BAYESIAN AUTOMATING FITTING FUNCTIONS FOR SPATIAL PREDICTIONS

Monica Palaseanu-Lovejoy

Geography, School of Environment & Development,
The University of Manchester, Manchester, UK
Correspondence to: mpal_lovejoy@hotmail.com

A Bayesian predictive model for automating mapping of background radiation has the advantage of fully accounting for all uncertainties in the inferred data. Ten training datasets of background radiation were used to set up the model. The model is robust for data containing only close outliers but fails to accurately predict values when the input data is contaminated with extreme outliers, which are the result of a different random underlying process than the background data. For an integrated decision support system for automating mapping when data contamination is expected, a two stage approach is required in which background data are modeled with one set of equations and the contaminated data with a different set of equations.

1. INTRODUCTION

The SIC2004 exercise required to set up a prediction model using ten training datasets representing background radiation data from Germany. The model should be robust regarding outliers, and optimal for automatic mapping.

This paper describes a Bayesian model of inference using the geoR routine (Ribeiro and Diggle 2001; Diggle and Ribeiro 2002) for the statistical package R (R Development Core Team 2003) and test its suitability in automatic mapping. The training datasets used to set up the inference equations were the result of only one spatial underlying process. This paper demonstrates that the proposed Bayesian model is robust as long as it models similar spatial processes, but fails to accurately make predictions when the basic spatial process is "contaminated" by a second spatial process with a different spatial correlation range than the first one. The second contaminating spatial process is seen as the result of an environmental accident, or point source pollution. None of the training sets provided information on such contamination.

For automatic mapping of environmental data when environmental accidents may occur randomly, a two stage approach is proposed. The input data is tested for extreme outliers likely to be the outcome of a contaminating process. If these outliers are not detected, then proceed with the Bayesian inference model trained on background data, otherwise use a different model that makes no assumption about data. It is argued that in cases of environmental pollution a decision maker may be more interested in the probability that a contaminant value is above a certain action threshold than in the predicted absolute value. This paper presents only the Bayesian model used to infer background radiation data. The same model indicates the existence of a point source pollution process

when used with the contaminated data, but fails to accurately predict the magnitude of the contamination. If the model is used only for visually indicating if accidents occurred or not, without taking in consideration the absolute value of the predictions, then one and the same model may be, arguably, appropriate for automatic mapping in decision making.

2. METHODOLOGY

In general, prediction uncertainty represents the difference between the real value and the predicted value according to a specified model, and usually is equated with error analysis. Interpreted strictly in the statistical (frequentist) sense, uncertainty calculates, by means of a specific confidence level (e.g. 95%), the likelihood or probability that a particular predicted or measured value is within a certain interval, usually centred on that value.

Three main factors contribute to the uncertainty in the predicted values: 1. input uncertainty; 2. simulation variability; and 3. structural uncertainty (McKay et al. 1999). Input uncertainty is linked to the uncertainties already imbedded in the measured values. Simulation variability is linked to stochastic processes and simulations, and represents local uncertainty in a true probabilistic sense. A stochastic simulation generates an array of values with equal probabilities of realization. The realizations' degree of variability describes the prediction uncertainty. The actual extent of the localized variability in the spatial data is defined by all the possible realizations for each coordinate location, and can be statistically described by the maximum, minimum, mean, variance, and skewness values. Structural uncertainty occurs from our assumptions regarding data distribution and behaviour, and from our choice of a particular mathematical paradigm to model the phenomenon under study (Draper 1995; McKay et al. 1999).

In terms of probability statements about parameter θ given data y, the Bayesian rule can be written as (Gelman et al. 1998):

$p(\theta \mid y) = p(y \mid \theta) \bullet p(\theta) / p(y)$, where:

$p(\theta \mid y)$ = the likelihood, or probability that the model is generating the data;

$p(y \mid \theta)$ = probabilistic statement of belief about the model after obtaining the data, or sampling distribution, posterior probability or output;

$p(\theta)$ = probabilistic statement of belief about the model before obtaining the data, the prior probability or expert input;

$p(y)$ = a standardizing constant, independent of the model.

To make inferences about an unknown measurement y, first a prior predictive distribution of y, named also a marginal distribution of x, is considered, such as (Gelman et al. 1998):

$p(y) = \sum_{\theta} p(\theta) p(y \mid \theta)$ for discrete θ, or

$p(x) = \int p(\theta) p(y \mid \theta) d\theta$ in case of continuous θ

The true value θ belongs to a population of measurements x_i with a mean μ and a variance σ^2. After a series of measurements for y, y_1, y_2 … y_n, a new prediction for y, $\tilde{y}$,

can be computed through a posterior predictive distribution of $\tilde{x}$ which is based on previous observations $y_1, y_2 \ldots y_n$ (Gelman et al., 1998):

$$p(\tilde{y}, y) = \int p(\tilde{y}, \theta \mid y) d\theta = \int p(\tilde{y} \mid \theta, y)\, p(\theta \mid y) d\theta = \int p(\tilde{y} \mid \theta)\, p(\theta \mid y) d\theta$$

The second and the third equalities define the posterior predictive distribution as an average of conditional predictions over the posterior distribution of θ (Gelman et al. 1998). Usually, the posterior predictive distribution is computed using simulations, such as Monte Carlo (Gibbs or Marcov Chains) or bootstrapping.

Information about prior distributions can be either obtained from past data or expert knowledge, or from a function of parameters.

