



Johns Hopkins University, Dept. of Biostatistics Working Papers

1-6-2008

GEOSTATISTICAL INFERENCE UNDER PREFERENTIAL SAMPLING

Peter J. Diggle

Lancaster University and the Johns Hopkins Bloomberg School of Public Health, p.diggle@lancaster.ac.uk

Raquel Menezes

University of Minho

Ting-li Su

Lancaster University

Suggested Citation

Diggle, Peter J.; Menezes, Raquel; and Su, Ting-li, "GEOSTATISTICAL INFERENCE UNDER PREFERENTIAL SAMPLING" (January 2008). *Johns Hopkins University, Dept. of Biostatistics Working Papers*. Working Paper 162. <http://biostats.bepress.com/jhubiostat/paper162>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

Geostatistical Inference Under Preferential Sampling

Peter J Diggle

(Lancaster University and Johns Hopkins University School of Public Health),

Raquel Menezes

(University of Minho)

and

Ting-li Su

(Lancaster University)

January 6, 2008

Abstract

Geostatistics involves the fitting of spatially continuous models to spatially discrete data (Chilès and Delfiner, 1999). Preferential sampling arises when the process that determines the data-locations and the process being modelled are stochastically dependent. Conventional geostatistical methods assume, if only implicitly, that sampling is non-preferential. However, these methods are often used in situations where sampling is likely to be preferential. For example, in mineral exploration samples may be concentrated in areas thought likely to yield high-grade ore. We give a general expression for the likelihood function of preferentially sampled geostatistical data and describe how this can be evaluated approximately using Monte Carlo methods. We present a model for preferential sampling, and demonstrate through simulated examples that ignoring preferential sampling can lead to seriously misleading inferences. We describe an application of the model to a set of bio-monitoring data from Galicia, northern Spain, in which making allowance for preferential sampling materially changes the inferences.

Key words: environmental monitoring; geostatistics; marked point processes; Monte Carlo inference; preferential sampling; spatial statistics.



1 Introduction

The term *geostatistics* describes the branch of spatial statistics in which data are obtained by sampling a spatially continuous phenomenon $S(x) : x \in \mathbb{R}^2$ at a discrete set of locations $x_i : i = 1, \dots, n$ in a spatial region of interest $A \subset \mathbb{R}^2$. In many cases, $S(x)$ cannot be measured without error. Measurement errors in geostatistical data are typically assumed to be additive, possibly on a transformed scale. Hence, if Y_i denotes the measured value at the location x_i , a simple model for the data takes the form

$$Y_i = \mu + S(x_i) + Z_i : i = 1, \dots, n \quad (1)$$

where the Z_i are mutually independent, zero-mean random variables. We adopt the convention that $E[S(x)] = 0$ for all x , hence in (1) $E[Y_i] = \mu$ for all i . The model (1) extends easily to the regression setting, in which $E[Y_i] = \mu_i = d_i' \beta$, with d_i a vector of explanatory variables associated with Y_i . The objectives of a geostatistical analysis typically focus on prediction of properties of the realisation of $S(x)$ throughout the region of interest A . Targets for prediction might include, according to context: the value of $S(x)$ at an unsampled location; the spatial average of $S(x)$ over A or sub-sets thereof; the minimum or maximum value of $S(x)$; or sub-regions in which $S(x)$ exceeds a particular threshold. Chilès and Delfiner (1999) give a comprehensive account of classical geostatistical models and methods.

Diggle, Moyeed and Tawn (1998) introduced the term *model-based geostatistics* to mean the application of general principles of statistical modelling and inference to geostatistical problems. In particular, they added Gaussian distributional assumptions to the classical model (1) and re-expressed it as a two-level hierarchical linear model, in which $S(x)$ is the value at location x of a latent Gaussian stochastic process and, conditional on $S(x_i) : i = 1, \dots, n$, the measured values $Y_i : i = 1, \dots, n$ are mutually independent, Normally distributed with means $\mu + S(x_i)$ and common variance τ^2 . Diggle, Moyeed and Tawn (1998) then extended this model, retaining the Gaussian assumption for $S(x)$ but allowing a generalized linear model (McCullagh and Nelder, 1989) for the mutually independent conditional distributions of the Y_i given $S(x_i)$.

As a convenient shorthand notation to describe the hierarchical structure of a geostatistical model, we use $[\cdot]$ to mean “the distribution of,” and write $S = \{S(x) : x \in \mathbb{R}^2\}$ and $Y = (Y_1, \dots, Y_n)$. Then, the Diggle, Moyeed and Tawn (1998) model has the simple structure $[S, Y] = [S][Y|S] = [S][Y_1|S(x_1)][Y_2|S(x_2)] \dots [Y_n|S(x_n)]$. Furthermore, in (1) the $[Y_i|S(x_i)]$ are univariate Gaussian distributions with means $S(x_i)$ and common variance τ^2 .

As presented above, and in almost all of the geostatistical literature, models for the data treat the sampling locations x_i either as fixed by design or otherwise stochastically independent of the process $S(x)$, and hence of Y . Admitting the possibility that the sampling design may be stochastic, and writing $X = (x_1, \dots, x_n)$, the structure of the model then becomes $[X, S, Y] = [X][S][Y|S]$, from which it is clear that conditioning on X does not affect inferences about S or Y . We refer to this as *non-preferential sampling* of geostatistical data. Conversely, *preferential sampling* refers to any situation in which $[X, S, Y] \neq [X][S, Y]$.

We contrast the term *non-preferential* with the term *uniform*, the latter meaning that, beforehand, all locations in A are equally likely to be sampled. Examples of designs which are

both uniform and non-preferential include completely random designs and regular lattice designs (strictly, in the latter case, if the lattice origin is chosen at random). An example of a non-uniform, non-preferential design would be one in which sample locations are an independent random sample from a prescribed non-uniform distribution on A . Preferential designs can arise either because sampling locations are deliberately concentrated in sub-regions of A where the underlying values of $S(x)$ are thought likely to be larger (or smaller) than average, or more generally when X and Y are the joint outcome of a marked point process in which there is dependence between the points, X , and the marks, Y .

