Beyond kriging - Dealing with discontinuous spatial data fields using adaptive prior information and Bayesian Partition Modelling.

John Stephenson[1] (john.stephenson@imperial.ac.uk),
K. Gallagher[1] and C. C. Holmes [2]

13th November 2003

[1] *Dept. of Earth Science and Engineering*,
[2] *Dept. of Mathematics*,
Imperial College London,
South Kensington, London. SW7 2AS

Abstract

The technique of kriging is widely known to be limited by its assumption of stationarity, and performs poorly when the data involve localized effects such as discontinuities or nonlinear trends. The Bayesian Partition Model (BPM) of Denison et al. (2002) is compared with results from Ordinary Kriging for various synthetic discontinuous one dimensional functions, as well as for 1986 precipitation data from Switzerland. This latter data set has been analyzed during a comparison of spatial interpolation techniques, and has been interpreted as a stationary distribution and one thus suited to kriging. The results demonstrate that the BPM outperformed kriging in all of the data sets compared (when tested for prediction accuracy at a number of validation points), with improvements of up to a factor of 6 for the synthetic functions.

Abbreviated Title:	Dealing with discontinuities via the BPM
Number of Words:	6340
Number of Figures:	15
Number of Tables:	2
Number of References:	19

The technique of kriging is used extensively in spatial interpolation problems and has found many applications in geology as well as ecology and environmental sciences. For example, predicting ore grades in mining, building heterogeneous reservoir models from well and seismic data, and inferring pollution levels from sparsely distributed data points.

The critical drawbacks to kriging lie in the assumptions that the spatial process results from a deterministic mean (either constant in the case of ordinary kriging or a linear combination of parameters as in universal kriging), combined with a zero mean spatially correlated stationary process. The principal work in kriging is defining this correlation process, which is usually achieved via modelling the semi-variogram. This framework does not allow for discontinuities in the spatial process, or other localized effects. The current Bayesian applications to this problem have revolved around optimal estimation of the covariance matrix, for example (Handcock and Stein, 1993).

A Bayesian Partition Model (BPM) (Denison et al., 2002a) avoids the kriging limitations. The BPM defines a set of disjoint regions, whose position and shape is defined using a Voronoi tessellation, such that the positions of the Voronoi centers form the parameters of the model. The choice of the Voronoi parametrization is not essential to the mechanics of the BPM, though it does provide an extremely simple implementation, that is applicable in any number of spatial dimensions. The data is then assumed to be stationary only within an individual partition, thus avoiding the key pitfall of kriging and can thereby deal gracefully with discontinuities. This parametrization allows for arbitrarily complex surfaces (in 2-D or 3-D, or indeed higher dimensions if required).

The model space is explored using Markov Chain Monte Carlo (MCMC) and a Bayesian framework, so that each iteration will either move, remove or create one partition. The prior information is fairly loosely defined, through hyperpriors on the mean level of the variables and regression variance within a partition, and the hyperpriors are adapted as the algorithm proceeds. The model is naturally biased towards simpler models (fewer regions), and there is no need to specify how many regions the model will have (thus providing models with varying dimension). Once convergence is achieved, the desired point predictions are then made by sampling from the Markov chain, and taking the average over all the samples. Using a Bayesian framework leads directly to the inclusion of prior information into the modelling framework, and provides immediate access to the uncertainties of the predictions.

In this contribution, we illustrate the application of Bayesian partition modelling to spatial data that would create problems using a conventional Kriging approach, as well as to stationary data where we would expect kriging to be well adapted.

1 Geostatistics and Kriging

The name synonymous with geostatistics is George Matheron, who, building on the foundational work of Daniel Krige in statistically estimating average ore grades, developed the now fundamental process of kriging (Matheron, 1963). The technique has found many diverse applications such as time lapse seismic noise characterization (Coleou et al., 2002) and image processing (Decenciere et al., 1998). We shall concentrate here simply on the spatial interpolation properties.

This introduction to kriging is not meant to be exhaustive, rather, we want to provide enough information to convey the key limitations of the process. For full coverage of geostatistics and kriging, we direct the reader either to (Cressie, 1993) for a theoretical approach, or to (Isaaks and Srivastava, 1989) for more accessible and practical methods.

1.1 Spatial Interpolation and Terms

The most obvious use of geostatistics involves using a data set to generate a map of values across a surface. Throughout this paper, we shall refer to the data set D, as comprising n data values $y_1 \dots y_n$, having been sampled at the corresponding irregularly spaced locations $\boldsymbol{x_1} \dots \boldsymbol{x_n}$ from a subset $\mathcal{D}$ of space $\mathbb{R}^d$ with dimension d. In the case of spatial statistics d is typically 2, e.g. northings and eastings, though the methods described are not limited by this. Traditionally, we then seek to estimate the new values on a regularly spaced grid across $\mathcal{D}$, so as to allow contour maps or surfaces to be generated.

To illustrate these approaches, we present a simple example using the heavily analyzed Walker Lake sample data set from (Isaaks and Srivastava, 1989), comprising values at 470 two dimensional spatial locations, drawn from an exhaustive data set of 78,000. The data values we use are referred to as set V in the text, and are derived from a digital elevation model of the Walker Lake region in California. Their positions are shown in figure 1. For full details of these data sets, see (Isaaks and Srivastava, 1989).

The task now is to use the data D to make a prediction of a new data value y_{n+1} at position $\boldsymbol{x_{n+1}}$. One approach is to estimate y_{n+1}, using a weighted sum of n available data, scaled by the weights $w_1 \dots w_n$.

$$y_{n+1} = \sum_{i=1}^{n} w_i y_i \tag{1.1}$$

Commonsense dictates that the best approximations to y_{n+1} will probably come from those points closest to it, and thus we expect to assign them higher weights. One way of achieving this

deterministically would be via the inverse distance weighting scheme i.e.

$$y_{n+1} = \frac{\sum_{i=1}^{n} \frac{1}{d_i} y_i}{\sum_{i=1}^{n} \frac{1}{d_i}} \tag{1.2}$$

where the distances $d_1 \dots d_n$, are given by $d_i = \|\boldsymbol{x}_i - \boldsymbol{x}_{n+1}\|$. It is common practice to place even greater emphasis on closer points by squaring these distances. Other simple deterministic weighting approaches include polygonal and triangulation methods. All of these methods suffer when the data is clustered (thus not accounting for redundant information when samples are close together) and make no allowance for noise in the data (they are exact interpolators) (Isaaks and Srivastava, 1989).

1.2 Kriging Approach

Kriging takes a more stochastic approach and, using the same predictor as in equation 1.1, seeks to choose optimal weights so as to minimize the variance of the estimation error. Before this is possible, many simplifications and assumptions must be made.

The starting point for Matheron was to assume that any observed data can be seen as the realization of a random variable, or more formally

$$\{Y(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}. \tag{1.3}$$

so that y will represent a realization from this random function, and $\hat{y}$ will be our estimate of y, which can never truly be known.

From this formulation, we are able to write the estimation error as another random variable R.

$$R(\boldsymbol{x_{n+1}}) = \hat{Y}(\boldsymbol{x_{n+1}}) - Y(\boldsymbol{x_{n+1}}) \tag{1.4}$$

$$= \sum_{i=1}^{n} w_i Y(\boldsymbol{x_i}) - Y(\boldsymbol{x_{n+1}}) \tag{1.5}$$

It is usual at this stage to split equation 1.3 further into two separate parts so that

$$Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \delta(\boldsymbol{x}) \tag{1.6}$$

where μ is a deterministic trend, while δ is a zero mean stochastic process with an unknown covariance matrix. The various forms of kriging differ only in how they treat the trend. For

example ordinary kriging assumes a constant mean μ, whilst universal kriging seeks a linear combination of known functions. For the purposes of this article, we shall only consider the mathematically simpler technique of ordinary kriging.

1.3 Stationarity

Before we can continue, we must make further assumptions about the stochastic process $\delta(\boldsymbol{x})$. To make the mathematics tractable, and in order to estimate a covariance function from the data D, we must assume that δ is second order stationary, such that its covariance matrix can be written as

$$Cov(\delta(\boldsymbol{x}_i), \delta(\boldsymbol{x}_j) = C(\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{1.7}$$

$$= C(\boldsymbol{h}_{ij}) \tag{1.8}$$

Thus the covariance of two random variables is a function only of the vector between them. Furthermore, if we assume isotropy, C becomes a function of $h = ||\boldsymbol{h}||$, i.e. the distance between them. We are now in a position to use the data to calculate this covariance function $C(h)$.

