Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2010)
59, *Part* 2, *pp.* 191–232

Geostatistical inference under preferential sampling

Peter J. Diggle,

Lancaster University, UK, and Johns Hopkins University School of Public Health, Baltimore, USA

Raquel Menezes

University of Minho, Braga, Portugal

and Ting-li Su

Lancaster University, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Environmental Statistics Section *on Wednesday, September 23rd, 2009, the President*, Professor D. J. Hand, *in the Chair*]

Summary. Geostatistics involves the fitting of spatially continuous models to spatially discrete data. 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 that are 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 by using Monte Carlo methods. We present a model for preferential sampling and demonstrate through simulated examples that ignoring preferential sampling can lead to misleading inferences. We describe an application of the model to a set of biomonitoring data from Galicia, northern Spain, in which making allowance for preferential sampling materially changes the results of the analysis.

Keywords: Environmental monitoring; Geostatistics; Log-Gaussian Cox process; Marked point process; Monte Carlo inference; Preferential sampling

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 $x_i, i = 1, \ldots, n$, 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

$$Y_i = \mu + S(x_i) + Z_i, \qquad i = 1, \ldots, n, \tag{1}$$

where the Z_i are mutually independent, zero-mean random variables with variance τ^2, often in this context called the *nugget variance*. One interpretation of the Z_i in model (1) is as measure-

Address for correspondence: Peter J. Diggle, School of Health and Medicine, Lancaster University, Bailrigg, Lancaster, LA1 4YB, UK.
E-mail: p.diggle@lancaster.ac.uk

 0035–9254/10/59191

ment errors in the Y_i. Another, which explains the more colourful terminology, is as a device to model spatial variation on a scale that is smaller than the shortest distance between any two sample locations x_i. We adopt the convention that $E[S(x)] = 0$ for all x; hence in model (1) $E[Y_i] = \mu$ for all i. 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 realization 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 subsets thereof, the minimum or maximum value of $S(x)$ or subregions in which $S(x)$ exceeds a particular threshold. Chilès and Delfiner (1999) have given a comprehensive account of classical geostatistical models and methods.

Diggle *et al.* (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 $S(x_i)$, $i = 1, \ldots, n$, the measured values Y_i, $i = 1, \ldots, n$, are mutually independent, normally distributed with means $\mu + S(x_i)$ and common variance τ^2. Diggle *et al.* (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 $S = \{S(x) : x \in \mathbb{R}^2\}$, $X = (x_1, \ldots, x_n)$, $S(X) = \{S(x_1), \ldots, S(x_n)\}$ and $Y = (Y_1, \ldots, Y_n)$. Then, the model of Diggle *et al.* (1998) implicitly treats X as being deterministic and has the structure $[S, Y] = [S][Y|S(X)] = [S][Y_1|S(x_1)][Y_2|S(x_2)] \ldots [Y_n|S(x_n)]$. Furthermore, in model (1) the $[Y_i|S(x_i)]$ are univariate Gaussian distributions with means $\mu + S(x_i)$ 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 $S(x)$. Admitting the possibility that the sampling design may be stochastic, a complete model needs to specify the joint distribution of S, X and Y. Under the assumption that X is independent of S we can write the required joint distribution as $[S, X, Y] = [S][X][Y|S(X)]$, from which it is clear that for inferences about S or Y we can legitimately condition on X and use standard geostatistical methods. We refer to this as *non-preferential sampling* of geostatistical data. Conversely, *preferential sampling* refers to any situation in which $[S, X] \neq [S][X]$.

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 subregions 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 together form a marked point process in which there is dependence between the points X and the marks Y.

We emphasize at this point that our definition of preferential sampling involves a stochastic dependence, as opposed to a functional dependence, between the process S and the sampling design X. For example, a model in which the mean of S and the intensity of X share a dependence

on a common set of explanatory variables does not constitute preferential sampling. In most geostatistical applications it is difficult to maintain a sharp distinction between the treatment of variation $S(x)$ as deterministic or stochastic because of the absence of independent replication of the process under investigation. Our pragmatic stance is to represent by a stochastic model the portion of the total variation in S that cannot be captured by extant explanatory variables.

Curriero *et al.* (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. They concluded that the non-ergodic estimators 'possess no clear advantage' over the traditional estimators that we describe in Section 3.1. Schlather *et al.* (2004) developed two tests for preferential sampling, which treat a set of geostatistical data as a realization of a marked point process. Their null hypothesis is that the data are a realization of a *random-field model*. This model assumes that the sample locations X are a realization of a point process $\mathcal{P}$ on A, that the mark of a point at location x is the value at x of the realization 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 following idea. Assume that S is stationary, and let $M_k(h) = E[S(x)^k | x, x+h \in \mathcal{P}]$. Under the null hypothesis that sampling is non-preferential, the conditioning on $x+h \in \mathcal{P}$ is irrelevant; hence $M_k(h)$ is a constant. Schlather *et al.* (2004) proposed as test statistics the empirical counterparts of $M_1(h)$ and $M_2(h)$, and implemented the resulting tests by comparing the observed value of each chosen test statistic with values calculated from simulations of a conventional geostatistical model fitted to the data on the assumption that sampling is non-preferential. Guan and Afshartous (2007) avoided the need for simulation and parametric model fitting by dividing the observation into non-overlapping subregions 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 inferences to be misleading when conventional geostatistical methods are applied to preferentially sampled data. Section 4 discusses likelihood-based inference using Monte Carlo methods and suggests a simple diagnostic for the fitted model. 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 of the same region, one of which is preferentially sampled, the other not. Section 6 is a concluding discussion.

The data that are analysed in the paper can be obtained from

`http://www.blackwellpublishing.com/rss`

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 incorrect 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 that was considered

by Rathbun (1996), in which S and X are stochastically dependent but measurements Y are taken only at a different, prespecified set of locations, i.e. independently of X. Then, the joint distribution of S, X and Y takes the form

$$[S, X, Y] = [S][X|S][Y|S]. \tag{2}$$

It follows immediately on integrating equation (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 that are analogous to equation (2) have also been proposed in a longitudinal setting, where the analogues of Y and X are a time sequence of repeated measurements at prespecified times and a related time-to-event outcome respectively. See, for example, Wulfsohn and Tsiatis (1997) or Henderson *et al.* (2000).

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

$$[S, X, Y] = [S][X|S][Y|X, S]. \tag{3}$$

Even when, as is typical in geostatistical modelling, equation (3) takes the form

$$[S, X, Y] = [S][X|S][Y|S(X)], \tag{4}$$

so that the algebraic form of $[Y|X, S]$ reduces to $[Y|S(X)]$, an important distinction between preferential and non-preferential sampling is that in equation (4) the functional dependence between S and X in the term $[Y|S(X)]$ cannot be ignored, because the implicit specification of $[S, Y]$ resulting from equation (4) is non-standard. Conventional geostatistical inferences which ignore the stochastic nature of X are therefore potentially misleading. The longitudinal analogue of equation (4) arises when subjects in a longitudinal study provide measurements at time points which are not prespecified as part of the study design; see, for example, Lipsitz *et al.* (2002), Lin *et al.* (2004) or Ryu *et al.* (2007).

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

Assumption 1. S is a stationary Gaussian process with mean 0, variance σ^2 and correlation function $\rho(u; \phi) = \mathrm{corr}\{S(x), S(x')\}$ for any x and x' a distance u apart.

Assumption 2. Conditional on S, X is an inhomogeneous Poisson process with intensity

$$\lambda(x) = \exp\{\alpha + \beta S(x)\}. \tag{5}$$

Assumption 3. Conditional on S and X, Y is a set of mutually independent Gaussian variates with $Y_i \sim N\{\mu + S(x_i), \tau^2\}$.

It follows from assumptions 1 and 2 that, unconditionally, X is a log-Gaussian Cox process (Møller *et al.*, 1998). If $\beta = 0$ in equation (5), then it follows from assumptions 1 and 3 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 $r_{ij} = \rho(||x_i - x_j||; \phi)$.

Ho and Stoyan (2008) discussed essentially the same construction as assumptions 1–3 viewed as a model for a marked point process of locations X and marks Y, and derived its first- and second-moment properties.

We do not suggest that this model will be adequate for all applications. However, it is sufficiently rich to provide a vehicle for investigating the consequences of preferential sampling, and for the application that is described in Section 5 of the paper.

3. Effect of preferential sampling on geostatistical inference

To illustrate how preferential sampling affects the performance of standard geostatistical methods, we have conducted a simulation experiment as follows. For each run of the experiment, we first simulated an approximate realization of a stationary Gaussian process on the unit square by simulating its values on a finely spaced lattice and treating the spatially continuous process $S(\cdot)$ as constant within each lattice cell. We then sampled the values of $S(\cdot)$, 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, we first generated a realization of S, then a realization of $X=\{x_i : i=1,\ldots,n\}$ conditional on S by using the model that is defined by equation (5) with parameter $\beta=2$, and finally a realization of $Y=\{y_i : i=1,\ldots,n\}$ conditional on S and on X by using the conditional model that is defined by assumption 3 above. For the *clustered* design, we used the same procedure to generate a realization of X, but then generated a realization of Y on the locations X by using a second, independent, realization of S, so that the resulting Y is a realization of the standard geostatistical model (1). This gives a non-preferential design with the same marginal properties for X and Y 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, we set the nugget variance, $\tau^2=0$; hence the data y_i consisted of the realized values of $S(\cdot)$ as the sample locations x_i.

The Matérn (1986) class of correlation functions takes the form

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

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, $\rho(u;\phi)=\exp(-u/\phi)$.

Fig. 1 shows a realization of each of the three sampling designs superimposed on a single realization of the process S. The preferential nature of the sampling in Fig. 1(b) results in the sample locations falling predominantly within the darker shaded areas.

3.1. Variogram estimation

The theoretical variogram of a stationary spatial process $Y(x)$ is the function $V(u)=\frac{1}{2}\,\mathrm{var}\{Y(x)-Y(x')\}$ 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 $(x_i, y_i), i=1,\ldots,n$, 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 estimate of $V(u_{ij})$, where u_{ij} is the distance between x_i and x_j. A scatterplot of v_{ij} against u_{ij} is called the *variogram cloud*. A smoothed version of the variogram cloud can be used to suggest appropriate parametric models for the spatial covariance structure of the data; in what follows, we use simple binned estimators. For more information on variogram estimation, see for example Cressie (1985), Cressie (1991), section 2.4, Chilès and Delfiner (1999), section 2.2, or Diggle and Ribeiro (2007), chapter 5.

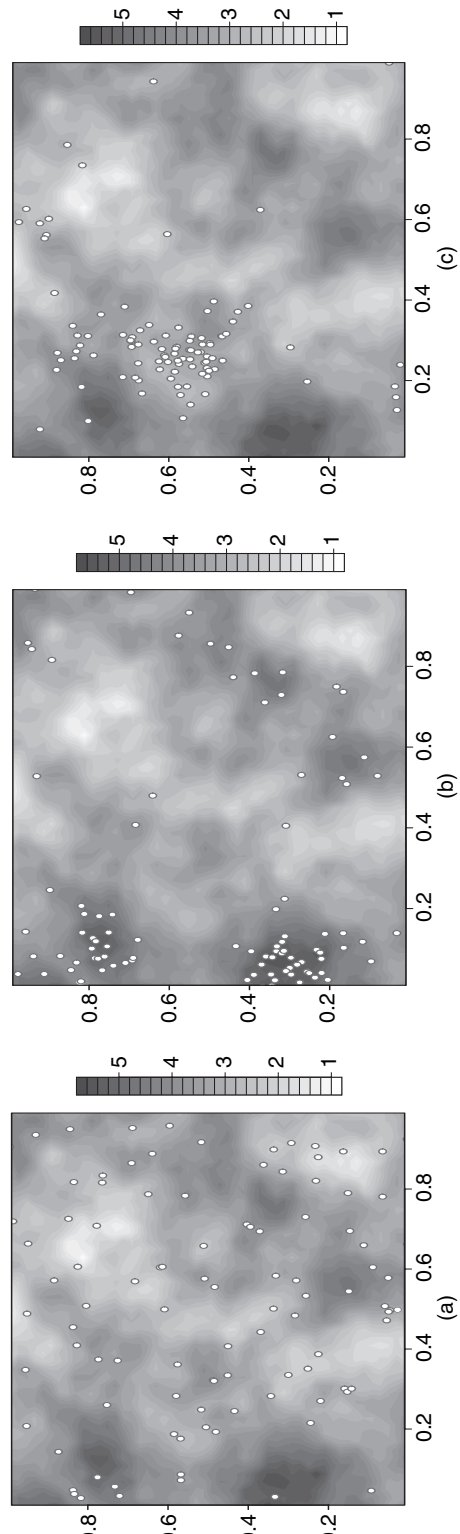

Fig. 1. Sample locations and underlying realizations of the signal process for the model that was used in the simulation study (in each case, the background image represents the realization of the signal process $S(x)$ that was used to generate the associated measurement data; the model parameter values are $\mu = 4$, $\sigma^2 = 1.5$, $\phi = 0.15$, $\kappa = 1$, $\beta = 2$ and $\tau^2 = 0$): (a) completely random sample; (b) preferential sample; (c) clustered sample

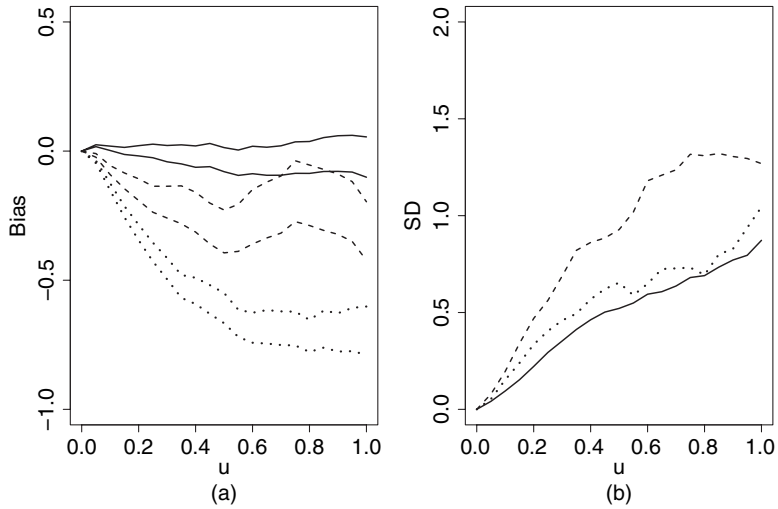

Fig. 2. Estimated bias and standard deviation of the sample variogram under random (———), preferential (· · · · · · ·) and clustered (- - - - - - -) sampling (see the text for a detailed description of the simulation model): (a) pointwise means plus and minus two pointwise standard errors; (b) pointwise standard deviations

Fig. 2 shows simulation-based estimates of the pointwise 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 approximate unbiasedness of the empirical variogram ordinates; although smoothing the empirical variogram ordinates does induce some bias, this effect is small in the current setting. In contrast, under preferential sampling the results show severe bias. With regard to efficiency, Fig. 2(b) illustrates that clustered sampling designs, whether preferential or not, are also less efficient than uniform sampling. The bias that is induced by preferential sampling is qualitatively unsurprising; the effect of the preferential sampling is that sample locations predominantly span a reduced range of values of $S(x)$, which in turn reduces the expectation of pairwise squared differences at any given spatial separation. Note, incidentally, that the sample variogram has substantially smaller variance under preferential than under non-preferential clustered sampling. However, this is of little practical interest in view of its severe bias under preferential sampling. In general, 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 under preferential sampling differs from the unconditional variance; see, for example, Wälder and Stoyan (1996) or Schlather (2001).

3.2. Spatial prediction

Suppose that our target for prediction is $S(x_0)$, which is the value of the process S at a generic location x_0, given sample data $(x_i, y_i), i = 1, 2, \ldots, n$. The widely used ordinary kriging predictor estimates the unconditional expectation of $S(x_0)$ by generalized least squares, but using plug-in estimates of the parameters that define the covariance structure of Y. In classical geostatistics, these plug-in estimates are obtained either subjectively (Chilès and Delfiner (1999), section 2.6) or by non-linear least squares (Cressie (1991), section 2.6.2). We used maximum likelihood estimation under the assumed Gaussian model for Y.

Each simulation yields an estimate of the bias $\hat{S}(x_0) - S(x_0)$ and the mean-square error $\{\hat{S}(x_0) - S(x_0)\}^2$, for the ordinary kriging predictor $\hat{S}(x_0)$. The first two rows of Table 1 show

Table 1. Effect of sampling design on the bias and mean-square error of the ordinary kriging predictor $\hat{s}(x_0)$, when $x_0 = (0.49, 0.49)$ and each sample consists of 100 locations on the unit square†

Model	*Parameter*	*Confidence intervals for the following sampling designs:*		
		Completely random	*Preferential*	*Clustered*
1	Bias	(−0.014, 0.055)	(0.951, 1.145)	(−0.048, 0.102)
1	Root-mean-square error	(0.345, 0.422)	(1.387, 1.618)	(0.758, 0.915)
2	Bias	(0.003, 0.042)	(−0.134, −0.090)	(−0.018, 0.023)
2	Root-mean-square error	(0.202, 0.228)	(0.247, 0.292)	(0.214, 0.247)

†Each entry is an approximate 95% confidence interval calculated from 500 independent simulations. See the text for a detailed description of the simulation models 1 and 2.

approximate 95% confidence intervals, calculated as means plus and minus 2 standard errors over 500 replicate simulations, for the bias and root-mean-square error at the prediction location $x_0 = (0.49, 0.49)$.

The bias is large and positive under preferential sampling with $\beta > 0$. 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 oversampling 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.

In a second set of simulations, we set the values of the model parameters to correspond to the maximum likelihood estimates that were obtained in the analysis of the 1997 Galicia biomonitoring data reported in Section 5 below; hence $\mu = 1.515, \sigma^2 = 0.138, \phi = 0.313, \kappa = 0.5, \beta = -2.198$ and $\tau^2 = 0.059$. The results are qualitatively as expected, but the differences among the three sampling designs are much smaller for two reasons. Firstly, the degree of preferentiality is much weaker; a measure of this is the product $\beta\sigma$, which takes the values 3 and 0.815 for the first and second simulation models respectively. Secondly, the effect is further diluted by the inclusion of a non-zero nugget variance.

