Robust automatic mapping algorithms in a network monitoring scenario

Ben Ingram[1], Dan Cornford[1], and Lehel Csató[2]

[1] Neural Computing Research Group, Aston University, Aston Street, Birmingham. B4 7ET. United Kingdom. B.R.Ingram@aston.ac.uk

[2] Faculty of Mathematics and Informatics, Universitatea BABES-BOLYAI, Str. Mihail Kogalniceanu, Nr. 1 RO-400084 Cluj-Napoca, Romania. lehel.csato@cs.ubbcluj.ro

1 Abstract

Automatically generating maps of a measured variable of interest can be problematic. In this work we focus on the monitoring network context where observations are collected and reported by a network of sensors, and are then transformed into interpolated maps for use in decision making. Using traditional geostatistical methods, estimating the covariance structure of data collected in an emergency situation can be difficult. Variogram determination, whether by method–of–moment estimators or by maximum likelihood, is very sensitive to extreme values. Even when a monitoring network is in a routine mode of operation, sensors can sporadically malfunction and report extreme values. If this extreme data destabilises the model, causing the covariance structure of the observed data to be incorrectly estimated, the generated maps will be of little value, and the uncertainty estimates in particular will be misleading.

Marchant and Lark [2007] propose a REML estimator for the covariance, which is shown to work on small data sets with a manual selection of the damping parameter in the robust likelihood. We show how this can be extended to allow treatment of large data sets together with an automated approach to all parameter estimation. The projected process kriging framework of Ingram et al. [2007] is extended to allow the use of robust likelihood functions, including the two component Gaussian and the Huber function. We show how our algorithm is further refined to reduce the computational complexity while at the same time minimising any loss of information.

To show the benefits of this method, we use data collected from radiation monitoring networks across Europe. We compare our results to those obtained from traditional kriging methodologies and include comparisons with Box–Cox transformations of the data. We discuss the issue of whether to treat or ignore extreme values, making the distinction between the robust methods which ignore outliers and transformation methods which treat them as part of the (transformed) process. Using a case study, based on an extreme radiological events over a large area, we show how radiation data collected from monitoring networks can be analysed automatically and then used to generate reliable maps to inform decision making. We show the limitations of the methods and discuss potential extensions to remedy these.

2 Introduction

Choosing an appropriate overall model is an important part of interpolating and analysing observations collected from sensor networks. The model should be based on assumptions about the underlying process that generated the observations. Practically speaking, it is almost impossible to exactly specify the correct model which introduces difficulties when attempting to estimate parameters within the model. In this paper we consider the concept of robust geostatistics. By applying robust geostatistical methods we aim to limit the effects of observations that do not correspond to our chosen model. Robust models are frequently employed with datasets where outliers are present, as might often be the case in an automatic monitoring scenario.

The idea of robust geostatistics is not new and has been studied in geostatistics for many years [Cressie and Hawkins, 1980]. In this paper we avoid parameter estimation techniques using method–of–moments based estimators such as those described by Genton [1998] and instead focus on likelihood based approaches such as those proposed by Marchant and Lark [2007]. In this paper, we show how a fast Bayesian projected process kriging framework can be used for robust parameter estimation to generate accurate maps of an area of interest. Using this framework allows the efficient utilisation of most commonly used likelihood functions without having to resort to computationally expensive Markov Chain Monte Carlo (MCMC) sampling techniques as used in other Bayesian methods [Diggle et al., 1998]. As a result, we can experiment with a number of robust likelihood models, in an near real-time framework.

In this paper, by applying a variety of non–Gaussian likelihood models that have heavier tails which help to account for outliers, we compare a number of robust methods. Specification of an appropriate robust likelihood model could be specific to the domain to which it is being applied; our results are particularly relevant to environmental monitoring of radioactivity.

3 Gaussian process

Model based geostatistics makes the assumption that any finite collection of random variables is jointly Gaussian. Here we assume that the data takes the form:

$$(x_i, y_i) : i = 1, \dots, n, \tag{1}$$

where we denote spatial location by x_i and observations at the location x_i are denoted by y_i. Each observation, y_i, is assumed to be a realisation of a random variable Y_i which is dependant on the value of an unobserved random process $S(x)$ [Diggle, Tawn, and Moyeed, 1998].