Frequentist statistics can be seen as a particular case of Bayesian statistics when all the prior probabilities are equal (flat or noninformative prior), and the posterior probability, or beliefs about the model after obtaining the data, is proportional with the likelihood that the model is generating the data.

Bayesian inference assumes the mean and the covariance parameters to be random variables, and integrates over the parameters' space to get the predictive distribution of any measurement (Ribeiro and Diggle 1999). The formal solution to any estimation or prediction translates into the inference of the conditional distribution of the unknown parameters based on what has been observed. The parameters' prior distribution sums up our knowledge and beliefs about the underlying spatial process. The parameters' posterior distribution is obtained by updating the prior distribution throughout data measurements. Consequently, the Bayesian predictive distribution corresponds to the average of classical predictive distributions for known values of the parameters' underlying spatial process, weighted by the parameters' posterior distribution (Ribeiro and Diggle, 1999). In this case, prediction variance will account for the uncertainty in the parameters' estimations.

In practice, it is often impractical to explicitly develop a mathematical equation for the Bayesian predictive distribution. To overcome this difficulty, the common practice is to generate a random sample from the Bayesian predictive distribution using Monte Carlo sampling algorithms, and calculate from these simulations any statistical parameters of interest. The estimate for the Bayesian point prediction can be either the mean or the median of the simulations, depending of its distribution.

Ribeiro and Diggle (1999) have demonstrated that ordinary or universal kriging methods produce an analytic Bayesian predictive distribution when only the mean parameter is considered unknown and random variable. When all parameters are unknown (i.e. mean, variance and range or correlation parameter) inference by Monte Carlo simulations are used to define the joint posterior distribution and thus the Bayesian predictive distribution.

Diggle and Ribeiro (2002) implemented the geoR routine for the statistical package R using the following algorithms in which the continuous distribution for Φ is replaced by a discret approximation, and where Φ is the scale parameter of the correlation function, β is the mean model parameter, and σ^2 is the data variance:

Algorithm for sampling the parameters' joint distribution:

1. Choose an array of values for the correlation parameter Φ in accordance with our prior knowledge about the underlying spatial process, and assign a discrete uniform prior distribution for Φ based on the previously chosen values;
2. Obtain the Φ posterior probabilities by defining a discrete posterior distribution with probability $p(\Phi|y)$;
3. Sample a value of Φ from the discrete distribution $p(\Phi|y)$;
4. Plug-in the sampled value Φ in the distribution $[\beta, \sigma^2|y, \Phi]$, and sample from this distribution, where β is the mean parameter and σ^2 is the data variance;
5. Repeat points 3 and 4 as many times as desired. The resulting sample (β, σ^2, Φ) represents a sample from the joint posterior distribution.

Algorithm for the Bayesian predictive distribution:

1. Follow steps 1 – 3 from the above algorithm;
2. Plug-in the sampled value Φ into $[y_o|y, \Phi]$ and sample from it to get realizations of the predictive distribution $y_o = S(x_o)$;
3. Repeat the above steps as many times as needed to obtain samples from the predictive distribution.

Different priors for the correlation parameter may be chosen. A flat or noninformative prior implies all Φ values are equally plausible, while a decreasing prior with the possibility to choose the shape but not the scale hyperparameter, or an asymmetric prior representing a certain distribution such as Gamma, or log-normal reflect our belief that certain values are more probable than others (Ribeiro and Diggle 1999). For correlation functions such as the exponential and spherical functions, the Φ parameter represents the "range" parameter with the same significance as in kriging. The Bayesian simulation counterpart of a 95% confidence interval is the difference between 97.5 quintile and 2.5 quintile values for each point simulations.

Theoretically, the kriging prediction variance is smaller than the Bayesian prediction variance, but in reality this is not necessarily the case, especially for prediction values of the unobserved random variables. This can be explain by the fact that kriging estimator is based on available but not necessarily relevant information about the underlying spatial process, while the Bayesian estimator is based on assumptions about the relevant information summarized in a stochastic model (Lantuejoul 2002). Obviously, the more accurate the model describes the physical reality of the underlying spatial process, the better the resulting estimation.

It was demonstrated that the Bayesian methods are superior to classical kriging methods when the spatial process is defined through a relatively small number of measurements with weak spatial correlation and outliers accounting up to 20% of the data points (Palaseanu-Lovejoy 2005). However, Moyeed and Papritz (2002) have distinguished little difference between kriging and Bayesian predictions when the data consist of a very large number of measurements. In the latter case, it seems that a very large dataset makes the Bayesian method insensitive to the choice of prior and gives a robust base for correlation function inference so that the uncertainty introduced by the plug-in parameters values in the kriging methods is minimal.

2.1 USE OF PRIOR INFORMATION

On the website of SIC2004 ten data sets of daily radioactivity measurements were given as training datasets. The mean, median, standard deviation and variance were calculated for each dataset. The "box and whisker" method (Tukey 1977) was employed to check for potential outliers. This method considers the first, second (median) and third quartiles of a data set, knowing that 50% of the data will lie between the first and the third quartiles, which represents the "box". The "whiskers" represent the percentile of the most extreme data-point that is no more than 1.5 (or 3) times the interquartile range from the box. As a result, the upper outliers are considered values larger or equal to the sum of the third quartile and the interquartile range multiplied by 1.5 for "close" outliers, or multiplied by 3 for "extreme" outliers. It was observed that if some datasets had very few close outliers, none of them had extreme outliers. This observation suggested that the employment of robust statistical methods to compute parameters such as mean and variance would suffice to fully characterize the prior distributions. None of the prior datasets gave any information I could successfully use in my analysis to be able to cope with extreme outliers as results of a different spatial process. The lack of such prior data hindered a potential better design for data prediction.