4. Fitting the shared latent process model

4.1. Monte Carlo maximum likelihood estimation

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

$$L(\theta) = [X, Y] = E_S[[X|S][Y|X, S]], \tag{6}$$

where θ represents all the model parameters and the expectation is with respect to the unconditional distribution of S. Evaluation of the conditional distribution $[X|S]$ strictly requires the realization of S to be available at all $x \in A$. However, and as previously noted in Section 3.1, we approximate the spatially continuous realization of S by the set of values of S on a finely spaced 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 equation (6) by a sample average over simulations S_j. This strategy fails when the measurement error variance τ^2 is 0, 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 rewrite the exact likelihood (6) as the integral

$$L(\theta)=\int [X|S][Y|X,S]\frac{[S|Y]}{[S|Y]}[S]\,\mathrm{d}S. \tag{7}$$

Now, write $[S]=[S_0][S_1|S_0]$ and replace the term $[S|Y]$ in the denominator of expression (7) 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, equation (7) becomes

$$\begin{aligned} L(\theta) &= \int [X|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0][S|Y]\,\mathrm{d}S \\ &= E_{S|Y}\left[[X|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0]\right] \end{aligned} \tag{8}$$

and a Monte Carlo approximation is

$$L_{\mathrm{MC}}(\theta)=m^{-1}\sum_{j=1}^{m}[X|S_j]\frac{[Y|S_{0j}]}{[S_{0j}|Y]}[S_{0j}], \tag{9}$$

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 antithetic pairs of realizations, i.e. for each $j=1,\ldots,m/2$ set $S_{2j}=2\mu_{\mathrm{c}}-S_{2j-1}$, where μ_{c} denotes the conditional mean of S given Y. The use of conditional simulation in equation (9) bypasses the difficulty with the naive strategy by guaranteeing that the simulated realizations of S are compatible with the data Y.

To simulate a realization from $[S|Y]$, we use the following construction. Recall that the data locations $X=\{x_1,\ldots,x_n\}$ constitute a subset of the $N\geqslant n$ prediction locations, $X^*=\{x_1^*,\ldots,x_N^*\}$ say. Define C to be the $n\times N$ matrix whose ith row consists of $N-1$ 0s and a single 1 to identify the position of x_i within X^*. Note that, unconditionally, $S\sim\mathrm{MVN}(0,\Sigma)$ and $Y\sim\mathrm{MVN}(\mu,\Sigma_0)$ with $\Sigma_0=C\Sigma C'+\tau^2 I$. Then, if Z denotes an independent random sample of size n from $N(0,\tau^2)$, S is a random draw from $\mathrm{MVN}(0,\Sigma)$ and y denotes the observed value of Y, it follows that

$$S+\Sigma C'\Sigma_0^{-1}(y-\mu+Z-CS) \tag{10}$$

has the required multivariate Gaussian distribution of S given $Y=y$ (Rue and Held (2005), chapter 2, and Eidsvik *et al.* (2006)). Hence, for conditional simulation when N is large, we need a fast algorithm for unconditional simulation of S. We have used 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 equation (10).

4.2. Goodness of fit

We have already noted in Section 1 the availability of tests for preferential sampling; see, for

example, Schlather *et al.* (2004) or Guan and Afshartous (2007). Here, we suggest a way of assessing the goodness of fit of the preferential sampling model that was described in Section 2, by comparing the sample locations with realizations of the fitted Cox model for their unconditional distribution.

A standard diagnostic tool for stationary spatial point processes is the reduced second-moment measure, or K-function (Ripley, 1977), that is defined by $\lambda K(s) = E[N_0(s)]$ where $N_0(s)$ denotes the number of points of the process within distance s of an arbitrary origin of measurement, conditional on there being a point of the process at the origin, and λ is the expected number of points of the process per unit area. Under the preferential sampling model, the marginal model for the sample locations X is a log-Gaussian Cox process with stochastic intensity $\Lambda(x) = \exp\{\alpha + \beta S(x)\}$. For this process, the K-function is of the form

$$K(s) = \pi s^2 + 2\pi \int_0^s \gamma(u) u \, \mathrm{d}u, \tag{11}$$

where $\gamma(u) = \exp\{\beta^2 \sigma^2 \rho(u; \phi)\} - 1$ is the covariance function of $\Lambda(x)$ (Diggle (2003), section 5.5). To assess the goodness of fit informally, we compare the estimated K-function of the data with the envelope of estimates that is obtained from sets of sample locations generated from simulated realizations of the fitted model. For a formal Monte Carlo test, we use a goodness-of-fit statistic that measures the discrepancy between estimated and theoretical K-functions, as described in Section 5.2.2. Note that this aspect of the model is not considered explicitly in the fitting process that was described in Section 4.1.

5. Heavy metal biomonitoring in Galicia

Our application concerns biomonitoring of lead pollution in Galicia, northern Spain, by using the concentrations in moss samples, in micrograms per gram dry weight, as the measured variable. An initial survey was conducted in the spring of 1995 'to select the most suitable moss species and collection sites' (Fernández *et al.*, 2000). Two further surveys of lead concentrations in samples of the moss species (*Scleropodium purum*) took place in October 1997 and July 2000. Fig. 3 shows the sampling locations that were used in these two surveys. Note that some locations appear to lie outside Galicia, and others in the sea to the north. However, this is an artefact of the fact that the boundary is both approximate and imperfectly registered; it plays no role in the analysis and is included only to add context to the map.

In the 1997 survey, sampling was conducted more intensively in subregions where large gradients in lead concentrations were expected, in line with suggestions in Ruhling (1994). The resulting design was highly non-uniform and potentially preferential. The second survey used an approximately regular lattice design, which is therefore non-preferential; gaps in the lattice arose only where a different species of moss was collected. For further details, see Fernández *et al.* (2000) and Aboal *et al.* (2006). In particular, Fernández *et al.* (2000) studied the changes in heavy metal concentrations between the two years ignoring the corresponding spatial distributions. Our objective in analysing these data is to estimate, and compare, maps of lead concentrations in 1997 and 2000.

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 sources of pollutant, or with an overall reduction in levels of pollution over the 3 years between the two surveys. Also, the log-transformation eliminates an

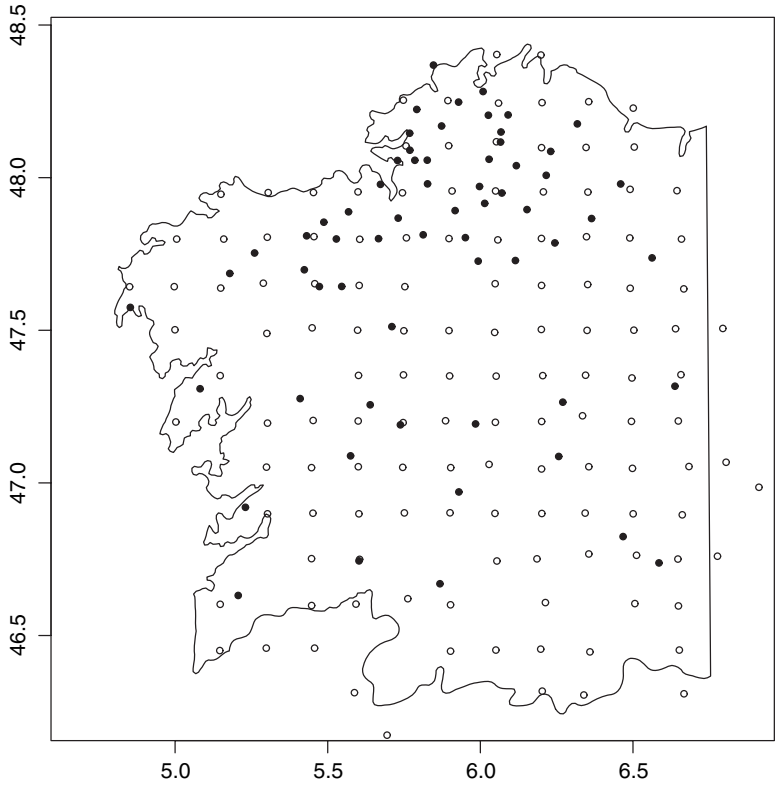

Fig. 3. Sampling locations for 1997 (•) and 2000 (∘): the unit of distance is 100 km; two outliers in the 1997 data were at locations (6.50,46.90) and (6.65,46.75)

apparent variance–mean relationship in the data and leads to more symmetric distributions of measured values (Fig. 4).

5.1. Standard geostatistical analysis

For an initial analysis, we assumed the standard Gaussian model (1) with the underlying signal $S(x)$ specified as a zero-mean stationary Gaussian process with variance σ^2 and Matérn correlation function $\rho(u;\phi,\kappa)$, and Gaussian measurement errors, $Z_i \sim N(0,\tau^2)$, and fitted this model separately to the 1997 and 2000 data.

Fig. 5 shows, for each of 1997 and 2000, smoothed empirical variograms and theoretical variograms with parameters fitted by maximum likelihood. On the basis of the general shape of the two empirical variograms, we used a fixed value $\kappa = 0.5$ for the shape parameter of the Matérn correlation function. The estimated variograms differ in some respects, notably the absence of a nugget component (i.e. $\hat{\tau}^2 \approx 0$) in the variogram that was estimated from the 2000 data; however, this parameter is poorly identified because of the lattice-like arrangement of the 2000 sampling design, whereas the inclusion of close pairs of locations in the 1997 sampling design enables better estimation of τ^2. Other features of the two fitted variograms are similar, e.g. the height of the asymptote (i.e. $\hat{\tau}^2 + \hat{\sigma}^2$) and the approximate range (i.e. $\hat{\phi}$). These observations support the idea that a joint model for the two data sets might allow at least some parameters in common between the two years. The generalized likelihood ratio test statistic (Cox and Hinkley (1974), section 9.3) to test the hypothesis of common σ, ϕ and τ, under the admittedly dubious

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

	Levels (μg (g dry weight)$^{-1}$) for the following scales and years:			
	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

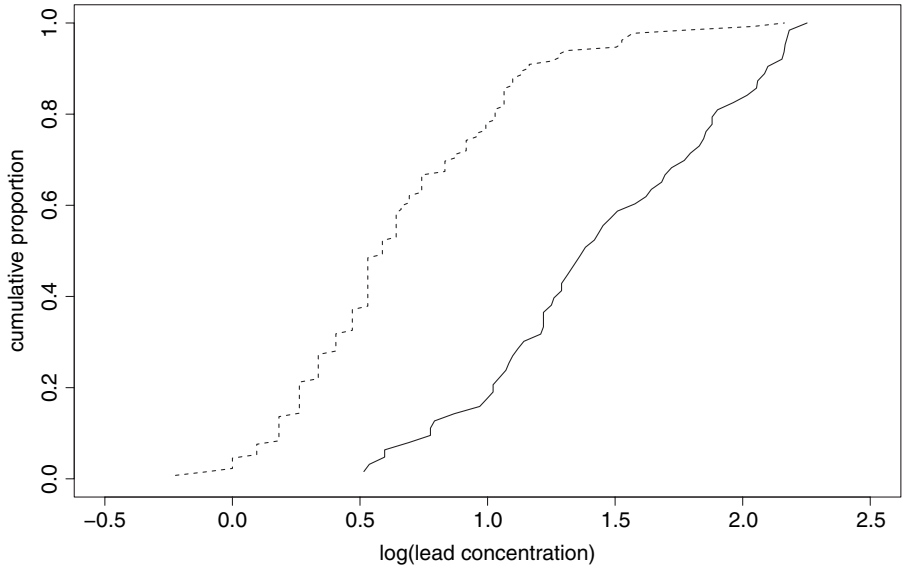

Fig. 4. Empirical distributions of log-transformed lead concentrations in the 1997 (———) and 2000 (-------) samples

assumption that neither sample is preferential, was 7.66 on 3 degrees of freedom ($p = 0.054$). We revisit this question in Section 5.2.

5.2. Analysis under preferential sampling

5.2.1. Parameter estimation

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 ratio was reduced to approximately 0.3 (the actual value varies over the parameter space) and the approximate generalized likelihood ratio test statistic to test $\beta = 0$ was 27.7 on 1 degree of freedom ($p < 0.001$).

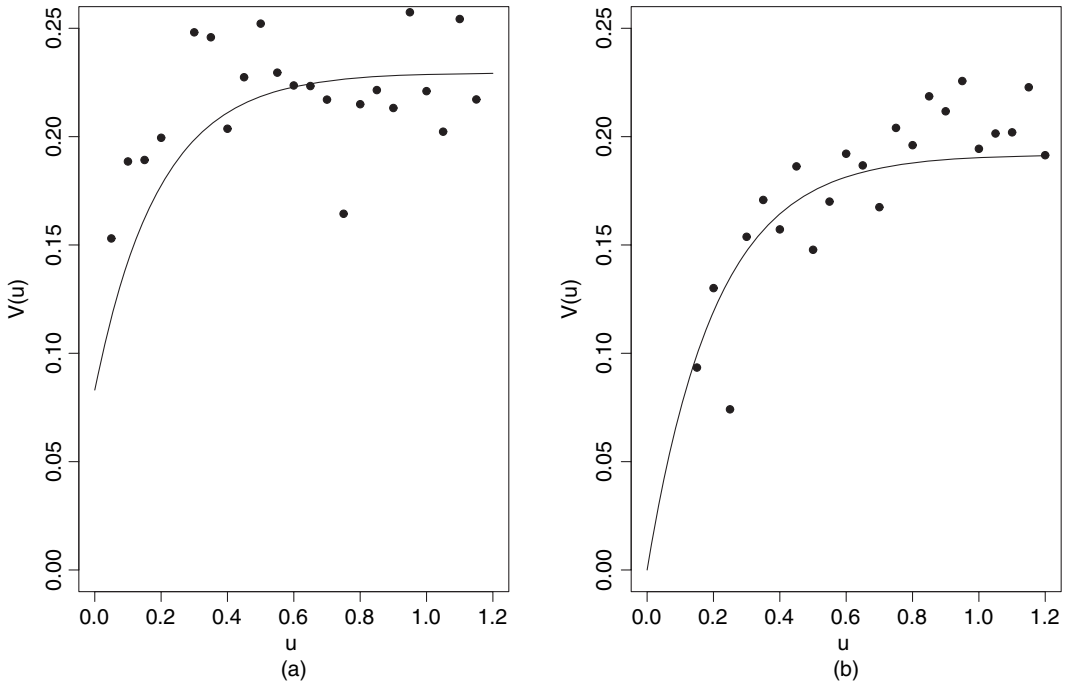

Fig. 5. Smoothed empirical (•) and fitted theoretical (———) variograms for (a) 1997 and (b) 2000 log-transformed lead concentration data

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 generalized likelihood ratio test statistic of 6.2 on 3 degrees of freedom ($p = 0.102$). 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.

Table 3 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 optimization algorithm. The quadratic surface was then refitted 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 $m = 10000$ 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, $\hat{\beta} = -2.198$. The latter finding is both counterintuitive, because the oversampled northern half of the region is more industrialized than the undersampled southern half, and critically dependent on our allowing the two mean

Table 3. Monte Carlo maximum likelihood estimates of parameters in the joint model for the 1997 and 2000 Galicia biomonitoring data†

Parameter	*Estimate*	*Standard error*	*Correlation matrix*					
μ_{97}	1.515	0.136	1.000	0.023	0.095	−0.243	−0.222	0.167
μ_{00}	0.762	0.110		1.000	0.230	−0.229	−0.281	0.342
$\log(\sigma)$	−0.992	0.049			1.000	−0.217	−0.744	0.469
$\log(\phi)$	−1.163	0.075				1.000	0.604	−0.675
$\log(\tau)$	−1.419	0.042					1.000	−0.652
β	−2.198	0.336						1.000

†Approximate standard errors and correlations are computed from a quadratic fit to the Monte Carlo log-likelihood surface (see the text for details).

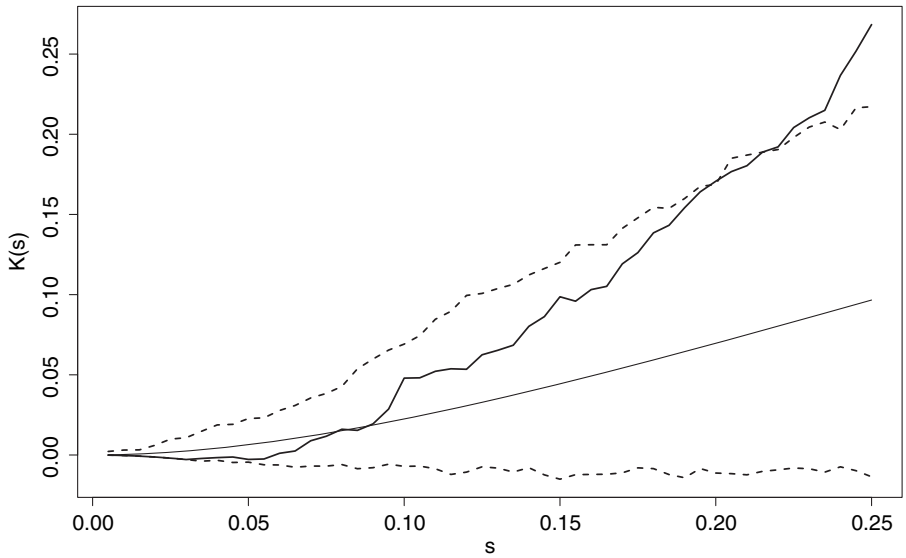

Fig. 6. Estimated K-function of the 1997 sample locations (———) and envelope from 99 simulations of the fitted log-Gaussian Cox process (-------)

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 oversampled northern half of the region ($n = 47$; mean log-concentration 1.38; standard deviation SD = 0.49) than in the undersampled southern half ($n = 16$; mean 1.62; SD = 0.40), which is consistent with a negative value of β.

5.2.2. Goodness of fit

Fig. 6 shows the estimated K-function for the 1997 sampling locations together with the envelope of 99 simulations of the fitted Cox process, and the theoretical K-function. The estimate lies within the simulation envelope for distances up to 0.22 (22 km). For a formal Monte Carlo goodness-of-fit test, we define the test statistic

$$T = \int_0^{0.25} \frac{\{\hat{K}(s) - K(s)\}^2}{v(s)} \mathrm{d}s$$

where $K(s)$ is given by equation (11) and $v(s)$ is the variance of $\hat{K}(s)$, estimated from the simulations of the fitted Cox process. This gives $p = 0.03$. The Cox model slightly underestimates the extent of spatial aggregation in the data locations.