We emphasise at this point that our definition of preferential sampling is as a stochastic phenomenon. A sampling design that deliberately focuses on sub-regions where the mean of $S(x)$, as opposed to its realised value, is atypically high, is not preferential. However, in most geostatistical applications it is difficult to maintain a sharp distinction between deterministic or stochastic variation in $S(x)$ because of the absence of independent replication of the process under investigation.

Curriero, Hohn, Liebhold and Lele (2002) evaluated a class of non-ergodic estimators for the covariance structure of geostatistical data, which had been proposed by Isaaks and Srivastava (1988) and Srivastava and Parker (1989) as a way of dealing with preferential sampling, but concluded that the non-ergodic estimators “possess no clear advantage” over the traditional estimators that we describe in Section 3.1 below. Schlather, Ribeiro and Diggle (2004) developed two tests for preferential sampling, which treat a set of geostatistical data as a realisation of a marked point process. Their null hypothesis is that the data are a realisation of a *random field model*. This model assumes that the sample locations X are a realisation of a point process \mathcal{P} on A , that the mark of a point at location x is the value at x of the realisation of a random field S on A , and that \mathcal{P} and S are independent processes. This is therefore equivalent to our notion of non-preferential sampling. Their test statistics are based on the idea that, under the null hypothesis that sampling is non-preferential, the low-order moment properties of pairs of measured values Y_i and Y_j should not depend on the distance between the corresponding sampling locations x_i and x_j , and each test is implemented by comparing the observed value of the chosen test statistics with values calculated from simulations of a conventional geostatistical model fitted to the data on the assumption that sampling is non-preferential. Guan and Afsharatus (2007) avoid the need for simulation and parametric model-fitting by dividing the observation into non-overlapping sub-regions that can be assumed to provide approximately independent replicates of the test statistics. In practice, this requires a large data-set; their application has a sample size $n = 4358$.

In this paper, we propose a class of stochastic models and associated methods of likelihood-based inference for preferentially sampled geostatistical data. In Section 2 we define our model for preferential sampling. In Section 3 we use the model to illustrate the potential for misleading inferences when conventional geostatistical methods are applied to preferentially sampled data. Section 4 discusses likelihood-based inference using Monte Carlo methods. Section 5 applies our model and methods to a set of biomonitoring data from Galicia, northern Spain in which the data derive from two surveys, one preferentially sampled the other not, of the same region. Section 6 is a concluding discussion.

2 A shared latent process model for preferential sampling

Recall that S denotes an unobserved, spatially continuous process on a spatial region A , X denotes a point process on A and Y denotes a set of measured values, one at each point of X . The focus of scientific interest is on properties of S , as revealed by the data (X, Y) , rather than on the joint properties of S and X , but we wish to protect against spurious inferences that might arise because of stochastic dependence between S and X .

To clarify the distinction between preferential and non-preferential sampling, and the inferential consequences of the former, we first examine a related situation considered by Rathbun (1996), in which S and X are stochastically dependent but measurements Y are taken only at a different, pre-specified set of locations, i.e. independently of X . Then, the joint distribution of S , X and Y takes the form

$$[S, X, Y] = [S][X|S][Y|S]. \quad (2)$$

It follows immediately on integrating (2) with respect to X that the joint distribution of S and Y has the standard form, $[S, Y] = [S][Y|S]$. Hence, for inference about S it is valid, if potentially inefficient, to ignore X , i.e. to use conventional geostatistical methods. Models analogous to (2) have also been proposed in a longitudinal setting, where the analogues of Y and X are a time-sequence of repeated measurements at pre-specified times and a related time-to-event outcome, respectively. See, for example, Wulfsohn and Tsiatis (1997) or Henderson, Diggle and Dobson (2000).

In contrast, if Y is observed at the points of X , the appropriate factorisation is

$$[S, X, Y] = [S][X|S][Y|X, S]. \quad (3)$$

Even when the algebraic form of $[Y|X, S]$ reduces to $[Y|S]$, an important distinction between (3) and (2) is that in (3) there is a functional dependence between S and X which cannot be ignored; typically, $[Y|S, X] = [Y|S_0]$, where $S_0 = S(X)$ denotes the values of $S(x)$ at all points $x \in X$. The implicit specification of $[S, Y]$ resulting from (3) is therefore non-standard, and conventional geostatistical inferences which ignore the stochastic nature of X are potentially misleading. The longitudinal analogue of (2) arises when subjects in a longitudinal study provide measurements at time-points which are not pre-specified as part of the study design; see, for example, Lipsitz, Fitzmaurice, Ibrahim, Gelber and Lipshultz (2002), Lin, Scharfstein and Rosenheck (2004) or Ryu, Sinha, Mallick, Lipsitz and Lipshultz (2007).

We now define a specific class of models through the following additional assumptions;

A1. S is a stationary Gaussian process with mean μ , variance σ^2 and correlation function $\rho(u; \phi) = \text{Corr}\{S(x), S(x')\}$ for any x and x' a distance u apart;

A2. conditional on S , X is an inhomogeneous Poisson process with intensity

$$\lambda(x) = \exp\{\alpha + \beta S(x)\}; \quad (4)$$

A3. conditional on S and X , Y is a set of mutually independent Gaussian variates with $Y_i \sim N(S(x_i), \tau^2)$.

It follows from A1 and A2 that, unconditionally, X is a log-Gaussian Cox process (Møller, Syversveen and Waagepetersen, 1998). If $\beta = 0$ in (4), then it follows from A1 and A3 that the unconditional distribution of Y is multivariate Gaussian with mean $\mu \mathbf{1}$ and variance matrix $\tau^2 I + \sigma^2 R$, where I is the identity matrix and R has elements $r_{ij} = \rho(\|x_i - x_j\|; \phi)$.

