Geostatistical Analysis Under Preferential Sampling

Peter J Diggle
(Lancaster University and Johns Hopkins University School of Public Health),
Raquel Menezes Da Mota Leite
(University of Minho)
and
Ting-li Su
(Lancaster University)

October 12, 2007

Abstract

Geostatistics involves the fitting of spatially continuous models to spatially discrete data (Chilès and Delfiner, 1999). Preferential sampling arises when the process which 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, geostatistical methods are often used in situations where sampling is likely to be preferential: for example, in a pollution monitoring network, monitors will typically be placed close to likely sources of pollution. We give a general expression for the likelihood function of preferentially sampled geostatistical data and discuss how this can be evaluated approximately using Monte Carlo methods. We present an idealised model for preferential sampling, and show through simulations how preferential sampling invalidates conventional geostatistical methods of inference. We discuss practical strategies for dealing with preferential sampling and describe two applications. The first application is to data from an air pollution monitoring network in California, USA, where the objective is to construct estimates of spatially averaged pollution levels. The second application is to a set of bio-monotoring data from Galicia, northern Spain, where the objective is to assess the impact of industry on the region-wide pollution surface.

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

1 Introduction

The term *geostatistics* describes the branch of spatial statistics in which data are obtained by sampling a spatially continuous phenomenon $S(x)$, where x denotes location, at a discrete set of locations $x_i : i = 1, ..., n$ in a spatial region of interest, A. Typically, and throughout this paper, $A \subset \mathbb{R}^2$. In many cases, $S(x)$ cannot be measured without error. In classical geostatistics, measurement errors are 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 would take the form

$$Y_i = S(x_i) + Z_i : i = 1, ..., n \tag{1}$$

where the Z_i are mutually independent, zero-mean random variables. The objectives of a geostatistical analysis typically focus on prediction of properties of the realisation of $S(x)$ throughout the region of interest A. Targets for prediction might include, according to context: the value of $S(x)$ at an unsampled location; the spatial average of $S(x)$ over A or sub-sets thereof; the minimum or maximum value of $S(x)$; or sub-regions in which $S(x)$ exceeds a particular threshold. Chilès and Delfiner (1999) give a comprehensive account of classical geostatistical models and methods.

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

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

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

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

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

Our aims in this paper are to demonstrate the problems that can arise if preferential sampling is ignored in the analysis of geostatistical data, to suggest ways of adjusting standard methods of analysis to alleviate these problems and to apply the adjusted methods to two environmental monitoring data-sets. Preferential sampling is a common feature of environmental monitoring networks, in which context there is a natural inclination to place monitors in areas which are thought to be at high risk for pollution.

In Section 2 we introduce our two applications. The first concerns a routine monitoring network in which monitoring locations are generally concentrated in areas of high population density. The second concerns a two-stage biomonitoring study, in which the region of interest was first sampled preferentially, with sample locations somewhat concentrated around sources of industrial pollution, and later non-preferentially with a lattice design. Section 3 presents an idealised model for preferential sampling, and uses this model to demonstrate how geostatistical analyses which ignore preferential sampling can be misleading. Section 4 discusses likelihood-based inference using Monte Carlo methods. Section 5 considers practical analysis strategies for dealing with the preferential sampling problem. Section 6 describes an analysis of the data from each of our two examples. The paper ends with a short discussion.

2 Motivating examples: exploratory analysis

Exploratory analysis of the data can help to reveal the nature and extent of any preferential sampling. When the status of the sampling method is in doubt, it may also be worthwhile to conduct a formal test for the existence of preferential sampling.

2.1 A test for preferential sampling

Schlather, Ribeiro and Diggle (2004, henceforth SRD) develop two tests for preferential sampling, which operate by treating a set of geostatistical data as a marked point process. Their *random field model*, which is equivalent to our notion of non-preferential sampling, is that the sample locations X are a realisation of a point process $\mathcal{P}$ on A, that the mark of a point at locations x is the value at x of the realisation of a random field S on A, and that $\mathcal{P}$ and S are independent processes.

