Geographical & Environmental Modelling, Vol. 6, No. 2, 2002, 129–146

Carfax Publishing
Taylor & Francis Group

Bayesian Inference in Gaussian Model-based Geostatistics

PETER J. DIGGLE & PAULO J. RIBEIRO JR

ABSTRACT *In a geostatistical analysis, spatial interpolation or smoothing of the observed values is often carried out by a procedure known as kriging. In its basic form, kriging involves the construction of a linear predictor for an unobserved value of the process, and the form of this linear predictor is chosen with reference to the covariance structure of the data as estimated by a data-analytic tool known as the variogram. Often, no explicit underlying stochastic model is declared. We adopt a model-based approach to this class of problems, by which we mean that we start with an explicit stochastic model and derive associated methods of parameter estimation, interpolation and smoothing by the application of general statistical principles. In particular, we use Bayesian methods of inference so as to make proper allowance for the uncertainty associated with estimating the unknown values of model parameters. To illustrate the model-based approach we analyse data on precipitation levels in Paraná State, Brazil.*

1. Introduction

Geostatistical analysis has been used for spatial prediction and uncertainty assessment in many subject areas including geographical and environmental modelling. Standard methods are implemented in some GIS and surface fitting packages. Typically these methods do not use a fully specified stochastic model and do not take into account the uncertainty in parameter estimates when computing prediction intervals. In this paper we discuss a class of models for geostatistical problems and derive prediction results which take into account the uncertainty in the model parameters. In particular, we adopt the Bayesian paradigm to derive predictive distributions conditional on the data, and present an algorithm for sampling from this predictive distribution. The targets for prediction can either be the values of the variable on a grid (for construction of a map of the variable over the area) or any other linear or non-linear functional, for example the probability of exceedance of a given threshold.

A fundamental goal of geostatistical analysis is to predict the unobserved values of a spatial field $S(x)$ using data of the form $(Y_i, x_i) : i = 1, \ldots, n$, where Y_i is a noisy version of $S(x_i)$. As an example, consider rainfall data collected at recording stations

Peter J. Diggle, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK. Fax: + 44-(0)1524-592681 ; Email: p.diggle@lancaster.ac.uk
Paulo J. Ribeiro Jr, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK. E-mail: p.ribeiro@lancaster.ac.uk

1361-5939 print/1469-8323 online/02/020129-18
DOI: 10.1080/1365930220000294678

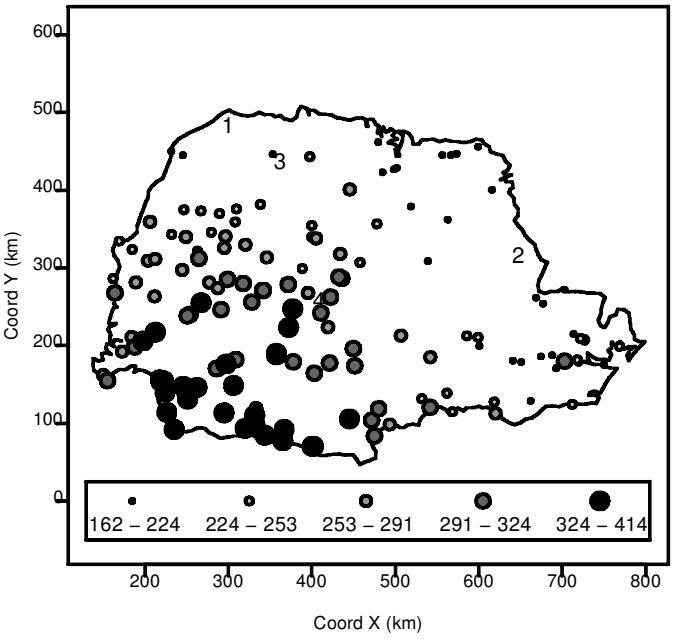

Figure 1. Map of Paraná showing the locations of the recording stations and the corresponding recorded rainfall values; numbers indicate particular prediction locations.

within a certain area, as illustrated by the data in Figure 1. We use the vector x_i to denote the spatial coordinates (latitude, longitude) which identify the data locations, and Y_i to denote the recorded value of the rainfall variable. This recorded value may be subject to measurement error, and we use $S(x)$ to denote the true rainfall value at an arbitrary location x. If there is no measurement error, then $Y_i = S(x_i)$. The surface $S(x)$ corresponds to the ideal map of true rainfall values over the area.

A common way to present the results of this kind of analysis is as a map of the predicted surface $\hat{S}(x)$ over the region of interest. It is then tempting to read off this map the values of quantities of interest in the particular application. This can be misleading if the quantity of interest is a non-linear property of $S(x)$. Suppose, for example, that we are interested in the maximum value of $S(x)$. Because the predicted surface is obtained by smoothing the data it is almost certain that the maximum of $\hat{S}(x)$ will under-estimate the true maximum, perhaps substantially so. An important secondary consideration is the precision with which any quantity of interest can be predicted. In practice, there are at least two components to prediction uncertainty: the inherent uncertainty in the true value of $S(x)$ when the stochastic mechanism which generates the data is known exactly; and the additional uncertainty when the generating mechanism is unknown. Conventional geostatistical methods ignore the second of these.

In this paper we shall describe a model-based approach to problems of this kind. This approach begins by specifying a stochastic model for the field $S(x)$ and the data. We shall use a linear Gaussian model which is a special case of the models considered in Diggle *et al.* (1998). We assume that the locations x_i associated with the noisy values Y_i are known exactly. Thus, the stochastic model can be specified by a sub-model for $\{S(x): x \in A\}$ where A is the region of interest, and a sub-model for $Y = (Y_1, \ldots, Y_n)$ conditional on $\{S(x): x \in A\}$.

Once the model has been specified, we use general statistical principles to fit it to the observed data and to make predictions. In particular, we use Bayesian methods as in Kitanidis (1986) and Handcock and Stein (1993), so that the predictive distribution which we attach to any quantity of interest makes proper allowance for the uncertainty inherent in estimating model parameter values from the data. Very briefly, classical and Bayesian model-based prediction proceeds as follows. Let S denote the underlying spatial process, Y the data and θ the set of parameters which define the model. We use the notation $[\cdot]$ to mean the distribution of the quantity within square brackets, and a vertical bar to indicate conditioning. Our hierarchical model specifies the two conditional distributions $[S \mid \theta]$ and $[Y \mid S,\theta]$, and hence the joint distribution $[Y,S \mid \theta]$. In this specification of the model, θ is an unknown constant, but we include it in the conditioning set for the joint distribution of Y and S so as to clarify the comparison between the classical and Bayesian approaches. Assuming that the value of θ is known, the classical predictive distribution for S is the corresponding conditional distribution $[S \mid Y,\theta]$, which is obtained by a standard application of Bayes' theorem:

$$[S \mid Y, \theta] = [Y, S \mid \theta]/[Y] = [S \mid \theta][Y \mid S, \theta]/\int [S \mid \theta][Y \mid S, \theta]\, dS.$$

