, A. E. (2003). A bayesian corregionalization approach for multivariate pollutant data, *Journal of Geophysical Research – Atmospheres* **108 (D24)**: 8783.

Schmidt, A. M. & O'Hagan, A. (2003). Bayesian inference for nonstationary spatial covariance structures via spatial deformations, *Journal of the Royal Statistical Society, Series B* **65**: 743–758.

Serra, J. (1980). Boolean model and random sets, *Computer Graphics and Image Processing* .

Serra, J. (1982). *Image Analysis and Mathematical Morphology*, Academic Press, London.

Spruill, T. B. & Candela, L. (1990). Two approaches to design of monitoring networks, *Ground Water* **28**: 430–442.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.

Takougang, I.and Meremikwu, M., Wanji, S., Yenshu, E. V., Aripko, B., Lamlenn, S., Eka, B. L., Enyong, P., Meli, J., Kale, O. & Remme, J. H. (2002). Rapid assessment method for prevalence and intensity of *loa loa* infection, *Bulletin of the World Health Organisation* **80**: 852–858.

Tanner, M. (1996). *Tools for Statistical Inference*, Springer, New York.

Thomson, M. C., Connor, S. J., D'Alessandro, U., Rowlingson, B. S., Diggle, P. J., Cresswell, M. & Greenwood, B. M. (1999). Predicting malaria infection in Gambian children from satellite data and bednet use surveys: the importance of spatial correlation in the interpretation of results, *American Journal of Tropical Medicine and Hygiene* **61**: 2–8.

Thomson, M. C., Obsomer, V., Kamgno, J., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Remme, J. H., Molyneux, D. H. & Boussinesq, M. (2004). Mapping the distribution of *loa loa* in cameroon in support of the african programme for onchocerciasis control, *Filaria Journal* **3**: 7.

Van Groenigen, J. W., Pieters, G. & Stein, A. (2000). Optimizing spatial sampling for multivariate contamination in urban areas, *Environmetrics* **11**: 227–244.

Van Groenigen, J. W., Siderius, W. & Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance, *Geoderma* **87**: 239–259.

Van Groenigen, J. W. & Stein, A. (1998). Constrained optimisation of spatial sampling using continuous simulated annealing, *Journal of Environmental Quality* **27**: 1076–1086.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM.

Waller, L. A. & Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*, Wiley, New York.

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

Warrick, A. & Myers, D. (1987). Optimization of sampling locations for variogram calculations, *Water Resources Research* **23**: 496–500.

Watson, G. S. (1971). Trend-surface analyis, *Mathematical Geology* **3**: 215–226.

Watson, G. S. (1972). Trend surface analyis and spatial correlation, *Geology Society of America Special Paper* **146**: 39–46.

Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method, *Biometrika* **63**: 27–32.

Whittle, P. (1954). On stationary processes in the plane, *Biometrika* **41**: 434–49.

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**: 974–74.

Winkels, H. & Stein, A. (1997). Optimal cost-effective sampling for monitoring and dredging of contaminated sediments, *Journal of Environmental Quality* **26**: 933–946.

Wood, A. T. A. & Chan, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$, *Journal of Computational and Graphical Statistics* **3**: 409–432.

Wood, S. N. (2003). Thin plate regression splines, *Journal of the Royal Statistical Society B* **65**: 95–114.

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

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, *Journal of the American Statistical Association* **99**: 250–261.

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

Zimmerman, D. L. & Homer, K. E. (1991). A network design criterion for estimating selected attributes of the semivariogram, *Environmetrics* **4**: 425–441.

Zimmerman, D. L. & Zimmerman, M. B. (1991). A comparison of spatial semivariogram estimators and corresponding kriging predictors, *Technometrics* **33**: 77–91.