We assume observations have the following relationship to the underlying process:

$$Y_i = S(x_i) + Z_i, \tag{2}$$

where Z_i is an additive, potentially non–Gaussian, error on the observations that is assumed to be independent for each observation. Equation 2 defines an *arbitrary likelihood function*, $p(Y_i|S(x))$, which we will generally assume has heavy tails to model the outlying observations.

3.1 Gaussian process approximations

We adopt a Bayesian framework for our iterative algorithm. Our aim is to infer the posterior distribution of the underlying random process $S(x)$ given the observed data, $Y = \{Y_i\}_{i=1..n}$. This has the standard form:

$$p(S(x)|Y, \theta) = \frac{\left[\prod_i p(Y_i|S(x))\right] p(S(x)|\theta)}{\int \left[\prod_i p(Y_i|S(x))\right] p(S(x)|\theta) dS(x)} \tag{3}$$

where the posterior is the product of the likelihood terms and the Gaussian process prior, divided by a normalising constant, often called the marginal likelihood, $p(Y|\theta)$.

3.2 Parametrisation of posterior moments

Since we allow for arbitrary likelihood models, in this case robust likelihood models, an exact solution would require the application of MCMC sampling from this very high dimensional posterior distribution, which will be prohibitively computationally expensive for large datasets in our real-time setting. Our approach is to approximate the true non-Gaussian posterior by the *optimal* Gaussian process posterior that minimises the Kullback–Leibler (KL) divergence measure between the *true* posterior distribution and the approximating posterior distribution. By minimising the KL divergence, we match the first two moments of the two distributions [Csató and Opper, 2002].

To enable the use of the arbitrary likelihoods, Equation 2, we represent the Gaussian process by a parametrisation of the posterior moments. The posterior mean is parametrised as:

$$\mu_{posterior}(x) = \mu_{prior}(x) + \sum_{i}^{m} \alpha_i c(x, x_i), \tag{4}$$

where $c(x, x_i)$ is the (*a priori*) covariance function between the point x and the points x_i used in the approximation. We write the covariance between two spatial locations as $c(x, x_i) = cov(x, x_i)$. $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1..n}$ is then the vector of the parameters of the posterior mean of the process. The posterior variance is parametrised as:

$$c_{posterior}(x, x') = c_{prior}(x, x') + \sum_{i,j=1}^{m} c(x, x_i)C_(i, j)c(x_j, x') \tag{5}$$

where $\mathbf{C} = \{C_{i,j}\}_{i,j=1..n}$ is a matrix of parameters for the posterior covariance.

Given the above parametrisation of the posterior moments, we now show how these parameters $\boldsymbol{\alpha}$ and $\mathbf{C}$ can be updated in an iterative algorithm. It was shown in Csató and Opper [2002] that the parametrisation can be applied recursively to give an iterative update rule:

$$\mu_{t+1} = \mu_t + q_{t+1}c_t(x, x_{t+1}), \tag{6}$$

$$c_{t+1}(x, x') = c_t(x, x') + r_{t+1}c_t(x, x_{t+1})c_t(x_{t+1}, x') \tag{7}$$

where t indicates the pseudo–time step in the algorithm or iteration, and x_{t+1} is the spatial location of the new observation being included at iteration $t+1$. The scalar coefficients q_{t+1} and r_{t+1}, which update the model at each iteration can be computed analytically or numerically. The analytic update equations derived in Csató and Opper [2002] are given by:

$$q_{t+1} = \frac{\partial}{\partial[S(x)]} log\langle p(Y_{t+1}|S(x))\rangle_t, \tag{8}$$

$$r_{t+1} = \frac{\partial^2}{\partial[S(x)]^2} log\langle p(Y_{t+1}|S(x))\rangle_t, \tag{9}$$

where the derivatives are with respect to the mean function at time $t+1$ and the expectations, denoted $\langle\cdot\rangle_t$, are taken with respect to the posterior Gaussian process at algorithm pseudo–time t. These update equations essentially process the observations one at a time and update the posterior parametrisation by matching the moments of the updated parametrised posterior to the true, potentially non-Gaussian posterior. Further details can be found in Csató and Opper [2002].