1.4 Covariance Function

For computational reasons, the covariance function is usually calculated by using the semi-variance estimation, which can be approximated using,

$$\hat{\gamma}(h) = \frac{1}{2N_h} \sum_{(ij)|h_{ij} \approx h} (y_i - y_j)^2 \tag{1.9}$$

where N_h represents the discrete number of point combinations that roughly correspond to the distance h, $\pm$ a specified tolerance. For example, in figure 2(a), the $470^2/2$ data pairs from the Walker Lake data set were grouped by distance into bins with interval 5 and tolerance ± 2.5. The semi-variance is related to the covariance by $C(h) = \gamma(\infty) - \gamma(h)$, and can be seen graphically in figure 2(b).

Equation 1.9 approximates the semi-variance by grouping together points a similar discrete distance h apart, and then finding their mean square difference. When the data are irregularly spaced (as is usually the case), we are forced to model the semi-variance to provide the required values of $\gamma(h)$ not covered by the approximation in equation 1.9. The model must be chosen so as to provide a positive definite covariance matrix. This is achieved by plotting a semi-variogram

(h against $\hat{\gamma}(h)$) and then fitting one of several well known semi-variogram models which can be combined linearly to give nested approximations to $\gamma(h)$.

Some common semi-variogram models include

$$Gaussian: \quad \gamma(h) = 1 - exp\left(-\frac{h^2}{a^2}\right) \tag{1.10}$$

$$Exponential: \quad \gamma(h) = 1 - exp\left(-\frac{h}{a}\right) \tag{1.11}$$

$$Spherical: \quad \gamma(h) = \frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3 \tag{1.12}$$

where a is a parameter commonly called the range, which reflects the degree of correlation between the data. It should be noted that many resources, including (Isaaks and Srivastava, 1989), include an extra factor 3 so that a reflects the distance at which the data is uncorrelated. Here we have adopted the form consistent with the GSTAT software (Pebesma and Wesseling, 1998) to be used later in prediction analysis. These models are then scaled by another value, termed the sill, which is effectively the value of the semi-variogram at infinite h (see figure 2(a)). Three different variogram models are shown in figure 3 using the Walker lake data sets. Each of these models also include a nugget effect which effectively prevents kriging from being an exact interpolator, thus making allowance for noise in the sample measurements (see figure 2(a)). This corresponds to a non-zero semi-variance at $h = 0$.

1.5 Prior Information

It cannot be overstated that the quality of prediction, generated using kriging, depends entirely on how the semi-variogram is modelled. For example, prior information regarding the smoothness of a function would lead to a choice of a gaussian model, whilst a noisy data set will usually require a nugget effect to prevent exact interpolation. Whilst computer fitting methods do exist, such as Cressie's Weighted Least Square (Cressie, 1993) or restricted maximum likelihood (REML) (Kitanidis, 1983), they are rarely relied upon. It is usual that a great deal of user information is built into the choice of models and parameters, and the process has been called more of a black art than a science.

Anisotropy can be included by limiting $\boldsymbol{h}$ to only being in either the major or minor anisotropic direction, modelling these semi-variograms separately, and then transforming the distance vector appropriately.

By using the model structure defined by equations 1.3, 1.6 and 1.8, combined with the properties of a weighted sum of random variables, allows us to write explicitly the estimation error

variance σ_R^2. The weights w_i from equation 1.1 are then calculated by finding the partial derivatives of σ_R^2 with respect to each of the unknown weights w_i, and then setting each equation equal to zero. This linear system of equations can then be solved analytically to give the optimum set of weights $\boldsymbol{w}$, to predict y_{n+1} at position $\boldsymbol{x}_{n+1}$, and the minimized error variance σ_{OK}^2. For details of this process, we direct the reader to the standard text (Cressie, 1993).

As an example, figure 4 demonstrates the kriging prediction surface for values of $\boldsymbol{x}_{n+1}$ that cover a grid with density 1, using the Walker lake data and the spherical covariance function shown in figure 2. Figure 5 presents the values of the kriging variance, σ_{OK}^2, that have been minimized for each of the desired locations. As we would expect, it is evident that the kriging variance is much lower close to sample locations, reflecting our increased certainty in these areas (see figure 1 for the sample locations).

Kriging is known as the best linear unbiased estimator (BLUE), where it is linear in the sense that the predictor is a linear combination of the available data, unbiased as it sets the mean error equal to zero, and best in that it minimizes the error variance.

1.6 Other Kriging Methods

It should be noted that numerous methods have been put forward to evade the inherent stationarity problems of kriging. Notable methods include lognormal, disjunctive kriging, indicator kriging and kriging of residuals (Cressie, 1993), as well as more recently, model based kriging (Fuentes, 2001). A good empirical assessment of some of these methods can be found in (Moyeed and Papritz, 2002), and their method of assessing prediction quality has been adopted in this work. Their assessment of prediction accuracy on the data set of (Mcbratney et al., 1982), concluded that the accuracy of ordinary kriging was comparable if not better than any of the tested non-linear methods. For this reason and to minimize complexity, we shall only compare the BPM to ordinary kriging.

2 Bayesian partition model

The aims of the BPM are similar to those of universal kriging, i.e. to reduce the influence of the deterministic trend μ in equation 1.6 by fitting a more complex model to it. Universal kriging fails, however, due to its reliance on simple linear polynomial models across the whole surface, whereas the underlying surface may contain various localized effects or clustering not taken into account, e.g. discontinuities from a fault in a digital elevation model. By modelling the trend in such a simple way, the assumption of stationarity becomes invalid, and will lead to spurious

results.

The BPM works around these limitations by fitting a far more flexible nonlinear trend to the data, and then treating the errors as strictly stationary, rather than second order stationary (i.e. regression). Again here we present an introduction to the techniques used, and for full details and other examples, we refer the reader to (Denison et al., 2002b) and (Denison et al., 2002a).

2.1 Model Description

The BPM seeks to separate the domain $\mathcal{D}$ into disjunct regions, within which the data D is modelled using linear regression. This simple framework allows a non-linear trend to be formed in any number of dimensions. For example, in the one-dimensional case, the trend could be modelled as a series of discontinuous linear regression fits to the data (see figure 6 for an example).

However, a single model, even after optimization over all the parameters, will struggle to fit the smoothly varying parts of the function in figure 6, and will introduce extra discontinuities not present in the true function. We adopt a prior distribution on models that reflects our relative beliefs on the accuracy of each. This leads to a model averaging approach to prediction, i.e. we can use many models from the MCMC sampling to produce a final model. By taking enough samples, a smoothly varying final model is possible, while still allowing for sharp discontinuities if required by the data, as these will be present in most, if not all, accepted models over which we average.

2.2 Model Parameters

The first step in the BPM is to parameterize the partitions in terms of Voronoi tessellations. This requires a set of k centers $\boldsymbol{\theta} = \boldsymbol{t}_1 \dots \boldsymbol{t}_k$. Every point in the domain $\mathcal{D}$ is then assigned to its closest center, creating disjoint regions whose boundaries are the perpendicular bisectors with the surrounding centers. Hence the vector $\boldsymbol{y}_j$ would comprise the subset of n_j points from D that are closest to the center $\boldsymbol{t}_j$, and so belong to the region R_j. One and two dimensional examples of these are given in figures 6 and 7 respectively. It is not essential for the Voronoi tessellation to be used, however it does reduce the number of model parameters, and can be applied to systems of any dimensionality.

A linear regression fit is then calculated within each of these regions so that when $\boldsymbol{x}$ lies within region R_j

$$y|\boldsymbol{x} \in R_j = \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\beta}_j + \delta \tag{2.1}$$

where as before δ represents a random error, $\boldsymbol{\beta}_j$ a $p \times 1$ vector of unknown scalars (regression

parameters) specific to the partition j, and $\boldsymbol{B}(\boldsymbol{x})$ the $1 \times p$ vector of values of the known basis function $\boldsymbol{B}$. For example, in 1D linear regression, $\boldsymbol{\beta}_j$ will contain an intercept and a slope, and $\boldsymbol{B}(\boldsymbol{x}) = [1\ x]$, thus giving p a value of 2. For ease of notation later, we assign $\boldsymbol{B}_j$ the $n_j \times p$ basis matrix comprising the value of $\boldsymbol{B}$ evaluated for the n_j data locations that lie within region j.

In contrast to the approach in universal kriging, where a similar regression is used to approximate the trend, here the error δ (see equation 1.6) is assumed to be strictly stationary with a covariance function of $\boldsymbol{I}\sigma^2$. σ^2 represents the regression variance and is assumed to be the same within all partitions, and $\boldsymbol{I}$ is the identity matrix. By making these assumptions, we only ever have to invert $p \times p$ matrices.

This fills the full complement of unknown parameters, which we will group as $\boldsymbol{\phi} = [\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2]$. It should be noted that there is no fixed number of partitions, and the value of k will vary from model to model, thus allowing the dimensionality of the model space to vary accordingly.