3 Impact of preferential sampling on geostatistical inference

We have conducted a simulation experiment in which we simulated data on A the unit square from an underlying stationary Gaussian process which we then sampled, with additive Gaussian measurement error, either non-preferentially or preferentially according to each of the following sampling designs. For the *completely random* sampling design, sample locations x_i were an independent random sample from the uniform distribution on A . For the *preferential* design, the x_i were generated from the model defined by equation (4), with parameter $\beta = 2$. For the *clustered* design, we used the same model, but with one realisation of S to generate the data Y and a second, independent realisation of S to generate X , thereby giving a non-preferential design with the same marginal properties as the preferential design.

The model for the spatial process S was stationary Gaussian, with mean $\mu = 4$, variance $\sigma^2 = 1.5$, and Matérn correlation with scale parameter $\phi = 0.15$ and shape parameter $\kappa = 1$. In each case, the data y_i consisted of the realised value of $S(x_i)$ plus an independent Gaussian measurement error with mean zero and variance $\tau^2 = 0.25$.

The three panels of Figure 1 show a realisation of each of the three sampling designs superimposed on a single realisation of the process S . The preferential nature of the sampling in the central panel of Figure 1 is clear.

3.1 Variogram estimation

The theoretical variogram of a stationary spatial process $Y(x)$ is the function $V(u) = \text{Var}\{Y(x) - Y(x')\}$ where u denotes the distance between x and x' . Non-parametric estimates of $V(u)$ are widely used in geostatistical work, both for exploratory data analysis and for diagnostic checking.

Consider a set of data $(x_i, y_i) : i = 1, \dots, n$, where x_i denotes a location and y_i a corresponding measured value. The *empirical variogram ordinates* are the quantities $v_{ij} = (y_i - y_j)^2/2$. Under non-preferential sampling, each v_{ij} is an unbiased estimator for $V(u_{ij})$, where u_{ij} is the distance between x_i and x_j . A scatterplot of v_{ij} against u_{ij} or, more usefully, a smoothed version of this scatterplot, can be used to suggest appropriate parametric models for the spatial covariance structure of the data. For more information on variogram estimation, see for example Cressie (1985; 1991, Chapter 2), Chilès and Delfiner (1999) or Diggle and Ribeiro

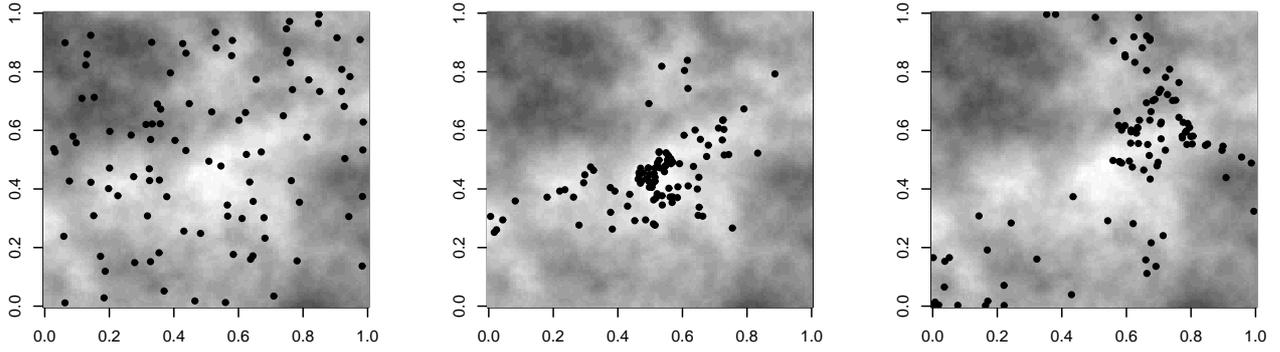


Figure 1: Sample locations and underlying realisations of the signal process for the model used in the simulation study. The left-hand panel shows the completely random sample, the centre-panel the preferential sample and the right-hand panel the clustered sample. In each case, the grey-scale image represents the realisation of the signal process, $S(x)$, used to generate the associated measurement data. The model parameter values are $\mu = 4$, $\sigma^2 = 1.5$, $\phi = 0.15$, $\kappa = 1$, $\tau^2 = 0.25$, $\beta = 2$

(2007, Chapter 5).

The two panels of Figure 2 show simulation-based estimates of the point-wise bias and standard deviation of smoothed empirical variograms, derived from 500 replicate simulations of each of our three sampling designs. With regard to bias, the results under both uniform and clustered non-preferential sampling designs are consistent with the unbiasedness of the empirical variogram ordinates; although smoothing the empirical variogram ordinates does induce some bias, this effect is negligible in the current setting. In contrast, under preferential sampling the results show severe bias. With regard to efficiency, the right-hand panel of Figure 2 illustrates that clustered sampling designs, whether preferential or not, are also less efficient than uniform sampling. The bias induced by preferential sampling is qualitatively unsurprising. The implicit estimand of the empirical variogram is the variance of $Y(x) - Y(x')$ conditional on both x and x' belonging to X , which in general will differ from the unconditional variance; see, for example, Walder and Stoyan (1996) or Schlather (2001).

3.2 Spatial prediction

Suppose that our target for prediction is $S(x_0)$, the value of the process S at a generic location x_0 , given sample data $(x_i, y_i), i = 1, 2, \dots, n$. The widely used ordinary kriging predictor estimates the unconditional expectation of $S(x_0)$ by generalised least squares, but using plug-in estimates of the parameters that define the covariance structure of Y . Traditionally, these plug-in estimates would be obtained by matching theoretical and empirical variograms in some way; we used maximum likelihood estimates under the assumed Gaussian model for Y .