4 Robust likelihood models

Robust likelihood models facilitate the estimation of the variogram parameters in the case where outlying observations are present in the data. If likelihoods which model a ‘robust’ error distribution are used within a traditional model based geostatistical approach then sampling from a potentially high dimensional distribution is required and can be very time consuming.

The method we presented earlier in this paper allows for the specification of arbitrary likelihoods without the large computational overhead that comes with existing MCMC based model based geostatistics. We now present and discuss some robust likelihoods that can be used and compare them to some existing techniques for treating data with outliers.

4.1 Two component Gaussian

We could assume that the observations come from separate processes: a routine process and an extreme process. One approach that seems intuitive is to introduce two components into the likelihood model, one component to model the routine observations and another component to model the extreme observations or outliers. We need not necessarily restrict ourselves to a two component Gaussian likelihood model, but for the purposes of this paper we employ a mixture of two components. Assuming that the routine observations follow a Gaussian distribution is a common hypothesis although this is often an approximation. However assuming that the extreme or outlier observations follow a Gaussian distribution with a large variance could be debated; empirically we have found it works well, although there is little theoretical justification.

The two component Gaussian mixture is constructed by summing two weighted Gaussian distributions to create the mixture likelihood:

$$p(Y_i|S(x_i)) = \beta \mathcal{N}_a(Y_i|S(x_i)) + (1-\beta)\mathcal{N}_b(Y_i|S(x_i)) \tag{10}$$

where β gives the weight of the mixture, or the fraction of the observations that belong to the routine process $\mathcal{N}_a(Y_i|S(x_i))$. We set the variance or noise σ_a^2, of the routine process to model our assumptions about the error in the observation process. The extreme process is denoted by $\mathcal{N}_b(Y_i|S(x_i))$ and a much larger noise σ_b^2 is defined, which represents our beliefs about the extreme process. Alternative mixtures of likelihoods could be considered, but in this paper we will only look at the case where the likelihood models are summed.

4.2 Laplace

A alternative approach which makes yields a cruder robust likelihood model is to assume that the likelihood function has a Laplace distribution. The Laplace distribution has the probability density function:

$$Laplace(x|\mu, b) = \frac{1}{2b} exp\left(-\frac{|x-\mu|}{b}\right) \tag{11}$$

where μ is the location parameter and b is a scale parameter. The Laplace distribution is also known as the double sided exponential distribution.

4.3 Huber functions

One approach to determining robust likelihood models was presented by Marchant and Lark [2007]. Here the Huber function is used in the likelihood term. The Huber function is given by:

$$\rho\left(d\right) = \begin{cases} \frac{1}{2}d^2 & \text{if } |d| \le c \\ c\left|d\right| - \frac{1}{2}c^2 & \text{otherwise} \end{cases} \tag{12}$$

where c is a constant determining the robustness of the estimator. In the case $c = \infty$ the model is equivalent to the standard maximum likelihood estimator, with a Gaussian likelihood model. Rather than optimising the parameter c, here we choose a number of values for c and see which gives the best results. Future work will investigate the selection c using alternative methods.

5 Box–Cox transformations

A standard alternative that is commonly used when a dataset is contaminated with outliers or at least when the dataset is assumed to be non–Gaussian distributed is that of the Box–Cox transformation Box and Cox [1964]. The data is transformed to be approximately Gaussian distributed using:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if} \lambda \neq 0 \\ \log(y) & \text{if} \lambda = 0 \end{cases} \tag{13}$$

and thus the effect of outliers can be reduced, but not completely removed. In this paper we try a number of values for λ to identify which is the most appropriate for the given data. We should note that the Box–Cox approach is very different in character to the preceding approaches, since in the previous methods we have assumed that the outliers arise because of a local corruption to observations, whereas in the Box-Cox approach we transform the entire field, albeit in a manner that attempts to maximise the (marginal) Gaussianity of the observations.