In practice, the value of θ is unknown and is replaced by an estimate as if this estimate were the true value, thus ignoring the uncertainty in estimation of θ. The Bayesian paradigm treats θ as a random variable and the Bayesian predictive distribution is $[S \mid Y]$. To evaluate this distribution we need additionally to specify the marginal, or prior, distribution $[\theta]$. A second application of Bayes' theorem then gives the conditional distribution $[\theta \mid Y]$, called the posterior distribution of θ, and it follows that

$$[S \mid Y] = \int [S \mid Y, \theta]\, [\theta \mid Y]\, \mathrm{d}\theta. \tag{1}$$

Thus, the Bayesian predictive distribution is an average of classical predictive distributions for particular values of θ, weighted according to the posterior distribution of θ. The effect of the averaging in (1) is typically to make the predictions more conservative, in the sense that the variance of the distribution (1) is usually larger than the variance of the distribution $[S \mid Y, \hat{\theta}]$ obtained by plugging an estimate of θ into the classical predictive distribution.

In practice, it is often necessary to implement Bayesian predictive inference by drawing a random sample from the predictive distribution (1), rather than by obtaining an explicit mathematical expression for this distribution. Indeed, the availability of feasible Monte Carlo sampling algorithms for a very wide range of statistical models has been an essential step in the development of practical Bayesian inference. See, for example, Gilks *et al.* (1996).

Having obtained a random sample from $[S \mid Y]$, if we are interested in a particular property T of the surface S, for example the region within which $S(x)$ exceeds a critical value, we then obtain a random sample from the required distribution $[T \mid Y]$ by simply computing the value of T directly from each sampled value of S.

To motivate and illustrate this general approach, we use a running example consisting of rainfall data collected at recording stations throughout Paraná state, Brazil. Figure 1 shows a map of the 143 locations at which data on average winter (dry season) rainfall are available. Note that the distance scale is in kilometres. The amount of winter rainfall has direct implications for the viability of particular kinds

of agricultural activity. Different dot sizes and shadings correspond to the data divided according to the quintiles of the empirical distribution of recorded values. The numbers 1 to 4 identify particular locations for which we shall later describe the prediction results in greater detail.

In a full analysis of these data, we would treat the problem as one of space-time modelling, since year-to-year variation is important. Here, we consider only average rainfall during May and June over the median period of 33 years covered by the data. We shall obtain a predicted map of the spatial variation in average dry-season precipitation and, more specifically, make a predictive inference for the region within which average dry-season precipitation is at least 300 mm.

We shall also use these data to illustrate the important differences which can arise between the results of our inferential methods and those of conventional geostatistics.

2. Stochastic Model

Our stochastic model for S is that the surface $\{S(x): x \in \mathbb{R}\}$ is a realization of a stationary Gaussian process with mean zero, variance σ^2 and correlation function $\rho(x - x') = \mathrm{Corr}\{S(x), S(x')\}$. In what follows, we shall assume that $S(x)$ is isotropic, so that $\rho(x - x') = \rho(u)$ where u is the distance between x and x'. Furthermore, we shall specify $\rho(\cdot)$ as a member of the Matérn family (Matérn, 1986)

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

where $K_{\kappa}(\cdot)$ denotes the modified Bessel function of order κ.

As discussed in Stein (1999), this two-parameter family has a very useful flexibility in that the integer part of the parameter κ determines the number of times the process $S(x)$ is mean-square differentiable, whilst the parameter α measures the scale (in units of distance) on which the correlation decays. Figure 2(a) shows the function $\rho(u; \alpha, 1.5)$ for some values of α, whilst Figure 2(b) compares $\rho(u; 0.1, \kappa)$ for $\kappa = 0.5$, 1.5 and 2.5. For $\kappa = 0.5$, the Matérn family reduces to the exponential, $\rho(u; \alpha) = \exp(-u/\alpha)$, which is the correlation function of a mean-square continuous but non-differentiable process $S(x)$. Values of $\kappa = 1.5$ and 2.5 correspond to processes which are mean-square differentiable and twice differentiable, respectively.

Our model for the data conditional on S is that the $Y_i: i = 1, \ldots, n$ are conditionally independent given S, with

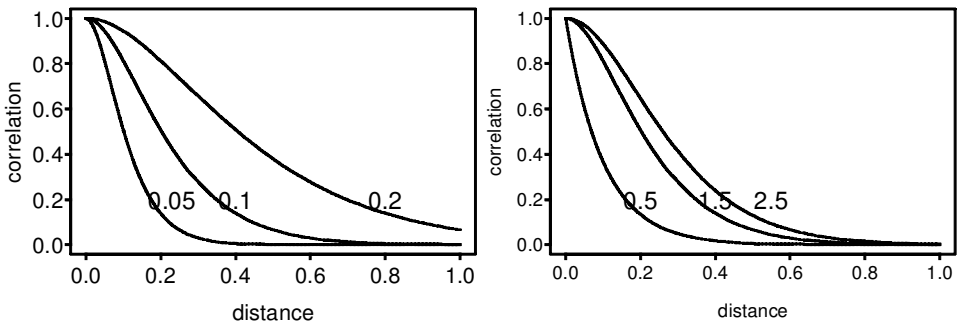

Figure 2. Examples of the Matérn correlation function: (a) $\kappa = 1.5$, varying α; (b) $\alpha = 0.1$, varying κ.

$$Y_i \mid S \sim N(\mu(x_i) + S(x_i), \tau^2). \quad (3)$$

The mean value surface $\mu(x)$ can be specified as a trend surface, or as a regression model based on spatially referenced explanatory variables, or it may simply be assumed constant; for the Paraná precipitation data we shall use a linear trend surface. The variance τ^2 is the classical 'nugget effect'. Strictly, this represents the variance of independently repeated determinations of Y at the same location, i.e. measurement error. In practice, the estimated value of τ^2 may represent the combined effect of measurement error and spatial variation at scales smaller than the smallest observed distance between sampling locations. In practice this distinction is relevant only when predicting at sample locations. If the nugget is considered as micro scale variation the predicted value at sampled locations coincides with the observed value.

This basic model can be extended in various ways according to the needs of particular applications. A very simple extension is to assume that the linear Gaussian model holds after a suitable transformation of the measurement scale. For example, when analysing measurements which are inherently non-negative it is sometimes useful to model $\log Y$ rather than Y itself. Note, in this context, that the underlying spatial process of interest is then likely to be $\exp\{S(x)\}$, and that in general if $\hat{S}(x)$ is the optimal predictor for $S(x)$, according to some stated criterion such as mean-square error, then $\exp\{\hat{S}(x)\}$ is not the optimal predictor for $\exp\{S(x)\}$. See, for example, Cressie (1993, Section 3.2.2) and De Oliveira *et al*. (1997) for a Bayesian approach to the transformation problem.