Table 2 shows 95% coverage intervals for the resulting biases and mean square prediction

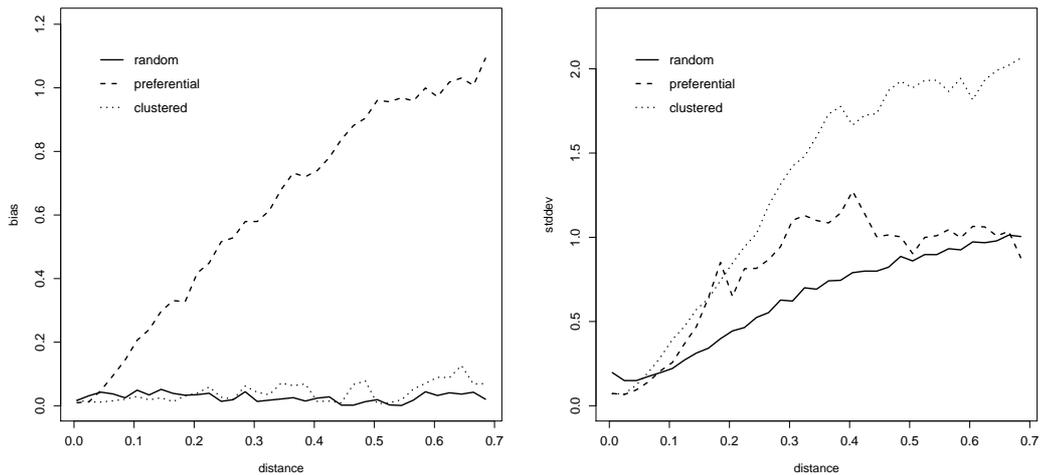


Figure 2: Bias and standard deviation of the sample variogram under random, preferential and clustered sampling. See text for detailed description of the simulation model.

Table 1: Impact of sampling design on the bias and mean square error of the ordinary kriging predictor $\hat{S}(x_0)$, when $x_0 = (0.5, 0.5)$ and each sample consists of 100 locations on the unit square. Each entry in the table is a 95% coverage interval calculated empirically from 500 independent simulations. See text for detailed description of the simulation model.

	Sampling design		
	Completely random	Preferential ($\beta = 2$)	Clustered
bias	(-0.081, 0.059)	(1.290, 1.578)	(-0.082, 0.186)
mean square error	(0.268, 0.354)	(2.967, 3.729)	(0.948, 1.300)

errors of the ordinary kriging predictor $\hat{S}(x_0)$, where $x_0 = (0.5, 0.5)$, in each case evaluated empirically over 500 replicate simulations.

The bias is large and positive under preferential sampling. This prediction bias is a direct consequence of the bias in the estimation of the model parameters, which in turn arises because the preferential sampling model leads to the over-sampling of locations corresponding to high values of the underlying process S . The correct predictive distribution for S is $[S|Y, X]$ which, with known parameter values, takes a standard multivariate Gaussian form whether or not sampling is preferential. The two non-preferential sampling designs both lead to approximately unbiased prediction, as predicted by theory. The substantially larger mean square error for clustered sampling by comparison with completely random sampling reflects the inefficiency of the latter, as already illustrated in the context of variogram estimation.

4 Monte Carlo maximum likelihood estimation

For the shared latent process model (3), the likelihood function for data X and Y can be expressed as

$$L(\theta) = [X, Y] = E_S [[X|S][Y|X, S]], \quad (5)$$

where the expectation is with respect to the unconditional distribution of S . Evaluation of the conditional distribution $[X|S]$ strictly requires the realisation of S to be available at all $x \in A$. In practice, we approximate the spatially continuous realisation of S by the set of values of S on a fine lattice to cover A , and replace the exact locations X by their closest lattice points. We then partition S into $S = \{S_0, S_1\}$, where S_0 denotes the values of S at each of n data-locations $x_i \in X$, and S_1 denotes the values of S at the remaining $N - n$ lattice-points.

To evaluate $L(\theta)$ approximately, a naive strategy would be to replace the intractable expectation on the right hand side of (5) by a sample average over simulations S_j . This strategy fails when the measurement error variance τ^2 is zero, because unconditional simulations of S will then be incompatible with the observed Y . It also fails in practice when the measurement error is small relative to the variance of S , which is the case of most practical interest.

We therefore re-write the exact likelihood (5) as the integral

$$L(\theta) = \int [X|S][Y|X, S] \frac{[S|Y]}{[S|Y]} [S] dS. \quad (6)$$

Now, write $[S] = [S_0][S_1|S_0]$ and replace the term $[S|Y]$ in the denominator of (6) by $[S_0|Y][S_1|S_0, Y] = [S_0|Y][S_1|S_0]$. Note also that $[Y|X, S] = [Y|S_0]$. Then, (6) becomes

$$\begin{aligned} L(\theta) &= \int [X|S] \frac{[Y|S_0]}{[S_0|Y]} [S_0][S|Y] dS \\ &= E_{S|Y} \left[[X|S] \frac{[Y|S_0]}{[S_0|Y]} [S_0] \right] \end{aligned} \quad (7)$$

and a Monte Carlo approximation is

$$L_{MC}(\theta) = m^{-1} \sum_{j=1}^m \left[[X|S_j] \frac{[Y|S_{0j}]}{[S_{0j}|Y]} [S_{0j}] \right], \quad (8)$$

where now the S_j are simulations of S conditional on Y . Note that when Y is measured without error, $[Y|S_{0j}]/[S_{0j}|Y] = 1$. To reduce the Monte Carlo variance, we also use anti-thetic pairs of realisations, i.e. for each $j = 1, \dots, m/2$ set $S_{2j} = 2\mu_c - S_{2j-1}$, where μ_c denotes the conditional mean of S given Y .