2.3 Model Selection

We are now in a position to make inferences of the value y_{n+1} at a new position $\boldsymbol{x}_{n+1}$ given our data samples D. From the general representation theorem, we represent the density of this new value as the integral

$$p(y_{n+1}|D) = \int_{\Phi} p(y_{n+1}|\boldsymbol{\phi})p(\boldsymbol{\phi}|D)d\boldsymbol{\phi}, \tag{2.2}$$

over the entire parameter space Φ. This represents model averaging over every possible set of model parameters. The first part of the integrand in 2.2 is the likelihood of the new data value, and is easily defined once $\boldsymbol{\phi}$ has been chosen. The second part represents the posterior distribution of model parameters, and is analytically intractable.

It is at this point we turn to Bayesian statistics and MCMC techniques to approximate 2.2. We write equation 2.2 as the Monte Carlo integral

$$p(y_{n+1}|D) \approx \frac{1}{M}\sum_{i=1}^{M} p(y_{n+1}|\boldsymbol{\phi}_i) \tag{2.3}$$

so that as the number of samples, M, increases, the summation will converge to the integral in 2.2, providing that we choose a sampling method that draws the parameter samples $\boldsymbol{\phi}_i$ from their posterior distribution $p(\boldsymbol{\phi}|D)$. This distribution is ensured by creating a Markov Chain, whose stationary distribution is the desired posterior density.

2.4 The posterior distribution via Bayes' Theorem

To begin this process, we must define the posterior distribution in terms of Bayes' Theory so that

$$p(\boldsymbol{\phi}|D) \propto p(D|\boldsymbol{\phi})p(\boldsymbol{\phi}) \tag{2.4}$$

or using common terminology

$$Posterior \propto Likelihood \times Prior.$$

We will now deal with each of the likelihood and prior definitions in turn

2.4.1 Likelihood

The likelihood is a density that is returned by considering the errors of our model δ as coming from a normal distribution. In principle, we can relax this assumption, but it does provide practical advantages in terms of evaluating the integrals (as we demonstrate below). Hence following from equation 2.1, the likelihood pdf of the data D_j within the partition R_j is given by

$$p(D_j|\boldsymbol{\beta}_j, \boldsymbol{\theta}_j, \sigma^2) = N(\boldsymbol{B}_j\boldsymbol{\beta}_j, \sigma^2) \tag{2.5}$$

so that, as each partition is independent, we can write the overall likelihood of the data, for all k partitions as

$$p(D|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \prod_{j=1}^{k} N(\boldsymbol{B}_j\boldsymbol{\beta}_j, \boldsymbol{I}\sigma^2). \tag{2.6}$$

This form for the likelihood, leads directly to our choice of function for the prior distribution over our parameters $\boldsymbol{\phi}$. We are able to facilitate the mathematics of the problem greatly by choosing a prior that is conjugate to the normal distribution. These conjugate priors are of a such a form that the resulting posterior distribution will be from the same distribution family as the prior (Bernardo and Smith, 1994).

2.4.2 Priors

We must place a joint prior on all of the unknown parameters in $\boldsymbol{\phi}$, so here we seek $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$. The prior placed on $\boldsymbol{\theta}$ is chosen to be flat, so that all possible positions and numbers of partitions are equally probable, and hence allows the dimension of $\boldsymbol{\theta}$ to be variable, up to some set maximum

number of partitions K. This allows us to express the joint prior as

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = p(\sigma^2)p(\boldsymbol{\beta}|\boldsymbol{\theta}, \sigma^2) \tag{2.7}$$

By analogy with the Bayesian linear model (Bernardo and Smith, 1994), the conjugate prior for the normal likelihood is the normal inverse gamma distribution, so that for each partition we get

$$p(\boldsymbol{\beta}_j|\boldsymbol{\theta}, \sigma^2) = N(\mathbf{0}, \sigma^2\boldsymbol{V}) \tag{2.8}$$

with $\boldsymbol{V} = v\boldsymbol{I}$ where v is a scalar value, giving the prior variance over the regression parameters. As the dimensions of $\boldsymbol{\beta}$ are fixed, $\boldsymbol{V}$ will be the same for each partition. Taking into account an inverse gamma prior over the common regression variance $p(\sigma^2) = IG(a, b)$ where a and b are the shape parameters of the distribution, as well as the independence of partitions, we get a full prior via

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = p(\sigma^2)p(\boldsymbol{\beta}|\boldsymbol{\theta}, \sigma^2) \tag{2.9}$$

$$= IG(a, b)\prod_{i=1}^{k} N(\mathbf{0}, \sigma^2\boldsymbol{V}) \tag{2.10}$$

so that in effect, we are left with three prior parameters a, b and v, which must be given values at the start. The parameter with the greatest influence is v, so this too is given a set of priors (or hyperpriors), and is updated throughout the MCMC run.

2.4.3 The Posterior

The power of using conjugate priors can be seen in evaluating the posterior density, so that by using the result from the standard Bayesian linear model, and taking a model with k partitions, we can give the posterior distribution as

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|D) = IG(a^*, b^*)\prod_{j=1}^{k} N(\boldsymbol{m}_j^*, \sigma^2\boldsymbol{V}_j^*) \tag{2.11}$$

where the prior values of a, b, m and V have been updated by information in the likelihood to give

$$\boldsymbol{m}_j^* = (\boldsymbol{V}^{-1} + \boldsymbol{B}_j'\boldsymbol{B}_j)^{-1}(\boldsymbol{B}_j'\boldsymbol{y}_j) \tag{2.12}$$

$$\boldsymbol{V}_j^* = (\boldsymbol{V}^{-1} + \boldsymbol{B}_j'\boldsymbol{B}_j)^{-1} \tag{2.13}$$

$$a^* = a + n/2 \tag{2.14}$$

$$b^* = b + \{\boldsymbol{y}'\boldsymbol{y} - \sum_{j=1}^{k}(\boldsymbol{m}_j^*)'(\boldsymbol{V}_j^*)^{-1}\boldsymbol{m}_j^*\} \tag{2.15}$$

This updating property is the primary reason for the choice of the conjugate prior, as we have evaded the need to evaluate the integral

$$p(D) = \int_{\Phi} p(D|\boldsymbol{\phi})p(\boldsymbol{\phi})d\boldsymbol{\phi} \tag{2.16}$$

which is the required constant of proportionality in equation 2.4, and provided a posterior distribution that we can easily integrate using standard integrals when we later consider the marginal likelihood.

2.5 Reversible Jump MCMC

The remarkable feature of MCMC, is its ability to converge to a desired stationary distribution, provided the probability of transition to a new state in the chain is correctly defined. The technique used here is the reversible jump Metropolis-Hastings of (Green, 1995) , which is capable of handling high and variable dimensional models in a very efficient manner.

Starting from a model with one randomly placed partition, the chain moves forward, around the model space by one of either of three steps with equal probability:

Birth A new Voronoi center is created at a random position within the domain $\mathcal{D}$, thus increasing the dimensionality of the model by one.

Death A randomly selected center is removed from $\boldsymbol{\theta}$, thus reducing the dimensionality of the model by one.

Move A randomly selected center within $\boldsymbol{\theta}$ is given a new random position within $\mathcal{D}$.

The acceptance probability for our model under reversible jump algorithm is given by

$$\alpha = min\left\{1, \frac{p(D|\boldsymbol{\beta}_{m+1}, \boldsymbol{\theta}_{m+1}, \sigma^2)}{p(D|\boldsymbol{\beta}_m, \boldsymbol{\theta}_m, \sigma^2)}\right\} \tag{2.17}$$

which is a ratio of the likelihoods of the data given two different model choices, and in turn are called marginal likelihoods. The models compared are the current member of the chain m and the proposed sample $m+1$. Note that this ratio is the same as the Bayes Factor, when we assume that the prior probabilities of each set of model parameters are the same. See (Bernardo and Smith, 1994).

Figure 8 demonstrates this process of model selection and model averaging (see equation 2.3), with 10 different samples taken from the posterior distribution via reversible jump MCMC. This figure also demonstrates the varying dimensionality of the model from sample to sample, as seen by the differing numbers of partitions used.

2.6 Marginal Likelihood

The marginal likelihood of a particular model $\boldsymbol{\phi}_i$, required in equation 2.17, can easily be defined by rearranging Bayes theorem to give (Denison et al., 2002a)

$$p(D|\boldsymbol{\phi}_i) = \frac{p(D|\boldsymbol{\beta}_i, \boldsymbol{\theta}_i, \sigma^2)p(\boldsymbol{\beta}_i, \sigma^2)}{p(\boldsymbol{\beta}_i, \boldsymbol{\theta}_i, \sigma^2|D)} \tag{2.18}$$

Because of the nature of the conjugate priors we have chosen previously, it becomes possible to integrate out the unknown model parameters $\boldsymbol{\beta}$ and σ^2, which for the BPM as described before, gives the following

$$p(D|\boldsymbol{\phi}_i) = \frac{(b)^a \Gamma(a^*)}{(b^*)^{a^*}\Gamma(a)} \pi^{-n/2} \prod_{j=1}^{k} \frac{|\boldsymbol{V}_j^*|^{1/2}}{|\boldsymbol{V}_j|^{1/2}} \tag{2.19}$$

where b, a, V and their starred updates are given in equation 2.12 to 2.15, and are specific to this ith model. This means that the marginal likelihood is dependant on only the position of the Voronoi centers, defined for model i by $\boldsymbol{\theta}_i$, and the pre-defined parameters a, b and v.