A richer extension, as discussed by Diggle *et al*. (1998), is to embed the linear Gaussian model within a wider distributional framework analogous to the embedding of the classical linear model within the generalized linear model (McCullagh & Nelder, 1989). This would include, for example, Poisson log-linear conditional models for count data and Bernoulli logistic-linear models for binary data.

Other kinds of extension may be context-specific. For example, in soil science non-linear partial differential equation models are used to relate soil transmissivity to the piezometric head, implying that the bivariate spatial field representing soil transmissivity and piezometric head cannot be jointly Gaussian. A description and a comparison of different approaches to this problem are presented by Zimmerman *et al*. (1998).

In practice, there are major limitations in the extent to which specific modelling assumptions can be validated by the data available in a particular application. However, in our view this is not an argument for abandoning a model-based approach but rather for combining the data with contextual information and honestly acknowledging model uncertainty in making predictive inferences, which is what our approach seeks to achieve. In particular, we would argue that the linear prediction methods used in standard (ordinary or universal) kriging are only natural within a linear Gaussian modelling framework, and the Gaussian modelling assumption simply makes explicit what is implicitly being assumed.

3. Approaches to Inference

3.1. Parameter Estimation and Prediction

From a non-Bayesian perspective, we can distinguish between two different kinds of inferential problem: *estimation* problems are concerned with finding out about the fixed but unknown values of model parameters (testing problems address similar

concerns, but are of limited relevance here); *prediction* problems are concerned with finding out about the realized values of random variables included in the model. Thus, finding out about the underlying surface $S(x)$ is a prediction problem.

From the Bayesian perspective which we shall ultimately take, model parameters are also random variables, whose marginal distributions are referred to as *prior distributions*, and the distinction between estimation and prediction vanishes. However, in order to compare our approach with others it is helpful initially to consider the two problems separately.

3.2. Non-Bayesian Inference

Classical (non-Bayesian) inference about a generic target T is based on the distribution $[T \mid Y,\theta]$, i.e. conditioning on the data and the model parameters. In practice the vector parameter θ is usually unknown and replaced by an estimate $\hat{\theta}$. Methods for parameter estimation and prediction based on $[T \mid Y, \hat{\theta}]$ are now discussed.

3.2.1. Parameter estimation by curve-fitting. The variogram is a useful and very well-known diagnostic for geostatistical data. For data $(Y_i, x_i): i = 1, \ldots, n$, the theoretical variogram is the set of values $V(u_{ij}) = \frac{1}{2}\mathrm{Var}(Y_i - Y_j)$, where u_{ij} is the distance between x_i and x_j. For the linear Gaussian model defined in Section 2

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

The corresponding empirical variogram is based on observed squared differences between pairs of values Y_i or, if a non-constant mean function $\mu(x)$ is included in the model, on squared differences between pairs of residuals after fitting the mean function. In the former case, the empirical variogram cloud is a scatterplot of u_{ij} against $v_{ij} = \frac{1}{2}(y_i - y_j)^2$. Each v_{ij} is an unbiased estimator for $V(u_{ij})$, but its relatively large sampling variance severely limits the direct interpretability of a variogram cloud. A more easily interpretable result is obtained by averaging all values of v_{ij} corresponding to the same, or approximately the same, value of u_{ij}. We call the resulting set of triplets $(u_k, v_k, n_k): k = 1, \ldots, m$, where each u_k identifies the midpoint of a chosen distance interval, v_k is the corresponding average of the v_{ij}, and n_k is the number of v_{ij} contributing to v_k, the *sample variogram*.

The approximate point-wise unbiasedness of the sample variogram has led to its being used for parameter estimation by matching the sample variogram to a theoretical variogram family $V(u;\theta)$, choosing the value of θ to optimize some curve-fitting criterion. This curve-fitting is sometimes done 'by eye', sometimes by ordinary least squares, sometimes by weighted least squares with the n_k and/or the theoretical $V(u_k;\theta)$ determining the weights, and sometimes by 'robust' alternatives to least squares. For a review, see Cressie (1993, Section 2.6).

There are at least two potential objections to this practice. Firstly, because the different v_k are highly correlated, the appearance of the complete sequence of values of v_k can be very markedly different from the underlying function $V(u;\theta)$. Secondly, within the Gaussian distributional framework which, we argue, is implicit in standard linear kriging methods where estimated variograms are widely used, more efficient estimation methods are available based on the likelihood function. Unlike *ad hoc* curve-fitting methods, these benefit from the well established and very widely applicable optimality properties of likelihood-based methods of parameter estimation.

3.2.2. Maximum likelihood (ML), and restricted maximum likelihood (REML), parameter estimation. The method of maximum likelihood (ML) is a generic estimation method which is easily stated: given a model in the form of a joint probability distribution $f(y;\theta)$ for data y dependent on a parameter θ, estimate θ to maximize $\log f(y;\theta)$ at the observed value of y. Under very general circumstances, this results in estimators for θ which, in large samples, are unbiased with the smallest possible variance amongst all unbiased estimators (Cox & Hinkley, 1974). The function $L(\theta) = \log f(y;\theta)$, with y held fixed at its observed value, is called the *log-likelihood* for θ.

In the context of the correlated data which arise naturally in geostatistical applications, this so-called asymptotic optimality needs to be interpreted cautiously. Firstly, as discussed in Stein (1999), for spatial data we could consider at least two different kinds of asymptotic regime: increasing the number of observations within a fixed region or increasing the size of the study region with the number of observations per unit area held fixed. These two regimes lead to *different* theoretical properties of estimators. Secondly, for strongly correlated data the effective sample size is less than the nominal sample size, and there are no general theoretical results concerning the optimality of ML estimation in small samples. Nevertheless, ML and other methods of estimation based on the likelihood function are central to modern statistical methodology and usually out-perform more *ad hoc* methods.

We shall consider the model defined by (2) and (3), with the additional assumption that the mean surface $\mu(x)$ is linear in a set of spatially referenced explanatory variables, i.e. $\mu(x) = \Sigma_{j=1}^{p} \beta_j z_j(x)$, where the values of $z_j(x)$ are observed without error. In this situation, let $\beta = (\beta_1, \ldots, \beta_p)$, write Z for the n by p matrix with ijth element $Z_j(x_i)$ and let $\mu = (\mu(x_1), \ldots, \mu(x_n)) = Z\beta$. Also, let $V = \tau^2 I + \sigma^2 R$, where I is the n by n identity matrix and $R = R(\alpha,\kappa)$ is the n by n matrix with ijth element $\rho(u_{ij})$. Then, writing $Y = (Y_1, \ldots, Y_n)$, our model specifies that Y follows a multivariate Gaussian distribution with mean vector $Z\beta$ and variance matrix V. It follows that the log-likelihood for $\theta = (\beta, \tau, \sigma, \alpha, \kappa)$ is

$$L(\beta, \tau, \sigma, \alpha, \kappa) \propto -0.5\{\log |V| + (y - Z\beta)'V^{-1}(y - Z\beta)\}. \tag{4}$$

Maximization of (4) yields the ML estimates of the model parameters. To simplify the computation involved in finding the ML estimates, note that for any fixed V, the log-likelihood is maximized by taking

$$\hat{\beta}(V) = (Z'V^{-1}Z)^{-1}Z'V^{-1}y. \tag{5}$$

Substitution of (5) into (4) then gives a function of $(\tau, \sigma, \alpha, \kappa)$ which must be maximized numerically. It is convenient to re-parameterize the model to $\nu^2 = \tau^2/\sigma^2$ and re-express V as $V = \sigma^2 V_0$ where $V_0 = \nu^2 I + R$. Then, for fixed $\phi = (\nu, \alpha, \kappa)'$, we obtain an explicit expression

$$\hat{\sigma}^2 = n^{-1}(y - Z\hat{\beta})'V_0^{-1}(y - Z\hat{\beta}), \tag{6}$$

and numerical maximization is required over only the three dimensions of ν, α and κ.