6 Covariance selection

We follow the methodology of Ingram et al. [2005] for determining the covariance function used in the experiments. We use a nested covariance model which has a linear sum of a Gaussian and exponential covariance function components:

$$c_{mix}(u) = \pi \sigma^2_{gau} exp\left(\frac{u^2}{\phi}\right) + (1-\pi)\sigma^2_{exp} exp\left(\frac{u}{\phi}\right). \tag{14}$$

We assume that the exponential component models the short range rough process and that the Gaussian component models the smoother properties of the process at longer lag separations, which is consistent with a belief that at short ranges the radioactivity field is dominated by turbulent mixing processes, while at longer range large scale weather, soil and geological differences dominate.

7 Datasets

To demonstrate the various methods discussed previously, we will use a radiation data collected over the German monitoring network. Radiation data for most countries in Europe is available from the EURDEP (EUropean Radiological Data Exchange Platform) website [3]. We use a dataset with a simulated release of radiation into the environment prepared by BfS[4], which uses the real EURDEP observed background radiation with an added deposition generated from a radiation dispersion model. The simulated release represents some kind of disaster that could potentially take place. The event in this case is not a serious disaster, but rather a small release into the environment over a large area. The release is dispersing in the E–W direction more rapidly than the N–S direction. Anisotropy in the contamination process will present problems for the models. In total there are 1900 observations. We divide this into two sets, a set for estimating the model parameters (1200 observations) and a prediction set for cross validation (700 observations).

8 Results

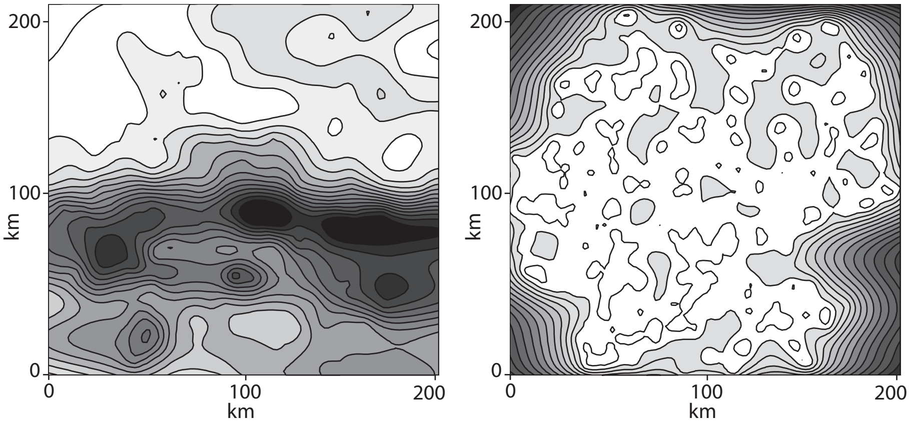

Fig. 1. Contour plot of (left) mean predictions and (right) variance estimates for default (Gaussian likelihood) model.

Contour maps have been produced to show the mean predictions and estimates for the kriging variance. These can be seen in Figures 1–5. The first