5.2.3. Prediction

What effect does the acknowledgement of preferential sampling make on the predicted 1997 pollution surface? Fig. 7 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 that are based on the preferential sampling model have a substantially wider range, over the lattice of prediction locations, than those that assume non-preferential sampling (1.310–7.654 and 1.286–5.976 respectively). The difference surface also covers a relatively large range (from -0.715 to 3.693) and shows strong spatial structure. The size of the difference between the two predicted surfaces is at first sight surprising, as both are partially constrained by the observed concentrations. However, in the presence of a nugget effect the predictions are not constrained to interpolate the data. Also, ignoring the preferential nature of the sampling leads to biased parameter estimates. Finally, the effect of the back-transformation from log-concentrations to concentrations of lead, this being the scale on which predictions are required, is to amplify the differences.

Using the conditional expectation as a point predictor is conventional, but questionable when, as here, the measurement process has a highly skewed distribution. As an alternative summary, Fig. 8 compares pointwise 5%, 50% and 95% limits of the plug-in predictive distribution of lead concentrations under preferential and non-preferential modelling assumptions, holding the model parameters fixed at their estimated values. The differences between the two are smaller than in Fig. 7, but still non-negligible.

Finally, in Fig. 9, we show kernel density estimates of the plug-in predictive distributions for the areal proportion of Galicia in which 1997 lead concentrations exceed 3, 5 or 7 μg (g dry weight)$^{-1}$. In all three cases, recognition of the preferential sampling results in a pronounced shift in the predictive distribution. Note, however, that these plug-in predictive distributions do not account for parameter uncertainty.

Our overall conclusion is that the preferential sampling has made a material difference to our predictive inferences for the 1997 pollution surface.

6. Discussion

We have shown that conventional geostatistical models and associated statistical methods can lead to 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 class of models proposed. The resulting methods are computationally intensive, each run taking several hours of central processor unit time. The computations that were reported in the paper were run on a Dell Latitude D620 laptop personal computer, using the R software environment (R Development Core Team (2008); see also `www.r-project.org`) and associated Comprehensive R Archive Network packages `fields`, `geoR` and `splancs`. The data

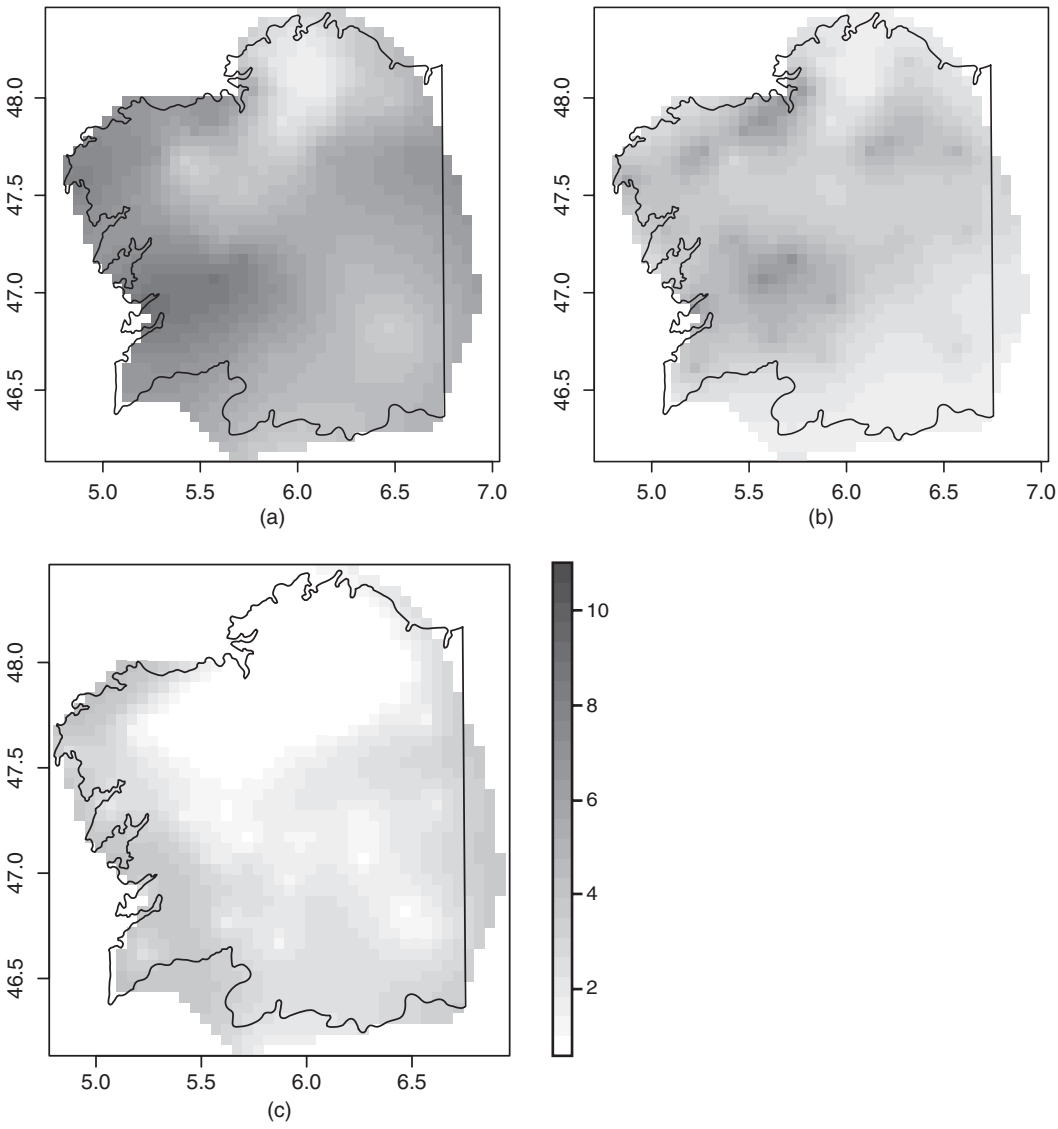

Fig. 7. Predicted surface of lead concentrations in 1997 under (a) preferential and (b) non-preferential assumptions, together with (c) the pointwise difference between the two: all three surfaces are plotted on a common scale, as shown

and R code are available from `www.lancs.ac.uk/staff/diggle/`. There is undoubtedly very considerable scope to improve the efficiency of the authors' code. In particular, we did not seek to automate the progressive increase in the number of Monte Carlo samples as we explored the log-likelihood surface.

The computation of the Monte Carlo likelihood uses direct simulation, as in Diggle and Gratton (1984), rather than Markov chain Monte Carlo sampling. Hence, issues concerning convergence of the simulations do not arise, and the variablity 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 biomonitoring data from Galicia,

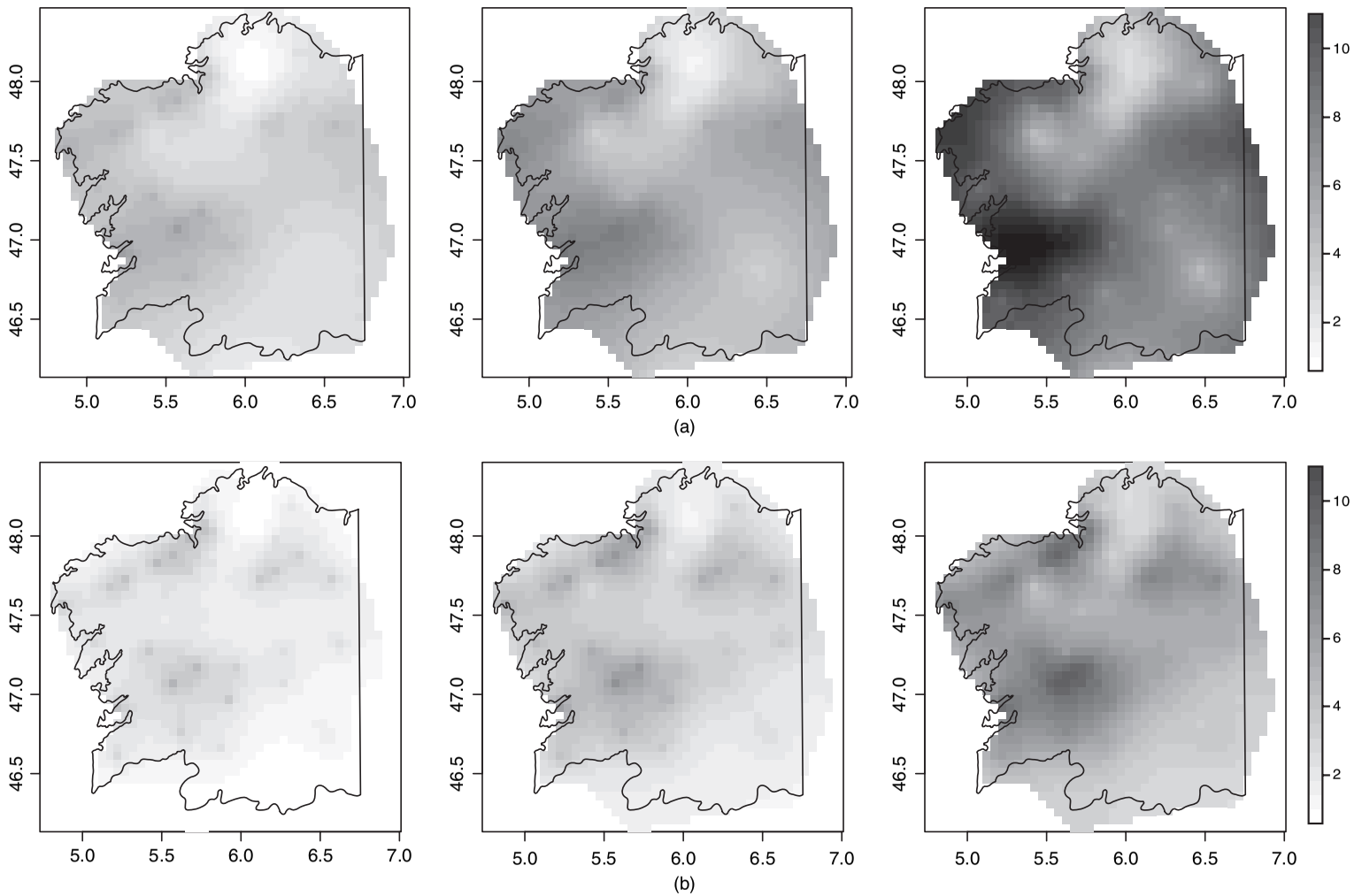

Fig. 8. Pointwise 5%, 50% and 95% limits of the predictive distribution of lead concentrations in 1997 under (a) preferential and (b) non-preferential assumptions: all six surfaces are plotted on a common scale, as shown

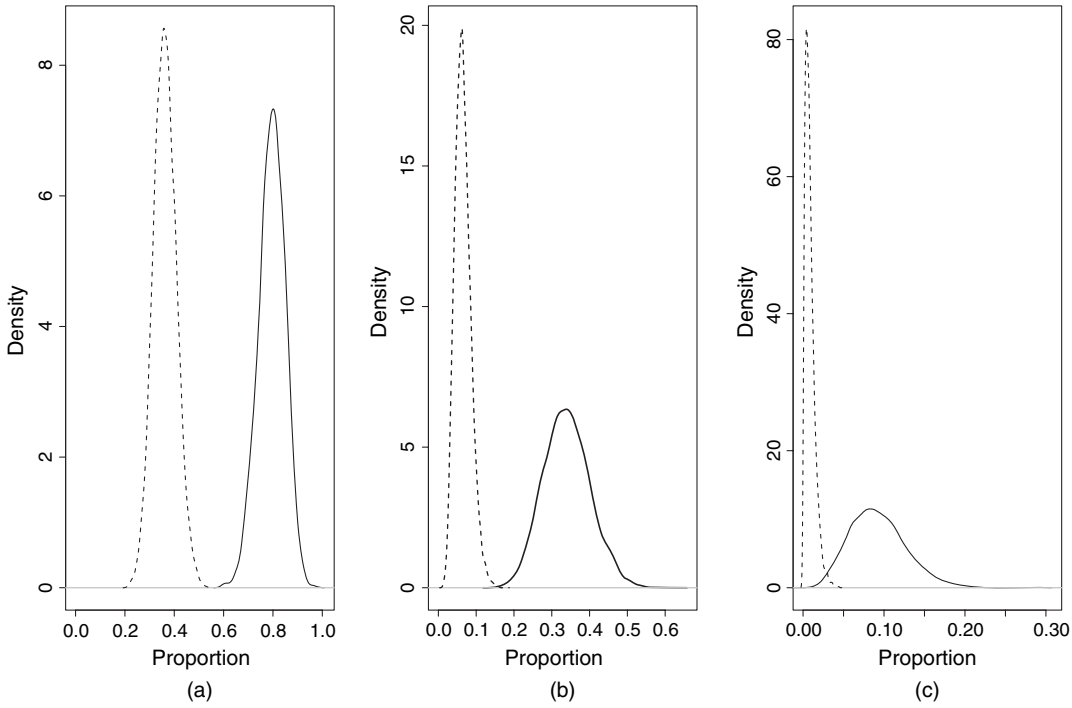

Fig. 9. Predictive distributions for the areal proportion of Galicia in which 1997 lead concentrations exceed (a) 3, (b) 5 and (c) 7 μg (g dry weight)$^{-1}$ under preferential (———) and non-preferential (-------) assumptions

northern Spain. An important feature of these data is that they are derived from two surveys of the region of interest, the first of which used a spatially irregular set of sampling locations and has been shown to be preferentially sampled. In the second survey the sampling locations formed a nearly regular grid over the study region and we have therefore taken it to be non-preferentially sampled. 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 maximized 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 10000 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 0, 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 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

$$[X, S] = \int [X|U][S|U][U]\,\mathrm{d}U, \tag{12}$$

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 marginalization, 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. In this context, any of the existing tests for preferential sampling can be applied, albeit approximately, to residuals after fitting a regression model for the mean response.

The value of seeking relevant explanatory variables to contribute to a spatial statistical model cannot be overstated. We hold the view that, in most geostatistical applications, spatial correlation reflects, at least in part, smooth spatial variation in relevant, but unobserved, explanatory variables rather than being an inherent property of the phenomenon being studied; an example to the contrary would be the spatial distribution of the prevalence of an infectious disease during an epidemic where, even for a uniformly distributed population in a completely homogeneous environment, the process of transmission from infectious to susceptible individuals would induce spatial correlation in the prevalence surface. This in turn leads us to emphasize that our paper is not a plea for uniform sampling, but rather for ensuring that any model for a set of data should respect whatever sampling design has been used to generate the data; for a thorough discussion that also uses point process models of sampling designs, albeit in a very different setting, see McCullagh (2008).

Returning to the geostatistical setting, and specifically to the application that was described in Section 5, Fernández *et al.* (2005) gave a practitioner's perspective on the ways in which different sampling designs can materially affect any analysis of spatial variation. To meet their primary objective of mapping concentration surfaces, they favoured regular lattice designs but noted that, for other purposes, 'additional sampling in areas of anomalously high concentrations of contaminants makes good sense'. We agree. Our paper formalizes this idea, while acknowledging that it necessarily complicates the subsequent modelling and inference.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council through the award of a Senior Fellowship to Peter Diggle.

We thank the Ecotoxicology Group, University of Santiago de Compostela, for permission to use the Galicia data and, in particular, José Angel Fernández, for helpful discussions concerning the data.

We also thank Håvard Rue for advice on efficient conditional simulation of spatially continuous Gaussian processes.

References

Aboal, J. R., Real, C., Fernández, J. A. and Carballeira, A. (2006) Mapping the results of extensive surveys: the case of atmospheric biomonitoring and terrestrial mosses. *Sci. Total Environ.*, **356**, 256–274.

Altham, P. M. E. (1984) Improving the precision of estimation by fitting a model. *J. R. Statist. Soc.* B, **46**, 118–119.

Chilès, J.-P. and Delfiner, P. (1999) *Geostatistics*. New York: Wiley.
Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.
Cressie, N. A. C. (1985) Fitting variogram models by weighted least squares. *J. Int. Ass. Math. Geol.*, **17**, 563–586.
Cressie, N. A. C. (1991) *Statistics for Spatial Data*. New York: Wiley.
Curriero, F. C., Hohn, M. E., Liebhold, A. M. and Lele, S. R. (2002) A statistical evaluation of non-ergodic variogram estimators. *Environ. Ecol. Statist.*, **9**, 89–110.
Diggle, P. J. (2003) *Statistical Analysis of Spatial Point Patterns*, 2nd edn. London: Arnold.
Diggle, P. J. and Gratton, R. J. (1984) Monte Carlo methods of inference for implicit statistical models (with discussion). *J. R. Statist. Soc.* B, **46**, 193–227.
Diggle, P. J. and Ribeiro, P. J. (2007) *Model-based Geostatistics*. New York: Springer.
Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299–350.
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.
Fernández, J. A., Real, C., Couto, J. A., Aboal, J. R. and Carballeira, A. (2005) The effect of sampling design on extensive bryomonitoring surveys of air pollution. *Sci. Total Environ.*, **337**, 11–21.
Fernández, J. A., Rey, A. and Carballeira, A. (2000) An extended study of heavy metal deposition in Galicia (NW Spain) based on moss analysis. *Sci. Total Environ.*, **254**, 31–44.
Guan, Y. and Afshartous, D. R. (2007) Test for independence between marks and points of marked point processes: a subsampling approach. *Environ. Ecol. Statist.*, **14**, 101–111.
Henderson, R., Diggle, P. and Dobson, A. (2000) Joint modelling of measurements and event time data. *Biostatistics*, **1**, 465–480.
Ho, L. P. and Stoyan, D. (2008) Modelling marked point patterns by intensity-marked Cox processes. *Statist. Probab. Lett.*, **78**, 1194–1199.
Isaaks, E. H. and Srivastava, R. M. (1988) Spatial continuity measures for probabilistic and deterministic geostatistics. *Math. Geol.*, **20**, 313–341.
Lin, H., Scharfstein, D. O. and Rosenheck, R. A. (2004) Analysis of longitudinal data with irregular, outcome-dependent follow-up. *J. R. Statist. Soc.* 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.
Matérn, B. (1986) *Spatial Variation*, 2nd edn. Berlin: Springer.
McCullagh, P. (2008) Sampling bias and logistic models (with discussion). *J. R. Statist. Soc.* B, **70**, 643–677.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Møller, J., Syversveen, A. and Waagepetersen, R. (1998) Log Gaussian Cox processes. *Scand. J. Statist.*, **25**, 451–482.
Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization. *Comput. J.*, **7**, 308–313.
Rathbun, S. L. (1996) Estimation of Poisson intensity using partially observed concomitant variables. *Biometrics*, **52**, 226–242.
R Development Core Team (2008) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Ripley, B. D. (1977) Modelling spatial patterns (with discussion). *J. R. Statist. Soc.* B, **39**, 172–212.
Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall.
Ruhling, A. (1994) *Atmospheric Heavy Metal Deposition in Europe: Estimation Based on Moss Analysis*. Helsinki: Nordic Council of Ministers.
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. *J. Am. Statist. Ass.*, **102**, 952–961.
Schlather, M. (2001) On the second-order characteristics of marked point processes. *Bernoulli*, **7**, 99–117.
Schlather, M., Ribeiro, Jr, P. J. and Diggle, P. J. (2004) Detecting dependence between marks and locations of marked point processes. *J. R. Statist. Soc.* B, **66**, 79–93.
Srivastava, R. M. and Parker, H. M. (1989) Robust measures of spatial continuity. In *Geostatistics*, vol. 1 (ed. M. Armstrong), pp. 295–308. Boston: Kluwer.
Wälder, O. and Stoyan, D. (1996) On variograms of point process statistics. *Biometr. J.*, **38**, 895–905.
Wood, A. T. A. and Chan, G. (1994) Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Computnl Graph. Statist.*, **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.