A variation of ML, introduced in a completely different context by Patterson & Thompson (1971), is restricted maximum likelihood, or REML. This method of estimation corresponds to applying ML in the residual space of dimension $n - p$ orthogonal to the space of dimension p spanned by the linear model for $\mu(x)$, as

follows. Let $\varepsilon = y - Z\tilde{\beta}$, where $\tilde{\beta}$ is the ordinary least squares estimate of β. Then, under our assumed model, ε has a singular multivariate Gaussian distribution with mean zero and variance matrix which does not depend on β. Writing $f^*(\varepsilon; \sigma, \nu, \alpha, \kappa)$ for the pdf of ε, the restricted log-likelihood is:

$$L^*(\sigma, \nu, \alpha, \kappa) = \log f^*(\varepsilon; \sigma, \nu, \alpha, \kappa)$$

$$\propto -0.5\{\log|V| + (y - Z\hat{\beta})'V^{-1}(y - Z\hat{\beta}) + \log|Z'V^{-1}Z|\}$$

and maximization of $L^*(\cdot)$ yields the REML estimates of $(\sigma, \nu, \alpha, \kappa)$.

In many applications, the difference between ML and REML estimation is small. Where the methods do differ, the REML estimators tend to be less biased; for example, in a linear model with independent errors, the REML estimate of the error variance is the unbiased residual mean square with divisor $n - p$, as used in classical analysis of variance, whereas the ML estimator uses the divisor n with no adjustment to allow for estimation of the p regression parameters. However, in the present context it must be emphasized that the bias-reduction property of REML estimation relies on the correct specification of the model for the mean, $\mu = Z\beta$, and in a geostatistical context this can be problematic. Our experience has been that when the mean model is specified pragmatically, for example as a polynomial trend surface, ML can give better results in practice than REML. Note, however, that in the geostatistical context, REML estimation under the stationary Gaussian model is closely related to ordinary kriging, in which prediction variances are inflated to allow for the uncertainty in estimating the mean parameter. REML also has connections with the Bayesian approach which we shall discuss in Section 3.3.1.

3.2.3. Variogram parameter estimates for the Paraná data. We now show the estimates of the variogram parameters obtained by curve-fitting and by ML estimation applied to the Paraná state precipitation data.

For the mean surface, we fit a linear trend surface in two dimensions, thus $p = 3$ corresponding to an intercept parameter and a slope parameter for each of latitude and longitude. Figure 3 shows the sample variogram of the ordinary least squares residuals from the linear trend surface together with three fitted models of the form

$$V(u) = \sigma^2\{1 - \exp(-u/\alpha)\},$$

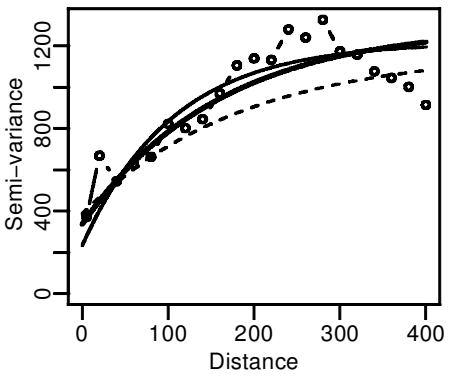

Figure 3. Fitted variograms, using three different methods of estimation: (a) curve-fitting (thin line); (b) ML (dashed line); (c) posterior mode (thick line).

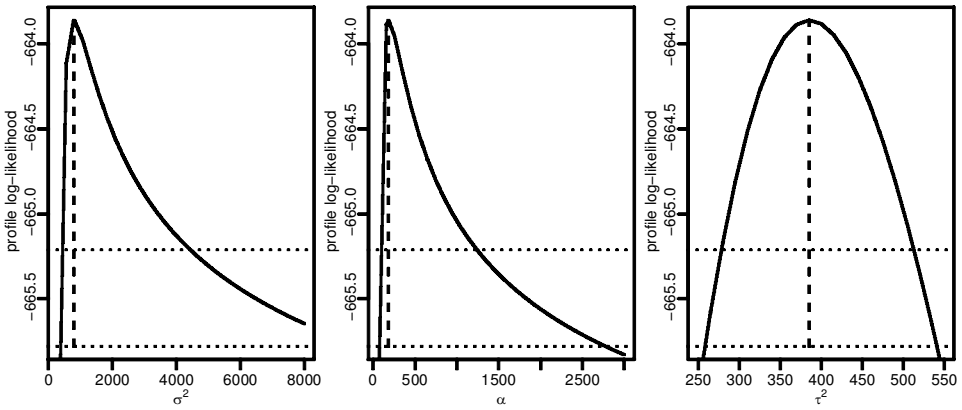

Figure 4. Profile log-likelihood plots for covariance parameters. In each case the upper and lower horizontal dashed lines identify the endpoints of 90 and 95% confidence intervals.

where the parameter estimates are obtained by three different methods: curve-fitting; ML; and Bayesian posterior mode (see Section 3.3.3 for an explanation of the last of these).

For the curve-fitting method of estimation, the precise specification of the criterion to be optimized requires a number of subjective choices. The result shown was obtained by weighted least squares with weights n_k, based on the choice of bins shown in Figure 3. In contrast, the ML method is automatic. The ML estimate of $V(u)$ differs markedly from the curve-fitting estimate. Figure 3 also shows a Bayesian estimate which combines the log-likelihood function with a prior distribution for the model parameters, as discussed in Section 3.6 below. In this example, the Bayesian and weighted least squares estimates are closer to the sample variogram than is the ML estimate, although this is not always so, and in any case begs the question of whether visual closeness between fitted and sample variograms is an appropriate estimation criterion.

