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

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.

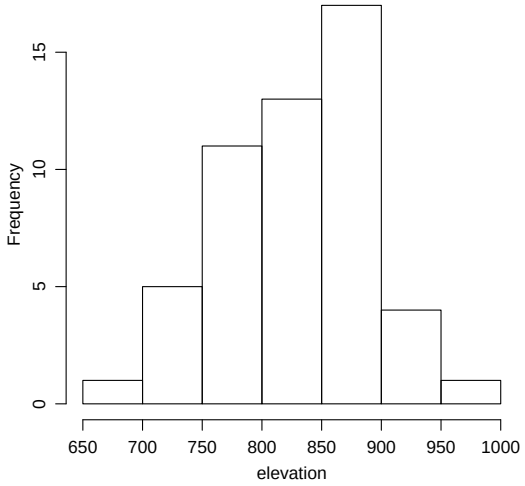

Figure 2.1. Histogram of the surface elevation data.

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.

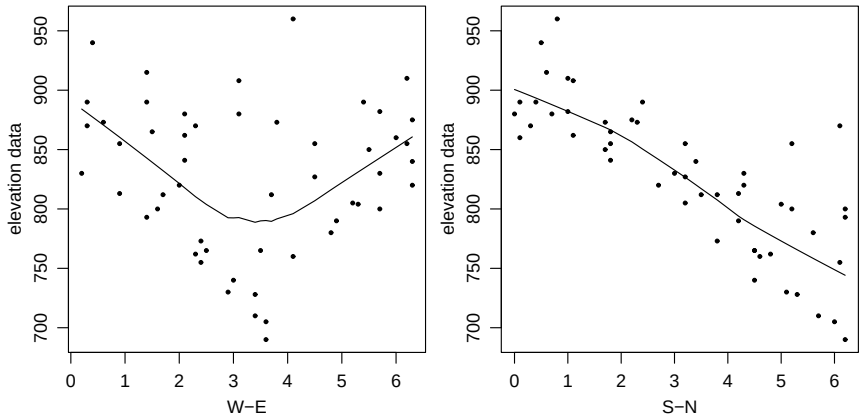

Figure 2.2. Elevation data against the coordinates.

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

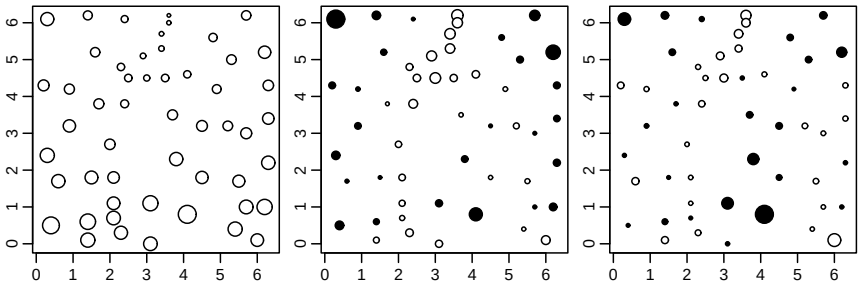

Figure 2.3. Circle plot of the surface elevation data. The left-hand panel shows the original data. The center and right-hand panels show the residuals from first-order (linear) and second-order (quadratic) polynomial trend surfaces, respectively, using empty and filled circles to represent negative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

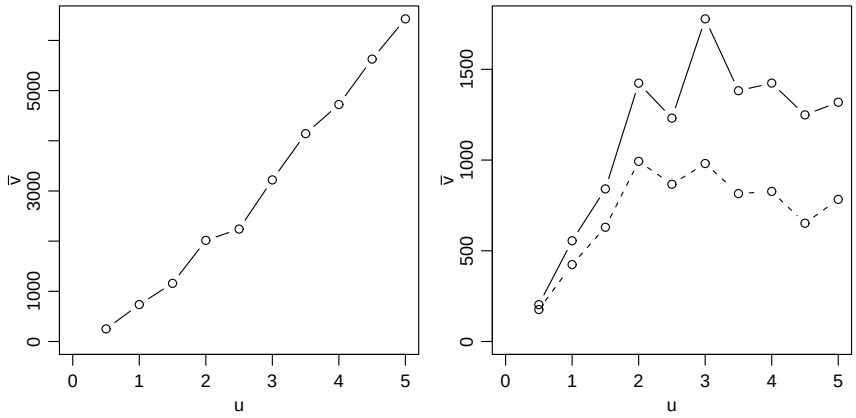

Figure 2.4. Empirical variograms for the original data (left-panel) and for residuals (right panel) from a linear (solid lines) or quadratic (dashed lines) trend surface. In all three cases, empirical variogram ordinates have been averaged in bins of unit width.