To simulate a realisation from $[S|Y]$, we use the following construction. Recall that the data-locations $X = \{x_1, \dots, x_n\}$ constitute a sub-set of the $N \geq n$ prediction locations, $X^* = \{x_1^*, \dots, x_N^*\}$ say. Define A to be the n by N matrix whose i th row consists of $N - 1$ zeros and a single 1 to identify the position of x_i within X^* . Note that, unconditionally, $S \sim \text{MVN}(0, \Sigma)$

and $Y \sim \text{MVN}(\mu, \Sigma_0)$ with $\Sigma_0 = A\Sigma A' + \tau^2 I$. Then, if Z denotes an independent random sample of size n from $N(0, \tau^2)$ and y denotes the observed value of Y , it follows that

$$S_c = S + \Sigma A' \Sigma_0^{-1} (y - \mu + Z - AS) \quad (9)$$

has the required multivariate Gaussian distribution of S given $Y = y$ (Rue and Held, 2005, Chapter 2; Eidsvik, Martino and Rue, 2006). Hence, for conditional simulation when N is large, we need a fast algorithm for unconditional simulation of S , for which we use the circulant embedding algorithm of Wood and Chan (1994) applied to a rectangular region containing the region of interest, A . The subsequent calculations for S_c then involve only the relatively straightforward inversion of the $n \times n$ matrix Σ_0 and simulation of the n independent Gaussian random variables that make up the vector Z in (9).

5 Heavy-metal bio-monitoring in Galicia

Our application concerns bio-monitoring of lead pollution in Galicia, northern Spain. The data consist of two spatial surveys of lead concentrations in moss samples, taken in 1997 and 2000. In the first survey, the sampling design was highly non-uniform and potentially preferential, whereas the second survey used a regular lattice design which is therefore non-preferential. For further details, see Fernández, Rey and Carballeira (2000) and Aboal, Real, Fernández and Carballeira (2005). One objective of analysing these data is to estimate, and compare, maps of lead concentrations in 1997 and 2000. Figure 3 shows the sampling locations for the two surveys.

Table 2: Summary statistics for lead pollution levels measured in 1997 and 2000.

	untransformed		log-transformed	
	1997	2000	1997	2000
Number of locations	63	132	63	132
Mean	4.72	2.15	1.44	0.66
Standard deviation	2.21	1.18	0.48	0.43
Minimum	1.67	0.80	0.52	-0.22
Maximum	9.51	8.70	2.25	2.16

The measured lead concentrations included two gross outliers in 2000, each of which we replaced by the average of the remaining values from that year's survey. Table 2 gives summary statistics for the resulting 1997 and 2000 data. Note that the mean response is higher for the 1997 data than for the 2000 data, which would be consistent either with the former being preferentially sampled near potential pollutant sources, or with an overall reduction in pollution levels over the three years between the two surveys. Also, the log-transformation eliminates an apparent variance-mean relationship in the data and leads to more symmetric distributions of measured values (Figure 4).

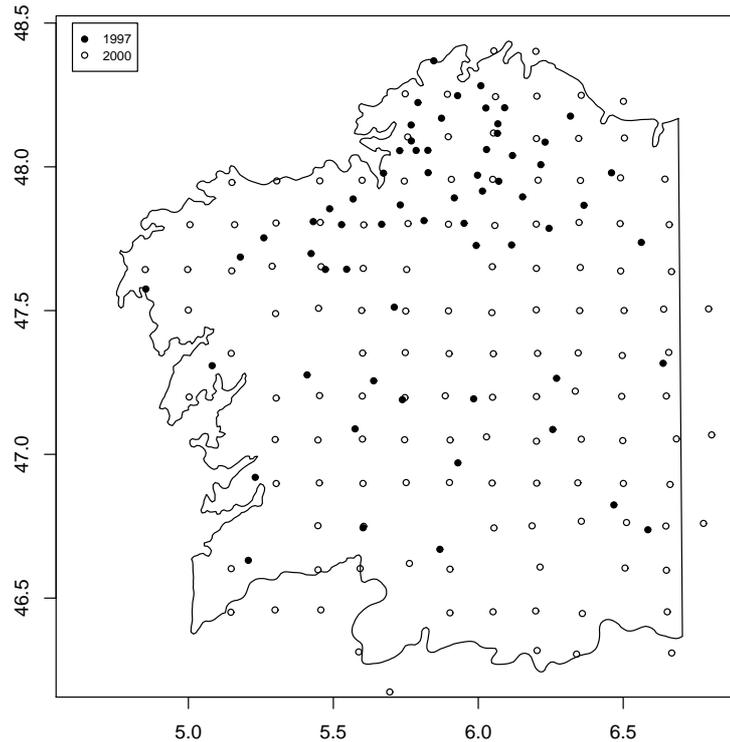


Figure 3: Sampling locations for 1997 (solid dots) and 2000 (open circles). The unit of distance is 100km.

5.1 Standard geostatistical analysis

For an initial analysis, we assume a standard linear Gaussian model for the underlying signal $S(x)$, with mean μ , variance σ^2 , Matérn correlation function $\rho(u; \phi, \kappa)$ and measurement error variance τ^2 , and fit this model separately to the 1997 and 2000 data. The Matérn (1986) class of correlation functions takes the form

$$\rho(u; \phi, \kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^\kappa K_\kappa(u/\phi) : u > 0,$$

where $K_\kappa(\cdot)$ denotes the modified Bessel function of the second kind, of order $\kappa > 0$. This class is widely used because of its flexibility. Although κ is difficult to estimate without extensive data, the integral part of κ determines the degree of mean square differentiability of the corresponding process $S(\cdot)$, giving both a nice interpretation and, in at least some contexts, a rationale for choosing a particular value for κ . The special case $\kappa = 0.5$ gives an exponential correlation function, $\rho(u; \phi) = \exp(-u/\phi)$.