Discussion on the paper by Diggle, Menezes and Su

Clive Anderson (*University of Sheffield*)

This paper shows innovative thinking about a generic problem in environmental statistics. It proposes a

new model, presents an ingenious inference procedure and will give impetus to further development in the area.

I would like to comment on aspects of the model, on the inference procedure and on the general basis of the approach.

The paper's conditional Poisson model represents the possibility that the locations chosen for observation depend on the variable of interest, $\mu + S(\mathbf{x})$. Initially therefore we might expect that the conditional Poisson sampling intensity would be of the form $\lambda(\mathbf{x}) = \lambda_0 \exp[\beta\{\mu + S(\mathbf{x})\}]$ for a parameter β and baseline intensity λ_0. The paper, of course, has this form, but with $\alpha = \beta\mu + \log(\lambda_0)$, which is a more economical parameterization that recognizes that λ_0 and $\beta\mu$ would not otherwise be separately estimable and corresponds to the commonsense consideration that the overall sample size, which will be largely governed by α, is likely to have been influenced by factors other than characteristics of the target variable—economic and organizational factors, for example—which will not usually be of direct interest themselves. In fact, it might therefore be reasonable in inference to treat the sample size as fixed and to argue conditionally on its value, eliminating α from consideration. I suspect that this is what the authors did.

The effect of selection on the values of $S(\mathbf{X})$ (and therefore of Y) obtained at the sampled locations may be seen by the following approximate argument. Conditionally on a realization $S(A) = \{S(u) : u \in A\}$ of the target process over the region A and given the number of points $\mathbf{X}$ that are generated in the resulting inhomogeneous Poisson process, the points correspond to an independent sample from A with probability density function

$$\lambda(x) \Big/ \int_A \lambda(u)\,\mathrm{d}u = \exp\{\beta S(x)\} \Big/ \int_A \exp\{\beta S(u)\}\,\mathrm{d}u.$$

The (conditional) marginal distribution function of selected values of S is therefore

$$\Pr\{S(X) \leqslant s\} = \frac{\displaystyle\int_{\{u:S(u)\leqslant s\}} \exp\{\beta S(u)\}\,\mathrm{d}u}{\displaystyle\int_A \exp\{\beta S(u)\}\,\mathrm{d}u} = \frac{\displaystyle\int_{-\infty}^{s} \exp(\beta w)\,\mathrm{d}\tilde{G}(w)}{\displaystyle\int_{-\infty}^{\infty} \exp(\beta w)\,\mathrm{d}\tilde{G}(w)}, \tag{13}$$

where $\tilde{G}$ is the empirical distribution function of the values of $S(A)$. The probability in equation (13) comes from the random selection of sampling point X and therefore the choice of the value $S(X)$ from $S(A)$. However, under the assumption of a Gaussian model for S we can expect that $\tilde{G}$ will approximate to a normal $\mathcal{N}(0, \sigma^2)$ distribution function (though there will be random variations around the exact Gaussian form, depending on the spatial dependence structure). Approximately therefore, from equation (13), the probability density function of the sampled values of S has value at s that is proportional to $\exp(\beta s - s^2/2\sigma^2)$, from which it follows that the distribution is approximately normal $\mathcal{N}(\beta\sigma^2, \sigma^2)$. The effect of selection is therefore to shift the mean S by $\beta\sigma^2$. Though the argument here is conditional on the realized field $S(A)$, the approximation is the same across realizations and so can be expected to give a guide to selection bias generally for this model. A similar argument should work also in non-Gaussian cases.

The authors' ingenious bypass of the difficulties of the naive strategy for Monte Carlo likelihood evaluation can be seen as an example of importance sampling. In this direction one wonders whether use of further techniques from the Monte Carlo toolkit—regression methods, for example—in conjunction with the paper's antithetic variables might help to reduce the long computation times.

The device for generating samples from the conditional distribution $[S|\mathbf{Y}]$ is similarly ingenious, though the appearance of C in equation (10) suggests that $[S|\mathbf{Y}, \mathbf{X}]$ is in the offing somewhere. Is there more going on in the Monte Carlo implementation than meets the eye (an extra simulation cycle, possibly)? Clarification would be valuable.

The authors show how a likelihood approach can be made to work for their model. An obvious alternative is Bayesian inference. Whether implementation would be any easier, the fact is that a Bayesian interpretation of the stochasticity of the target process S, that it describes the analyst's beliefs about the concentration field rather than (directly) the outcome of physical processes, may be regarded by some as a more natural basis on which to formulate the problem.

The broad aim of the paper is to make sense of data whose collection was potentially influenced by the variable under investigation. The model that is presented here describes the selection process only in generalized terms. In the future we might look forward to more tailored models that build in specific knowledge of the mechanism of selection of the spatial process itself. It is surely in this direction that (with due attention to robustness to assumptions) long-run scientific progress lies. Though an ideal in

these developments might be a model that could, as suggested in Section 6 of the paper, be analysed on the assumption that sampling is non-preferential, the user would need reassurance that such an analysis would be safe, and alternatives if it were not. More general models that admit preferentiality will therefore be needed. The paper gives a very welcome springboard for their development.

We have very good reason to be grateful to the authors, and it gives me great pleasure to propose a vote of thanks.

E. Marian Scott (*University of Glasgow*)

It is always a pleasure to comment on an 'interesting read', and this paper offers much food for thought. For me, it is a paper of many parts, combining an extension to existing methodology, computational challenges, an interesting application and a potentially important practical message. At the end, once all is revealed, the message—the practical relevance and potentially pervasive nature of preferential sampling and its effect on inference—seems entirely natural. The solution that is presented in the paper is technical, but it will become clear that in some ways a simpler solution exists when we have precise and specific information about the sampling scheme (use of covariates and stratification), although there remain many situations where such information is not available.

I would like to focus my comments on the sampling aspects and practical relevance of the comments and approach that are presented in the paper.

Statistical sampling is, as we all well know, a process that allows *inferences* to be made, and the use of valid statistical sampling techniques increases the chance that the sample is collected in a manner that is *representative* of the population. In the context of environmental sampling, our approach would be to stipulate the objectives and to summarize our knowledge of the environmental context, so that the sampling design makes use of this expert knowledge. This knowledge may be difficult to quantify, but it will include the nature of the population such as the physical or biological material of interest, its spatial extent, its temporal stability and other important characteristics, such as the expected pattern and magnitude of variability in the observations.

This could lead to preferential sampling but is a case where we might in fact have some understanding of the choices being made in the sampling design.

As an example, suppose that you wanted to estimate the inventory of ^{60}Co in the sediments of an estuary whose boundaries have been clearly defined. From our scientific knowledge, we know that ^{60}Co is particle reactive and we have a map of sediment type in the estuary. If there is knowledge of different strata over the sampling domain (such as sediment type), the use of a stratified sample would be recommended and a random sample of locations would be selected within each stratum. So this could lead to a preferential sample, but we have the explanatory information in the shape of the sediment types and distribution across the estuary floor and so can use this knowledge in our modelling.

Preferential sampling means that there is a stochastic dependence between the sampling locations and the spatial process that we are interested in. This would seem appropriate in the situation where we believe that sampling is preferential but cannot define other than in the very vaguest of terms what knowledge has defined the sampling strategy. How likely is this? Perhaps it is more than we think. In many regulatory situations, environmental monitoring is prescribed and potentially preferential (either by intent or by implicit design). As an example of regulatory monitoring, in Scotland, the environment protection agency, for purposes of the water framework directive, describes their strategy as risk driven (i.e. it focuses on assessing the greatest risks on water bodies) (`www.sepa.org.uk/water/monitoring_and_classification/scottish_monitoring_strategy`). Further, in terms of their monitoring network, they classify sites in three ways, surveillance, operational and investigative but, importantly under operational, choice of site is driven by risk assessments based on pressure information and located in areas of known risk, whereas investigative envisages a more variable network that is responsive to unplanned events and emerging risks, where the source of the risk (the pressure) is not always well understood.

So what effect might ignoring the preferential nature of the sampling have? The answer is fairly obvious, namely biased estimates and incorrect inferences from our geostatistical model including the variogram estimator and predictions. (The degree of bias may depend on the strength of the preferential nature of the sampling).

The presence of bias seems natural; the sites chosen are unlikely to be representative of the population as a whole but may be representative of a subset of the population, so our mistake would be in assuming representativeness. In that sense do the tests and the $\beta\sigma$-term in the paper allow us to assess this? If so, then this is extremely useful since, in the applied context, often the question is 'how representative?' is my sample.

The paper then proceeds to discuss a procedure to assess whether the sampling is preferential and also provides a modelling approach to accommodate preferential sampling through a marked point process approach. There is much to recommend this paper, not least the computational aspects.

In summary then, is preferential sampling broadly equivalent to judgemental sampling which we so often urge our collaborators not to use, since ultimately our samples are not generally representative of the population as a whole? I suspect that this is a widespread issue, which is most likely insufficiently recognized, so this paper has at the very least raised awareness. Does the test for preferential sampling generalize to a measure of representativeness? If so that would be extremely useful.

It is my great pleasure to thank the authors for a thought-provoking paper and to second the vote of thanks.

The vote of thanks was passed by acclamation.

Janine B. Illian (*University of St Andrews*)

Geostatistics and point process methodology have developed largely independently, yet they are strongly related. Combining point processes with random fields through Cox processes has been the topic of a substantial body of recent work, including Ho and Stoyan (2008), Illian *et al.* (2008), Møller and Waagepetersen (2004), Myllymäki and Penttinen (2009) and Schlather (2001), which aims at modelling a point pattern on the basis of a latent random field. Diggle and his colleagues now nicely consider Cox processes from the point of view of geostatistics with the aim of modelling a random field, both making the applied user aware of point processes and demonstrating that the standard geostatistical model is biased under preferential sampling.

In many applications, only preferentially sampled data from a single year are available. With this in mind we again note two distinct sampling areas in the Galician lead data from 1997; Fig. 10. The pair correlation functions for these patterns both reveal inhibition but up to different distances (Fig. 11). This indicates two very different sampling strategies—differing both in intensity and in spacing. Ignoring that the assumption of a Poisson distribution of points given the underlying random field has been violated might not be a serious issue when aiming at predicting a random field. However, the Cox process model for preferential sampling still appears inappropriate here. More specifically, the sampling pattern is not continuous in space and does not reflect the underlying random field S but the researcher's limited knowledge and resulting crude sampling strategy.

To take this further, in applications the researcher is likely to decide consciously to sample more intensely in some areas than in others. Hence tests for preferential sampling are only relevant if there have been issues of accessibility to sampling locations. In many cases, however, the scientist samples preferentially since knowledge on the spatial distribution of the variable of interest from previous studies or covariates

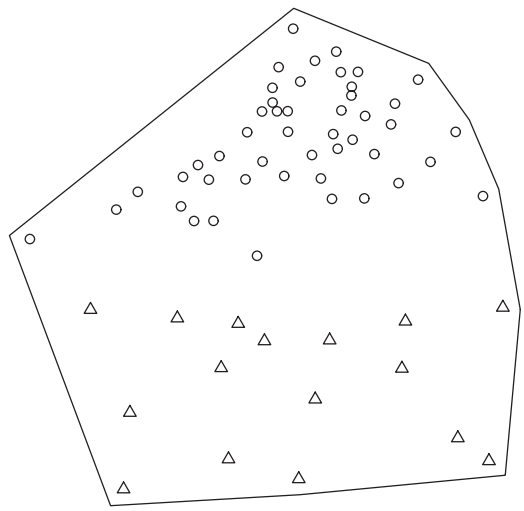

Fig. 10. Sampling locations for the Galician data (1997)

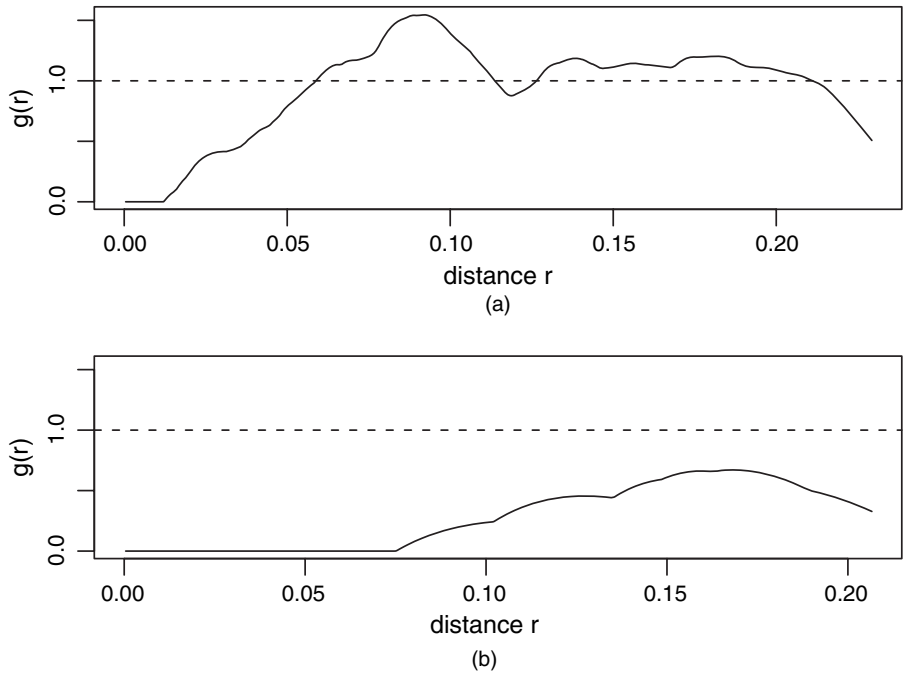

Fig. 11. Pair correlation functions for (a) the top and (b) the bottom part of the lead data from Galicia for the year 1997, split as indicated by the circles and triangles in Fig. 10

is available. These can then be explicitly included in a model. If prior knowledge is unavailable, one may still see virtue in preferential sampling. However, lacking sufficient information the researcher will crudely consider distinct sampling areas and—similarly to what we saw for the Galician data—apply different sampling regimes in each area. An appropriate model would incorporate these different sampling regimes.

In summary, this raises the question whether the model that is described by Diggle and his colleagues is a model for realistic applied data sets, or, in other words, whether assuming a different but still inappropriate model allows the user to predict the random field with less bias.

Stephen L. Rathbun (*University of Georgia, Athens*)

I congratulate the authors for addressing the influence of preferential sampling on geostatistical inference, which is an issue that has been neglected by geostatisticians but is of considerable importance in environmental applications. For example, empirical investigations of global climate change involve data from weather stations that are not uniformly distributed over the globe (e.g. Hansen and Lebedeff (1987) and Vinnikov *et al.* (1990)). The highest densities occur in the USA and Europe, there are more stations in the northern than in the southern hemisphere and little effort goes into sampling the oceans. Moreover, the distribution of weather stations has changed over time. Is the distribution of weather stations preferential and, if so, how does this affect empirical estimates of global climate change?

The effect of sampling design on statistical modelling has long been considered in the survey sampling literature; for a review, see Chambers and Skinner (2003). Subjects may be sampled according to complex sampling designs, often with unequal sampling weights. Parameters may be estimated as solutions to weighted score equations, weighted by the inverse of sampling probabilities (Skinner, 1989). Data for members of the target population are generated from independent realizations of a superpopulation model, from which samples are obtained by using a known probability sampling design. In the geostatistical setting that is considered by Diggle and co-workers, however, only a single realization of a random field is used to generate the data for the population of points in the study region. Nevertheless, weighting by the inverse of the sampling intensity may be used to obtain unbiased estimators for various quantities of interest without resorting to computationally intensive Monte Carlo algorithms.