[3] http://eurdep.jrc.it/

[4] German Federal Office for Radiation Protection

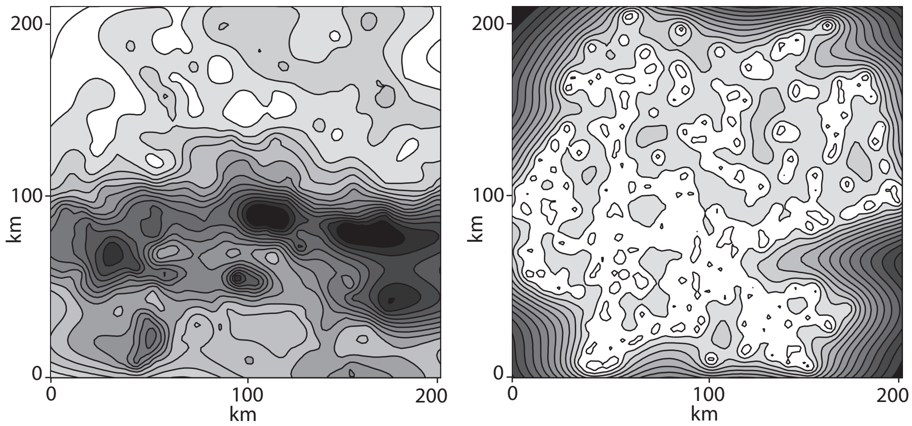

Fig. 2. Contour plot of (left) mean predictions and (right) variance estimates for mixture likelihood model.

thing to note is that each seems to capture the features of the simulated contaminant along the lower middle section of the area. Looking at the Gaussian range for the default method shows how this parameter has become extremely large and this effect can be seen as over smoothing the effect of the contamination. The Huber function and Box–Cox transformation model also suffer somewhat from over estimating the Gaussian range parameter in the E–W direction, but to a lesser degree. The Gaussian Mixture and Laplace models further improve, but anisotropy in the estimation is still marked, which is realistic.

The summary statistics show that the Gaussian Mixture has the lowest error (both MAE and RMSE) of all the methods investigated. The predictions are also more correlated with the observations. The Huber function, Laplace and Box–Cox transformation methods all improve on the default method where no robust assumptions are made, however the improvement is quite small.

The variance plot for the mixture Gaussian likelihood (Figure 2) indicates that the parameters estimated are a good model since there are lower kriging variances than with the other methods, and this is consistent with the observed errors. The mean plot shows how the contaminant has a distinct pattern which cannot be observed in the plot using the default model (Figure 1).

All experiments were carried out on a Pentium 4 2Ghz PC. Since the main difference between these methods was in the specification of the likelihood term, the computational performance was roughly identical across the methods. The computational time, for parameter estimation and prediction was approximately 2 minutes per model.

Table 1. Covariance parameters for the different methods considered. Sill gives the overall sill, summing both components.

Method	Nugget	Sill	Gaussian range	Exp. range	MAE	RMSE	R
Default	0.32	1.35	(530.30, 1.67)	(0.23, 0.12)	0.0010	0.0210	0.83
Gaussian Mixture	0.11	0.67	(48.36, 7.60)	(0.09,0.07)	0.0004	0.0131	0.87
Laplace	0.16	0.75	(58.23, 5.42)	(0.13, 0.19)	0.0006	0.0175	0.86
Huber function	0.22	1.01	(148.37,0.09)	(0.43, 0.09)	0.0009	0.0192	0.86
Box–Cox	0.19	0.90	(136.89, 0.60)	(0.82, 0.77)	0.0007	0.0190	0.86

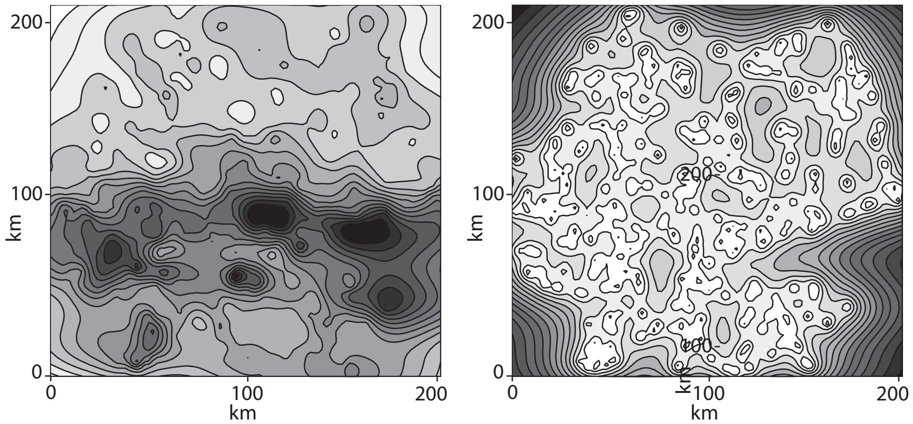

Fig. 3. Contour plot of (left) mean predictions and (right) variance estimates for Laplace likelihood model.