2.2 TUNING THE ALGORITHMS

All computations were done in the package R using the geoR library (Ribeiro and Diggle 2001). The empirical semi-variogram vas computed for 100 lags of 6000 meters each. Same results would be obtained using 20 lags of 30000 meters each. The first option was used for graphical esthetic purposes. The Bayesian model considered all parameters as being unknown. Empirical trials suggested that a transformed Gaussian model, with a circular correlation structure, was the best option.

The priors for the parameters mean (β), variance (σ^2) and range (ϕ) were defined as:

Mean: β = noninformative prior

Range: $prior(\phi) \propto 1/\phi$ with 51 support points of the discrete prior in the interval defined by (0, varr*1000), where varr = robust.variance (data) + robust.mean(data).

Variance: $p(\sigma^2) \propto 1/\sigma^2$

Relative nugget parameter: τ^2/σ^2 = c(0.3, 0.5, 0.19, 0.01) a vector of probabilities of a discrete distribution with support points provided by the vector τ^2 = c(t, 2t, 3t, 4t), where t=-0.0017*classic.mean(data) + 0.1977, and c is the R notation for vector.

To make the parameters more "personal" to the different distributions of the training datasets, their respective classical and robust mean and robust variance were incorporated into the parameters variance (σ^2) and range (ϕ) definitions. These definitions were established through successive trials and errors in the effort to characterize as well as possible all 10 available datasets.

Ten thousands random samples were drawn from the posterior and predictive distributions following the algorithms described above in section 2. The posterior distributions for model parameters are plotted in Figure 1. The data shown belongs to the first training set of data.

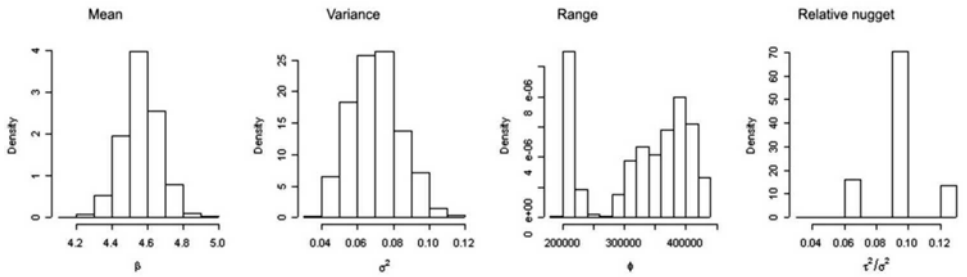

Figure 1

Posterior distributions of model parameters (β, σ^2, ϕ, τ^2/σ^2), first training dataset

The described Bayesian model (mean of simulations) was fitted to all ten training datasets in order to assess its compatibility (Figure 2).

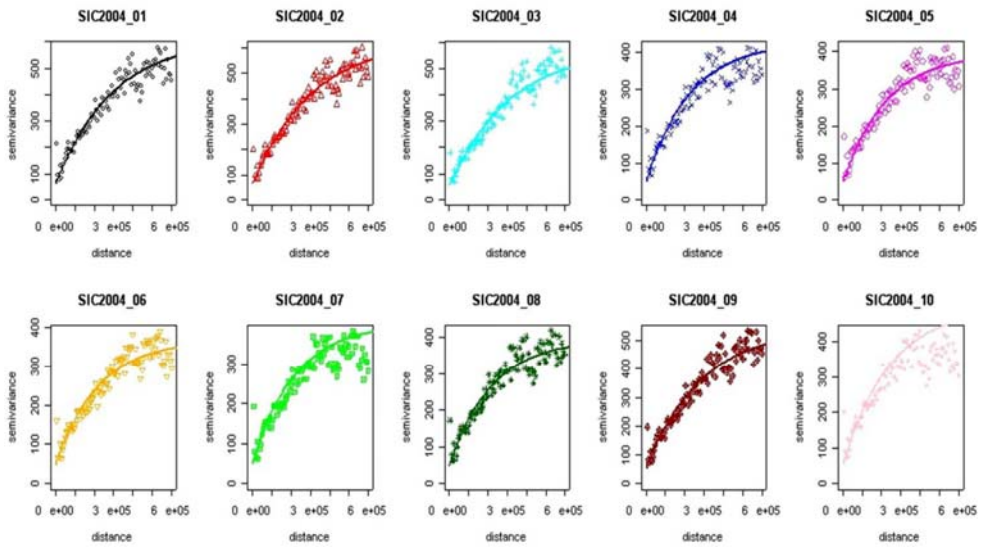

Figure 2

Training datasets' empirical semi-variograms and the Bayesian fitted model.

Without any other additional data available to simulate a radioactive accident in order to assess the model in the presence of extreme outliers, this model was considered robust enough to predict background radioactivity in an automatic way.

3. RESULTS

The Bayesian model described above was used to predict radioactivity values in 808 locations using the first and second datasets provided for validation. The posterior parameters distributions for both datasets are plotted in Figure 3.