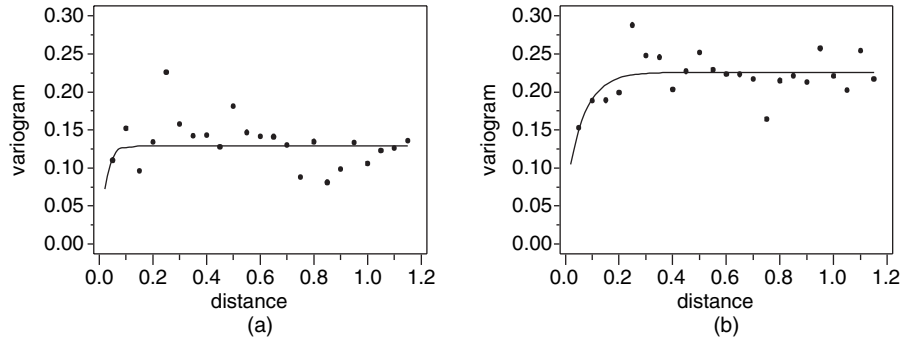

Fig. 12. (a) Unweighted and (b) weighted estimates for the variogram together with fitted exponential variograms for 1997

Consider, for example, the variogram. The empirical variogram estimates that are presented in Fig. 5 are consistent for the variogram under non-preferential sampling, and for the mark variogram (Wälder and Stoyan, 1996) under preferential sampling (Guan *et al.*, 2007); see Ho and Stoyan (2008) for the mark variogram under preferential sampling. Weighting by the inverse of the sampling intensity, we can obtain the following bias-corrected estimator for the variogram:

$$2\hat{\gamma}(r) = \frac{\sum_{i \neq j} \kappa(r - \|\mathbf{x}_i - \mathbf{x}_j\|/h)\, \lambda^{-1}(\mathbf{x}_i)\, \lambda^{-1}(\mathbf{x}_j)(y_i - y_j)^2}{\sum_{i \neq j} \kappa(r - \|\mathbf{x}_i - \mathbf{x}_j\|/h)\, \lambda^{-1}(\mathbf{x}_i)\, \lambda^{-1}(\mathbf{x}_j)},$$

where $\kappa(\cdot)$ is a kernel density function and $h > 0$ is the bandwidth. When the data are measured with error, the intensity $\lambda(\mathbf{x})$ depends on the unobserved value of $S(\mathbf{x})$, but $\lambda^{-1}(\mathbf{x}_i)$ can be replaced by the bias-corrected estimator $\exp(-\alpha - \beta y_i - \frac{1}{2}\beta^2 r^2)$ When applied to the 1997 data, the weighted estimator for the variogram shows a lower sill and a shorter range of spatial variation than the unweighted empirical variogram estimator (Fig. 12).

Serge Guillas (*University College London*)

I congratulate the authors for their very convincing paper. My first comment is that the authors have done an analysis that is both necessary and long overdue. Many research studies in geostatistics that are now published may have not taken into account the preferential sampling issue that is raised by the authors here. The examples from Section 3 indeed show that large prediction biases can result from this shortcoming. It is terrifying to realize that some environmental regulations might have been inspired by such biased studies.

Another extension of this work would be in the understanding of preferential sampling for space–time models (Stein, 2005). It may be that one must vary the level of preferential sampling over time periods, if say sampling strategies are seasonal or driven by political motivations.

The Monte Carlo method that is used here is elegant, as demonstrated by the various tricks in the approximation of the likelihood, and is much faster than a careful Markov chain Monte Carlo approach. However, as the authors recognize in their discussion, some problems remain: fitting a quadratic surface to the log-likelihoods to retrieve standard errors and correlation is not perfect. In addition, the authors make use of antithetic pairs of realizations to reduce the Monte Carlo variance but do not quantify the resulting improvement.

Nevertheless, it is important that scientists who employ spatial statistics in their research become aware of the preferential sampling issue and are taking care of it. An implementation of the technique that is discussed here in common statistical packages would serve that purpose. In particular, measures of the level of preferential sampling, say through the parameter β, may help scientists to pin down the explanatory variables that could readily improve the overall statistical representation.

A. P. Dawid (*University of Cambridge*)

Preferential sampling is defined as non-independence between the sites X at which we choose to observe and the underlying process S. But, if we ourselves choose the sites, we must do so on the basis of prior,

not future, knowledge and, so long as all variables—including, necessarily known, prior assessments of unknown outcomes and processes—that are taken into account in our selection process are appropriately included in our analysis, there can be no further dependence on how the still unknown process S will actually turn out. So 'preferential sampling' is misnamed: it is not an attribute of the sampling scheme (we cannot choose to conduct preferential sampling), but of our knowledge of that scheme. In particular, no amount of inspection of maps can tell us whether the sampling is preferential or not—we would, for example, expect to see a relationship between sampled sites X and the process S if X were based on prior expectations of S, but that would disappear if appropriate allowance were made for that prior knowledge, and it does not constitute preferential sampling, since where the sites are positioned will still not depend on the actual, rather than the predicted, process S.

In the case of sequential studies, where selection of sites in a second study is based on the finding in the first study, we shall thereby induce dependence between the studies, which needs to be appropriately taken into account, but even this does not constitute preferential sampling. In particular, any likelihood-based analysis of the two studies together would be entirely unaffected by this complication, yielding exactly the same results as if we had two independent studies, each with preassigned sampling sites.

I see a close analogy with the issue of confounding in observational studies, which is likewise untestable from data. If the statistician knew all the variables the doctor knew when deciding how to treat a patient, she could build a model allowing for these, and thereby hope to obtain reliable estimates of the effects of treatments on outcomes: confounding occurs when these variables are not known (or not included or appropriately handled in our model).

Modelling confounding is generally regarded as a poor and unreliable second best, with emphasis placed rather on conducting studies and including variables in such a way as to avoid the problem if at all possible. Are there lessons to be drawn here?

Jonathan Rougier and Li Chen (*University of Bristol*)

This very interesting paper has prompted us to to think harder about the role of additional information in spatial statistics. The numerical findings are clear cut. If sampling is preferential, the sites will mostly be at the high points (or low points, but let us stick with the former) of the process, and the process will appear, from the observations, to be high and flat. Hence standard estimators will be biased but have small variance. In contrast, if sampling is clustered but non-preferential, then the sites will sometimes mostly be at the high points of the process, sometimes mostly at the low points and often somewhere in between. In this case standard estimators will be unbiased but have large variances. This is exactly as shown in the numerical experiment. Any more precise characterization (e.g. concerning the mean-squared error in the two cases) would depend on the nature of the preferential sampling, and this would be application specific.

What is less clear cut is how we should respond to the possibility of preferential sampling, and this comes down to how additional information is incorporated in the statistical model. The authors incorporate the potential for preferential sampling in terms of a parametric model. This fairly intractable treatment involves a clever Monte Carlo approach to estimating the likelihood, using a couple of neat devices: circulant embedding and conditional kriging. This parametric model is designed to represent a lack of knowledge concerning how the site locations were actually chosen, but a suspicion that additional information was involved. But surely such a lack of knowledge would be rare. If we asked the scientists who performed the 1997 and 2000 experiments why they preferentially sampled the north, they would probably reply 'Because the north is more industrialized than the south, and lead concentrations are likely to be higher there, so there is more danger that they will breach the maximum safe dose that is specified by an ecotoxicological assessment'. If we want to incorporate this additional information in our inference about lead concentrations, then it is natural and simple to do so in terms of a spatially varying mean response.

It is difficult to imagine that information of this type would not be forthcoming, either from the scientists who did the experiment or from other scientists who would be familiar with the field. And therefore our concern is that this paper obscures the benefits of spending more time talking to the scientists and less time hunched over the computer.

The following contributions were received in writing after the meeting.

Julian Besag (*Bristol University*)

In the limited space available, I shall concentrate mainly on the prevalence of the mathematically seductive Matérn correlation function in contemporary geostatistics. Its apparent flexibility derives from the shape

parameter κ but this can be difficult to estimate. Indeed, despite their claim in Section 3.2 to use maximum likelihood, the authors fix $\kappa = 0.5$, a spatially unappealing exponential correlation, in their data analysis. The variogram plots in Fig. 5 are unsatisfactory: the left-hand cloud is impenetrable and the right-hand one not only lacks any evidence of a sill, which is rather typical, but there are also six points below the fitted curve and 15 above it. Incidentally, the replacement of two aberrant observations by the overall mean denies the very ethos of spatial statistics and could lead to serious consequences in the fitted variogram.

Interestingly, McCullagh and Clifford (2006) used the Matérn family to analyse an extensive catalogue of uniformity trials for many different crops, and in every case they found that $\kappa = 0$ does about as well as any $\kappa > 0$ in fitting plot yields via likelihood methods. The value $\kappa = 0$ is denied by some but was not by Matérn himself and corresponds to the de Wijs process, defined with respect to averages on arbitrary areas rather than on points. One might require a moment's reflection to convince oneself that, for many, perhaps most, spatial variables, including core samples, an areal formulation is strictly more appropriate, though that does not preclude processes defined on $\mathcal{R}^2$ as working approximations. One interpretation is that the de Wijs process, with its conformal invariance, should play a pivotal role in modern geostatistics, as it did for mining engineers in the 1950s. This would be convenient for me because of the close links between intrinsic Gaussian Markov random fields and the de Wijs process (Besag and Mondal, 2005). However, my guess is that the simple stationarity of the Matérn family when $\kappa > 0$ is its downfall for environmental phenomena. Perhaps a new family should be sought or perhaps it should be admitted that statistical modelling with few data points provides little more than a gateway to numerical analysis.

A different, as yet unexplored, approach might use the Voronoi tessellation and Delaunay triangulation of the x_is to define an intrinsic spatial prior for locally averaged (log-) concentrations, with a single unknown scale parameter. Prediction areas could be constructed by augmenting a roving point x^* in the study area, for example. Fast methods of updating are available via the Green and Sibson (1978) algorithm and Markov chain Monte Carlo importance sampling. Approximate self-consistency of the formulation as x^* moves should exist, again via links with the de Wijs process.

Paul Fearnhead (*Lancaster University*) **and Omiros Papaspiliopoulos** (*Universitat Pompeu Fabra, Barcelona*)

We congratulate the authors for this stimulating paper. First, we would like to ask for some clarification. It is stated that with known parameter values the correct predictive distribution for S is Gaussian. Do the authors mean that the predictive for S_1 given S_0 is Gaussian? Is this true under preferential sampling?

Our comment focuses on the Monte Carlo estimation of the likelihood. The main difficulty comes from $[X|S]$:

$$\left[\prod_{i=1}^{n} \exp\{\alpha + \beta\, S(x_i)\}\right] \exp\left[-\int_{\mathcal{A}} \exp\{\alpha + \beta\, S(x)\}\, \mathrm{d}x\right]. \tag{14}$$

The authors simulate the Gaussian process on a lattice and then approximate the integral in expression (14) by a sum. This leads to biases in the estimate of both the likelihood and the parameters. Although these can be reduced by refining the lattice, it is difficult to quantify such biases.

It may be possible to adapt recent Monte Carlo methods for estimating likelihoods for diffusions (Beskos *et al.*, 2006; Fearnhead *et al.*, 2008) to estimating the likelihood (or the score) of the models that are considered in this paper unbiasedly. This would imply that Monte Carlo sampling is the only source of error, and there are simple approaches to quantifying it.

For simplicity we focus on estimates of expression (14) and write $I = \int [c - \exp\{\alpha + \beta\, S(x)\}]\, \mathrm{d}x$ and $|A| = \int_A \mathrm{d}x$ where c is an arbitrary constant. We expand the exponential in a power series to obtain

$$\exp\left[-\int_{\mathcal{A}} \exp\{\alpha + \beta\, S(x)\}\right] = \exp(-c|A|)\exp(I) = \exp(-c|A|)\sum_{k=0}^{\infty} \frac{I^k}{k!}.$$

Then, we can use importance sampling to produce estimates of the sum on the right-hand side of expression (14) as follows.

Step 1: simulate k from a Poisson distribution with mean $\gamma|A|$.
Step 2: simulate $x'_1, \ldots, x'_k$ independently and uniformly on A.
Step 3: simulate $S(x_1), \ldots, S(x_n), S(x'_1), \ldots, S(x'_k)$.
Step 4: estimate expression (14) by

$$\left[\prod_{i=1}^{n} \exp\{\alpha + \beta\, S(x_i)\}\right] \exp\{-(c+\gamma)|A|\}\gamma^{-k} \prod_{i=1}^{k} [c - \exp\{\alpha + \beta\, S(x'_i)\}].$$

Simulation of $S(x)$ at the set of points $x_1, \ldots, x_n, x'_1, \ldots, x'_k$ in step 3 can be computationally costly. These points will be different for each Monte Carlo iteration and will no longer be located on a lattice. However, experience suggests that the mean value of k can often be much smaller than the number of lattice points that is used.

Montserrat Fuentes (*North Carolina State University, Raleigh*)

I congratulate Diggle, Menezes and Su for a very well-written paper, that without a doubt will have a very significant influence on the field of geostatistics. They present an elegant approach to improve spatial prediction and geostatistical inference under informative sampling. The model for preferential sampling that they adopted, which was previously introduced by Ho and Stoyan (2008), is applied to lead data from Galicia.

The shared parameter model, which is an alternative approach to that presented by Diggle and his colleagues, accounts for informative missingness by introducing random effects that are shared between the missing data process and the measurement process. Conditioned on the random effects, the missing data and measurement processes are assumed to be independent. Fuentes *et al.* (2008) introduced a spatial shared parameter model to predict precipitation data by using satellite data with missing values due to clouds. The framework that was presented by Fuentes *et al.* (2008), discretizing the spatial domain, is easy to implement and computationally efficient, because there is no Cox process and everything becomes conjugate. This approach could be adopted for spatial inference and prediction under preferential sampling.

Diggle and his colleagues assume that the latent spatial process is Gaussian. For non-spatial data, in the context of shared parameter models, several researchers have proposed methods that avoid assuming that the shared random effects are Gaussian (Lin *et al.*, 2000; Song *et al.*, 2002; Beunckens *et al.*, 2008; Tsonaka *et al.*, 2009). These approaches could be extended to the setting that is presented by Diggle and his colleagues by replacing the Gaussian spatial model for S with a non-Gaussian spatial model (e.g. Gelfand *et al.* (2005), Griffin and Steel (2006) and Reich and Fuentes (2007)).

It would be interesting to discuss the role of covariates a little. It is possible that, once we have accounted for some relevant covariates in the mean of the lead concentrations, then the preferential sampling might not improve the prediction any more. In this study, some relevant covariates could be the shortest distance from any given site to the Galician Rias, and maybe population density.

A Bayesian approach for estimation and prediction could be adopted, not only to help with the identifiability problem for the covariance parameters, but also to characterize the uncertainty of the estimated covariance parameters in the spatial prediction. As a side-note, it would have been helpful to calculate Euclidean distances and the range of correlation in kilometres rather than degrees. In Galicia, a degree longitude is different from a degree latitude, and that could introduce some artificial anisotropy that makes it more difficult to interpret the semivariogram.

A. E. Gelfand and A. Chakraborty (*Duke University, Durham*)

The authors have illuminated a problem which arises in various guises and opens up many paths for future exploration. We confine our comments to modelling issues. First, within the authors' setting we can take responses as equivalent to marks. Then, we have the customary two conditional modelling options—[responses|locations] or [locations|marks]. It is tricky to speak about the joint distribution of responses and locations; what would we mean by a marginal distribution for responses or a marginal distribution for all locations in say a set A? Additionally, it is important to note that the authors are linking locations to 'latent' responses, not to observed responses. There is work in the literature on the latter and it may be more preferred in some applications.

It is useful to note an aggregated version of their problem, i.e. to counts that are associated with grid cells. Now, preferential sampling presumes that counts for the cells are driven by stochastic integration of $S(x)$ over cells. Now, we must avoid the ecological fallacy that is associated with such integration (see, for example, Wakefield and Shaddick (2006)). So, it does not seem that computation for the aggregated problem will be simpler.

There is an interesting data confidentiality problem which has a similar flavour to preferential sampling. The objective is to create synthetic data sets which have essentially the same structure as a given data set so that the former can be made available for public release. For data sets with spatial identifiers, this requires 'relocating' individuals. Evidently, the new locations will depend on the response, as well as covariates that are associated with the individual. We are exploring novel preliminary models to accomplish this.

Finally, there are obvious connections of preferential sampling to spatial design issues when the goal is to sample where exposure levels are, say, high. Illustrative recent work includes Xia *et al.* (2006), who

introduced a suitable utility function for such sampling, and Loperfido and Guttorp (2008), who proposed the use of a closed skew normal distribution.

Peter Guttorp (*University of Washington, Seattle, and Norwegian Computing Center, Oslo*) **and Paul D. Sampson** (*University of Washington, Seattle*)

When studying health effects of air pollution, it is quite common to use opportunistic sampling. Health data often come from hospital registries, whereas air quality data (at least in the USA) come from air quality regulation compliance networks. In the state of Washington, the ozone network is frequently modified to eliminate monitors with low readings in favour of locations with high ozone values in populated areas (Reynolds *et al.*, 1998). Furthermore, since the health effects of ozone are quite small (Bates, 2005), issues of bias due to study design, exposure estimates etc. become very important. If locations are chosen to achieve high readings, the exposure estimates will tend to be too high, and the health effects potentially underestimated. This is particularly serious since the estimated health effects are used to set air quality standards (Guttorp, 2006). The size of the exposure estimate bias can be calculated for Gaussian processes by using skew normal calculations (Loperfido and Guttorp, 2008).

The proposed model for preferential sampling appears relevant to some environmental sampling problems (although it does not accord with the sampling for high gradients in the example that was presented), but it is almost certainly too simplistic for most air quality monitoring networks designed to satisfy multiple objectives. Sites may be intentionally located to represent

(a) 'background' (low) levels of pollution outside urban areas,
(b) air quality levels in residential areas and
(c) air quality concentrations near pollutant sources.

The locations of monitors are almost certainly made in terms of observable explanatory variables ('land use covariates') that are also used in spatial modelling. Thus variation in $S(\mathbf{x})$ that is not explained by these explanatory variables may not be related to sampling through a model like that of assumption 2.

There is little theoretical work on robustness issues in geostatistics. A framework for such theory must develop general classes of outliers that one wants to protect against. Martin and Yohai (1986) described such classes for time series, and Assunção and Guttorp (1999) for point processes. Cressie and Hawkins (1980) and Genton (1998) proposed robust variogram estimators, that guard against particularly large squared differences at a given distance. One type of outliers is grossly misspecified locations. In this case the locations do not depend on the value of the process, but location errors clearly can affect both variogram estimates and geostatistical predictions. Can analyses similar to those in this paper be applied to misspecified locations as well?

Duncan Lee (*University of Glasgow*)

I thank the authors for an interesting and valuable paper, which will be useful in many application areas. One such area is spatiotemporal modelling of air pollution concentrations, particularly over an urban environment. Pollution modelling of this type can be used

(a) to construct representative indices of overall air quality,
(b) to produce daily high resolution air pollution maps and
(c) to create a proxy measure of exposure in health impact studies.

In these contexts, incorporating the authors' preferential sampling model to allow for the possibly non-representative nature of the pollution monitors would potentially reduce the bias in the pollution concentrations that are estimated by existing models.