The more powerful of the two tests considered in SRD, as indicated by the results of a simulation study, used an empirical counterpart of the function $E(h) = \mathrm{E}[S(o)|o, x \in \mathcal{P}$ where h denotes the distance between an aribtrary origin, o, and the point x. Under the random field hypothesis, the conditioning on o and x belonging to $\mathcal{P}$ does not affect any property of the random field S, hence $E(h)$ is constant. The intutive interpretation of a non-constant $E(h)$ is that if there is a positive association between S and the conditional intensity of $\mathcal{P}$ given S, then pairs of points of $\mathcal{P}$ at small separation distances h will, on average, be associated with larger than average values of S and conversely if the association is negative, whereas the values of S at widely separated locations will typically be independent, irrespective of any assocation between S and $\mathcal{P}$, and $E(h)$ will therefore approach the unconditional mean of S as h becomes large.

In practice, the test must be applied to the noisey measurement data Y, rather than to S; this does not affect the validity of the test, but will reduce its power because one effect of the measurement component in (1) is to dilute any association beteen $\mathcal{P}$ and S. To implement a formal test, SRD suggest fitting a stationary Gaussian process to Y, after transformation if necessary, and comparing the empirical function $\hat{E}(h)$ for the (transformed) data with versions of $\hat{E}(h)$ obtained using independent simulations from the fitted random field model, holding the data-locations fixed.

2.2 Air pollution monitoring in California

Our first example concerns data giving the locations of 84 PM2.5 monitors operational during the year 2000 in the state of California, USA, and the corresponding year-long average measured values of the pollutant at each monitor.

One simple question posed by these data is how we should estimate the spatial average of PM2.5 over a region of interest, for example a zip-code or county. One currently used estimator is a simple arithmetic average of values from all monitors that lie within the region of interest (CHECK FUENTES ET AL PAPER). Both this and, to a lesser extent, more sophisticated estimators based on classical geostatistical methods, for example kriging (Chilès and Delfiner, 1999, Chapter 3), are potentially misleading if the sampling is preferential. The rationale for the locations of the monitors is unknown but appears to be related to population density. Also, we shall show in Section 7.1 that the local density of monitors is related to a number of socio-economic descriptors at zip-code level.

Figure 1 shows the locations of the 84 monitors, superimposed on point predictions of average concentrations obtained under the assumption that sampling is non-preferential. The underlying geostatistical model for the spatial prediction was a stationary Gaussian process for square-root transformed PM2.5 values, with a linear model for the mean, and Matérn spatial correlation function; see Section 7.1 for details.

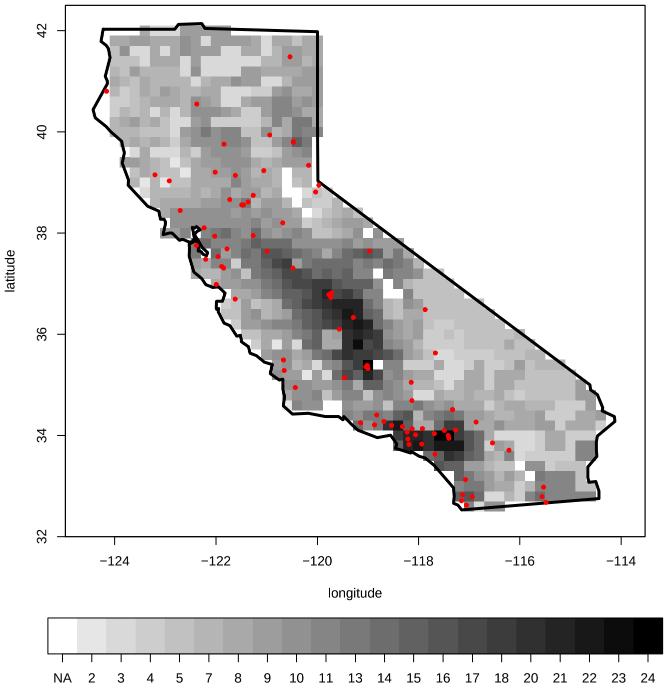

Figure 1: Locations of 84 $PM2.5$ monitors in the state of California, USA, operational in the year 2000. Grey-scale image shows point predictions of average 2000 concentrations throughout California (see text for details).