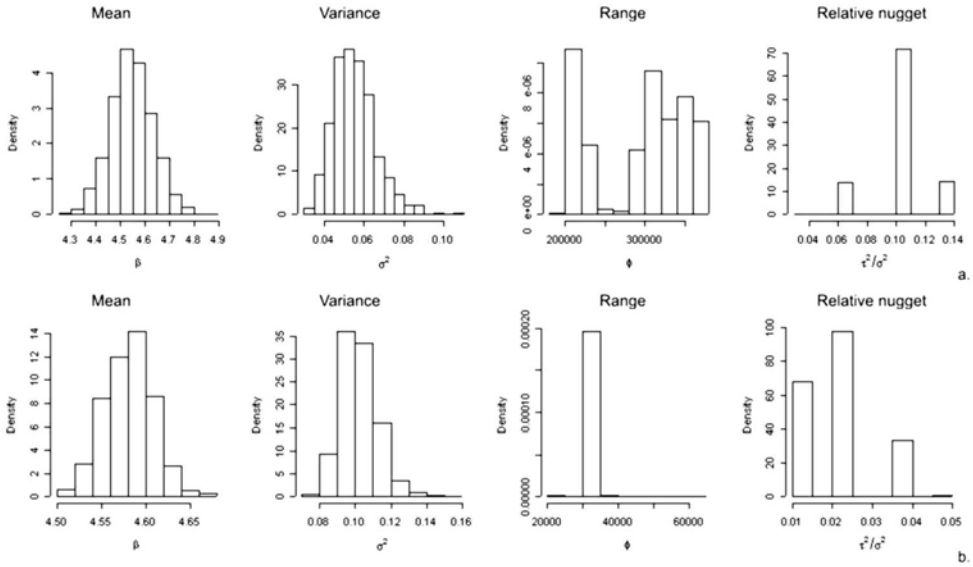

Figure 3
The posterior parameters distributions for the first (a) and the second (b) datasets.

If the first dataset behaves very similar with the training datasets, the second dataset has clearly a very different distribution with extreme outliers. The Bayesian model performs quite well for the middle range values, but clearly fails to predict values close to the extreme outliers. These observations are depicted also by the two datasets empirical semi-variograms and the Bayesian model semi-variograms (Figure 4.).

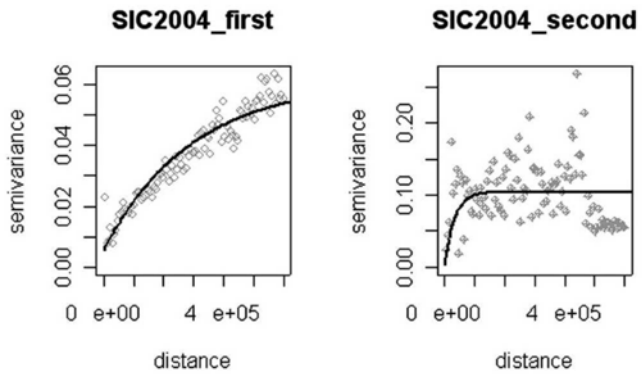

Figure 4
Empirical and Bayesian model semi-variograms for the first and second datasets.

The first dataset is moderately well fitted by the Bayesian model semi-variogram which will slightly sub-estimate high values. The first dataset distribution is very similar with the training datasets 03 and 09 distributions. The Bayesian model will fail to predict

values as high as the extreme outliers present in the second dataset but it will predict satisfactorily the median range values.

3.1. OVERALL RESULTS

The comparisons between the 808 estimated values and the observed values for both datasets are given in the following table:

N = 808	Min.	Max.	mean	median	std. dev.
Observed (first data set)	57	180	98.018	98.8	20.022
Estimates (first data set)	66.6	128.41	96.62	98.52	15.26
Observed (second data set)	57	1528,2	105.42	98.95	83.71
Estimates (second data set)	68.74	419.46	103.09	102.61	24.91

Table 1
Comparison of the estimated and measured values (nSv/h)

The Bayesian method predicts reasonably well the mean and the median of the first and the second datasets, but the Bayesian prediction datasets have a much smaller standard deviation than the measured datasets, especially for the second dataset. Also, if for both datasets the minimum values were slightly overestimated, the maximum value is grossly underestimated for the second dataset.

The comparison of the two datasets' errors and the Pearson's r coefficient of correlation between the estimated and the true values are given in Table 2.

Data sets:	MAE	ME	Pearson's *r*	RMSE
First data set	9.05	1.4	0.79	12.46
Second data set	19.76	2.33	0.5	74.54

Table 2
Comparison of the errors.