As the authors note, the model for the monitor intensity $\lambda(x)$ may not be adequate for all applications, and in this context two generalizations are needed. Firstly, pollution monitoring networks are typically constructed piecemeal, and the reasons for the choice of monitoring location are generally unclear. The monitors could be located where pollution levels are expected to be either low or high, where the first case would improve the chances of pollution legislation being met, whereas the second would monitor the worst case scenario. It is likely that monitors have been placed for both these reasons, meaning that the log-linear relationship that is implied by equation (5) may not be appropriate in this context. A simple extension would be to model the intensity function $\lambda(x)$ as a low order polynomial or natural cubic spline of the pollution concentrations $S(x)$, although care would have to be taken to ensure that the data contained enough information to estimate the additional parameters reliably.

Secondly, the intensity function $\lambda(x)$ would have to be extended to a spatiotemporal setting, because the size and composition of pollution monitoring networks change over time. Therefore a spatiotemporal intensity function of the form

$$\lambda(x,t) = \exp\{\alpha + f(x) + g(t) + h(x,t)\}$$

would be required, where x indexes space and t indexes time. Here $f(\cdot)$ represents the effects of preferential sampling, $g(\cdot)$ the change in the size of the monitoring network over time and $h(\cdot)$ is an interaction that allows the degree of preferential sampling to change over time. A simple model would be to represent these functions as linear terms, whereas low order polynomials or splines could be used if more flexibility was required.

Peter McCullagh (*University of Chicago*)

Biased sampling, or preferential sampling, is a phenomenon that may arise when sites are selected or units are generated in such a way that the covariate configuration $\mathbf{x}$ is not independent of the response process Y. This paper demonstrates clearly that the phenomenon is not peculiar to logistic models, but may also occur in Gaussian models.

Near the end of Section 2, the authors consider a model in which events are generated in $A \times \mathcal{R}$ by a Cox process with intensity

$$\lambda(x,y) = \exp\{\alpha + \beta S(x)\}\tau^{-1}\phi\left\{\frac{y - \mu(x) - S(x)}{\tau}\right\}, \tag{15}$$

where $S \sim N(0, K)$ is a Gaussian process, and $\phi(\cdot)$ is the standard normal density. The marginal point process $\mathbf{x} \subset A$ is also a Cox process, and the observation consists of the pair $(\mathbf{x}, Y[\mathbf{x}])$. Evidence is presented that standard variogram estimates and spatial predictions may have appreciable bias under this sort of sampling. The technical project is advanced by the development of numerical techniques for likelihood computation for the point process density

$$p(\mathbf{x},\mathbf{y}) = E_\lambda[\exp\{-\Lambda(A)\}\prod_{\mathbf{x}}\lambda(x_i,y_i)].$$

One rationale for a model of this type in biostatistical work is that it associates with each individual i a random effect $S(x_i)$, which governs both the probability of recruitment and the response distribution (McCullagh, 2008). For obvious reasons, this argument is less compelling for laboratory experiments and agricultural field trials than it is for clinical trials where each recruit is an eligible volunteer. The biostatistical rationale may be persuasive in other areas such as studies of the effect of economic incentives on individual behaviour, but its relevance to environmental sampling is less obvious. Sequential sampling, interpreted as $\mathbf{x}^{(r+1)} \perp Y|Y[\mathbf{x}^{(r)}]$ following the discussion in Section 6, has no effect on the likelihood or on predictive distributions.

The observation generated by the Cox process (15) is a random measure in $A \times \mathcal{R}$ whose moments can be computed directly from those of λ without resorting to the joint density. The single-event moment density $E\{\lambda(x,y)\}$ is such that, for each $x \in \mathbf{x}$, the conditional density of $Y(x)$ is a ratio of expected values, $E\{\lambda(x,y)\}/E\{\lambda(x)\}$, i.e. Gaussian with mean $\mu(x) + \beta K(x,x)$ and variance $K(x,x) + \tau^2$. For distinct k-tuples $\mathbf{x}' \subset \mathbf{x}$ the conditional distribution of $Y[\mathbf{x}']$ is Gaussian with mean $\mu[\mathbf{x}'] + \beta K[\mathbf{x}']\mathbf{1}$ and covariance matrix $K[\mathbf{x}'] + \tau^2\delta[\mathbf{x}']$. One complication for $\beta \neq 0$ is that the difference

$$\Delta(x,x') = E\{Y(x)|x,x' \in \mathbf{x}\} - E\{Y(x)|x \in \mathbf{x}\} = \beta K(x,x')$$

is non-zero, a characteristic of interference. For arbitrary functions $h : \mathcal{R}^2 \to \mathcal{R}$, the additive estimating function $T = \Sigma_{\mathbf{x}} h(x)\{Y(x) - \mu(x) - \beta K(x,x)\}$ has zero mean. Formula (15) of McCullagh (2008) for the covariance of two such functions simplifies to

$$\begin{aligned}\operatorname{cov}(T,T') = &\int_A h(x)h'(x)\{\tau^2 + K(x,x)\}m_1(x)\,\mathrm{d}x \\ &+ \int_{A^2} h(x)h'(x')\{K(x,x') + \Delta^2(x,x')\}m_2(x,x')\,\mathrm{d}x\,\mathrm{d}x'\end{aligned}$$

where $m_1(x) = E\{\lambda(x)\}$ and $m_2(x,x') = E\{\lambda(x)\,\lambda(x')\}$. Both integrals can be estimated by summation over the observed values. The occurrence of the interference term shows how the theory of estimating functions for preferential samples differs from the theory for regression models.

Donald E. Myers (*University of Arizona, Tucson*)

Spatial statistical methods assume that there is at least a perceived spatial correlation in the values of the phenomenon of interest as opposed to randomness or constant distribution. Thus there is always the potential for preferential sampling but there is still the difference between intentional and accidental preference. For the former it is necessary to know or assume knowledge of the spatial distribution. The particular example that is presented in this paper appears to fall in that category and hence it is of interest to consider the effect of this knowledge. In some cases this knowledge is gained during sampling; for example, exploratory drilling for an ore body requires considerable time and hence it is possible to analyse drill cores before choosing new sample locations. In other cases one may have knowledge of the deposition process and hence some knowledge of the spatial distribution, e.g. soil pollution emanating from a known source. This might be incorporated in a Bayesian model as in Cui *et al.* (1995). Air pollution is both spatial and temporal but the number of spatial locations is usually quite small; hence those locations will almost always be preferentially selected. The notion of preferential sampling seems to be much broader than the model that is described in the paper will allow. The model $[S, X] \neq [S][X]$ seems to imply only a stochastic dependence whereas in some cases it may be more deterministic. Having said all this the presentation is interesting and has raised some interesting issues.

A few observations, however, are relevant. Geostatistics developed largely outside statistics and in particular was motivated by problems in mining, hydrology and petroleum. Subsequently it spread to soil science, environmental monitoring and later to many other areas. In most of those applications the assumption of a multivariate Gaussian distribution is completely unrealistic and it seldom appears in the applications literature. A univariate transformation such as the Box–Cox transformation will not ensure a multivariate distribution. Particularly in mining the data are nearly always non-point support and most often the objective is to estimate 'block' averages. The derivation of the ordinary kriging equations does not require any distribution assumptions. The model in equation (l) is not the model that is typically used in geostatistics since this corresponds to a smoothing estimator rather than an exact interpolator. The use of data with non-point support would complicate the use of maximum likelihood estimation.

Mari Myllymäki (*University of Jyväskylä*) **and Felix Ballani and Dietrich Stoyan** (*Technische Universität Bergakademie Freiberg*)

We appreciate that the authors highlight the importance of taking into account preferential sampling. Our discussion concerns Sections 2 and 3 of the paper.

The class of models that is defined through assumptions 1–3 is a particular case of the geostatistical model for preferential sampling in Ho and Stoyan (2008). The sampling locations, or points x_i, are from the log-Gaussian Cox process and the Gaussian variates Y_i of the paper correspond to the marks $m(x_i) = \mu + S(x_i) + \varepsilon(x_i)$, where $\varepsilon(x_i) \sim N(0, \tau^2)$.

In the context of models such as that considered in the paper, it makes sense to speak both about the 'field variogram' $\gamma(u)$ and the 'mark variogram' $\gamma_m(u)$, which belongs to the corresponding marked point process; see for example Illian *et al.* (2008), page 344. The empirical variogram based on Y_i observed at x_i estimates $\gamma_m(u)$, whereas, in geostatistics, a statistician hopes to estimate $\gamma(u)$.

For the model of the paper,

$$\gamma(u) = \gamma_m(u), \tag{16}$$

when $\tau^2 = 0$ (as in Section 3); see Ho and Stoyan (2008), equation (19). This seems to be in contradiction to the words of the paper on page 196, lines 8–11.

We made simulations of the model. Fig. 13 shows the results for the model with the random-field parameters of Section 3 (mean 0), $\alpha = 3.0$ and $\beta = 1.5$ (about l00 points on average in a unit square). We simulated full randomness whereas a form of conditional simulation was used in the paper.

We think that the bias occurs since the estimator for the mark variogram is only asymptotically ratio unbiased. We are surprised that the bias is so large, but it decreases along an increasing window size. For large windows the simulations become extremely long; the variability of the model is high.

Because of equation (16), the model is not a good example for problems with preferential sampling. Models for which equation (16) is not true are discussed in Ho and Stoyan (2008) and Myllymäki and Penttinen (2009) in terms of marked point processes.

Håvard Rue and Sara Martino (*Norwegian University for Science and Technology, Trondheim*), **Debashis Mondal** (*University of Chicago*) **and Nicolas Chopin** (*Ecole Nationale de la Statistique et de l'Administration Economique, Paris*)

We thank the authors for illustrating how the issues of *preferential sampling* can be accounted for in a

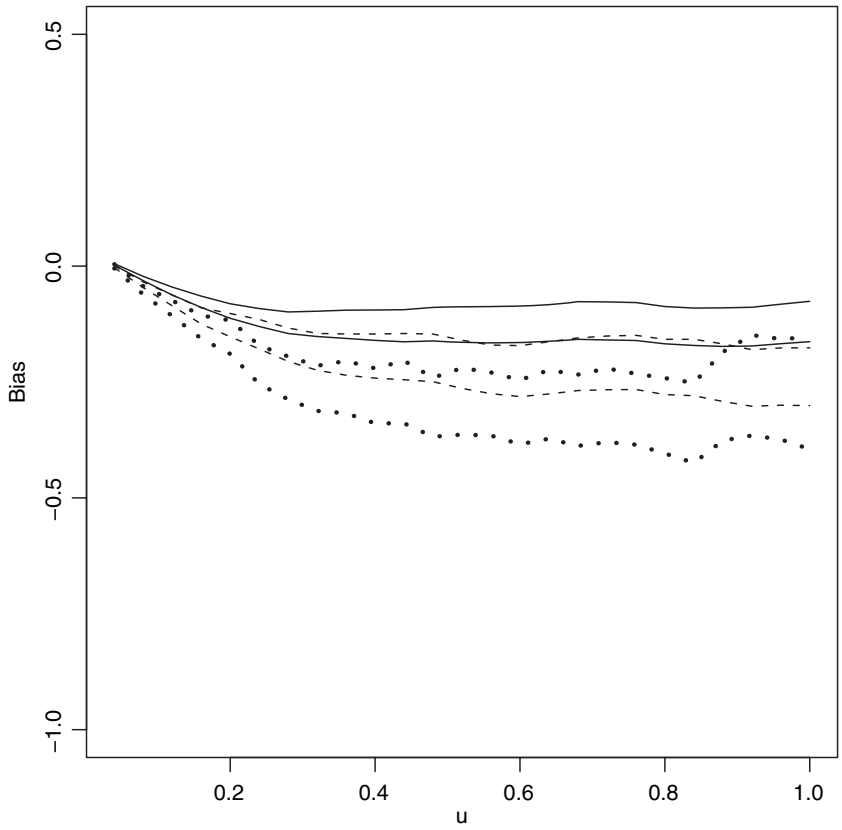

Fig. 13. Estimated bias (pointwise mean plus and minus two pointwise standard errors) of the mark variogram derived from 500 replicate simulations in windows of size [0,1] × [0,1] (·····), [0,2] × [0,2] (-------) and [0,5] × [0,5] (———)

simplified setting of geostatistical analysis. Preferential sampling is often regarded as part of a better study design when gaining knowledge on certain *excursion sets* of underlying spatial random fields is of primary concern. In mineral exploration, for example, these excursion sets can form the body of ore that has a grade higher than that which supports profitable mining; see Veneziano and Kitanidis (1982) for another kind of preferential sampling in such settings. There is also a connection between preferential sampling and data missing not at random (e.g. Diggle and Kenward (1994)), which hints at a possible lack of parametric robustness in some settings.

Our main comments concern the spatial model considered and the inferential methods. First, estimation of variance and range parameters is troublesome for a Matérn covariance model in a fixed domain, particularly when the range parameter is large compared with the region itself. Zhang (2004) showed that, under infill asymptotics, only σ^2/ϕ^{2k} can be estimated consistently. This imposes ridge-like behaviour on the log-likelihood, and a second-order expansion of the log-likelihood around the mode may not provide correct estimates of the variability. Second, the model that is presented here is yet another example of a latent Gaussian model for which Bayesian inference based on *integrated nested Laplace approximations* is both swift and efficient; see Rue *et al.* (2009). The integrated nested Laplace approximation accounts automatically for the strong dependence between the spatial variance and the range, provides marginal posterior densities rather than point estimates and is faster than Monte-Carlo-based estimation methods. R code that redoes the simulation example and reanalyses the Galicia data is available at `www.r-inla.org`. We have used Markov random-field representations of Matérn covariance models and thin plate splines, for which there is no range parameter; see Lindgren and Rue (2007), Lindgren's discussion of Rue *et al.* (2009) and Besag and Mondal (2005).

We indicate some remarks about adjusting spatial models that respect the boundary between sea and land which can be somewhat problematic for the exponential covariance model that is used by the authors; see Wood *et al.* (2008) who discuss soap film smoothing. Lindgren and Rue's (2007) approach based on stochastic partial differential equations leads to more appropriate geostatistical covariance models on complex domains; work along these lines is on going. Further challenges remain in modelling space–time environmental monitoring data with more complex preferential sampling and with covariates that are also included preferentially.

Michael L. Stein (*University of Chicago*)

This paper clearly shows the effect of preferential sampling based on prior information about S. When the preferential sampling is based on observed values of S, the effect may be much smaller. Indeed, as I demonstrated in an unpublished chapter of my dissertation (which is available from me), in a limited circumstance, there will be no effect at all. Specifically, let us assume model (1) with S and Z jointly Gaussian and the covariance structure of S and τ^2 known. Consider some initial set of fixed x_is and subsequent X_is that are a function of the contrasts of the previous $i-1$ observations (and the previous observation locations). Then the distribution of the error of any best linear unbiased predictor is unaffected by the sampling scheme. Note that this sampling rule allows preferential sampling near locations with large observed values, so designs qualitatively like those shown in Fig. 1(b) could result, even though the design rule is very different from equation (5).

In practice, the covariance structure would need to be estimated, but this effect might be small if the number of observations is sufficiently large, despite the results that are shown here on the effect of preferential sampling on variogram estimation. In particular, I do not think that maximum likelihood estimates of parameters controlling local behaviour of S should be greatly biased by either the kind of sequential sampling that is described here or the preferential sampling given by equation (5). For the Matérn model, the parameters controlling the local behaviour are κ and $\sigma^2/\phi^{2\kappa}$ (and the nugget). If these parameters are estimated reasonably well, we might be able to ignore the effect of sequential sampling on the error distribution of kriging predictions at sites that are not too far from existing observations. The authors show a substantial effect on prediction errors under sampling rule (5) (see Table 1), but it would be interesting to see how preferential sampling performs in those cases for which there are observations close to x_0.

Fig. 2 shows the substantial bias that preferential sampling can produce in the sample variogram even at short lags. For a differentiable process (S is not quite differentiable for $\kappa=1$), elementary calculus implies that squared increments at short lags should be small near local maxima. I would guess that, for $\kappa=0.5$ and sufficiently large ϕ, the (relative) bias at short lags would be substantially smaller.

Dietrich Stoyan (*Technische Universität Bergakademie Freiberg*)

Geostatistical inference under preferential sampling is a very important problem. Indeed, quite often sampling locations are deliberately concentrated in 'interesting' subregions where the spatial variable investigated has particularly large or small values. This is often so in geological exploration or in air pollution monitoring. A naive application of geostatistical methods can then fail. For example, estimated variograms may be misleading and may not give the information that is expected.

Therefore it is of great value that the authors consider and discuss this problem. Their model-based approach has the potential to demonstrate and solve some of the difficulties in this context. It may be a warning for a naive statistician. Under preferential sampling, the true variogram of the random field of interest may heavily deviate from the empirical variogram resulting from the classical estimation procedure if the sampling locations are correlated with the field.

For practical applications, a weak point of the ideas that are presented in the paper is the stationarity assumption of the basic model. We can learn what may happen in preferential sampling, but we have only a rather unrealistic model. For my taste also the very interesting example with lead pollution and moss suffers from this disadvantage. The point pattern of the black dots in Fig. 3 does not look like a sample of a stationary point process, and for the lead pollution in Galicia I expect some source and thus a clear inhomogeneous distribution, also. Therefore, I have doubts about the use of Ripley's K-function and the statistics that are based on it. By the way, it would be interesting to know the thoughts of the Spanish biologists who took the moss samples, when they hear that they have produced a stationary Cox process sample.

Perhaps in situations like those given in the lead pollution example it is better to use spatial interpolation methods without a model in the background and perhaps without any stochastic thinking, e.g. inverse distance weighting.

The models for marked point processes that are mentioned in Section 2 and those in Myllymäki and Penttinen (2009) have some potential for spatial statistics and modelling. They may find applications in situations where for example an initial distribution of matter leads to nucleation and particles, which are represented by the points, where the marks characterize for example particle size, or in forest statistics where the random field describes the growth conditions and controls both tree density and tree sizes.

Paul Switzer (*Stanford University*)

I had some difficulty visualizing how the model of Diggle and his colleagues for preferential site selection could practically operate where the site selector has no information whatever about the realized spatial function S. All she knows is that she has a spatially stationary stochastic model $\mathbf{S}^*$ that generated S and, from her viewpoint, there is no basis for preferential sampling. If she does indeed have some information about S, e.g. side information about where high values are likely to be found, then the appropriate model for $\mathbf{S}^*$ should be a non-stationary model that incorporates or conditions on this side information. It is then in the context of the non-stationary model where accounting for preferential sampling in parameter estimation should be addressed.