2.7 Adaptive Priors

In order to ensure greater independence from the prior regression parameter variance v, the reversible jump method allows the data to determine suitable values. This comes about by assuming that v itself has a distribution of values, and as such can be assigned a prior (termed a hyperprior) and consequently a posterior distribution from which suitable values can be drawn.

Full discussion can be found in (Denison et al., 2002b), but we shall give the result here.

$$p(v^{-1}) = Ga(\alpha_1, \alpha_2) \tag{2.20}$$

$$p(v^{-1}|D, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = Ga\left(\alpha_1 + \frac{p \times k}{2}, \alpha_2 + \frac{1}{2\sigma^2}\sum_{j=1}^{k} \boldsymbol{\beta}'\boldsymbol{\beta}\right) \tag{2.21}$$

Here, Ga represents the usual gamma distribution, and α_1 and α_2 are two constants that define the distribution of the hyperprior $p(v^{-1})$ (in this case a noninformative hyperprior was used, by giving α_1 and α_2 a value of 0.1). First however, as we have integrated out σ^2 and $\boldsymbol{\beta}$, we need to obtain up to date values for these parameters for use in equation 2.21, so they are drawn from their updated distributions, $IG(a^*, b^*)$ and $N(\boldsymbol{m}^*, \sigma^2\boldsymbol{V}^*)$ respectively.

2.8 Computational Advantages

The key computational advantage of partition models over any of the forms of kriging, is that for each sample of parameters $\boldsymbol{\phi}_i$, to calculate the marginal likelihood it is only necessary to invert k $(p \times p)$ matrices where, as before, k is the number of partitions in the model, and p the order of $\boldsymbol{B}$ in equation 2.1. This value will depend on the spatial dimension we are working in, and the order of the regression being fitted. For example, in a two dimensional spatial system with coordinates x_1 and x_2, and fitting linear planes inside each partition, $\boldsymbol{B}$ will take values of $\boldsymbol{B} = [\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2,]$, and thus give p a value of 3.

This is in stark contrast to kriging, where we must invert one large $n \times n$ matrix. This is another of the highlighted drawbacks to kriging, and typically leads to only kriging subsets of the data for each prediction point. The drawback using the BPM, is that we must iterate the process many times to provide a large enough number of samples in the Monte Carlo integral (equation 2.3), though these calculations take little time due to the size of the necessary inversions. It should also be mentioned that many applications in the oil industries now require that kriging itself be iterated many times, for example, sequential gaussian simulation in reservoir characterization (Hohn, 1999).

2.9 Prediction Uncertainties

Another positive feature of the BPM is the ease with which prediction uncertainties can be assigned. These are generated by considering all of the predictions at each validation point generated over the course of the MCMC sampling procedure, and then returning the 95% credible intervals of their distributions.

With kriging, in the ordinary kriging form we have presented here, uncertainty can be adjudged by use of the minimized kriging variance σ^2_{OK}, as displayed in figure 5. By assuming multivariate normal distributions, 95% confidence intervals can be assigned in the usual way so that we get the range $\hat{y} \pm 2\sigma_{OK}$ (Isaaks and Srivastava, 1989). The results from such a treatment however are largely unsatisfactory and usually represent only a measure of the local sample distribution (Journel, 1986). In addition, kriging uncertainties are always underestimated, as the parameters of the covariance function are assumed to be known, despite they're having been estimated when modelling the semi-variogram. More realistic assessments of uncertainties are reviewed in (Goovaerts, 2001), and necessitate either the more complex approaches of indicator or disjunctive kriging, or finally, an iterative stochastic simulation approach.

3 The Data Sets

In order to demonstrate the main benefits of the BPM over ordinary kriging(OK), we have compared the two methods for various synthetic examples, as well as for a real data set. The former were chosen as examples best suited to the BPM (i.e. where the assumptions of stationarity are challenged), whilst the latter represents a stationary data set, and is thus suited to a kriging approach.

3.1 Synthetic Examples

Two synthetic one dimensional functions, A and B, were chosen both with large and significant discontinuities in their functions. For the following we define the boxcar step function $Box(c_1, c_2)$ as

$$Box(c_1, c_2) = \left\{ \begin{array}{ll} 1 & \text{if} \quad c_1 < x < c_2 \\ 0 & otherwise \end{array} \right\} \tag{3.1}$$

and the precise definitions of the two functions are given below.

Function A An upside down boxcar function

$$y(x) = 1 - Box(0.3, 0.7) \tag{3.2}$$

Function B An addition of a plane, gaussian and another upside down boxcar. Explicitly, this gives

$$y(x) = 0.6x + exp\left(-\frac{(x-0.9)^2}{(0.2)^2}\right) - Box(0.25, 0.6) \tag{3.3}$$

The sample data sets were formed by taking 100 random x positions within the range $0 < x < 1$, and then adding gaussian noise with standard deviations of 0.1 and 0.08 respectively to the y values of functions A and B. The validation data sets comprise 1000 and 500 data points for functions A and B respectively, along an equally spaced grid. The BPM and OK algorithms were then used to give predictions at the validation data set positions and compared. The data sets, together with the noise free functions, are shown graphically in figures 9 and 10.

3.2 Rainfall data from Switzerland

The rainfall data set comprises 100 measurements of daily rainfall in Switzerland on the 8th of May 1986, taken randomly from a full data set of 467 points. This data set is freely available, and formed the basis of the 1997 Spatial Interpolation Comparison (SIC97), a comparison of numerous geostatistical techniques. Indeed, to ensure the validity of the BPM, we compare it to the published results using OK and Indicator Kriging(IK) (Atkinson and Lloyd, 1998).

The values are in units of 1/10th of a mm. The importance of such a data set derives from rainfall being used as the prime measurement of radioactivity deposition, after the Chernobyl disaster in 1986.

Further information regarding elevation was also provided separately, but this information was not incorporated in either model. Figure 11 shows the relative densities and positions of sample and validation data sets.

3.3 Prediction Quality

We first define the criteria used in comparing OK to the BPM,

- Root Mean Square Error

$$RMSE = \left\{ \frac{1}{N_v} \sum_{i=1}^{N_v} (\hat{y}_i - y_i)^2 \right\}^{\frac{1}{2}} \tag{3.4}$$

- Mean Absolute Error

$$MAE = \frac{1}{N_v} \sum_{i=1}^{N_v} |\hat{y}_i - y_i| \tag{3.5}$$

- Relative Bias

$$rBIAS = \frac{\sum_{i=1}^{N_v} (\hat{y}_i - y_i)}{N_v \bar{y}_p} \tag{3.6}$$

- Relative Mean Separation

$$rMSEP = \frac{\sum_{i=1}^{N_v}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N_v}(\bar{y}_p - y_i)^2} \tag{3.7}$$

where $\hat{y}_i$ represents the prediction of the true function value y_i over the N_v validation data points and $\bar{y}_p$ is the arithmetic mean of the prediction data set. These tests are a combination of those used in (Atkinson and Lloyd, 1998) and (Moyeed and Papritz, 2002). In all cases, a better prediction model is one that produces a lower value of these measures.

4 Synthetic Function : Model Specification and Results

For each of the test functions, A and B, an OK and a BPM prediction were made. In the latter case, it was enough to give the data to the program and let it run, whilst for OK, numerous decisions had to be made to ensure a viable fit.

4.1 Ordinary Kriging

All of the kriging results were produced using GSTAT, a highly flexible program, and a standard geostatistical package (Pebesma and Wesseling, 1998). The first step for any kriging analysis is the modelling of the variogram. The assumptions and models made during this process influence the results greatly. The two semi-variograms are given in figures 12 and 13 for functions A and B respectively, along with their corresponding models. It should be noted that the parameters for these models assume there is not a factor 3 in the exponential (see section 1.4). The models were originally fitted with the weighted least squares method of Cressie, and then adjusted manually to provide a more consistent fit.

Although numerous more complicated nested models were considered, it was found that the best results were produced using simple nugget and gaussian model combinations. This is consistent with the known smoothness and noise characteristics of the two functions we were considering.

Predictions were then made for the validation data sets, again using GSTAT, with a maximum search radius of 0.25. This search radius was used to take into account our knowledge of the evident localized clustering in the sample data.

4.2 Partition Model

The only parameters required for the BPM were the number of samples to be taken, and the order of the regression fit within the partitions. For function A, 20,000 samples were taken, with

regressions of order 0, whilst for function B, 50,000 samples were taken with regressions of order 1. These choices reflect the differing complexities of the two functions.