9 Conclusions

In this paper we have presented four methods for treating outliers in datasets. We have shown that the projected process kriging framework with robust likelihoods can be used in the presence of outliers, and on quite large datasets. This is based on using maximum likelihood type II estimates of the parameters in the covariance functions. The overall computational time is under two minutes. This is using an unoptimised Matlab implementation and initial work on a C++ library suggests this can be reduced by an order of magnitude simply by changing the implementation language. Furthermore in other experiments, not shown here, we have processed over ten thousand observations in reasonable time. Employing a Bayesian framework, the Gaussian process prior allows us to make robust inference on the covariance function parameters despite the complex structure in the observations, with possible outliers, which would not be possible with standard method of moments estimators. The Bayesian approach taken here should be called an empirical Bayes (or

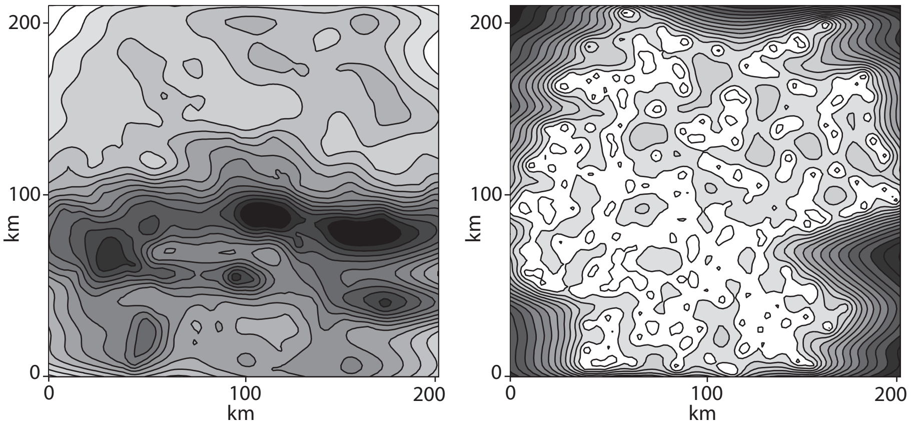

Fig. 4. Contour plot of (left) mean predictions and (right) variance estimates for Huber likelihood model.

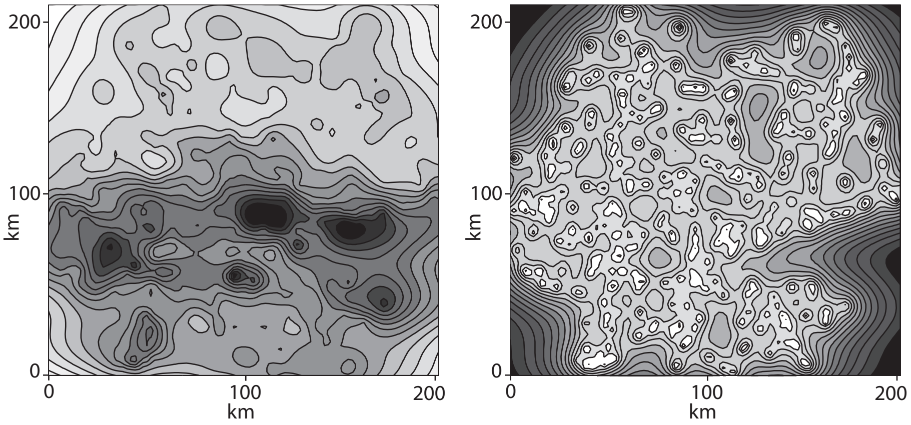

Fig. 5. Contour plot of (left) mean predictions and (right) variance estimates for Box–Cox transformation based model.

plug-in) method since maximum a-posteriori estimates of covariance function parameters are used; it would be interesting to assess the impact of sampling from (and then marginalising with respect to) the parameters in the covariance function. This would require far more computationally expensive sampling methods, but would give a clear indication of the role of parameter uncertainty in (posterior) predictive uncertainty.