Figure 5 shows, for each of 1997 and 2000, smoothed empirical variograms and theoretical variograms with parameters fitted by maximum likelihood. Based on the general shape of the two empirical variograms, we used a fixed value $\kappa = 0.5$ for the shape parameter of the

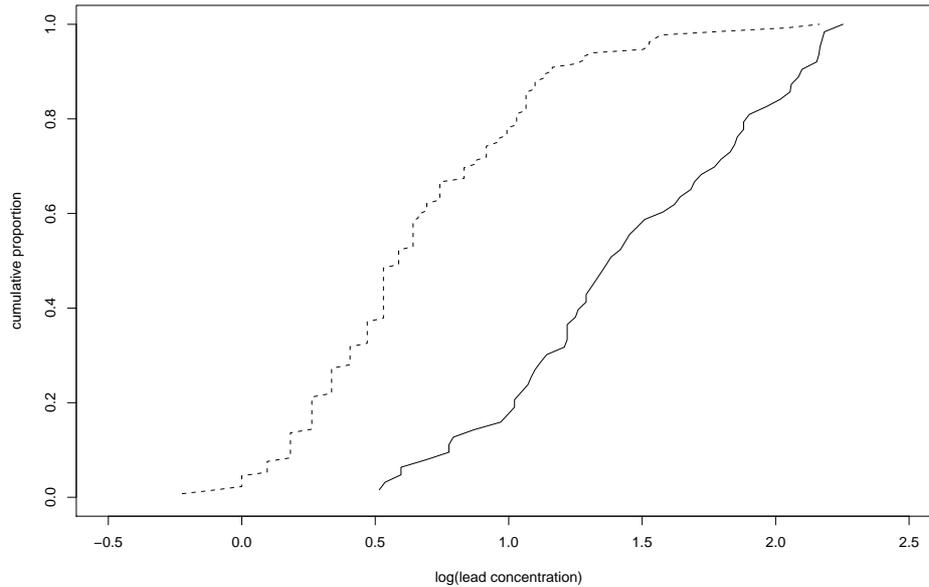


Figure 4: Empirical distributions of log-transformed lead concentrations in the 1997 (solid line) and 2000 (dashed line) samples.

Matérn correlation function. The similarity between the two fitted variograms supports the idea that a joint model for the two data-sets might allow some parameters in common between the two years. The generalised likelihood ratio test statistic (GLRTS) to test the hypothesis of common σ , ϕ and τ , under the dubious assumption that neither sample is preferential, was 7.66 on 3 degrees of freedom ($p = 0.054$). We re-visit this question in the next sub-section.

5.2 Analysis under preferential sampling

We now investigate whether the 1997 sampling is indeed preferential. We used the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) to estimate the model parameters, increasing the number of Monte Carlo samples, m , progressively to avoid finding a false maximum. With $m = 100,000$, the Monte Carlo standard error in the evaluation of the log-likelihood was reduced to approximately 0.3 (the actual value varies over the parameter space) and the GLRTS to test $\beta = 0$ was 27.68 on 1 degree of freedom ($p < 0.001$).

We then fitted a joint model to the two data-sets, treating the 1997 and 2000 data as preferentially and non-preferentially sampled, respectively. To test the hypothesis of shared values for σ , ϕ and τ , we fitted the model with and without these constraints, obtaining a GLRTS of 6.18 on 3 degrees of freedom ($p = 0.103$). The advantage of using shared parameter values when justified is that the parameters in the joint model are then estimated more efficiently and the model is consequently better identified (Altham, 1984). This is particularly important in the geostatistical setting, where the inherent correlation structure of the data reduces their information content by comparison with independent data having the same sample size.

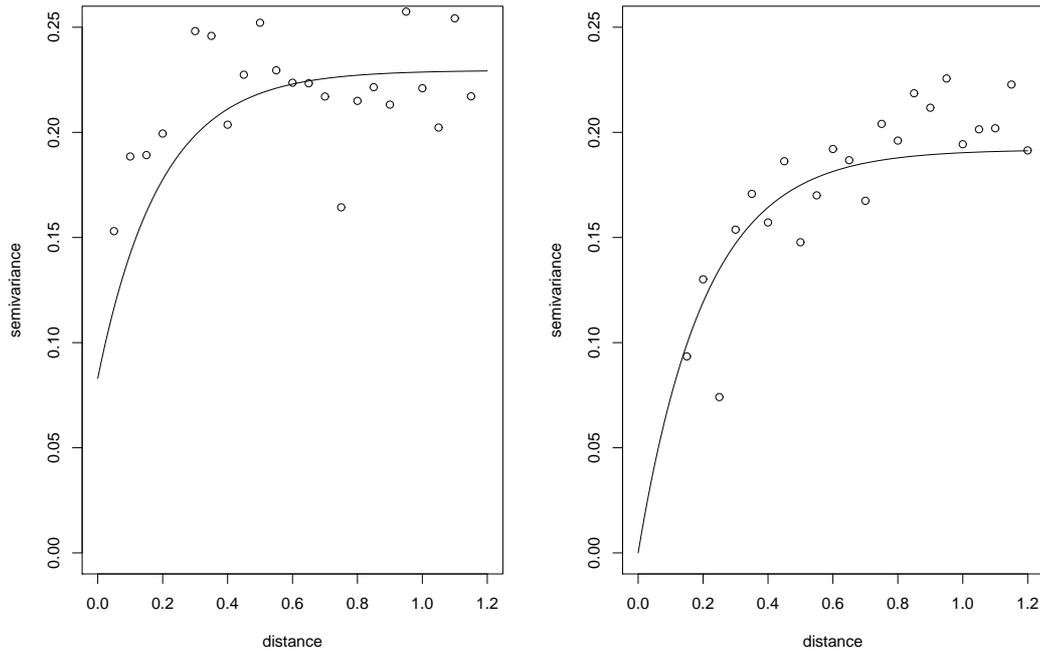


Figure 5: Smoothed empirical (open circles) and fitted theoretical (lines) variograms for 1997 (left-hand panel) and 2000 (right-hand panel) log-transformed lead concentration data.