The number of partitions k for function A ranged between 3 and 5, whilst for function B, values ranged from 3 to 9, with 5 being by far the most frequent. These numbers illustrate the model's desire to reduce complexity, as well as the fact that the dimension of the model space is allowed to vary from sample to sample. This property is built in to any Bayesian inference process, and is commonly called *Occam's Razor*. For further details of this effect with regard to the BPM, we refer the reader to (Denison et al., 2002b).

4.3 Results

We are now able to analyze the resulting predictions using both visual and the statistical values described previously in section 3.3. Comparative plots of the two techniques are given in figures 14 and 15 for functions A and B respectively.

It is evident that in terms of overall prediction, the BPM has outperformed OK for the given synthetic examples. Table 1. gives the prediction measures outlined and the improved performance of BPM is clear.

In both functions, the prediction from BPM is between 3 and 6 times better, and the bias is reduced. However, these simple measures do not get across the principal advantage of the BPM, that of dealing gracefully with discontinuities. This is best evidenced in figures 14 and 15, where for the BPM, the edge is sharply defined, whilst for OK, the values are smeared across the discontinuity. This is easily explained when considering a point on the discontinuity. The points either side of the discontinuity at similar distances, will be treated with equal weight, thus producing the smoothing effect. This effect is impossible to remove with OK, unless you implicitly tell it there are discontinuities.

5 Rainwater Results

5.1 Ordinary Kriging Implementation

The variogram models used for analysis of the rainwater data set, are taken directly from (Atkinson and Lloyd, 1998). These models take into account the evident anisotropies at 45 °, fitting two models to each of the major and minor axes of anisotropy. The models fitted are a nested combination of gaussian and spherical models (equations 1.10 and 1.12 respectively). The prediction values for their models were recreated using GSTAT, and verified against the values given

in Atkinson's work to ensure our correct implementation. Values are also given for the analysis of IK for two of the statistics we are concerned with (RMSE and MAE), and are included for comparison.

5.2 Partition Model Implementation

Again, implementing the BPM was merely a matter of giving data to the program and ensuring a high enough number of samples were taken. The regression order was taken to be 1, and thus linear planes were fitted within each partition.

5.3 Result comparison

Four different models were compared in how well they predicted on the 367 members of the validation set. These were OK, IK, Inverse squared Distance Weighting (IDW) (see equation 1.2), and the BPM.

The IDW estimate is displayed for a benchmark, and the entire sample data set was used in the calculation of each prediction point. This comparison is given in table 2.

It is clear again from the results, that in terms of prediction, the BPM has provided the best estimates, although the performance relative to OK is not as dramatic as a consequence of the data set being more suited to kriging than the synthetic examples considered earlier. In addition when compared to OK, the distribution statistics of the BPM were found to be closer to the true values of the prediction data set. The true mean of the data is given as 185.36, with OK predicting a value of 181.87 and BPM a value of 186.53. This is clear further when comparing the rBIAS values and demonstrate that OK has underestimated values in this case

6 Summary

The fundamental limitation of kriging is the assumption of stationarity. This is coupled with difficulties in choosing appropriate covariance function models as well as a common need to analyze only subsets of the data due to computational considerations. The synthetic one-dimensional results demonstrate effectively that OK performs poorly when faced with discontinuities which are common in geological data. The model is not flexible enough to incorporate the smooth underlying function, as well as the sharp cut off points of the discontinuities, resulting in a smearing over the edges. This was despite good data support close to the boundaries of the discontinuities.

For all of the synthetic functions, the BPM outperformed OK by several orders of magni-

tude, and identified the boundaries of each of the discontinuities to an accuracy of the scale of the validation grid. These data sets were well adapted to the properties of the BPM and the predictions produced left very little room for improvement. The most evident example of this is found in function B, where the model was able to regenerate both the smooth gaussian and plane combination towards the right of the function, as well as the sharp boundaries of the discontinuity.

Furthermore the BPM was able to model the stationary rainwater data. Despite the data lacking an underlying trend, the BPM was able to effectively model the validation set, as well as to give good estimates of its distributions. This demonstrates that BPM performs well in situations where OK performs well, but requires less model specification.

In summary, some of the advantages of the BPM over OK are:

- Easy to access prediction uncertainties. (see section 2.9 for a discussion regarding uncertainties with OK)
- No need to invert $n \times n$ matrices (for 2-d linear plane fitting in BPM, only ever need to invert 3×3 matrices). This is in stark contrast with the model based kriging method of (Fuentes, 2001), where large covariance matrices must be inverted at each step of the MCMC chain.
- Only limited knowledge of the data set is required in order to produce high quality predictions. This is possible due to the automatic updating of the priors via the hyperprior structure.
- Very flexible models can be produced.
- Able to define the model over the entire space. It is an important but subtle statistical point, that by being forced for computational reasons to krige using only a small number of closely surrounding points (a maximum of 16 in the case of the rainwater data), it means that the model is not defined over the entire region.

Some further improvements we are looking at implementing include using alternatives to the standard Voronoi tessellation for partitioning. This method, although extremely easy to implement, is found to be restrictive in that it precludes certain data configurations. For example data which evidently belong to the same partition are sometimes forced to be separated by the position of surrounding Voronoi centers. This results in regions being split into more partitions than necessary, and reduces the efficacy of the regression calculation.

Another of the current disadvantages of the model, is the inability to model microscale trends easily, and a relevant example may be boulders in a digital elevation model. These will probably

be treated as noise in the regression fit, and not modelled correctly. This can be taken into account by either letting the Markov chain run for longer periods of time, or perhaps starting to use some of the stationary properties of kriging within partitions.

The authors would like to thank the reviewers, Andrew Curtis and Alberto Malinverno, for their helpful and insightful comments on all aspects of this paper, as well as Paul Williamson of Total for his guidance and ongoing support. This research has been supported by the combined NERC and Total CASE award NER/S/C/2002/10625.

References

Atkinson, M. & Lloyd, D. 1998. Mapping precipitation in switzerland with ordinary and indicator kriging. *Journal of Geographic Information and Decision Analysis*, **2**, 65–76.

Bernardo, J. M. & Smith, A. F. M. 1994. *Bayesian theory.* Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, Chichester ; New York.

Coleou, T., Hoeber, H., & Lecerf, D. 2002. Multivariate geostatistical filtering of time-lapse seismic data for an improved 4d signature. In *Society of Exploration Geophysicists, international exposition and 72nd annual meeting; technical program, expanded abstracts with authors' biographies.* Society of Exploration Geophysicists. Tulsa, OK, United States.

Cressie, N. A. C. 1993. *Statistics for spatial data.* Wiley series in probability and mathematical statistics. Applied probability and statistics. J. Wiley, New York, rev. edition.

Decenciere, E., De-Fouquet, C., & Meyer, F. 1998. Applications of kriging to image sequence coding. *Signal Processing: Image Communication*, **13**, 227–49.

Denison, D. G. T., Adams, N. M., Holmes, C. C., & Hand, D. J. 2002a. Bayesian partition modelling. *Computational Statistics and Data Analysis*, **38**, 475–485.

Denison, D. G. T., Holmes, C. C., Mallick, B., & Smith, A. F. M. 2002b. *Bayesian methods for nonlinear classification and regression.* Wiley series in probability and statistics. Wiley, Chichester, England ; New York, NY.

Fuentes, M. 2001. A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, **12**, 469–483.

Goovaerts, P. 2001. Geostatistical modelling of uncertainty in soil science. *Geoderma*, **103**, 3–26.

Green, P. J. 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711–732.

Handcock, M. S. & Stein, M. L. 1993. A bayesian-analysis of kriging. *Technometrics*, **35**, 403–410.

Hohn, M. E. 1999. *Geostatistics and petroleum geology.* Kluwer, Dordrecht, 2nd edition.

Isaaks, E. H. & Srivastava, R. M. 1989. *Applied geostatistics.* Oxford University Press, New York.

Journel, A. G. 1986. Geostatistics - models and tools for the earth-sciences. *Mathematical Geology*, **18**, 119–140.

Kitanidis, P. K. 1983. Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, **19**, 909–921.

Matheron, G. 1963. Principles of geostatistics. *Economic Geology and the Bulletin of the Society of Economic Geologists*, **58**, 1246–1266.

Mcbratney, A. B., Webster, R., Mclaren, R. G., & Spiers, R. B. 1982. Regional variation of extractable copper and cobalt in the topsoil of southeast scotland. *Agronomie*, **2**, 969–982.

Moyeed, R. A. & Papritz, A. 2002. An empirical comparison of kriging methods for nonlinear spatial point prediction. *Mathematical Geology*, **34**, 365–386.

Pebesma, E. J. & Wesseling, C. G. 1998. Gstat: A program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, **24**, 17–31.