An example arises in two-stage sampling where second-stage sample sites are preferentially chosen according to what is revealed by the first-stage sample. The appropriate model before first-stage sampling might indeed have been a homogeneous $\mathbf{S}^*$-model but, conditional on the first-stage data, the model for $\mathbf{S}^*$ is no longer stationary and the analysis of the second-stage data will need to account for preferentiality of site selection via the conditioning. In this example one does not condition on all of S, as in the paper, but only on that part of S that was observed in the first stage.

As a simple illustration of the two-stage example suppose that $\mathbf{S}^*$ is stationary Gaussian with zero mean, unit variance and isotropic variogram function $V(d)$, where d is interpoint distance. A single observation yields the value S_1 and a second observation site is preferentially selected at distance $d = D(S_1)$ from the first observation, yielding the value S_2. If the same two sites had been selected in advance then the naive variogram estimator $(S_1 - S_2)^2$ has conventional expected value $V(d)$. However, the preferential second-stage site selection that depended on S_1 changes the expected value of the naive variogram estimator by an amount that is readily calculated and that depends on S_1.

None of this at all contradicts what Diggle and his colleagues clearly and usefully point out, i.e. that a naive analysis of spatial data derived from preferentially located sampling sites can lead to biases in parameter estimation and spatial interpolation when the interpolator depends on the model parameters, such as in kriging.

Adam A. Szpiro and Lianne Sheppard (*University of Washington, Seattle*)

In air pollution epidemiology, we use geostatistical methods to predict exposures at subject locations based on monitoring data (Jerrett *et al.*, 2005; Szpiro *et al.*, 2009a). This introduces measurement error that can adversely affect inference (Kim *et al.*, 2009). Recent work has introduced correction methods (Szpiro *et al.*, 2009b; Gryparis *et al.*, 2009), but these are based on uniform sampling. If the sampling is preferential then the implications for health effect inference are further complicated. It is plausible to account for preferential sampling in the geostatistical model or as part of the measurement error correction. Both are important directions for future research.

The preferential sampling modelled by Diggle and his colleagues does not account for all of the relevant considerations in air pollution monitoring. Data are often derived from regulatory monitors sited for policy objectives. Sometimes additional data are collected, and a major consideration in siting is the covariate design space (Cohen *et al.*, 2009). Preferential sampling in regions with high variability has also been proposed (Kanaroglou *et al.*, 2005).

We simulate effects of preferential sampling on health effect estimation. The exposure model is

$$Y = \mu + S(x) + Z$$

with intercept $\mu = 1.36$ and exponential variogram with range 90, sill 3.76 and nugget 1.34. On a (500×400)-unit domain (discretized to 50×50 cells), we sample Y at 200 locations and use kriging based on maximum likelihood to predict the exposure Y^* at $N^* = 300$ locations where the health outcome Z^* is measured. Assume

$$Z^* = \beta_0 + Y^* \beta_1 + \varepsilon$$

with ε independent and identically distributed and $\beta_1 = -0.322$ the parameter of interest. Consider five sampling schemes:

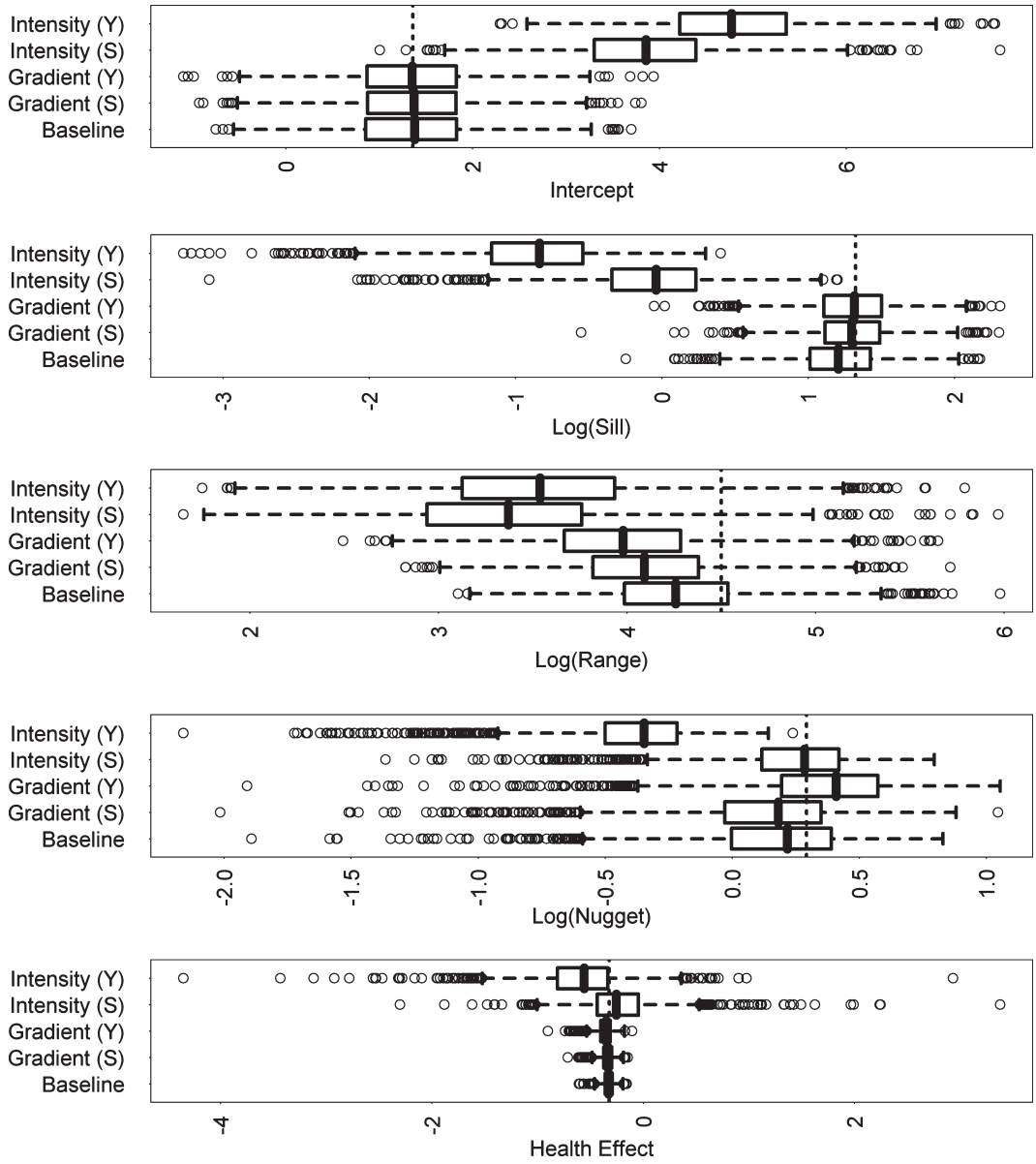

Fig. 14. Simulation results showing the effects of preferential sampling on geostatistical model parameters and the health effect parameter: ⋮, values used in generating the simulated data

(a) *baseline*, non-preferential uniform sampling;
(b) *intensity (S)*, preferential sampling based on $S(x)$ (as in the paper with $\alpha = 1$ and $\beta = 2$);
(c) *intensity (Y)*, preferential sampling based on $Y(x)$;
(d) *gradient (S)*, preferential sampling based on the variability of $S(x)$ (variance of 30 nearest neighbours);
(e) *gradient (Y)*, preferential sampling based on the variability of $Y(x)$.

Results for 2000 realizations are shown in Fig. 14 (excluding those with unsatisfactory maximum likelihood fits as in Szpiro *et al.* (2009b)). Sampling based on intensity results in bias of the mean and sill in the

geostatistical model (which is consistent with Diggle and his colleagues). There is also bias in the health effect estimates, with the sign depending on whether the sampling weights incorporate the nugget. With sampling based on local variance, bias in the mean goes away and bias in the sill and health effect are diminished. We conclude that preferential sampling can adversely affect health effect inference, but the magnitude of the problem depends on the specific sampling design.

K. F. Turkman (*University of Lisbon*)

I congratulate the authors on a paper which will stimulate further research in modelling spatial data.

When sampling design is determined by nature, marked point processes and the consequent inferential methods, as suggested by the authors, seem to be the right way to go. For example, while modelling wildfire sizes, ignoring locations of ignition may produce bias, as it is reasonable to assume that the ignition event and the consequent size of fire will depend on common factors.

If there is a fixed preferential sampling design before data collection, how should we include this design in the model? In many sampling designs, the sampling locations depend on a latent process S_1 (call it 'expert opinion') which stochastically depends on the latent process S, but is not equal to S. In this case, the likelihood for data X and Y can be expressed as

$$[X, Y] = E_{(S_1, S)}[[Y|X, S_1, S][X|S_1]].$$

For 'unbiased' preferential sampling schemes such as rank set sampling (which results in an unbiased estimator for the mean surface no matter who the expert is) it is reasonable to assume that Y depends on S_1 only through X and S, so that

$$[Y|X, S, S_1] = [Y|X, S].$$

In this case, the likelihood for data X and Y can be expressed as

$$\begin{aligned}[X, Y] &= E_{(S_1, S)}[[Y|X, S_1, S][X|S_1]] \\ &= \int_s \left(\int_{s_1} [X|S_1][S_1|S] \, \mathrm{d}s_1 \right) [Y|X, S][S] \, \mathrm{d}s \\ &= \int_s [X|S][Y|X, S][S] \, \mathrm{d}s \\ &= E_S[[Y|X, S][X|S]],\end{aligned}$$

and the likelihood will not depend on expert opinion. However, intuitively we would expect that the efficiency will depend on the quality of the expert opinion and hence dependence between S_1 and S will still matter and probably will enter through $[X|S]$.

When $[Y|X, S, S_1] \neq [Y|X, S]$, then no such simplification will exist in the likelihood, and we must take into account the sampling scheme S_1, as well as its relationship to S. I can envisage such a situation when the sampling is not a good sampling scheme, in the sense that it will not capture minimally the variability that is inherent in the process.

The exact sampling scheme for the l997 data set is not known, although the sampling was done at places where expert opinion thought concentration levels were higher. But apparently the expert was wrong from the beginning and sampled more where the concentration was lower. The model proposed captures it, by fitting a negative β, and adjusts the bias due to sampling accordingly. However, I hope that the model proposed will not be taken as a panacea for bad sampling designs, in the sense that, no matter how badly the sampling scheme is designed, it will be taken care of by the model.

Richard D. Wilkinson (*University of Nottingham*)

Preferential sampling occurs when sampling locations X and the process S are dependent. This implies that, when choosing X, the survey designer had prior knowledge of S which was used in some manner in the design. For example, the locations in the 1997 sample in the Galicia data set were chosen to lie in regions where the surveyors believed *a priori* that there were likely to be large gradients of lead concentrations. This dependence complicates the analysis as we can then no longer factorize the distribution as

$$[S, X, Y] = [Y|S(X)][X][S]. \tag{17}$$

My question concerns whether we can bypass the problem of preferential sampling by conditioning on the information that induced the dependence between S and X. Suppose that B is the surveyor's prior belief about S before observing Y and that this information includes the model for how X was chosen. For example, we might hope to elicit the prior expected response surface $\mathbb{E}\{S(x)\}$ and perhaps also the variance. This corresponds to specifying a mean and correlation function in *assumption 1* for the Gaussian process that is assumed in the paper. Models for how the locations X were chosen can be decided after discussion with the surveyor. Examples might include choosing X where we expected S to be large (e.g. $X \sim U[x : \mathbb{E}\{S(x)\} > \varepsilon]$), or where the derivative is large ($X \sim U[x : |\mathrm{d}\mathbb{E}\{S(x)\}/\mathrm{d}x| > \varepsilon]$) etc. The model stated in *assumption 2* in the paper, namely that X is a Poisson process with rate $\lambda(x) = \exp\{\alpha + \beta\, S(x)\}$, could never arise in practice as the surveyors do not know S. However, it is feasible that the rate could be $\lambda(x) = \exp[\alpha + \beta\, \mathbb{E}\{S(x)\}]$, for example.

For each of these models, S and X are dependent. However, they are conditionally independent given $\mathbb{E}\{S(x)\}$. Hence, if B contains the information that induced the dependence between S and X, then $[X|S, B] = [X|B]$. Thus, given prior knowledge B, the distributional relationship becomes

$$\begin{aligned}[S, X, Y|B] &= [Y|S(X), B][X|S, B][S|B] \\ &= [Y|S(X), B][X|B][S|B], \end{aligned} \tag{18}$$

which returns us to a non-preferential setting (see equation (17)). Typically, $[Y|S(X), B] = [Y|S(X)]$ if the prior information solely concerns S and X, and the distribution $[X|B]$ is the model that is used by the surveyors to choose X. The final term in equation (18), $[S|B]$, is tractable for many types of prior information if we assume that S is a Gaussian process. In conclusion, by using expert knowledge there may be an alternative approach to dealing with preferential sampling.

The **authors** replied later, in writing, as follows.

We are grateful to all the discussants for their interest, and for their perceptive and helpful comments. We agree with most of their comments, including some of the more critical ones. Our aim in writing the paper was to air what we believe is an important topic that has been neglected for too long. We do not claim to have provided a definitive solution, and many of the discussants have given useful suggestions for improvement.

We shall structure our reply around topics rather than individuals, and we apologize in advance if we have not captured every point raised.

Geostatistics and marked point processes

Formally, our model is a marked Cox process of a kind that was included (with acknowledgement of priority to Raquel Menezes's doctoral thesis) in Ho and Stoyan (2008). However, as noted by Professor Anderson, we fitted this model conditionally on the number of locations, at least in part because we do not literally believe the point process model, but rather consider it as a device within which a single parameter β (or, more accurately, the product $\beta\sigma$) controls the extent to which sampling is preferential. This conditioning turns out not to be entirely innocuous. For example, the result that was presented by Dr Myllymäki, Dr Ballani and Professor Stoyan in their equation (16) is striking and can be understood intuitively from the observation that, in a stationary Gaussian process $S(\cdot)$, the random variables $S(x) - S(x')$ and $S(x) + S(x')$ are independent. But the result does not hold conditionally on the number of locations in A; furthermore, their simulations agree with ours that the estimated variogram is biased. In this connection it would be interesting to understand the theoretical properties of the weighted variogram estimator that was suggested by Professor Rathbun; some results for a related proposal are given in Menezes *et al.* (2007).

Dr Illian makes the valid observation that, if modelling the point process of locations were of scientific interest, our marked Cox model would be found wanting as it fails to capture the small-scale inhibitory behaviour that was revealed by her analysis. However, our model does describe reasonably well the larger-scale spatial heterogeneity (Fig. 15).

Monte Carlo methods of inference

Professor Fearnhead and Dr Papaspiliopoulos emphasize the approximate nature of our likelihood calculations and suggest a way forward using their impressive work with Gareth Roberts and Alex Beskos, in a paper that also was read to the Society. Dr Guillas, and Professor Rue, Dr Martino, Professor Mondal and Dr Chopin, make the different point that a quadratic approximation to the likelihood may be unreliable as a basis for calculating approximate standard errors. Both of these issues could be addressed in an efficient

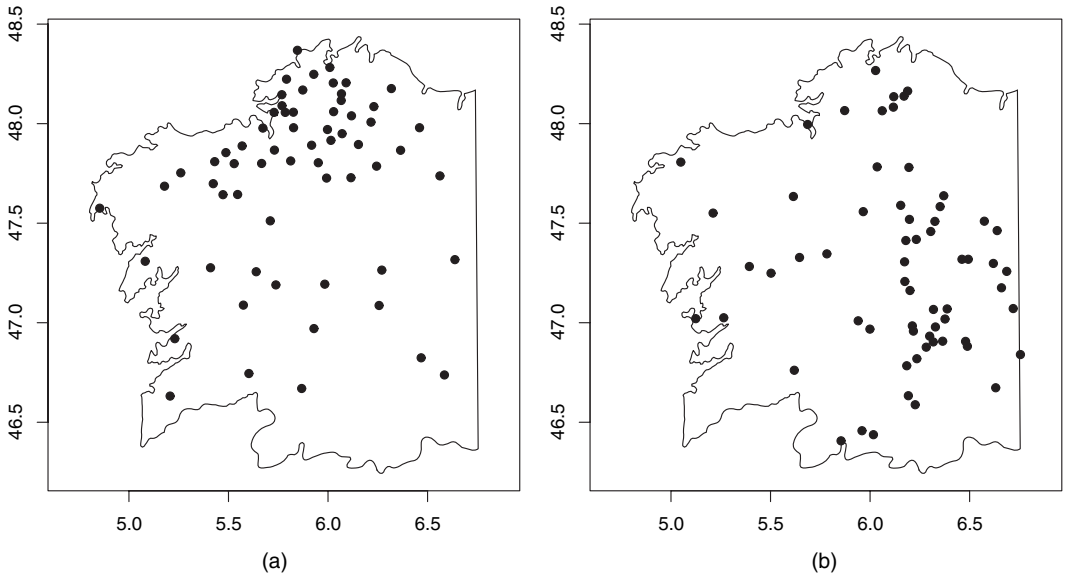

Fig. 15. (a) Galicia (1997) data locations and (b) realization of the fitted log-Gaussian Cox process

Bayesian implementation, which we freely admit we have not yet been able to construct; this also touches on Fearnhead and Papaspiliopoulos's questioning of our assuming a Gaussian predictive distribution, which strictly requires parameter values to be known rather than estimated. Here, as elsewhere, a Bayesian formulation would naturally incorporate parameter uncertainty into predictive probability statements.

The integrated nested Laplace approximation methodology that was mentioned by Rue, Martino, Mondal and Chopin is, in our opinion, an important development that is likely to have a major influence as it becomes more widely known. According to our understanding of the integrated nested Laplace approximation method, it does not remove the need for Monte Carlo methods to sample from the joint predictive distribution of all elements of the (suitably discretized) process $S(\cdot)$. However, provided that the method can deliver a good approximation to the joint posterior for the model parameters, it is then a straightforward exercise to generate Monte Carlo samples from a mixture of plug-in predictive distributions by using the joint posterior as the mixing distribution.