Figure 2 shows the empirical function $\hat{E}(h)$ as defined in Schlather, Ribeiro and Diggle (2004). For a formal test, we fit a Gaussian random field model to square-root transformed PM2.5 measurements and use the test statistic

$$T = \sum_i \sqrt{\frac{1}{\sum_{j=0}^{i-1} Ebin(j)}} |E(i) - E(0)|$$

. NOTE THAT $E(0)$ IS THE BIN THAT CONTAINS ZERO, AND $Ebin$ IS THE NUMBER OF POINTS WITHIN THE BIN. The visual impression of a decreasing trend in $\hat{E}(h)$ withy increasing h is confirmed by rejection of the random field hypothesis at the conventional 5% level, reinforcing the visual impression from Figure (13) that monitor intensity is positively associated with PM2.5 levels.

2.3 Heavy-metal bio-monitoring in Galicia

Our second example concerns bio-monitoring of heavy metal pollution in Galicia, northern Spain. The data consist of two spatial surveys of heavy metal concentrations in moss samples, taken in 1997 and 2000. In the first survey, sampling was deliberately concentrated in or near areas of industrial activity which are known to be sources of heavy metal pollution and

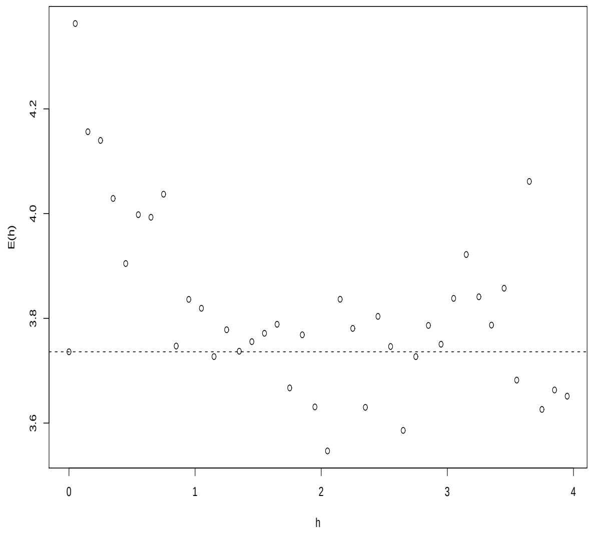

Figure 2: Test for independence between PM2.5 monitor locations and pollutant levels. The empirical function $\hat{E}(h)$ was calculated after applying a square-root transformation to PM2.5 levels; see text for details.

is therefore preferential, whereas the second survey used a regular lattice design which is therefore non-preferential. For further details, see Fernández, Rey and Carballeira (2000) and Aboal, Real, Fernández and Carballeira (2005). One objective of analysing these data is to estimate, and compare, maps of heavy metal concentrations in 1997 and 2000. No major sources of pollution were introduced between 1997 and 2000. Our working assumption is therefore that, whilst overall pollution levels may have changed during the five years between the two surveys, a common model for the underlying pollution field $S(x)$ is justifiable. A classical geostatistical analysis seems appropriate for the 2000 data, but less so for the 1997 data because of the preferential nature of the sampling.

Figure 3 shows the sampling locations for the two surveys, together with the locations of known major sources of heavy-metal pollution.

Figure 3: Sampling locations for 1997 (open circles) and 2000 (closed circles) surveys of heavy-metal pollution in Galicia. Triangles denote locations of known sources of pollution.

Levels of several different heavy metals were recorded in the study. We focus here on the concentrations of lead in each of 1997 and 2000. The data included two gross outliers in 2000, each of which we replaced by the average of the remaining values for the year 2000.