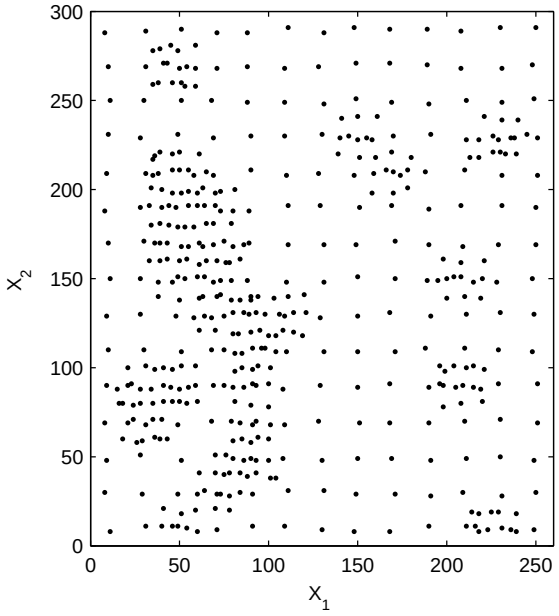

Figure 1: Spatial locations of the Walker Lake sample data set

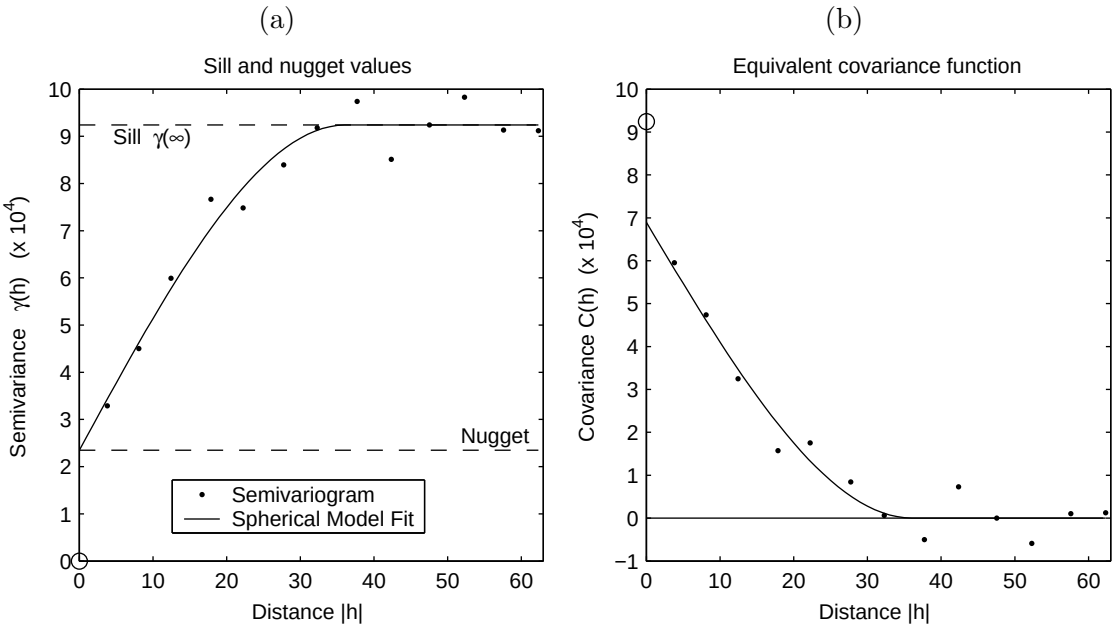

Figure 2: (a) Omnidirectional semi-variogram of the Walker Lake sample data set, with a spherical model fitted, to demonstrate the values of the sill and the nugget. (b) The covariance function that corresponds to the model in (a).

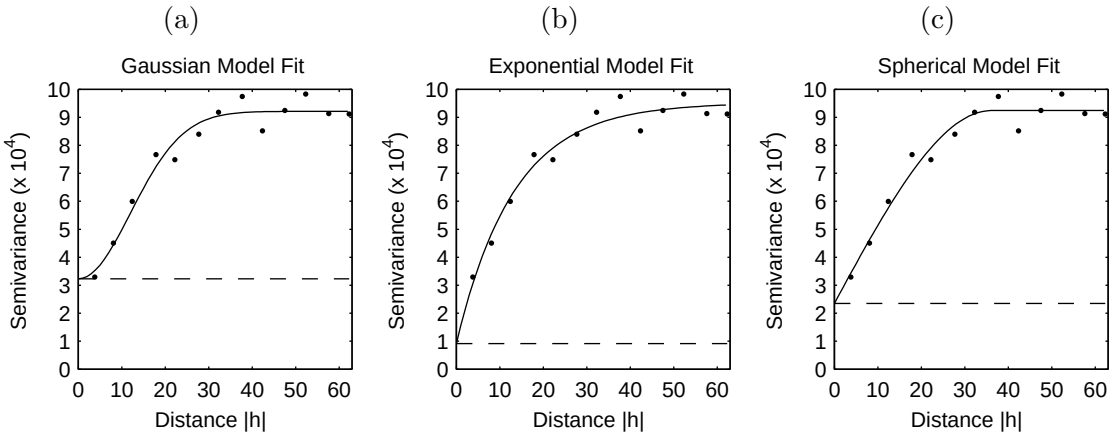

Figure 3: Three different models fitted to the Walker Lake sample semi-variance. (a) Gaussian, (b) Exponential and (c) Spherical. Each model also includes a nugget effect.

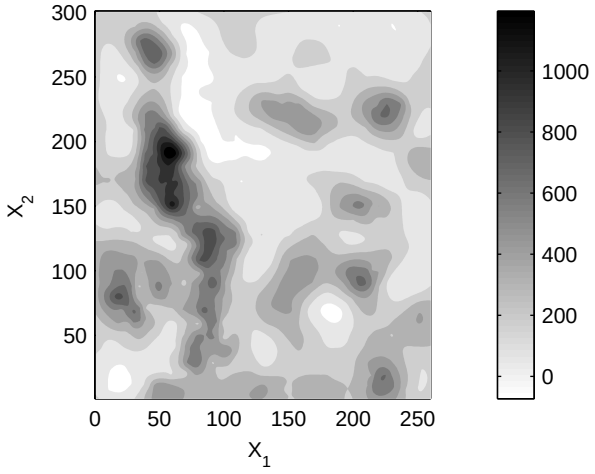

Figure 4: Contour plot of the kriging prediction for the Walker lake data set using a spherical covariance function. The predictions are made over a grid, with density 1.

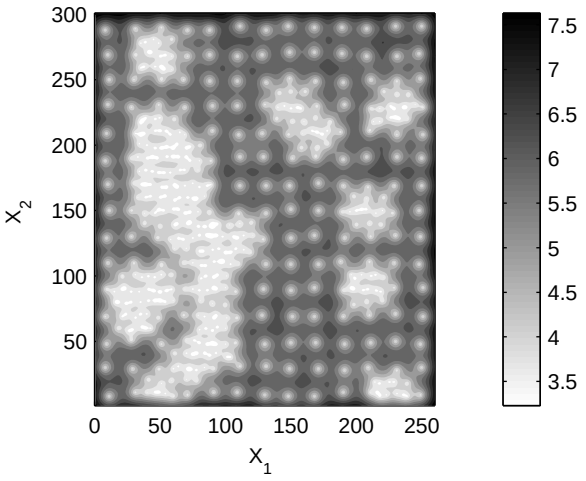

Figure 5: Contour plot of the ordinary kriging error variance σ^2_{OK} evaluated at each point on a grid with density 1 (units $\times 10^4$).

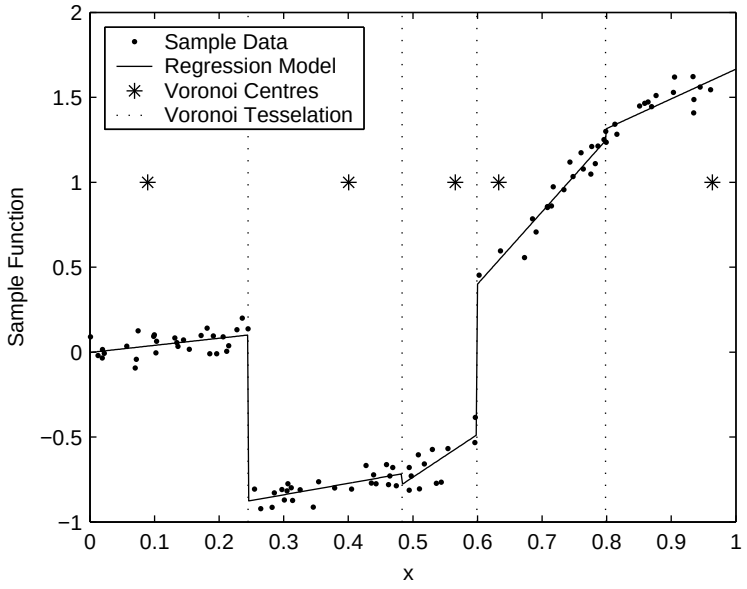

Figure 6: Construction of a single model via partitioning and regression fits to the data