The radiological dataset that we have used shows that all four 'robust' methods offer an improvement over standard kriging results, in terms of some standard metrics, however the Gaussian mixture likelihood seems to perform slightly better that other methods in this example. An explanation might be that the 2nd Gaussian component of the mixture likelihood seems to better model the contamination process, although we have not rigorously shown this. The contamination process is more than a few outlying observations, but rather a large number of observations from a second process. The other models may not be able to capture this 'second process' since they are based on heavy tailed distributions which, conceptually at least, arise as the result of a single process.

There is a difference between the robust likelihood methods and the Box–Cox transformation. The robust likelihoods all assume an underlying latent Gaussian process, with observations that are contaminated by heavy tailed, zero mean, symmetric, noise models; their aim is essentially to represent the underlying process filtering the noise appropriately. In the Box–Cox approach the observations are transformed such that their marginal distribution is approximately Gaussian, by a range of transformations from the identity to the log transform. Thus the robustness arises from the squashing affect of the transformation which reduces the impact of large observations (i.e. deals with the skew of the distribution) – the outliers. The key question to consider in choosing an appropriate method is probably more related to assumptions about the form of the noise on the observations together with assumptions about the distribution of the latent process. Note the Box–Cox transformation can only transform variables such that they are marginally Gaussian, not jointly so. In practice, to confirm ones beliefs, it seems that it will always be necessary to compare a range of methods using validation or cross validation to select the empirically best method, even when strong prior information is available.

There are a number of aspects to the modelling process that were only touched on and require further investigation. The selection of the parameters of the likelihood models, for example, estimating the mixing coefficient and variances for each component in the Gaussian mixture model, could be performed automatically rather than being specified *a-priori.* This would also be possible for the c parameter for the Huber function. This is not trivial however, as there is a conceptual difficulty in partitioning the observation errors without additional knowledge, and would probably require a Bayesian treatment, with the effort being applied to defining appropriate priors. So called Trans–Gaussian Kriging [Pilz et al., 2004] incorporates a method to estimate the Box–Cox transformation parameter, which could be incorporated into future models. It is interesting to speculate whether other approaches, such as indicator kriging or copula based methods might also be employed in circumstances where the underlying process has a skewed or otherwise non-Gaussian distribution, potentially also using robust likelihood models to account for the presence of outliers caused by a heavy tailed noise distribution. Finally, although we have not directly tackled this here, in some cases it might be preferable to remove the outlier prior to processing, for example in cases where the outlier represent failure of the observing system or some other catastrophic error.

Acknowledgements

This work is funded by the European Commission, under the Sixth Framework Programme, by the Contract N. 033811 with the DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and are not necessarily those of the European Commission.

References

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, 26(2):211–252, 1964.

N. Cressie and D.M. Hawkins. Robust estimation of the variogram: I. *Mathematical Geology*, 12(2):115–125, 1980.

L. Csató and M. Opper. Sparse online Gaussian processes. *Neural Computation*, 14 (3):641–669, 2002.

P. J. Diggle, J. A. Tawn, and R. A. Moyeed. Model-based geostatistics. *Applied Statistics*, 47:299–350, 1998.

M.G. Genton. Highly Robust Variogram Estimation. *Mathematical Geology*, 30(2): 213–221, 1998.

Ben Ingram, Lehel Csató, and David Evans. Fast spatial interpolation using sparse Gaussian processes. *Applied GIS*, 1(2):15:1–17, 2005.

Ben Ingram, Dan Cornford, and David Evans. Fast algorithms for automatic mapping with space–limited covariance functions. *Stochastic Environmental Research and Risk Assessment*, 2007.

B.P. Marchant and R.M. Lark. Robust estimation of the variogram by residual maximum likelihood. *Geoderma*, 140(1-2):62–72, 2007.

J. Pilz, P. Pluch, and G. Spoeck. Bayesian Kriging with lognormal data and uncertain variogram parameters. In *Proceedings of the Fifth European Conference on Geostatistics for Environmental Applications*. Springer Berlin Heidelberg, 2004.