Table 1 gives summary statistics for the 1997 and 2000 data. Note that the mean response is

Table 1: Summary statistics for lead pollution levels measured in 1997 and 2000.

	untransformed		log-transformed	
	1997	2000	1997	2000
Number of locations	63	132	63	132
Mean	4.72	2.05	1.44	0.64
Standard deviation	2.21	0.91	0.48	39
Minimum	1.67	0.80	0.52	2.25
Maximum	9.51	6.00	-0.22	1.79

Figure 4: Empirical distributions of log-transformed lead concentrations in the 1997 and 2000 samples.

higher for the 1997 data than for the 2000 data, consistent with the former being preferentially sampled near potential pollutant sources. Also, the log-transformation eliminates an apparent variance-mean relationship in the data and leads to more symmetric distributions of measured values. Figure 4 shows the empirical distributions of log-transformed lead concentrations in each of the two years, again showing the shift between 1997 and 2000, consistent with the preferential character of the 1997 sample.

Figure 5 shows scatterplots of the 1997 and 2000 log-tranformed measurements against distance from the nearest pollution source. These suggest a weak, negative association between lead concentrations and distance.

3 A model for preferential sampling

Our interest is in enabling valid inferences about the unobserved spatial process S when the placement of sampling locations is potentially preferential. As is the case in both of our motivating examples, this situation often arises because those conducting the survey exercise a degree of subjective judgement in choosing the sampling locations. This makes it difficult to justify formal modelling of the joint spatial distribution of locations and measured values at those locations. Nevertheless, we shall propose and investigate an idealised model and use this to demonstrate some of the ways in which preferential sampling affects the inferences which can be made from the data.

Figure 5: Scatterplot of log-transformed lead concentrations against distance to nearest pollution source for 1997 (open circles) and 2000 (solid dots) data

3.1 A common latent process model

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

To clarify the distinction between preferential and non-preferential sampling, and the inferential consequences of the latter, we first examine a situation considered by Rathbun (1996), in which S determines the conditional intensity of an inhomogeneous Poisson process X, whilst measurements Y are taken at a pre-specified set of locations, i.e. independently of X. Then, the joint distribution of S, X and Y takes the form

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

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

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

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

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

3.2 A simple parametric model

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

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

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

$$\lambda(x) = \exp\{\alpha + \beta S(x)\}. \tag{4}$$

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

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

In what follows, we shall use this special form of the common latent process model to investigate the impact of preferential sampling on conventional geostatistical methods of analysis.

4 Impact of preferential sampling on geostatistical inference

We have conducted a simulation experiment in which we simulated data on A the unit square from an underlying stationary Gaussian process which we then sampled, with additive Gaussian measurement error, either non-preferentially or preferentially according to each of the

Figure 6: Sample locations and underlying realisations of the signal process for the model used in the simulation study. The left-hand panel shows the completely random sample, the centre-panel the preferential sample and the right-hand panel the clustered sample. In each case, the grey-scale image represents the realisation of the signal process, $S(x)$, which was used to generate the associated measurement data. The model parameter values are $\mu = 4$, $\sigma^2 = 1.5$, $\phi = 0.15$, $\kappa = 1$, $\tau^2 = 0.25$, $\beta = 2$

following sampling designs. For the *completely random* sampling design, sample locations x_i are an independent random sample from the uniform distribution on A. For the *preferential* design, the x_i are generated by the model described in Section 3.2, with parameter $\beta = 2$. For the *clustered* design, we used the same model, but used one realisation of S to generate the data Y and a second, independent realisation of S to generate X, thereby giving a non-preferential design with the same marginal properties as the preferential design.

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

Figure 6 shows a realisation of each of the three sampling designs superimposed on a single realisation of the process S. The preferential nature of the sampling in the central panel of Figure 6 is clear.

4.1 Variogram estimation

The theoretical variogram of a stationary spatial process $Y(x)$ is the function $V(u) = \mathrm{Var}\{Y(x) - Y(x')\}$ 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. In this section, we illustrate the impact of preferential sampling on non-parametric variogram estimation.

Consider a set of data $(x_i, y_i) : i = 1, ..., n$, where x_i denotes a location and y_i a corresponding measured value. The *empirical variogram ordinates* are the quantities $v_{ij} = (y_i - y_j)^2/2$. Each v_{ij} is an unbiased estimator for $V(u_{ij})$, where u_{ij} is the distance between x_i and x_j. A scatterplot of v_{ij} against u_{ij} or, more usefully, a smoothed version of this scatterplot, can be used to suggest appropriate parametric models for the spatial covariance structure of the data; for more information on variogram estimation, see for example Cressie (1985; 1991, Chapter 2), Chilès and Delfiner (1999) or Diggle and Ribeiro (2007, Chapter 5).