Table 5.2 shows the Monte Carlo maximum likelihood estimates together with estimated standard errors and correlations for the model with shared σ , ϕ and τ . Standard errors and correlations were evaluated by fitting a quadratic surface to Monte Carlo log-likelihoods by ordinary least squares. Parameter combinations were initially set as a 3^6 factorial design centred on the Monte Carlo maximum likelihood estimates, with parameter values chosen subjectively after examining the trajectories through the parameter space taken by the various runs of the Nelder-Mead optimisation algorithm. The quadratic surface was then re-fitted after augmenting this design with a 2^6 factorial on a more closely spaced set of parameter values, to check the stability of the results. Each evaluation of the log-likelihood used $m = 10,000$ conditional simulations. The non-negative parameters σ , ϕ and τ are estimated on a log-transformed scale, to improve the quadratic approximation to the log-likelihood surface.

Note that the expectation of $S(\cdot)$ shows a substantial fall between 1997 and 2000, and that the preferential sampling parameter estimate is negative, $\hat{\beta} = -1.007$. The latter finding is critically dependent on our allowing the two mean parameters to differ. Otherwise, because the observed average pollution level is substantially higher in 1997 than in 2000, we would have been forced to conclude that the 1997 sampling was preferential with a positive value of β . One piece of evidence against this alternative interpretation is that, within the 1997 data, the observed pollution levels are lower in the over-sampled northern half of the region than in the under-sampled southern half, consistent with a negative value of β .

What impact does the acknowledgement of preferential sampling make on the predicted 1997

Table 3: Monte Carlo maximum likelihood estimates of parameters in the joint model for the 1997 and 2000 Galicia biomonitoring data. Approximate standard errors and correlations are computed from a quadratic fit to the Monte Carlo log-likelihood surface (see text for details)

Parameter	Estimate	Standard error	Correlation matrix						
μ_{97}	1.685	0.193	1.000	0.248	0.301	0.563	0.134	-0.017	
μ_{00}	0.735	0.095	0.248	1.000	0.097	0.255	0.107	-0.124	
$\log(\sigma)$	-0.936	0.044	0.301	0.097	1.000	0.181	-0.547	0.088	
$\log(\phi)$	-1.402	0.065	0.563	0.255	0.181	1.000	0.470	-0.188	
$\log(\tau)$	-1.478	0.040	0.134	0.107	-0.547	0.470	1.000	-0.230	
β	-1.007	0.212	-0.017	-0.124	0.088	-0.188	-0.230	1.000	

pollution surface? Figure 6 shows the predicted surfaces $\hat{T}(x) = E[T(x)|X, Y]$, where $T(x) = \exp\{S(x)\}$ denotes lead concentration on the untransformed scale, together with the pointwise differences between the two. Each surface is a Monte Carlo estimate based on $m = 10,000$ simulations, resulting in Monte Carlo standard errors of 0.026 or less. The predictions based on the preferential sampling model have substantially wider range than those that assume non-preferential sampling (0.836 to 8.358 and 1.273 to 5.989, respectively). The difference surface also covers a relatively large range (-0.756 to 4.221) and shows strong spatial structure. Acknowledgement of the preferential sampling therefore has made a material difference to the prediction of the 1997 pollution surface.

6 Discussion

In this paper, we have shown that conventional geostatistical models and associated statistical methods can lead to very misleading inferences if the underlying data have been preferentially sampled. We have proposed a simple model to take account of preferential sampling and developed associated Monte Carlo methods to enable maximum likelihood estimation and likelihood ratio testing within the proposed class of models. The resulting methods are computationally intensive, but comfortably within the capacity of a modern lap-top PC; all of the computations reported in the paper were run in this mode, using the R software environment and associated CRAN packages. The data and R code are available from the first author on request.

The computation of the Monte Carlo likelihood uses direct simulation, as in Diggle and Gratton (1984), rather than Markov chain Monte Carlo. Hence, issues of convergence do not arise, and the variability between replicate simulations gives a direct estimate of the size of the Monte Carlo error.

We have described an application to a set of environmental bio-monitoring data from Galicia, northern Spain. An important feature of these data is that they are derived from two spatial surveys of the region of interest, only one of which involved preferential sampling.

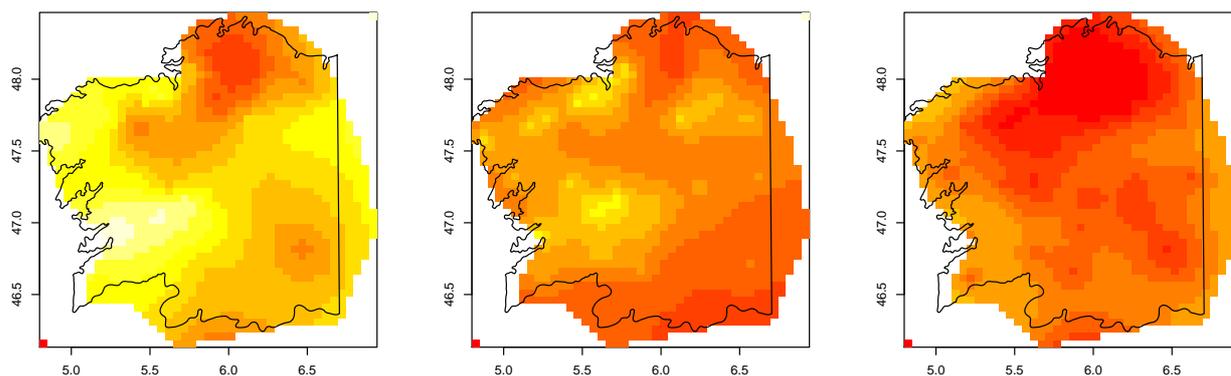


Figure 6: Predicted surface of lead concentrations in 1997 under preferential (left-hand panel) and non-preferential (centre panel) assumptions, together with the pointwise difference between the two (right-hand panel). All three surfaces are plotted on a common scale, from -0.756 (red) to 8.358 (white)

This, coupled with our finding that several of the model parameters can be assumed to take a common value for the two samples, led to a better identified joint model for the two surveys. To illustrate this point, we also fitted the preferential sampling model to the 1997 data alone. Although, as reported earlier, the value of the maximised log-likelihood was obtained relatively easily, the subsequent quadratic fitting method to estimate the standard errors of the maximum likelihood estimates proved problematic. Using a $3^5 + 2^5$ factorial design analogous to the earlier $3^6 + 2^6$ design for the model fitted to the 1997 and 2000 data jointly, and with 10,000 simulations for each log-likelihood evaluation as before, the quadratic fit explained only 72% of the variation in the Monte Carlo log-likelihoods, compared with 93% for the joint model, the implied estimate of $\partial^2 L / \partial \beta^2$ was not significantly different from zero, and the ratio of largest to smallest eigenvalues of the Hessian matrix was 34.5, compared with 22.3 for the joint model.