Figure 4 shows profile log-likelihoods for each of the three variogram parameters, σ^2, α and τ^2. The associated confidence intervals are very wide for both σ^2 and α, indicating that neither of these parameters is estimated with high precision. For a fuller explanation we need to consider the joint profile log-likelihood surface for (σ^2, α), which is shown in Figure 5 and indicates a diagonal ridge of near-constant values of the log-likelihood. In other words, neither parameter has its value well determined by the data, although certain combinations of individually feasible values are jointly infeasible. Different parametrizations are adopted by Stein (1999) and De Oliveira *et al.* (1997). However, under these reparametrizations the covariance parameters no longer have such a direct, intuitive geostatistical interpretation.

3.2.4. Plug-in prediction. If we treat the estimated model parameters as known, we can obtain predictions, $\hat{S}(x)$ say, of the underlying surface $S(x)$ at an arbitrary location x directly from the conditional distribution $[S(x) \mid Y, \hat{\theta}]$. In particular, the conditional mean of this distribution, which is linear in y and is identical to the simple kriging predictor, would be the minimum mean-square error predictor if $\hat{\theta}$ were replaced by the true parameter value θ. The conditional variance similarly corresponds to the simple kriging variance.

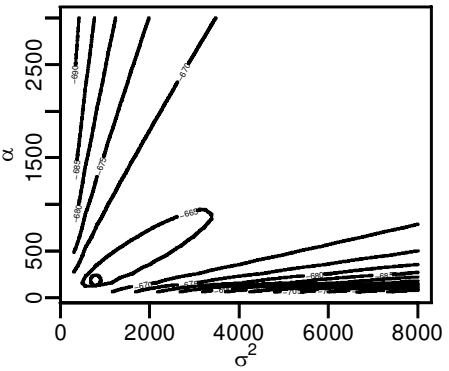

Figure 5. Contours of the joint profile log-likelihood for covariance parameters σ^2 and α.

The more widely used methods known as ordinary or universal kriging make explicit allowance for estimation of an unknown mean, using the estimator given by equation (wls), in which the mean is assumed to be constant (ordinary kriging) or to follow a trend surface model (universal kriging). Practical geostatistics often uses a variant of this linear predictor in which, for each prediction location x, only data-values at locations close to x are used to compute $\hat{S}(x)$ and its associated prediction variance.

All of these methods treat the estimated variogram parameters as if they were known beforehand. This often has a small effect on the predicted surface $\hat{S}(x)$ but, as we shall illustrate in Section 4, can seriously compromise the validity of nominal prediction variances.

3.3. Bayesian Inference for the Linear Gaussian Model

In the Bayesian approach, both the variable Y and parameters θ are considered to be random quantities with joint distribution $[Y,\theta] = [Y \mid \theta][\theta]$. Note that $[Y \mid \theta]$ is algebraically identical to $\exp\{L(\theta)\}$ as defined in Section 3.3, but now considered as a conditional distribution for Y given θ. The marginal, or *prior* distribution $[\theta]$ represents uncertainty about θ before the data are collected. The conditional distribution $[\theta \mid Y]$ is called the posterior distribution for θ, and represents the residual uncertainty about θ after Y has been observed. Using Bayes' theorem, we have that

$$[\theta \mid Y] \propto [\theta][Y \mid \theta].$$

Formally, the Bayesian solution to inference about θ is the conditional distribution $[\theta \mid Y]$. A Bayesian point estimate would be any convenient summary of this distribution's central tendency, for example the posterior mean or posterior mode. A Bayesian interval estimate would be the range spanned by selected quantiles of the posterior, for example the range from the 5th to the 95th percentile defines a central 90% credible interval.

The probability density function of the posterior for $\theta = (\beta, \tau, \sigma, \alpha, \kappa)$ in the model with log-likelihood function (4) is

$$f(\beta, \tau, \sigma, \alpha, \kappa \mid y) \propto f(\beta, \tau, \sigma, \alpha, \kappa) \mid V \mid^{-1/2} \exp\left\{ -\frac{1}{2}(y - Z\beta)' V^{-1} (y - Z\beta) \right\}.$$

The choice of priors can be a delicate issue in Bayesian inference. A prior which leads to a posterior in the same family of distributions is called a *conjugate prior*. Conjugate priors can be computationally convenient, although this alone should not justify their choice. Vague prior knowledge is expressed by giving the prior distribution a large variance. A pragmatic strategy is therefore to use a conjugate prior whenever one is available but to experiment with different values for the prior variance. Once the posterior distribution for θ has been obtained, Bayesian predictive inference then proceeds by weighting the classical predictive distribution according to the posterior distribution of θ, as in equation (1).

The material in the remainder of this section is a summary of the more detailed discussion in Ribeiro and Diggle (1999a). We consider two cases of particular interest. The first case is when only the β parameters which define the mean surface $\mu(x)$ are unknown. Whilst implausible, this turns out to have a direct relationship with ordinary or universal kriging methods, making precise the sense in which we claim that these methods are implicitly assuming the linear Gaussian model. The second case is when all model parameters are unknown, which in our view is the more realistic scenario. In the remainder of this section we will denote $\phi = (\nu^2, \alpha, \kappa)$ or simply $\phi = (\nu^2, \alpha)$ when κ is fixed.

3.3.1. Uncertainty only in the mean parameter. If we assume a Gaussian prior for the mean parameter, $\beta \sim \mathrm{N}(m_\beta; \sigma^2 V_\beta)$, where σ^2 is the (assumed known) variance of $S(x)$, the posterior is given by

$$[\beta \mid Y, \sigma^2, \phi] \sim \mathrm{N}((V_\beta^{-1} + Z'R^{-1}Z)^{-1}(V_\beta^{-1}m_\beta + Z'R^{-1}y); \sigma^2(V_\beta^{-1} + Z'R^{-1}Z)^{-1})$$

$$\sim \mathrm{N}(\tilde{\beta}; \sigma^2 V_{\tilde{\beta}}),$$

say. This shows that the Gaussian distribution is a conjugate prior for β, since it leads to a posterior for β which is again Gaussian. In the limit as all diagonal elements of $V_\beta \to \infty$, the posterior mean $\tilde{\beta} \to \hat{\beta}$ as given by equation (5).

Prediction of $Y_0 = S(x_0)$ at an arbitrary location x_0 is based on the posterior distribution $[Y_0 \mid Y; \sigma^2, \phi]$. The probability density of this predictive distribution is

$$[Y_0 \mid Y, \sigma^2, \phi] = \int [Y_0 \mid Y; \beta, \sigma^2, \phi][\beta \mid Y; \sigma^2, \phi] \, \mathrm{d}\beta.$$

The first probability distribution inside the last integral is the conditional distribution implied by the linear Gaussian model with β assumed known, whilst the second is the posterior distribution for β given by (7). The term inside the integral is a bivariate Gaussian density, and it follows that the predictive distribution $[Y_0 \mid Y; \sigma^2, \phi]$ is also Gaussian. The inclusion of σ^2 and ϕ in the conditioning set emphasises that these parameters are (temporarily) assumed to have known values. The mean and variance of the predictive distribution are, respectively

$$E[Y_0 \mid Y; \sigma^2, \phi] = (Z_0 - r'R^{-1}Z)(V_\beta^{-1} + Z'R^{-1}Z)^{-1}V_\beta^{-1}m_\beta$$

$$+ [r'R^{-1} + (Z_0 - r'R^{-1}Z)(V_\beta^{-1} + Z'R^{-1}Z)^{-1}Z'R^{-1}]Y,$$

$$\mathrm{Var}[Y_0 \mid Y\sigma^2, \phi] = \sigma^2[R_0 - r'R^{-1}r + (Z_0 - r'R^{-1}Z)(V_\beta^{-1} + Z'R^{-1}Z)^{-1}(Z_0 - r'R^{-1}Z)'],$$

where r is the vector of correlations between $S(x_0)$ and $S(x_i) : i = 1, \ldots, n$.