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-

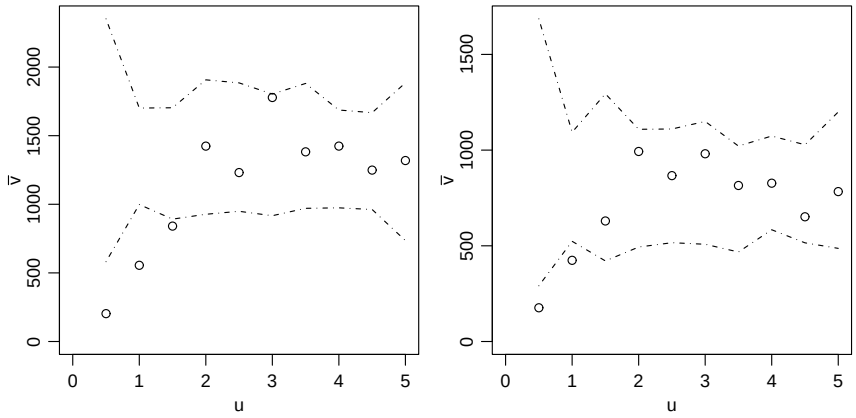

Figure 2.5. Monte Carlo envelopes for the variogram of ordinary least squares residuals of the surface elevation data after fitting linear (left-hand panel) or quadratic (right-hand panel) trend surface models.

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.