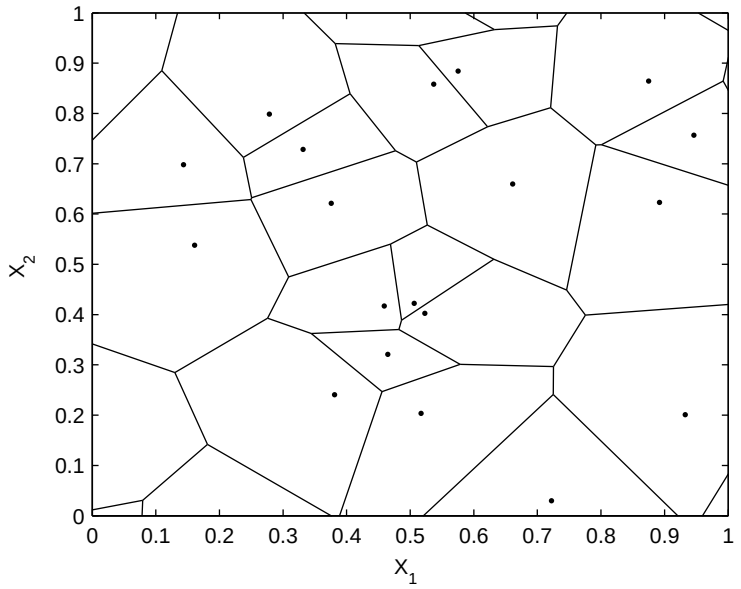

Figure 7: Example of a 2D Voronoi tessellation

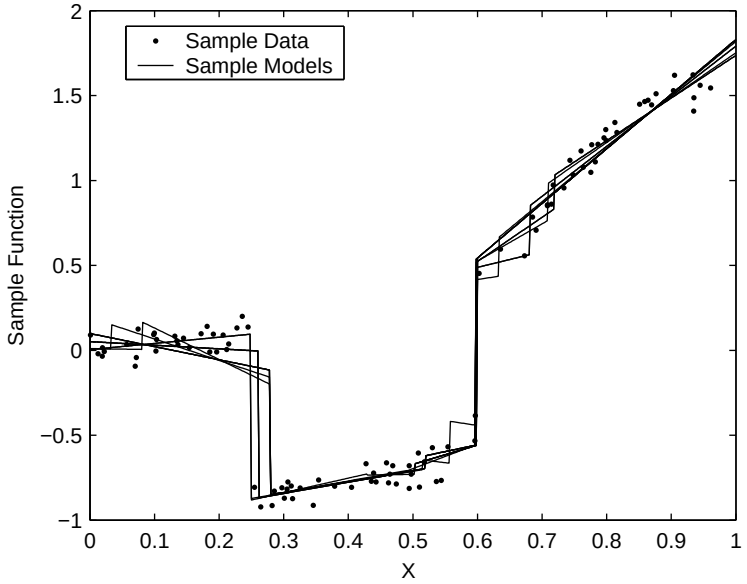

Figure 8: 10 different samples taken from the posterior distribution

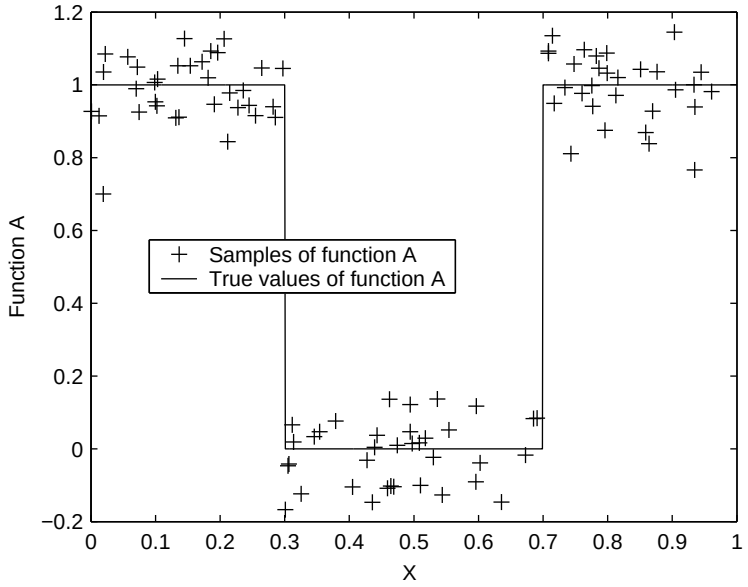

Figure 9: Sample and validation data sets for function A. Sample set comprises 100 random locations, with added gaussian noise $\sim N(0, 0.1)$. Validation set comprises 1000 regularly spaced locations

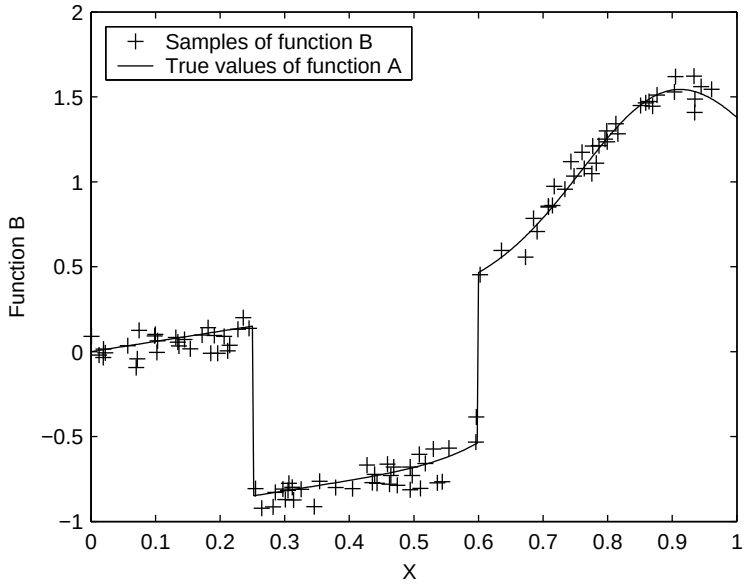

Figure 10: Sample and validation data sets for function B. Sample set comprises 100 random locations, with added gaussian noise $\sim N(0, 0.08)$. Validation set comprises 500 regularly spaced locations

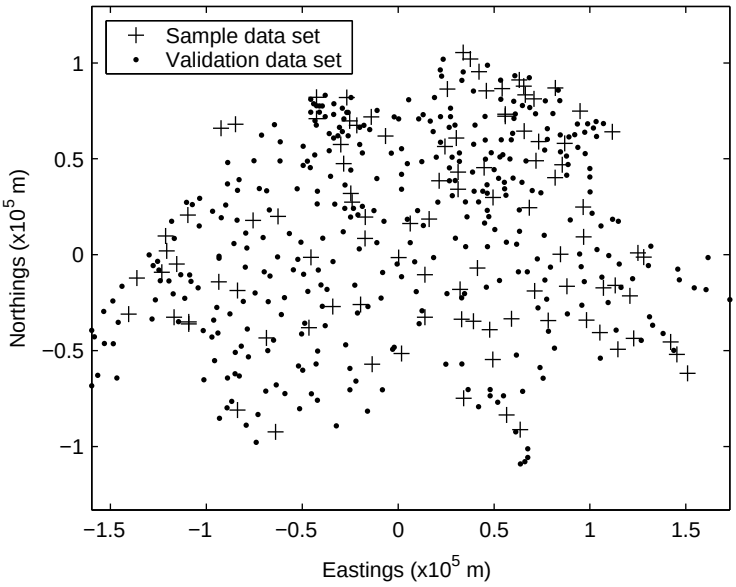

Figure 11: Map showing positions of the sample and validation data sets for the rainwater data sets

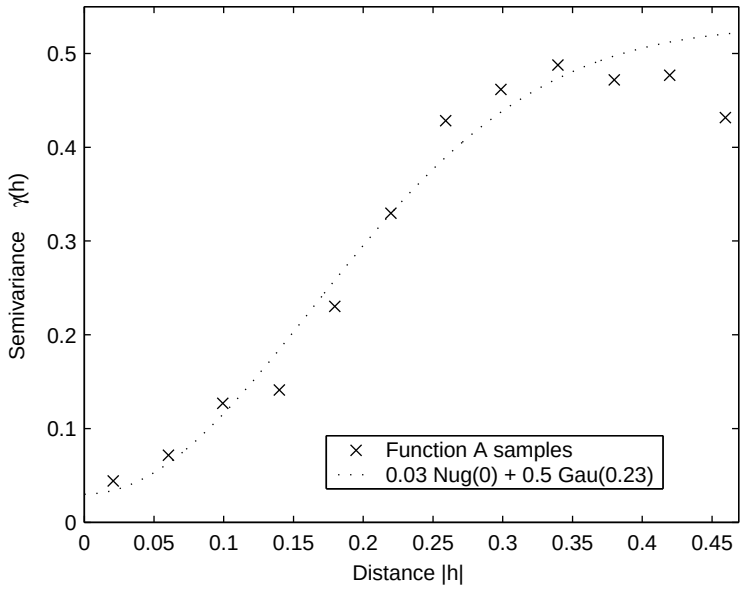

Figure 12: Omnidirectional variogram and fitted gaussian/nugget model for function A, (steplength 0.04)