The predictive variance above has three components. The first and second components represent the marginal variance for Y_0 and the variance reduction after observing Y, respectively, whilst the third component accounts for the additional uncertainty due to the unknown value of β. This last component reduces to zero if $V_\beta = 0$, since this formally corresponds to β being known beforehand, and the result then coincides with simple kriging.

In the limit as all diagonal elements of $V_\beta \to \infty$, these formulae for the predictive mean and variance correspond exactly to the universal kriging predictor and its associated kriging variance, which in turn reduce to the formulae for ordinary kriging if the mean value surface is assumed constant. For a non-Bayesian description of kriging methods see, for example, Journel and Huijbregts (1978), Isaaks and Srisvastava (1989), Goovaerts (1997) and Chilès and Delfiner (1999). Thus, ordinary and universal kriging can be interpreted as Bayesian prediction under a form of prior ignorance about the mean.

3.3.2. Uncertainty in all model parameters. If we allow for uncertainty in the parameter σ^2 and assign a conjugate prior for this parameter, we can again derive explicit expressions for the posterior and predictive distributions. The conjugate prior in this case assigns to σ^2 a scaled inverse chi-squared distribution. We say that σ^2 follows a scaled inverse chi-squared distribution, for which the shorthand notation is $\sigma^2 \sim \chi^2_{\text{Sc}I}(v, q)$, if vq/σ^2 is distributed as chi-squared on v degrees of freedom. This choice of prior corresponds to prior knowledge equivalent to v independent observations from a distribution with variance q (Gelman *et al.*, 1995).

The ability to derive explicit results by exploiting conjugacy is important in practice because this reduces the burden of computation associated with the full Bayesian implementation when all model parameters unknown. For the model considered in this paper, the additional parameters to be considered are α, κ and τ. Evaluation of the posterior and predictive distributions now requires Monte Carlo methods since explicit expressions cannot be derived.

For illustration of algorithms, we shall assume temporarily that only one parameter, say α, is unknown. In fact, this is not an entirely unrealistic assumption, since on the one hand, τ^2 is sometimes known to a good approximation using contextual knowledge, and on the other the Matérn shape parameter κ will often be chosen from a very limited set of discrete possibilities, say $\kappa = 0.5$, 1.5 or 2.5.

No specific prior will be assumed for α, because in this case there is no convenient conjugate family. The posterior distribution for the parameters is formally given by:

$$[\beta, \sigma^2, \alpha \mid Y] = [\beta \mid Y, \sigma^2, \alpha][\sigma^2 \mid Y, \alpha][\alpha \mid Y] \tag{8}$$

where $[\beta \mid Y, \sigma^2, \alpha]$ is Gaussian, $[\sigma^2 \mid Y, \alpha]$ follows a scale-inverse-χ^2 distribution and $[\alpha \mid Y]$ is obtained by the relation

$$[\alpha \mid Y] \propto \frac{[\beta, \sigma^2, \alpha][Y \mid \beta, \sigma^2, \alpha]}{[\beta \mid Y, \sigma^2, \alpha][\sigma^2 \mid y, \alpha]}. \tag{9}$$

For the case where the prior $f(\beta, \sigma^2 \mid \phi) \propto 1/\sigma^2$ is adopted, the posterior for the correlation function parameter is given by

$$[\phi \mid Y] \propto [\phi] \mid V_{\hat{\beta}} \mid^{-1/2} \mid R \mid^{-1/2} \left(\frac{n}{n-p} \hat{\sigma}^2 \right)^{-(n-p)/2}$$

where p is the number of components of the mean parameter β and $\hat{\sigma}^2$ is given by equation (6). However, this expression does not define a standard probability distribution, and we adopt a Monte Carlo inferential strategy, see, for example, Tanner (1996). Specifically, we generate samples from the posterior and predictive distributions, and use the resulting empirical distributions as the basis for inference and prediction, respectively.

In order to sample from the posterior distribution (8), we use the following algorithm.

Algorithm 1

1. Discretize the distribution $[\alpha \mid Y]$ by choosing a set of values for α in a sensible interval considering the problem in hand, and assigning a discrete uniform prior for α on the chosen support set.
2. Compute the posterior probabilities in this support using (9), so defining a discrete posterior distribution with probability mass function $\tilde{f}(\alpha \mid y)$, say, which is an approximation to $[\alpha \mid Y]$.
3. Sample a value of α from the discrete distribution $\tilde{f}(\alpha \mid y)$.
4. Attach the sampled value of α to $[\beta, \sigma^2 \mid Y, \alpha]$ and sample from this distribution.
5. Repeat steps (3) and (4) as many times as required; the resulting sample of triplets $(\beta, \sigma^2, \alpha)$ is a sample from the joint posterior distribution.

The size of the sample generated in this way should be large enough to permit stable estimation of the underlying distribution.

We now consider the corresponding algorithm for prediction of the underlying surface $S(x)$. The predictive distribution for $Y_0 = S(x_0)$, where x_0 is the prediction location, is given by

$$
\begin{aligned}
[Y_0 \mid Y] &= \int\int\int [Y_0, \beta, \sigma^2, \alpha \mid Y] \, \mathrm{d}\beta \, \mathrm{d}\sigma^2 \, \mathrm{d}\alpha \\
&= \int [Y_0 \mid Y, \alpha][\alpha \mid Y] \, \mathrm{d}\alpha .
\end{aligned}
$$

The resulting predictive distribution depends on the prior distribution adopted. Usually, it is not a standard probability distribution and the integral must be solved by numerical methods. We again use a Monte Carlo method and the algorithm proposed is similar to the previous one.

Algorithm 2

1. Follow steps 1–3 of Algorithm 1.
2. Attach the sampled value of α to $[Y_0 \mid Y, \alpha]$ and sample from it to obtain a realization from the predictive distribution of Y_0.
3. Repeat steps 3 and 4 as many times as required, thereby generating a sample from the predictive distribution of $Y_0 = S(x_0)$.

In the general case, where at least one of τ and κ are also treated as unknown parameters, we use algorithms of the same kind, except that we also need to specify a discrete prior distribution for τ and/or κ on a multi-dimensional grid of values. Each increase in dimensionality carries an associated increase in the computational load, but introduces no new principles.