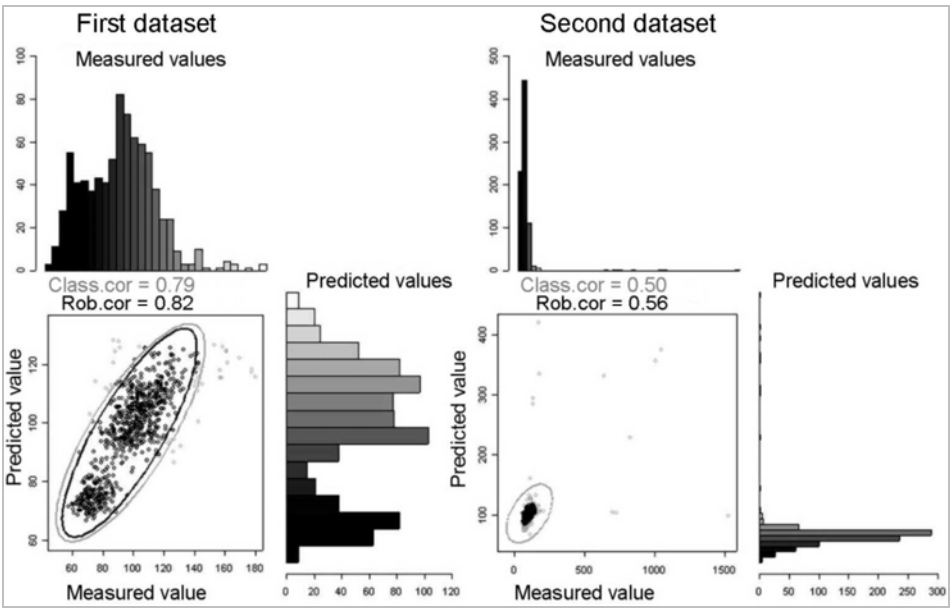

Figure 5
Multivariate analysis of measured and predicted values for first and second datasets.

Above is a multivariate analysis between the measured data and the predicted data for the two datasets (Figure 5).

For both datasets, the correlation between the measured and the prediction values increases when robust methods are used, indicating that extreme values are not as correlated as the majority of the values. In the case of the second dataset, in almost 50% of the cases the predicted values do not follow the distribution of the measured values while for the first dataset the divergence between measured and predicted values is in only 20% of the cases.

1) ESTIMATIONS AND UNCERTAINTY

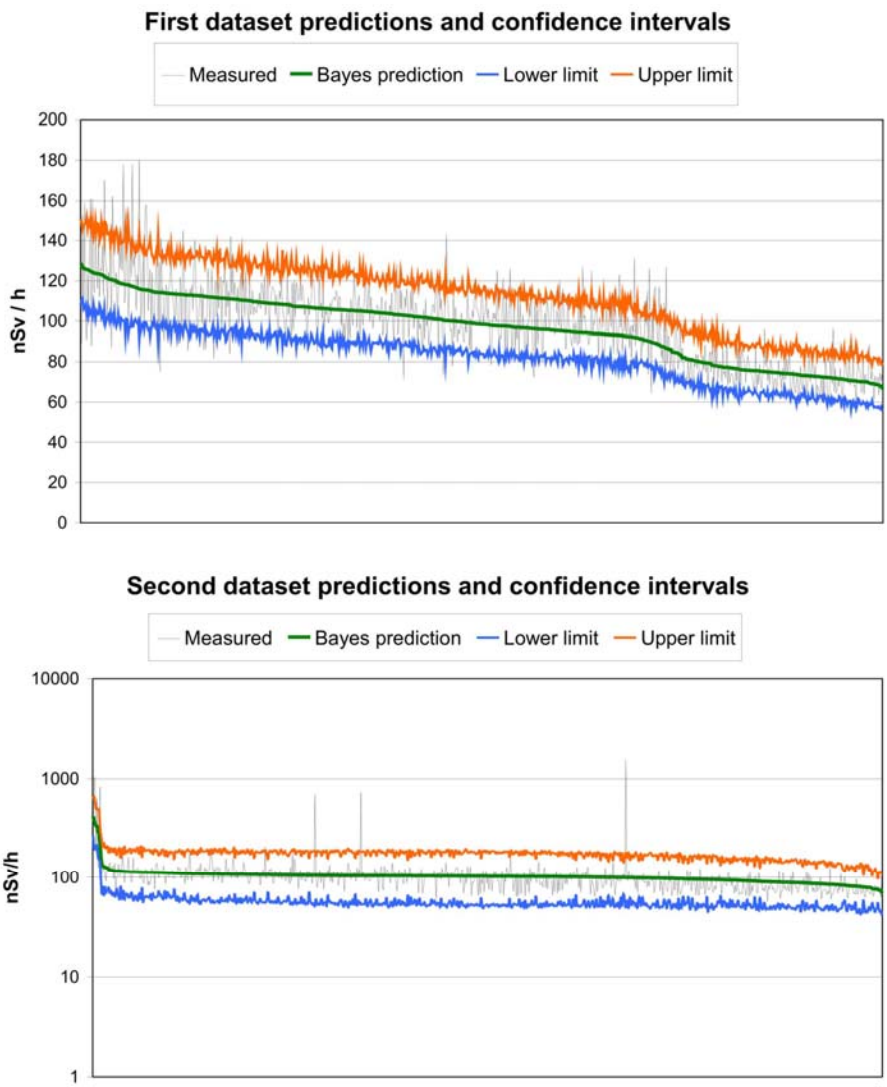

Figure 6
Comparison between the measured and the predicted values and their respective confidence intervals for both datasets.

The Bayesian uncertainty interval for a 95% level of confidence is calculated as the difference between the 97.5 percentile simulations value and 2.5 percentile simulations value, respectively (Ribeiro and Diggle 1999; Diggle and Ribeiro 2002). These intervals are not automatically centred on the mean prediction simulation value since the simulation distributions don't have with necessity a Gaussian distribution. Figure 6, above, shows a comparison between the measured and the predicted values and their respective confidence intervals for both datasets.