The two panels of Figure 7 show simulation-based estimates of the point-wise bias and standard deviation of smoothed empirical variograms, derived from 500 replicate simulations. With

Figure 7: Bias and standard deviation of the sample variogram under random, preferential and clustered sampling. See text for detailed description of the simulation model.

regard to bias, the results under both uniform and clustered non-preferential sampling designs are consistent with the unbiasedness of the empirical variogram ordinates; although smoothing the empirical variogram ordinates does induce some bias, this effect is negligible in the current setting. In contrast, under preferential sampling the results show severe bias. With regard to efficiency, the right-hand panel of Figure 7 illustrates that clustered sampling designs, whether preferential or not, are also less efficient than uniform sampling. The bias induced by preferential sampling is qualitatively unsurprising. The implicit estimand of the empirical variogram is the variance of $Y(x) - Y(x')$ conditional on both x and x' belonging to X, which in general will differ from the unconditional variance; see, for example, Walder and Stoyan (1996) or Schlather (2001).

4.2 Spatial prediction

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

Table 2 shows 95% coverage intervals for the resulting biases and mean square prediction errors of the ordinary kriging predictor $\hat{S}(x_0)$, where $x_0 = (0.5, 0.5)$, in each case evaluated empirically over 500 replicate simulations.

The bias is large and positive under preferential sampling, because the sampling model leads to a higher density of sample locations close to high values of the underlying process S. The other two sampling designs both lead to approximately unbiased prediction, as predicted by theory. The substantially larger mean square error for clustered by comparison with completely random sampling reflects the inefficiency of the latter, as previously seen in the context of variogram estimation.

Table 2: Impact of sampling design on the bias and mean square error of the ordinary kriging predictor $\hat{S}(x_0)$, when $x_0 = (0.5, 0.5)$ and each sample consists of 100 locations on the unit square. Each entry in the table is a 95% coverage interval calculated empirically from 500 independent simulations. See text for detailed description of the simulation model.

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

5 Likelihood-based analysis for the common latent process model

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

$$L(\theta) = [X, Y] = \mathrm{E}_S\left[[X|S][Y|X, S]\right], \tag{5}$$

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

To evaluate $L(\theta)$ approximately, a naive strategy would be to replace the intractable expectation on the right hand side of (5) by a sample average over simulations of S. This would give the crude Monte Carlo approximation

$$L_{MC}(\theta) = m^{-1}\sum_{j=1}^{m}[X|S_j][Y|S_j, X], \tag{6}$$

where the S_j are independent realisations of S. To reduce the Monte Carlo variance, we could use anti-thetic pairs of realisations, hence for each of $j = 1, ..., m/2$ we set $S_{2j} = 2\mu - S_{2j-1}$.

Both the crude Monte Carlo approximation and its anti-thetic variant fail when Y is measured without error, because in this case the term $[Y|S_j, X]$ in (6) will be zero with probability one. For essentially the same reason, the method fails in practice when the measurement error is small relative to the variance of S, yet this is the situation in which preferential sampling potentially has most impact on the analysis. We therefore modify (6) by introducing an importance sampler as follows.

Firstly, write the exact likelihood (5) as the integral

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

Now, write $[S] = [S_0][S_1|S_0]$ and replace the term $[S|Y]$ in the denominator of (5) 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, (7) becomes

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

and a Monte Carlo approximation is

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

where now the S_j are simulations of S conditional on Y. Note in particular that when Y is measured without error, $[Y|S_{0j}]/[S_{0j}|Y] = 1$, and that Y and S_0 have the same unconditional expectations.

To simulate a realisation from $[S|Y]$, we first simulate a realisation s from the unconditional distribution of S, using the circulant embedding algorithm of Wood and Chan (1994), and a realisation z consisting of an independent random sample from $\mathrm{N}(0, \tau^2)$.