Alternative strategies for dealing with poorly identified model parameters could include treating the preferential sampling parameter β as a sensitivity parameter, since its value is typically not of direct scientific interest, or using Bayesian methods with informative priors.

A natural response to a strongly non-uniform sampling design is to ask whether its spatial pattern could be explained by the pattern of spatial variation in a relevant covariate. Suppose, for the sake of illustration, that S is observed without error, that dependence between X and S arises through their shared dependence on a latent variable, U , and that the joint distribution of X and S is of the form

$$[X, S] = \int [X|U][S|U][U]dU, \quad (10)$$

so that X and S are conditionally independent given U . If the values of U were to be observed, we could then legitimately work with the conditional likelihood, $[X, S|U] = [X|U][S|U]$ and eliminate X by integration, exactly as is done implicitly when conventional geostatistical

methods are used. In practice, “observing” U means finding explanatory variables which are associated both with X and with S , adjusting for their effects and checking that after this adjustment there is little or no residual dependence between X and S . If so, the analysis could then proceed on the assumption that sampling is no longer preferential. Note, in this context, that any of the proposed tests for preferential sampling can be applied, albeit approximately, to residuals after fitting a regression model for the mean response.

References

- Aboal, J.R., Real, C., Fernández, J.A. and A. Carballeira (2006). Mapping the results of extensive surveys: the case of atmospheric biomonitoring and terrestrial mosses. *Science of the Total Environment*, **356**, 256–274.
- Altham, P.M.E. (1984). Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society*, B **46**, 118–119.
- Baddeley A, Møller J. and Waagepetersen R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* **54**, 329–350.
- Chilès, J-P and Delfiner, P. (1999). *Geostatistics*. New York : Wiley.
- Cox, D.R.(1972). The statistical analysis of dependencies in point processes. In *Stochastic Point Processes*, ed P.A.W. Lewis, 55–66. New York : Wiley.
- Cressie, N.A.C. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association of Mathematical Geology*, **17**, 563–86.
- Curriero, F.C., Hohn, M.E., Liebhold, A.M. and Lele, S.R. (2002). A statistical evaluation of non-ergodic variogram estimators. *Environmental and Ecological Statistics*, **9**, 89–110.
- Diggle, P.J. and Gratton, R.J. (1984). Monte Carlo methods of inference for implicit statistical models (with Discussion). *Journal of the Royal Statistical Society*, B **46**, 196–227.
- Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1998). Model-based geostatistics (with Discussion). *Applied Statistics* **47** 299–350.
- Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*. New York: Springer.
- Eidsvik, J., Martino, S. and Rue, H. (2006). Approximate Bayesian inference in spatial generalized linear mixed models. Technical Report, STATISTICS 2/2006, Norwegian University of Science and Technology, Trondheim, Norway.
- Fernández, J.A., Rey, A. and Carballeira, A. (2000). An extended study of heavy metal deposition in Galicia (NW Spain) based on moss analysis. *Science of the Total Environment*, **254**, 31–44.
- Guan, Y. and Afshartous, D.R. (2007). Test for independence between marks and points of marked point processes: a subsampling approach. *Environmental and Ecological Statistics*, **14**, 101–111.

- Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of measurements and event time data. *Biostatistics*, **1**, 465–480.
- Isaaks, E.H. and Srivastava, R.M. (1988). Spatial continuity measures for probabilistic and deterministic geostatistics. *Mathematical Geology*, **20**, 313–341.
- Lin, H, Scharfstein, D.O. and Rosenheck, R.A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society*, B **66**, 791–813.
- Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J.G., Gelber, R. and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, **58**, 621–630. McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (second edition). London : Chapman and Hall.
- Matérn, B. (1986). *Spatial Variation* (second edition). Berlin: Springer.
- Møller, J., Syversveen, A. and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, **25**, 451–82.
- Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308–313.
- Rathbun, S.L. (1996). Estimation of Poisson intensity using partially observed concomitant variables. *Biometrics*, **52**, 226–42.
- Ripley, B.D. (1976). The second order analysis of stationary point processes. *Journal of Applied Probability* **13**, 255–266.
- Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society B* **39**, 172–212.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall.
- Ryu, D., Sinha, D., Mallick, B., Lipsitz, S.R. and Lipshultz, S.E. (2007). Longitudinal studies with outcome-dependent follow-up: models and Bayesian regression. *Journal of the American Statistical Association*, **102**, 952–961.
- Schlather, M. (2001). On the second-order characteristics of marked point processes. *Bernoulli*, **7**, 99–117.
- Schlather, M., Ribeiro, P. J. and Diggle, P. J. (2004). Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society*, B **66**, 79–93.
- Srivastava, R.M., and Parker, H.M. (1989). Robust measures of spatial continuity. In *Geostatistics, Volume 1*, ed. M. Armstrong, 295–308. Boston: Kluwer.
- Wälder, O. and Stoyan, D. (1996). On variograms of point process statistics. *Biometrical Journal*, **38**, 895–905.
- Wood, A. T. A. and Chan, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *Journal of Computational and Graphical Statistics*, **3**, 409–432.

Wulfsohn, M.S. and Tsiatis, A.A (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