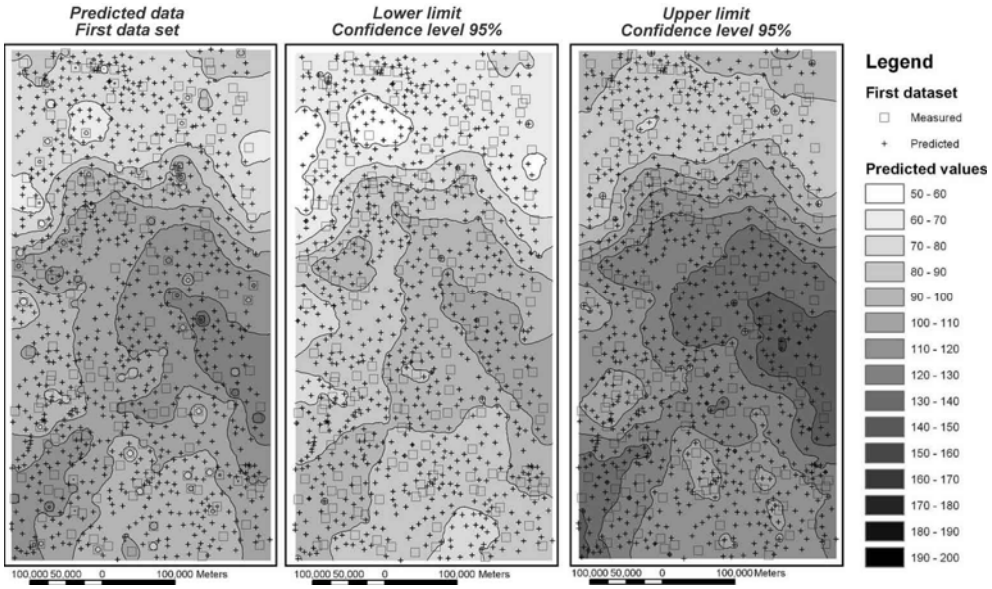

Figure 7
Predicted levels (nSv/h) for the first dataset and corresponding confidence limits.

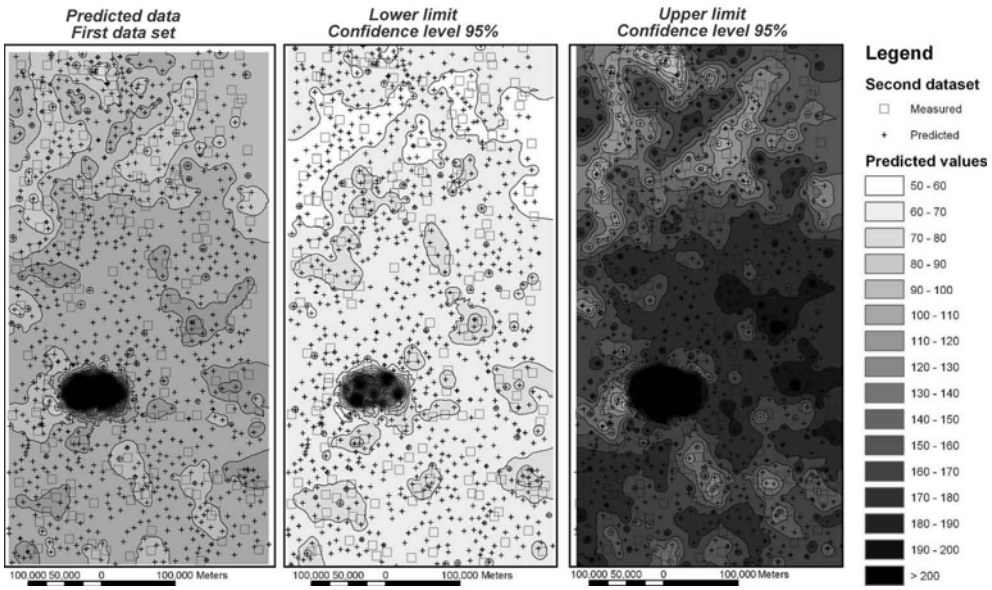

Figure 8
Predicted levels (nSv/h) for the second dataset and corresponding confidence limits.

The interpolated prediction map for the first dataset and its associated lower and upper confidence values are plotted in Figure 7. The next figure (Figure 8) represents the Bayesian predictions and their confidence limits values for the second dataset.

Even if the Bayesian model fits the first dataset better than the second dataset, only 84.4% of the first data measured values are inside the Bayesian confidence interval, while for the second dataset, 98.5 % of the measured values are actually inside the confidence interval.

3.2. DETECTING ANOMALIES AND OUTLIERS.

The extreme values predicted by the Bayesian model are displayed in a 3D map (Figure 9). The grey scale follows the values from the previous 2D maps, with white for values up to 50 nSv/h and black for values above 200 nSv/h. Even if the maximum value measured in the second dataset used for predictions was 1499 nSv/h, the Bayesian model predicted a maximum value of 412.7 nSv/h.