3.3.3. Application to the Paraná precipitation data. In the analysis reported here, we use the linear Gaussian model with a linear trend model for the mean surface (i.e. $p = 3$), an exponential correlation function (i.e. Matérn with $\kappa = 0.5$) and an unknown nugget variance. Our trivariate prior for β is an improper uniform distribution, corresponding to a conjugate Gaussian prior with arbitrarily large variances. Our prior for σ^2 is proportional to $1/\sigma^2$, which is equivalent to the limiting form of an inverse scaled chi-squared as the degrees of freedom tend to zero. In both cases, these choices can be interpreted as an expression of prior ignorance. For the correlation parameter α we use a discrete prior with support points between 0 and 600. For the relative nugget variance $\nu^2 = \tau^2/\sigma^2$ we use a discrete uniform prior with support points between 0 and 0.60. Prediction maps were obtained from a grid of 134×94 points covering the area. All the analyses were performed using the public domain software `geoR` described in Ribeiro and Diggle (1999b) and available at: http://www.maths.lancs.ac.uk/~ribeiro/geoR.html.

As discussed earlier, Figure 3 includes a comparison of the Bayesian posterior mode estimate of the variogram with alternative, non-Bayesian estimates. Figure 6 shows a sample of size 1000 drawn from the posterior distribution of the variogram parameters σ^2, α and τ^2 using Algorithm 1. The long upper tails of the posterior distributions for σ^2 and α are the Bayesian counterparts of the strong asymmetry in the corresponding profile log-likelihoods shown in Figure 4, although the chosen prior for α confines the posterior to a range much narrower than the profile-based confidence interval for α and this in turn leads to a smaller range for the posterior of σ^2. Figure 7 shows the effect of this on the joint posterior for (σ^2, α). The values of these two parameters are now much less dependent than was the case in the ML analysis.

Figure 8 shows two aspects of the Bayesian solution to the prediction problem: the point-wise posterior mean surface of $S(x)$ in Figure 8(a); and the corresponding posterior variances in Figure 8(b). These results were obtained following the Algorithm 2, drawing a sample of size 500 from the predictive distribution for each point of the prediction grid.

Figure 9 compares the Bayesian predictive distributions for $S(x)$ with the predictive distributions corresponding to ordinary kriging, where x refers in turn to each of the four numbered locations in Figure 1. These have been deliberately selected to

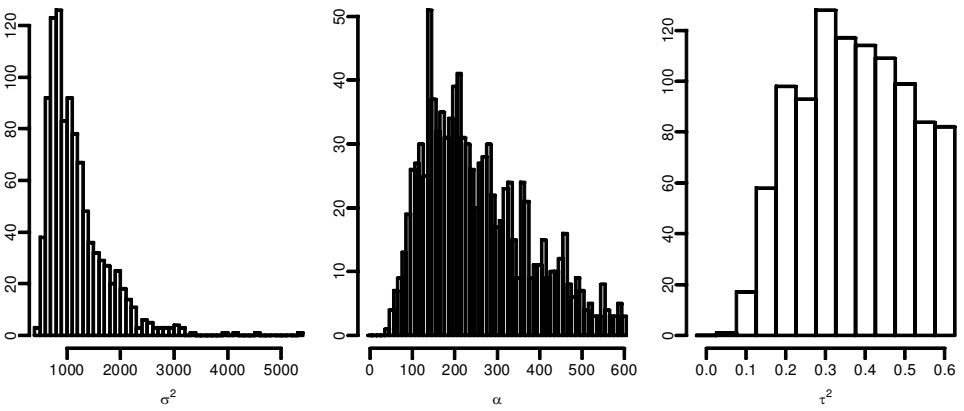

Figure 6. Samples of the marginal posterior distributions for covariance parameters.

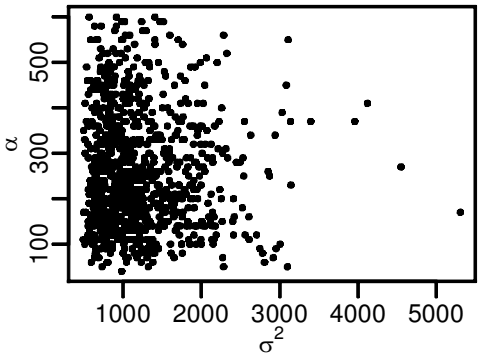

Figure 7. Samples from posterior of the parameters σ^2 and α.

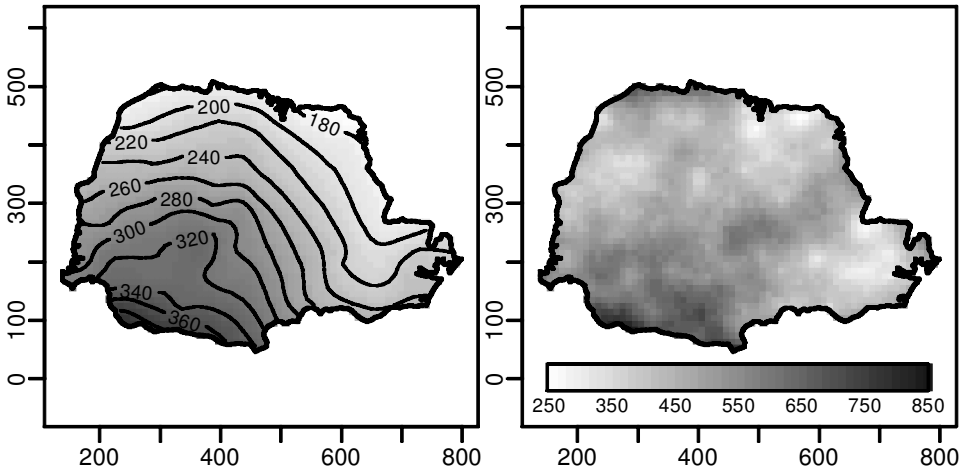

Figure 8. Predicted signal surfaces and associated measures of precision for the rainfall data: (a) posterior mean; (b) posterior variance.

have different spatial relationships to the data-locations and so emphasize that there is no simple relationship between the Bayesian and plug-in predictive distributions: the two solutions may differ in either or both of their location (point prediction) and scale (interval prediction), but not necessarily in the same way at all locations.

We now turn to the specific question raised earlier: what inference can we make about the random set $T = \{x : S(x) > 300\}$? Generating a random sample from the posterior distribution for T is straightforward; choosing an appropriate summary description of such a sample is not, but Figure 10 is an attempt to do so. This diagram partitions the study region into sub-regions consisting of all those points x for which the posterior probability of $S(x) > 300$ is greater than 0.90, between 0.90 and 0.5, between 0.5 and 0.10, and less than 0.10. The extent of the hinterland between the 0.90 and the 0.10 contours gives some indication of how imprecisely the data determine the boundary of T. This can easily be confirmed by generating an independent random sample from the posterior distribution of T, but the results are not easy to summarize in a single picture.