Now, let A denote the n by N matrix in which each row contains $N - 1$ zeros and a single element 1 that identifies the position of each data-location x_i amongst the N lattice-points of S. Also, let Σ_S denote the N by N variance matrix of S, and write $R = A\Sigma_S A' + \tau^2 I$. Note that our required $[S|Y]$ is multivariate Gaussian, with mean

$$\Sigma_S A' R^{-1}(Y - \mu_Y) \tag{10}$$

and variance matrix

$$\Sigma_S - \Sigma_S A' R^{-1} A \Sigma_S \tag{11}$$

It follows that

$$S_j = s + \Sigma_S A' R^{-1}(Y + z - As - \mu) \tag{12}$$

is a realisation from $[S|Y]$. Finally, we again use an anti-thetic variant of (9) as a variance reduction device.

ILLUSTRATIVE CALCULATION WITH SIMULATED DATA - ANTICIPATE FLAT LIKELIHOOD PROBLEMS

MAYBE ALSO SIMULATION SHOWING THAT POISSON MODEL IS ROBUST TO SMALL-SCALE REGULARITY IN PREFERENTIAL SAMPLING DESIGN.

6 Practical strategies to accommodate preferential sampling

The simulated example at the end of Section 5 indicates that likelihood-based inference for preferentially sampled data is liable to suffer from difficulties caused by poor identifiablity of model parameters. In some contexts, we may be willing to circumvent this difficulty problem by adopting a Bayesian approach with informative priors. In others, we may be willing to specify a realistic range for the degree of preferentiality and to treat β in the common latent process model, equation (4), as a sensitivity parameter. In this Section we consider two different strategies that do not impose any *a priori* constraints on the degree of preferentiality.

6.1 Shared covariate information

One possible strategy is to seek explanatory variables which eliminate, or at least reduce, the adverse effects of preferential sampling. Suppose, for the sake of illustration, that S is observed without error and that an unconditional dependence between X and S arises through their shared dependence on a latent variable, U, and that the joint distribution of X and S is of the form

$$[X, S] = \int [X|U][S|U][U]dU, \tag{13}$$

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

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, i.e. that sampling in no longer preferential.

For the California monitoring data, a reasonable hypothesis is that monitor placement is related to local demographic and socio-economic conditions, in which case the US census provides a number of candidate covariates. For the Galicia bio-monitoring data, the preferential sampling in 1997 arose because sampling locations were concentrated around industrial locations. Including some function of distance to the nearest industry as a covariate at each sampling location should therefore reduce the stochastic dependence between X and S.

This strategy is unlikely completely to eliminate dependence between S and X, but it may well reduce it to the point where its effects are innocuous. Note also that the Schlather, Ribeiro and Diggle (2004) test can be applied to residuals from a regression model and so provides a diagnostic check on the extent of any residual dependence.

6.2 Two-stage sampling

A second possibile strategy is to use a two-stage sample. The Galicia data provide an illustration. The 2000 data are non-preferentially sampled. If we were prepared to fit a single model, or at least a model with common covariance structure, to the underlying Gaussian process S in 1997 and in 2000, the 2000 data would give information about the parameters of S uncompromised by the effects of preferential sampling. Combining the information from the 1997 and 2000 data should therefore alleviate the identifiability problems which would arise if the 1997 data were analysed separately.

7 Motivating examples re-visited

7.1 PM2.5 monitoring in California

Recall from Section 2.1 that the Calfornia data show significant evidence of preferential sampling. Our analysis strategy for these data is to look for shared covariate information as described in Section 6.1 above and to ascertain whether adjustment for any such covariates approximately eliminates the effects of preferential sampling, thereby leading to more trustworthy predictive inference for spatially averaged pollution levels.

The US Census 2000 data include a number of socio-economic and demographic variables recorded at zip-code level. In the analysis reported here, we define the study-region A to be mainland California, with a total census population count of 33,867,596 distributed amongst 1709 zip-codes. The areas of these 1,709 zip-codes vary by orders of magnitude, from 0.01255 km^2 to 19,870 km^2 because of the similarly wide variation in population density, between 0 and 20,100 people per km^2 (26 zip-code areas, accounting for for 1.4% of the total area of mainland California, have no inhabitants). In what follows, we treat each census variable as a piece-wise constant surface at zip-code level, so as to define a value for every location in the study region. In this way, we construct the following potential explanatory variables:

Popdensity	log of population per km^2
Black	percentage of population who are black or African American
Hispanic	percentage of population who are Hispanic or Latino
College	percentage of population AGE 25 and over educated at least to college level
Income	median family income

We write $U(x)$ for the vector of covariate values at location x. Note that in the US census coding, “Hispanic or Latino” and “Black or African American” are not mutually exclusive.