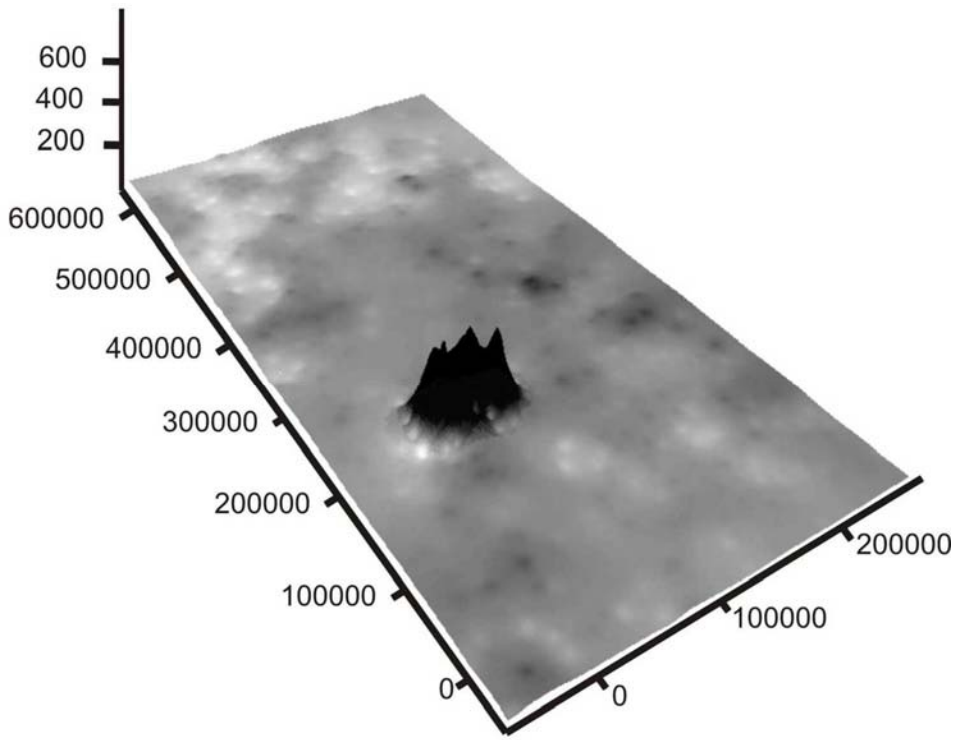

Figure 9
3D map showing extreme values found in the 2nd set (vertical scale in nSv/h).

The Bayesian model seems to be robust enough for datasets with only close outliers, but it fails in the presence of extreme outliers likely to be the outcomes of a contamination process with a different range of spatial correlation. The first dataset has one close outlier and no extreme outliers, while for the second dataset the identified closed outliers are also extreme outliers (Figure 10).

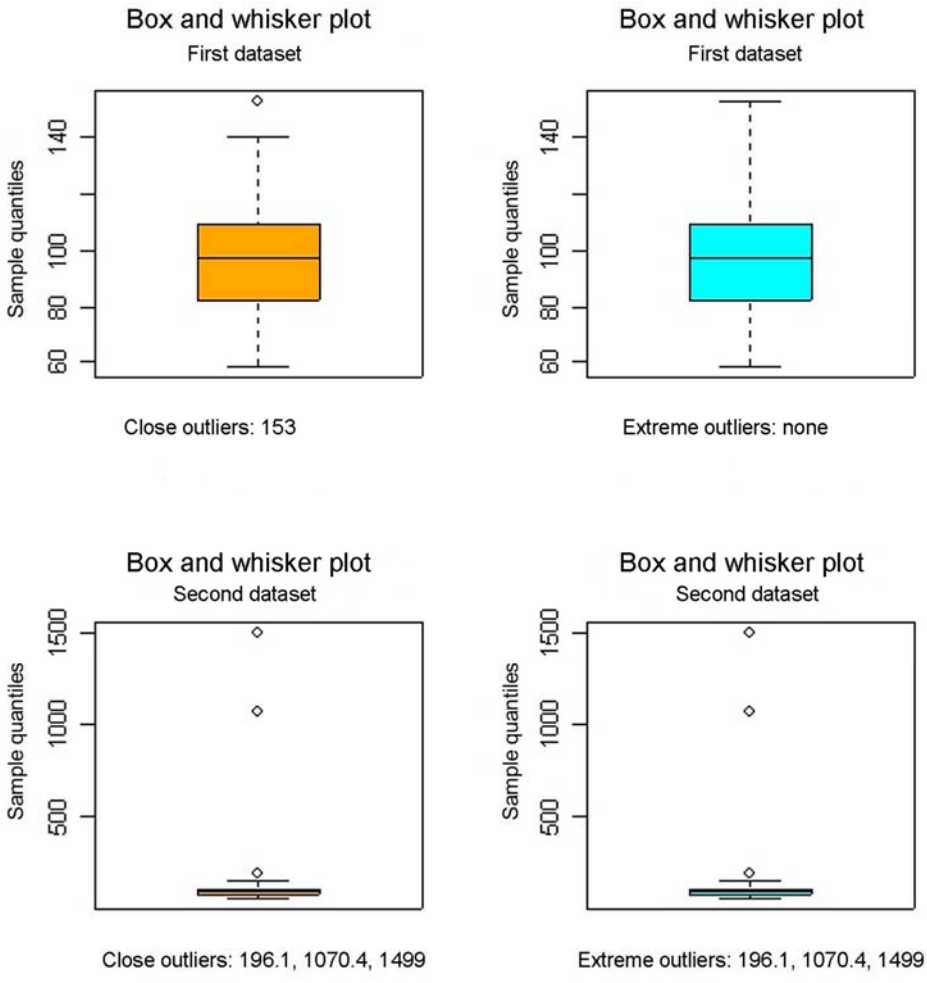

Figure 10
Identified close and extreme outliers for the first and second datasets.