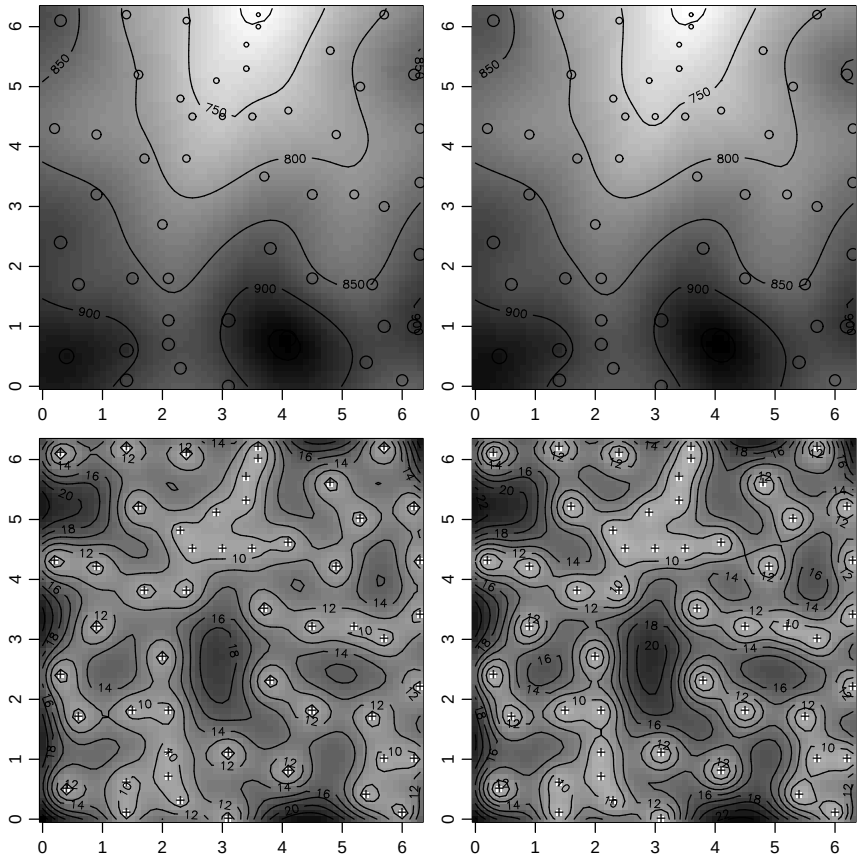

Figure 2.6. Simple kriging predictions for the surface elevation data. The top-left panel shows the simple kriging predictor as a grey-scale image and contour plot; sampling locations are plotted as circles with radii proportional to observed elevations. The bottom-left panel shows the prediction standard 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)
```

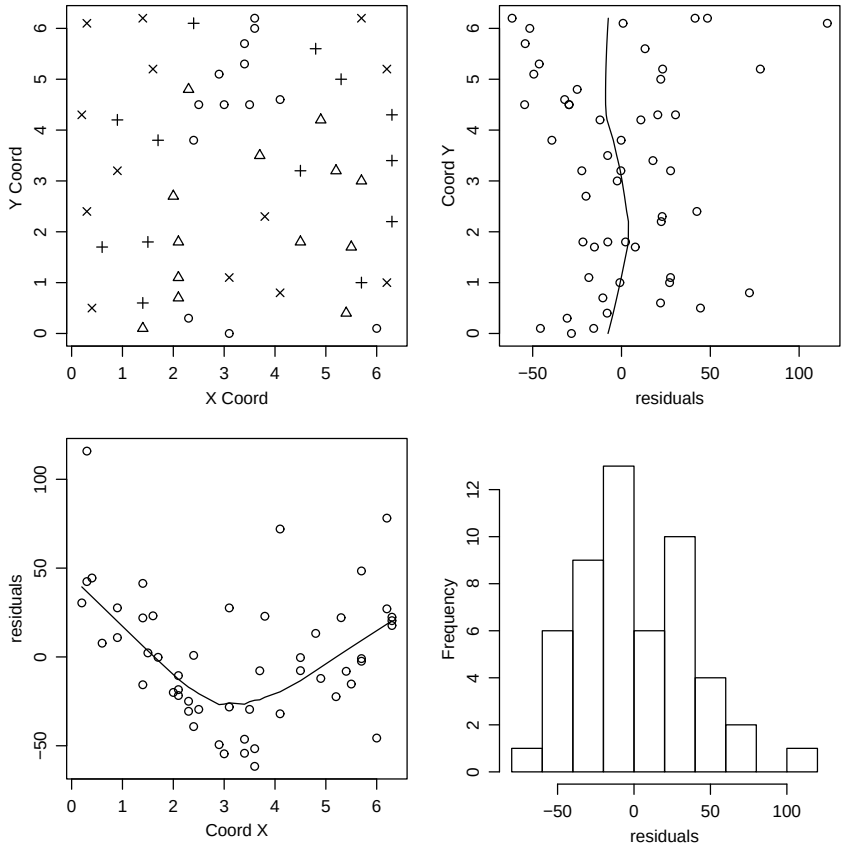

Figure 2.7. Output of `plot.geodata()` when setting the argument `trend=~coords[,2]`.

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

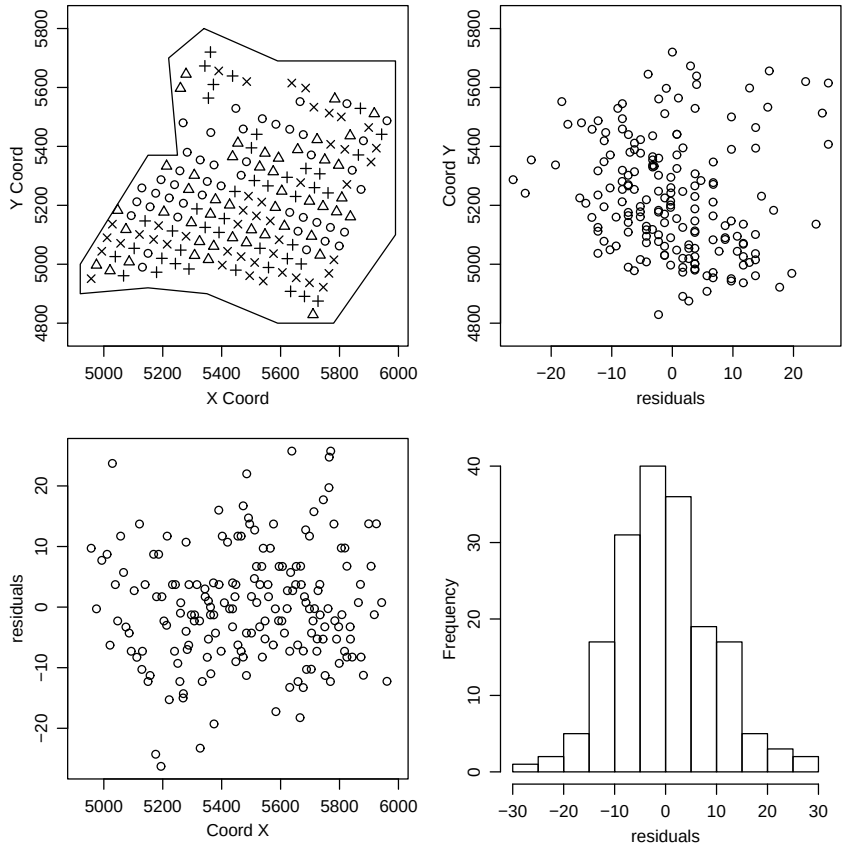

Figure 2.8. Exploratory plot for the `ca20` data-set obtained when setting `trend=~area`.

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.