To investigate covariate effects, we modify the model described in Section 3.2 as follows. Firstly, in A1 we replace the constant mean μ by a regression function, $U(x)'\gamma$. Secondly, in place of A2 we assume that the monitor locations $X = \{x_i : i = 1, ..., n\}$ form a partial realisation on A of an inhomogeneous Poisson process with intensity surface $\lambda(x)$, where

$$\log \lambda(x; \theta) = U(x)'\theta, \tag{14}$$

as in Cox (1972).

The likelihood for the data X and Y now consists of two independent terms, $[X|U][Y|U]$. The term corresponding to $[Y|U]$ takes the standard form of a linear Gaussian model with covariates (Diggle and Ribeiro, 2007, Section 5.4.2). The log-likelihood corresponding to $[X|U]$ is

$$L(\theta; X) = \sum_{i=1}^{i=n} \log \lambda(x_i; \theta) - \int_{x \in A} \lambda_\theta(x) dx. \tag{15}$$

To fit the model, we estimate its parameters by maximum likelihood, and use generalised likelihood ratio tests in conjunction with a forward search to select covariates for inclusion in the model. Note that the integral in (15) reduces to a finite summation because each element of $U(x)$ is piece-wise constant over A.

7.1.1 The measurement sub-model

The measurement sub-model is a standard linear Gaussian model applied to appropriately transformed PM2.5 levels. The model-fitting proceeds in three stages: selection of a transformation; preliminary identification the covariance structure of the transformed data; likelihood-based inference (see, for example, Diggle and Ribeiro, 2007).

For the first stage, we examined the profile log-likelihood for the transformation parameter λ in the Box-Cox family (Box and Cox, 1964), under the assumption that all five candidate explanatory variables are included into the trend surface and a Matérn family for the residual correlation is used. The profile log-likelihood is maximised at $\lambda \approx 0.5$, suggesting a square-root transformation (Figure 8, upper row). We therefore fix $\lambda = 0.5$, use the model specification above, and examine the marginal distribution of the residuals. They appears to be approximately Gaussian (Figure 8, bottom row).

For the second stage, we examined the smoothed empirical variogram of the residuals (Figure 9); this confirms the presence of residual spatial correlation, but does not clearly identify its parametric form.

We therefore proceed to formal likelihood-based inference to fit a linear model for the mean response in conjunction with a Matérn correlation structure for the residual signal $S(x)$. For the Matérn shape parameter κ, we considered the values 0.5, 1.5 and 2.5, obtaining maximised log-likelhoods -231.1, -229.5, -229.2; we therefore proceed fixing $\kappa = 2.5$. We then used a forward selection in conjunction with a standard likelhood ratio criterion, D, to decide which explanatory variables to include in the model. This led to our including Popdensity ($D = 2(237.9 - 232.6) = 10.6$, $p = 0.0011$) and College ($D = 2(232.6 - 229.6) = 6.0$, $p = 0.0143$). Anticipating the results of the next sub-section we also included Income, although its effect was not significant ($D = 2(229.6 - 229.2) = 0.8$, $p = 0.3711$).

Table 3 gives the maximum likelihood parameter estimates for the fitted model. For comparison, we also show Bayesian point estimates and 95% credible intervals. The estimated Model (I) for PM2.5 pollutant data is disaplyed in Figure 1; and the fitted Models (I') and (II') are shown as middle and bottom plots in Figure 10.

We use the residuals to check the independency between monitor locations and pollutant level after adjusting for the covariates. Figure 11 shows the empirical $\hat{E}(h)$ function. We do not reject the random-field model hypothesis for both Model (I) (P-value=> 0.5) and Model (II) (P-value= 0.42-0.43). (Figure 12 shows the empirical $\hat{V}(h)$ function. We do not reject the random-field model hypothesis for both Model (I) (P-value=0.33-0.34) and Model (II) (P-value= 0.16-0.17).