4. DISCUSSION

In Bayesian inference we expect that covariance parameters are random variables and not known as it is the case with the kriging inference. In order to estimate the predictive distribution of any measurement these parameters are integrated over the parameter's space of variance. The parameters' prior distribution is based on what has been observed and summarizes our understanding about the underlying spatial process. Consequently, the Bayesian predictive distribution represents the average of classical predictive distributions for known parameters' values, weighted by the parameters' posterior distribution (Ribeiro and Diggle 1999). In this case, the predictions variance is accounting for the uncertainty in the parameters' estimations and it is expected to be larger than the variance obtained only through the kriging inference process. Implementing a Bayesian model in which all parameters are considered random variables calls for considerable computing

time and resources, but if the original dataset is not a good approximation of the underlying random spatial process, Bayesian predictions are potentially a better estimate than the kriging predictions. To run a full Bayesian procedure for one dataset, as the one described, it takes about 30 minutes on a PC with 2.66 GHz and 2Gb physical internal memory.

The proposed Bayesian model proved to be robust for datasets with few close outliers and no extreme outliers, but fails to do a good prediction for datasets with extreme outliers. Hawkins (1980 p. 1) defined an outlier as *"an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"*. Rousseeuw and Leroy (1987) have described outliers as data having a large influence on the least squares estimator (LSE), that do not follow the trend of the majority of the observations, while Barnett and Lewis (1994 p. 7) consider an outlier to be *"an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"*. From these definitions it is obvious that if a set of data has extreme outliers it may be very difficult to find one single model to predict accurately both extreme outliers and background data. In some cases the task can be unsolvable.

On the other hand, for the purpose of automating interpolations, one may adopt a two stage approach. First, the data is checked for the presence of extreme outliers. If these are absent, then proceed with the Bayesian method described in this paper. If the extreme outliers are present, then use a different method specifically designed for this situation, such as a model that does not make any assumption about the underlying processes. Using this approach the automating interpolations may succeed in predicting more accurately an environmental accident, than using one single "fit" for "all sizes".

5. CONCLUSIONS

Extreme outliers generated by a different process than the one producing the majority of the data are difficult to predict using a model calibrated exclusively on background data. In theory, automatic interpolation is possible, but in practice it seems it is successful only when we try to interpolate data generated by very similar processes. Such a model will fail to predict accurate values when the underlying process is "contaminated" by a secondary process with a totally different range and variance than the first one. In order to cope with contamination situations, a two stage procedure is proposed instead that uses two sets of models, one calibrated solely on background data, and one calibrated to predict contaminated data without any assumptions regarding the spatial processes involved. It is believed that a Bayesian methodology of interpolation is usually superior to classical methods, since the simulation variance and confidence intervals account for the uncertainties introduced by the parameter's values estimations. In addition, if the original dataset does not properly approximate the underlying random spatial process, the Bayesian predictions are potentially a better estimate than the classical estimators.

ACKNOWLEDGMENTS

I would like to thank to Dr. P. Filzmoser for the help he has provided with the multivariate analysis R code.

REFERENCES

Barnett, V; Lewis, T. *Outliers in Statistical Data*. Chichester: John Wiley & Sons Ltd. 1994. 584 p. (Wiley series in probability and mathematical statistics).

Diggle, PJ; Ribeiro, PJ. 'Bayesian inference in Gaussian model-based geostatistics'. *Geographical and Environmental Modelling* 2002; 6 (2): 129–146.

Draper, D. 'Assessment and propagation of model uncertainty'. *J. R. Statist. Soc.* 1995; B 57 (1): 54–97.

Gelman, A; Carlin JB; Stern, HS, et al. *Bayesian Data Analysis*, 4th ed. London: Chapman & Hall; 1995. 526 p.

Hawkins, DM. *Identification of Outliers*. London: Chapman and Hall; 1980. 188 p (Monographs on Applied Probability and Statistics Series).

Lantuejoul, C. *Geostatistical simulation; Models and Algorithms*. London: Springer-Verlag; 2002. 256 p.

McKay, MD; Morrison, JD; Upton, SC. 'Evaluating prediction uncertainty in simulation models'. *Comp. Physics Communications* 1999; 117 (1 – 2): 44–51.

Moyeed, RA; Papritz, A. 'An empirical comparison of kriging methods for nonlinear spatial point prediction'. *Mathematical Geology* 2002; 34 (4): 365–386.

Palaseanu-Lovejoy, Monica. 'Alternative geostatistical approaches to the assessment of a contaminated site' [dissertation]. Manchester (UK): The University of Manchester; 2005 Jan. 297 p.

Ribeiro, PJ; Diggle PJ. (Lancaster University, UK). 1999. 'Bayesian inference in Gaussian model-based geostatistics'. Lancaster (UK): Department of Mathematics and Statistics; 1999. Technical report ST – 99 – 08.

Ribeiro, PJ; Diggle, PJ. 'geoR: A package for geostatistical analysis'. *R-News* 2001; 1 (2): 15-18.

Rousseeuw, PJ; Leroy, AM. *Robust Regression and Outlier Detection*. New York; Chichester: Wiley; 1987. 329 p.

The R Development Core Team [Internet]. 'The R environment for statistical computing and graphics: Reference Index'. Version 1.8.0. TU Wien (AU): Center for Computational Intelligence, c2000 – 2005 [cited 2003 Oct. 15]. Available from: http://www.r-project.org/.

Tukey, J W. *Exploratory Data Analysis*. London (etc.); Reading; Massachusetts: Addison-Wesley; 1977. 688 p.

Cite this article as: Palaseanu-Lovejoy, Monica. 'Bayesian automating fitting functions for spatial predictions'. *Applied GIS*, Vol 1, No 2, 2005. pp. 14-01 to 14-14. DOI: 10.2104/ag050014