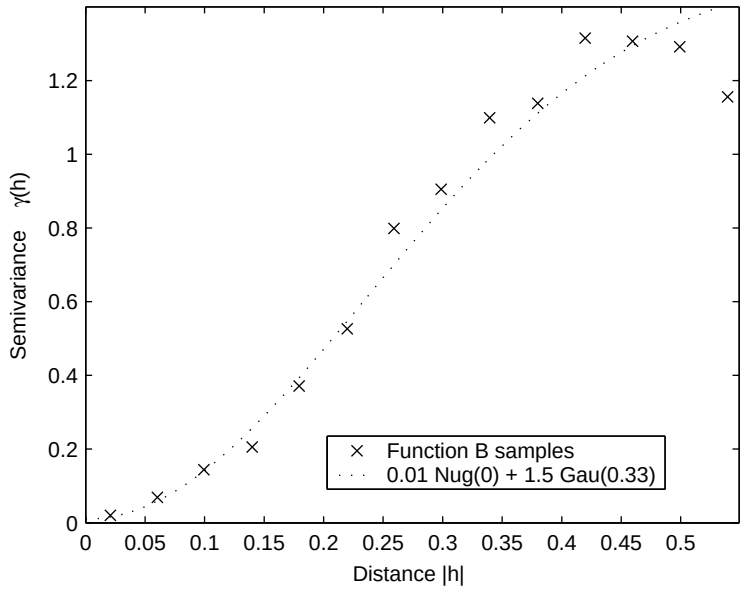

Figure 13: Omnidirectional variogram and fitted gaussian/nugget model for function B, (steplength 0.04)

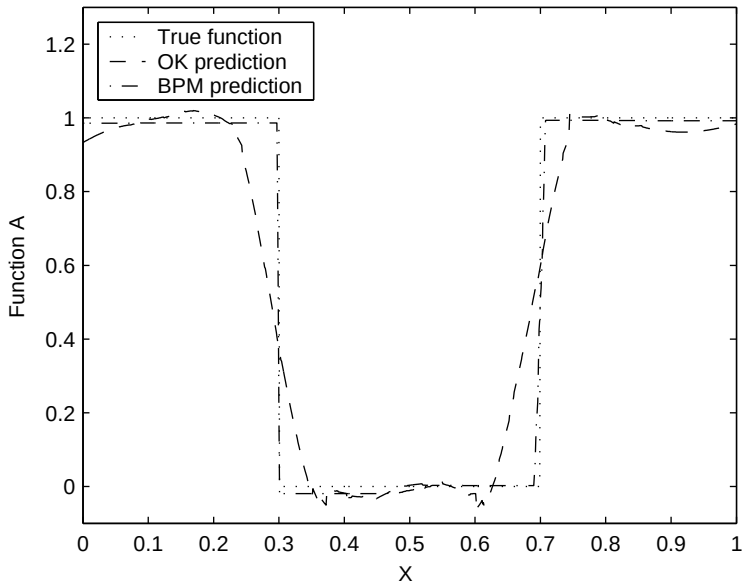

Figure 14: Predictions of function A test data via OK and BPM (20,000 samples)

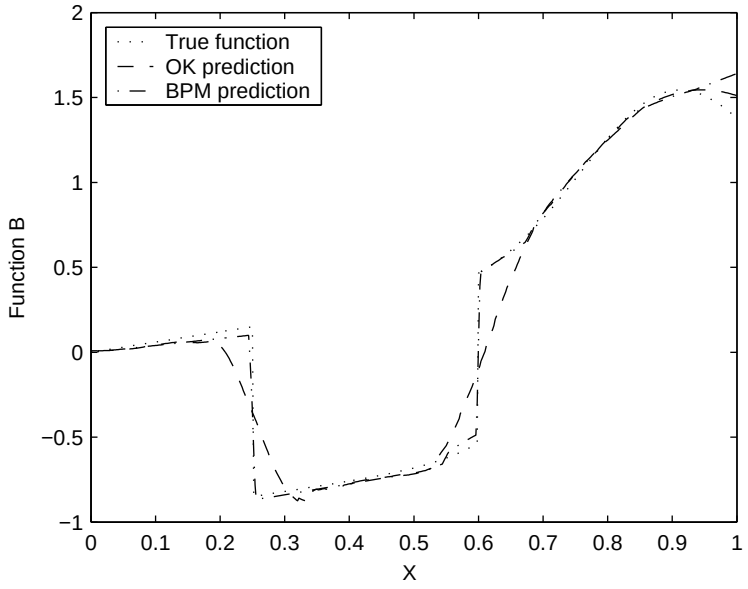

Figure 15: Predictions of function B test data via OK and BPM (50,000 samples)

Table 1: *Prediction errors for functions A and B*

	Function A		**Function B**	
Test	OK	BPM	OK	BPM
RMSE	0.1455	0.0507	-0.1062	-0.0254
MAE	0.0736	0.0165	0.0284	0.0048
rBIAS	-0.0208	-0.0184	0.1428	0.0585
rMSEP	0.0880	0.0107	0.0801	0.0336

Table 2: *Comparison of prediction errors for 4 different methods. The units of RMSE and MAE are 1/10s of mm*

Test	OK	IK	BPM	IDW
rBIAS	-0.0194	.	0.0065	0.0001
rMSEP	0.2887	.	0.2625	0.3824
RMSE	59.71	60.0	56.94	68.72
MAE	41.10	42.6	41.10	50.82

Figure Captions

Fig1_Sec2_walkerlake_positions.eps

Fig. 1: Spatial locations of the Walker Lake sample data set

Fig2a_Sec2_walkerlake_sill_1.eps, Fig2b_Sec2_walkerlake_sill_2.eps

Fig. 2: (a) Omnidirectional semi-variogram of the Walker Lake sample data set, with a spherical model fitted, to demonstrate the values of the sill and the nugget. (b) The covariance function that corresponds to the model in (a).

Fig3a_Sec2_walkerlake_fit_1.eps, Fig3b_Sec2_walkerlake_fit_2.eps, Fig3c_Sec2_walkerlake_fit_3.eps

Fig. 3: Three different models fitted to the Walker Lake sample semi-variance. (a) Gaussian, (b) Exponential and (c) Spherical. Each model also includes a nugget effect.

Fig4_Sec2_walkerlake_contours_1.eps

Fig. 4: Contour plot of the ordinary kriging prediction for the Walker lake data set using a spherical covariance function. The predictions are made over a grid, with density 1.

Fig5_Sec2_walkerlake_contours_2.eps

Fig. 5: Contour plot of the ordinary kriging error variance σ^2_{OK} evaluated at each point on a grid with density 1 (units $\times 10^4$).

Fig6_Sec3_BPM_example_one.eps

Fig. 6: Construction of a single model via partitioning and regression fits to the data

Fig7_Sec3_Voronoi.eps

Fig. 7: Example of a 2D Voronoi tessellation

Fig8_Sec3_BPM_example_many.eps

Fig. 8: 10 different samples taken from the posterior distribution

Fig9_Sec4_stepfunc.eps

Fig. 9: Sample and validation data sets for function A. Sample set comprises 100 random locations, with added gaussian noise N(0, 0.1). Validation set comprises 1000 regularly spaced

locations

`Fig10_Sec4_trench_hill.eps`

Fig.10: Sample and validation data sets for function B. Sample set comprises 100 random locations, with added gaussian noise N(0, 0.08). Validation set comprises 500 regularly spaced locations

`Fig11_Sec4_rainwater.eps`

Fig. 11: Map showing positions of the sample and validation data sets for the rainwater data sets

`Fig12_Sec5_stepfunc_variogram.eps`

Fig. 12: Omnidirectional variogram and fitted gaussian/nugget model for function A, (steplength 0.04)

`Fig13_Sec5_trench_hill_variogram.eps`

Fig. 13: Omnidirectional variogram and fitted gaussian/nugget model for function B, (steplength 0.04)

`Fig14_Sec5_stepfunc_models.eps`

Fig. 14: Predictions of function A test data via OK and BPM (20,000 samples)

`Fig15_Sec5_trench_hill_models.eps`

Fig. 15: Predictions of function B test data via OK and BPM (50,000 samples)

7 Table Captions

Table 1: Prediction errors for functions A and B

Table 2: Comparison of prediction errors for 4 different methods. The units of RMSE and MAE are 1/10s of mm