The importance of explanatory variables

Many of the discussants alluded to the importance of explanatory variables either directly or indirectly (Dr Rougier and Dr Chen, Professor Scott, Professor Fuentes, Professor Guttorp and Professor Sampson, Dr Illian, Professor Myers, Professor Dawid, Professor Switzer and Dr Wilkinson, with apologies for any omissions). We could not agree more strongly that modelling an association between sampling intensity and measured values by a shared dependence on extant explanatory variables is almost always preferable to modelling the association stochastically, which is what we were trying to say in Section 6, following equation (12). We think that it is still useful to be able to embed such a model within a preferential sampling framework, both to test its adequacy and to gain some understanding of the consequences should it prove inadequate. But we absolutely agree with the point made by several discussants, most succinctly by Rougier and Chen, that talking to scientists is preferable to spending time 'hunched over the computer'.

When does preferential sampling matter?

Over and above our response above with respect to the use of explanatory variables, we acknowledge that the effects of stochastic association between the sampling design and the process $S(\cdot)$ may be more subtle than can be captured by the model that is described in the paper. Professor Anderson's nice result reinforces the intuitive idea that the biggest effect is likely to be to bias the estimation of the mean. This relates indirectly to the choice between stationary and intrinsic models. Several years ago, Professor Myers suggested to one of us (PJD) that preferential sampling should not affect geostatistical inferences 'because geostatistics is based on differences' (we paraphrase). Comments now by Professor Stein on geostatistical prediction, by Professor Besag in advocating the use of intrinsic models, by Dr Szpiro and Professor Shep-

pard in their empirical study of different preferential sampling mechanisms, by Professor Dawid concerning sequential sampling schemes and by Professor Turkman in suggesting that the sampling intensity might be driven by a second latent process, correlated with but not identical to $S(\cdot)$, give specific examples of circumstances in which the effects of preferential sampling may be either diminished in size or completely eliminated.

The basic idea of biased sampling is, of course, neither new nor specific to spatial statistics. Professor Scott reminds us that we teach our students the importance of representative sampling designs but sometimes ignore our own advice when presented with complex data. Also, my impression is that in many degree syllabuses sampling and design topics feature less prominently than used to be so. Professor McCullagh's comments, following on from his paper read to the Society, redress the balance by setting out a general theoretical framework for joint modelling of a response process and a sampling design. Professor Rathbun makes explicit the connection of the paper to classical survey sampling methodology, in which context weighting schemes are often used to correct for the effects of sampling bias.

Spatiotemporal sampling

As pointed out by Professor Dawid, sequential sampling schemes in which the locations of later samples are allowed to depend on earlier measurements raise no particular difficulties for conventional methods of geostatistical inference provided that all the relevant conditioning variables are included; in this situation, the earlier measurements are converted through the modelling assumptions into extant explanatory variables.

Dr Lee describes a class of problems in which the underlying process $S(\cdot)$ has spatiotemporal structure. By the same argument as advanced by Professor Dawid, if the monitor locations are stochastically dependent on measured values at time $t=0$ then conditioning a spatiotemporal analysis on these initial values should remove the need to adjust for the preferential nature of the original sampling design.

Modelling the Galicia data

Professor Stoyan claims that the pattern of the 1997 data locations 'does not look like a sample of a stationary point process'. Fig. 15 compares a map of the data locations with a map of the fitted log-Gaussian Cox process. The areas of relatively high and low intensity are, of course, different but to our eyes the data and model show similar degrees of spatial heterogeneity.

Professor Besag suggests that our Fig. 5 shows a poor fit between the empirical and fitted variograms. In a sense we agree, and this is precisely why we do not advocate using the variogram to fit the model, but only as an exploratory tool to suggest a suitable family of parametric models, here $V(u)=\tau^2+\sigma^2\{1-\exp(-u/\phi)\}$; for further discussion see, for example, Diggle and Ribeiro (2007), chapter 5. The smooth curves in Fig. 5 are maximum likelihood estimates, albeit within the exponential correlation model rather than the wider Matérn class. In our experience, data sets of this size give little information about κ, a point that was mentioned indirectly by Professor Stein, and by Rue, Martino, Mondal and Chopin. Professor Besag also criticizes our crude imputation of the two outliers in the 2000 data and states that this 'could lead to serious consequences in the fitted variogram'. Indeed it could, but it does not. Figs 16(a) and 16(b) respectively show empirical and fitted (by maximum likelihood) variograms for four different imputations: the imputation that was used in the paper; both values imputed as the minimum of the other 61 values; both values imputed as the maximum; one as the minimum and one as the maximum.

Other modelling approaches

Professor Besag's more fundamental point, which was also raised in different ways by Fuentes and by Rue, Martino, Mondal and Chopin, is whether a stationary Gaussian process with Matérn correlation structure is appropriate at all. Within the Gaussian process framework, the most important practical difference between stationary and intrinsic models is that, when data are irregularly distributed over the spatial region of interest, the predictions from stationary models are mean reverting in sparsely sampled areas, whereas those from intrinsic models are not. Which of these properties is preferable could be context dependent. On balance, we agree that, in the absence of explanatory variables, allowing a global mean to have a major influence on local predictions is a questionable strategy. When explanatory variables are available, we are less convinced.

Professor Myers questions whether any form of Gaussian process is an appropriate model for geostatistical data. Our counter to this is that we think of the model-based approach as a device through which we determine a principled solution to a prediction problem. The Gaussian assumption leads naturally to linear prediction, whereas, in the classical geostatistical approach that is known as ordinary kriging, linearity is imposed as an *a priori* constraint. In similar vein, a model-based way to answer Professor Scott's

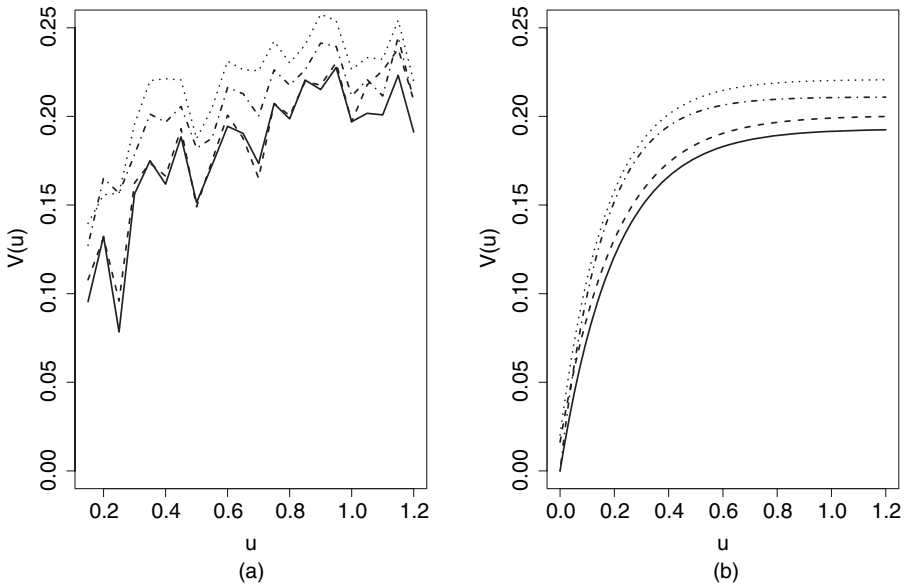

Fig. 16. (a) Empirical and (b) fitted (by maximum likelihood, assuming non-preferential sampling) variograms for the Galicia 1997 data: the different curve styles correspond to four different imputations to replace the two outliers (see the text for details)

request for a 'measure of representativeness is to derive the score statistic for the relevant parameter. For our model, if we assume that all parameters other than β are known, the score statistic to test $\beta = 0$ is

$$T = n^{-1} \sum_{i=1}^{n} Y_i - |A|^{-1} \int \hat{S}(x)\, \mathrm{d}x.$$

This has an obvious interpretation beyond the model assumed. It does not provide a useful test for preferential sampling because, as shown explicitly by Professor Anderson, when $\beta \neq 0$ fitting under the incorrect assumption that $\beta = 0$ leads to biased estimation of the mean. However, it provides a more tangible measure of non-representativeness than does the product $\hat{\beta}\hat{\sigma}$.

Professor Myers's comment about non-point support is well taken, but we reply that, if the spatially continuous process $S(\cdot)$ is Gaussian, then so is any spatially averaged process of the form $T(x) = \int w(u) S(x-u)\, \mathrm{d}u$. Whether one should then derive the correlation structure of $T(\cdot)$ from an assumed correlation structure for $S(\cdot)$ and a specified weighting function $w(\cdot)$ or, more pragmatically, specify the assumed correlation structure of $T(\cdot)$ directly is open to debate in view of the empirical status of most geostatistical models.

Somewhat analogous to the non-point support problem is the aggregated data problem that was raised by Professor Gelfand and Professor Chakraborty, in which observations Y_i relate to each of n subregions A_i that together form a partition of the study region. Most applications involving data in this form treat the study region as discrete and use Markov random-field models for the vector $Y = (Y_1, \ldots, Y_n)$ (Rue and Held, 2005). An alternative is to use a spatially continuous latent process $S(\cdot)$; for example, in an epidemiological setting where the Y_i are disease counts a starting point might be to assume that, conditional on $S(\cdot)$, the Y_i are independent Poisson variates with conditional means $\mu_i = N_i \int_{A_i} S(x)\, \mathrm{d}x$, where N_i denotes the number of people at risk in the subregion A_i. Note, however, that both approaches are susceptible to aggregation bias if spatial variation in population density within subregions is correlated with spatial variation in disease risk (Diggle and Elliott, 1995).

Professor Gelfand and Professor Chakraborty comment that different conditional modelling options lead, in practice if not in theory, to different overall models. To their two options, our approach adds a third, namely conditional independence of responses and locations given an underlying latent process. Professor Fuentes's 'shared parameter' models have essentially the same rationale as our equation (12), and we are happy to acknowledge her priority; incidentally, our co-ordinates are in units of 10^5 m, rather than degrees of latitude and longitude, so her comment about artificial anisotropy does not apply.

As Rue, Martino, Mondal and Chopin mention, similar ideas have been proposed in the context of longitudinal data analysis. However, we see a distinction in practice between informative missingness as discussed by Diggle and Kenward (1994), among many others, and informative observation as considered in the work that we cite in Section 2.

Professor Besag's suggestion of tessellation-based models is related to proposals in Heikkinen and Arjas (1998, 1999). It would be very interesing to develop this suggestion in the geostatistical setting.

References in the discussion

Assunção, R. M. and Guttorp, P. (1999) Robustness for point processes. *Ann. Inst. Statist. Math.*, **51**, 657–678.

Bates, D. V. (2005) Ambient ozone and mortality. *Epidemiology*, **16**, 427–429.

Besag, J. and Mondal, D. (2005) First-order intrinsic autoregressions and the de Wijs process. *Biometrika*, **92**, 909–920.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. R. Statist. Soc.* B, **68**, 333–382.

Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2008) A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics*, **64**, 96–105

Chambers, R. L. and Skinner, C. J. (2003) *Analysis of Survey Data*. New York: Wiley.

Cohen, M. A., Adar, S. D., Allen, R. W., Avol, E., Curl, C. L., Gould, T., Hardie, D., Ho, A., Kinney, P., Larson, T. V., Sampson, P. D., Sheppard, L., Stukovsky, K. D., Swan, S. S., Liu, L.-J. S. and Kaufman, J. D. (2009) Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Environ. Sci. Technol.*, **43**, 4687–4693.

Cressie, N. and Hawkins, D. M. (1980) Robust estimation of the variogram. *J. Math. Geol.*, **12**, 115–125.

Cui, H., Stein, A. and Myers, D. E. (1995) Extension of spatial information, Bayesian kriging and updating of prior variogram parameters. *Environmetrics*, **6**, 373–384.

Diggle, P. J. and Elliott, P. (1995) Disease risk near point sources: statistical issues for analyses using individual or spatially aggregated data. *J. Epidem. Commty Hlth*, **49**, 520–527.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.*, **43**, 49–93.

Diggle, P. J. and Ribeiro, P. J. (2007) *Model-based Geostatistics*. New York: Springer.

Fearnhead, P., Papaspiliopoulos, O. and Roberts, G. O. (2008) Particle filters for partially observed diffusions. *J. R. Statist. Soc.* B, **70**, 755–777.

Fuentes, M., Reich, B. and Lee, G. (2008) Spatial-temporal mesoscale modelling of rainfall intensity using gage and radar data. *Ann. Appl. Statist.*, **4**, 1148–1169.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Am. Statist. Ass.*, **100**, 1021–1035.

Genton, M. (1998) Highly robust variogram estimation. *J. Math. Geol.*, **30**, 213–221.

Green, P. J. and Sibson, R. (1978) Computing Dirichlet tessellations in the plane. *Comput. J.*, **21**, 168–173.

Griffin, J. E. and Steel, M. F. J. (2006) Order-based dependent Dirichlet processes. *J. Am. Statist. Ass.*, **101**, 179–194.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. and Coull, B. A. (2009) Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, **10**, 258–274.

Guan, Y., Sherman, M. and Calvin, J. A. (2007) On asymptotic properties of the mark variogram estimator of a marked point process. *J. Statist. Planng Inf.*, **137**, 148–161.

Guttorp, P. (2006) Setting environmental standards: a statistician's perspective. *Environ. Geosci.*, **13**, 261–266.

Hansen, J. and Lebedeff, S. (1987) Global trends of measured surface air temperatures. *J. Geophys. Res.*, **92**, 13345–13372.

Heikkinen, J. and Arjas, E. (1998) Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scand. J. Statist.*, **25**, 435–450.

Heikkenen, J. and Arjas, E. (1999) Modeling a Poisson forest in variable elevations: a nonparametric Bayesian approach. *Biometrics*, **55**, 738–745.

Ho, L. and Stoyan, D. (2008) Modelling marked point patterns by intensity marked Cox processes. *Statist. Probab. Lett.*, **78**, 1194–1199.

Illian, J. B., Penttinen, A., Stoyan, H. and Stoyan, D. (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester: Wiley.

Jerrett, M., Burnett, R. T., Ma, R., Pope, C. A., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. E. and Thun, M. J. (2005) Spatial analysis of air pollution mortality in Los Angeles. *Epidemiology*, **16**, 727–736.

Kanaroglou, P. S., Jerrett, M., Morrison, J., Beckerman, B., Arain, M. A., Gilbert, N. L. and Brook, J. R. (2005) Establishing an air pollution monitoring network for intraurban population exposure assessment: a location-allocation approach. *Atmos. Environ.*, **39**, 2399–2409.

Kim, S. Y., Sheppard, L. and Kim, H. (2009) Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology*, **20**, 442–450.
Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. and Clark, L. (2000) A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statist. Med.*, **19**, 1303–1318.
Lindgren, F. and Rue, H. (2007) Explicit construction of GMRF approximations to generalised Matern fields on irregular grids. *Technical Report 12*. Centre for Mathematical Sciences, Mathematical Statistics, Lund Institute of Technology, Lund University, Lund.
Loperfido, N. and Guttorp, P. (2008) Network bias in air quality monitoring design. *Environmetrics*, **19**, 661–671.
Martin, R. D. and Yohai, V. (1986) Influence functionals for time series. *Ann. Statist.*, **14**, 781–818.
McCullagh, P. (2008) Sampling bias and logistic models (with discussion). *J. R. Statist. Soc.* B, **70**, 643–677.
McCullagh, P. and Clifford, D. (2006) Evidence of conformal invariance for crop yields. *Proc. R. Soc. Lond.* A, **462**, 2119–2143.
Menezes, R., Garcia-Soidán, P. and Febrero-Bande, M. (2007) A kernel variogram estimator for clustered data. *Scand. J. Statist.*, **35**, 18–37.
Møller, J. and Waagepetersen, R. P. (2004) *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton: Chapman and Hall–CRC.
Myllymäki, M. and Penttinen, A. (2009) Conditionally heteroscedastic intensity-dependent marking of log Gaussian Cox processes. *Statist. Neerland.*, **63**, 450–473.
Reich, B. J. and Fuentes, M. (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Statist.*, **1**, 249–264.
Reynolds, J. H., Das, B., Sampson, P. D. and Guttorp, P. (1998) Meteorological adjustment of Western Washington and Northwest Oregon surface ozone observations with investigation of trends. *Technical Report 15.* Northwest Research Center for Statistics and the Environment, University of Washington, Seattle. (Available from `http://www.nrcse.washington.edu/pdf/trs15_doe.pdf`.)
Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall.
Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *J. R. Statist. Soc.* B, **71**, 319–392.
Schlather, M. (2001) On the second-order characteristics of marked point patterns. *Bernoulli*, **7**, 99–117.
Skinner, C. J. (1989) Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (eds C. J. Skinner, D. Holt and T. M. F. Smith), pp. 59–87. New York: Wiley.
Song, X., Davidian, M. and Tsiatis, A. A. (2002) A semiparametric likelihood approach to joint modeling of longitudinal and time to event data. *Biometrics*, **58**, 742–753.
Stein, M. (2005) Space-time covariance functions. *J. Am. Statist. Ass.*, **100**, 310–321.
Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D. and Kaufmann, J. D. (2009a) Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*, to be published, doi 10.1002/env.1014.
Szpiro, A. A., Sheppard, L. and Lumley, T. (2009b) Efficient measurement error correction with spatially misaligned data. *Working Paper 350*. Department of Biostatistics, University of Washington, Seattle.
Tsonaka, R., Verbeke, G. and Lesaffre, E. (2009) A semi-parametric shared parameter model to handle non-monotone nonignorable missingness. *Biometrics*, **65**, 81–87.
Veneziano, D. and Kitanidis, P. K. (1982) Sequential sampling to contour an uncertain function. *Math. Geol.*, **14**, 387–404.
Vinnikov, K. Y., Groisman, P. Y. and Lugina, K. M. (1990) Empirical data on contemporary global climate changes (temperature and precipitation). *J. Clim.*, **3**, 662–677.
Wakefield, J. and Shaddick, G. (2006) Health-exposure modelling and the ecological fallacy. *Biostatistics*, **7**, 438–455.
Wälder, O. and Stoyan, D. (1996) On variograms in point process statistics. *Biometr. J.*, **8**, 895–905.
Wood, S. N., Bravington, M. V. and Hedley, S. L. (2008) Soap film smoothing. *J. R. Statist. Soc.* B, **70**, 931–955.
Xia, G., Miranda, M. L. and Gelfand, A. E. (2006) Approximately optimal spatial design approaches for environmental health data. *Environmetrics*, **17**, 363–385.
Zhang, H. (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Statist. Ass.*, **99**, 250–261.