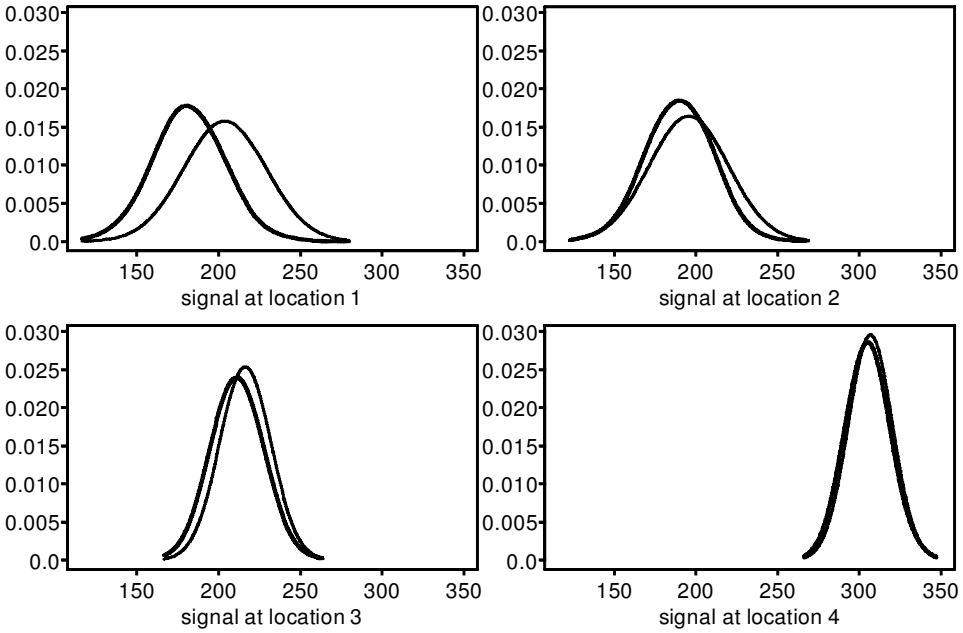

Figure 9. Ordinary kriging (thin lines) and Bayesian (thick lines) predictive distributions for average rainfall at selected locations.

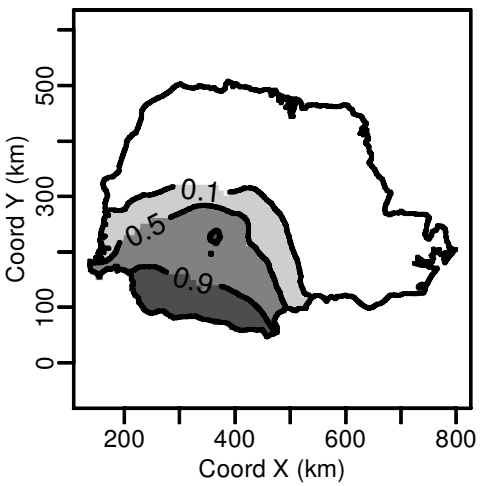

Figure 10. Posterior probability contours for levels 0.10, 0.50 and 0.90 for the random set $T = \{x : S(x) > 300\}$.

4. Discussion

The model-based approach described in this paper gives a coherent framework for geostatistical inference. The approach makes appropriate allowance for residual uncertainty in unknown quantities which impinge on the problem, whilst giving an opportunity to incorporate contextual knowledge in two ways: in the formulation of

the model to be fitted; and in assigning measures of prior uncertainty to the values of parameters within the model. Implementing the approach in practical problems requires a degree of pragmatism. Firstly, it is rare that the physical mechanisms which generate a given set of data would be so well understood as to point to a unique model. Secondly, the choice of a prior distribution is all too often made on grounds of convenience.

In our analysis of the Paraná data, we used priors which represent prior ignorance about β and σ^2, but are informative about α and ν^2. The informative prior for ν^2 has relatively little impact on the results because the data themselves provide strong information about this parameter. In contrast, the comparison between Figures 5 and 7 shows that the informative prior for α materially affects the inferences for α and σ^2.

Ideally, we would have preferred to incorporate more covariates and contextual knowledge to choose our priors. However, this knowledge is not currently available to us.

In general, we would argue that using a convenient prior which covers a realistic range of values for a true but unknown parameter is preferable to pretending that a point estimate of an unknown parameter is the truth; and that using procedures which follow from the application of general, theoretically justifiable inferential principles to a reasonable and explicitly declared model is preferable to a more *ad hoc* approach. This is not to deny that *ad hoc* methods can give excellent results in the hands of an expert. The formal statistical machinery which we have described in this paper is intended to supplement, rather than replace, the subject-matter expertise of the scientist.

Acknowledgements

We thanks Laura Regina Bernardes Kiihl, Instituto Agronômico do Paraná, Londrina, Brazil and Jacinta Loudovico Zamboti for organizing and providing the data. This research was partially supported by the EU TMR network ERB-FMRX-CT96-0095 on 'Computational and statistical methods for the analysis of spatial data'. The second author also thanks CAPES/Brazil, grant BEX 1676/96–2.

References

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

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

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

De Oliveira, V., Kedem, B. & Short, D. (1997) Bayesian prediction of transformed gaussian random fields. *JASA*, 92(440), 1422–1433.

Diggle, P., Tawn, J. & Moyeed, R. (1998) Model based geostatistics (with discussion). *Applied Statistics*, 47(3), 299–350.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (1995) *Bayesian Data Analysis.* London: Chapman & Hall.

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

Handcock, M. & Stein, M. (1993) A bayesian analysis of kriging. *Technometrics*, 35(4), 403–410.

Isaaks, E. & Srisvastava, R. (1989) *An Introduction to Applied Geostatistics.* New York: Oxford University Press.

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

Kitanidis, P. (1986) Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research*, 22(4), 499–507.

Matérn, B. (1986) *Spatial Variation*, 2nd edn. Berlin: Springer.

McCullagh, P. & Nelder, J. (1989) *Generalized Linear Models.* London: Chapman & Hall.

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

Ribeiro Jr, P. & Diggle, P. (1999a) Bayesian inference in gaussian model-based geostatistics. Technical Report ST-99-08, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Ribeiro Jr, P. & Diggle, P. (1999b) `geoR/geoS`: A geostatistical library for `r/s-plus`. Technical Report ST-99-09, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Stein, M. (1999) *Interpolation of Spatial Data*. New York: Springer.

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

Zimmerman, D., de Marsily, G., Gotway, L.A., Marietta, M.G., Axness, C.L., Beauheim, R., Bras, R., Carerra, J., Dagon, G., Davies, P.B., Gallegos, D.P., Galli, A., Gómes-Hernandez, J., Grindrod, P., Gutjahr, A.L., Kitanidis, P.K., Lavenue, A.M., McLaughlin, D., Neuman, S.P., Ramaros, P.S., Ravenne, C. & Rubin, Y. (1998) A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. *Water Resources Research*, 34(6), 1373–1413.