NOTE THAT SE FROM MLE IS GREATER THAN 1/4 OF CI FROM BAYESIAN.

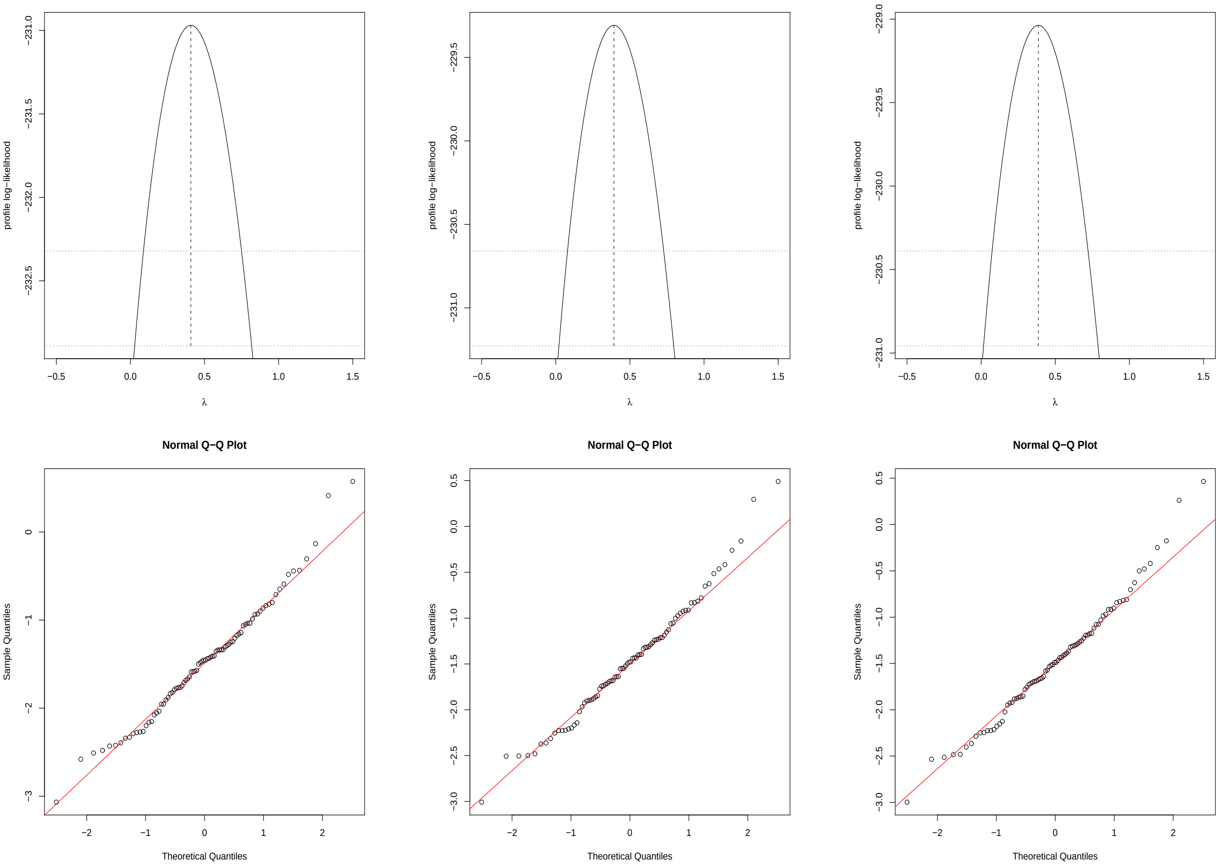

Figure 8: Selecting a transformation for the PM2.5 data: profile log-likelihood for the Box-Cox transformation parameter λ (top-row) while fixing κ at 0.5 (left), 1.5 (middle), 2.5 (right). They maximised at 0.41, 0.39, 0.39 (left to right). Plots on the bottom row are the Normal q-q plot of residuals while taking $\lambda = 0.5$.

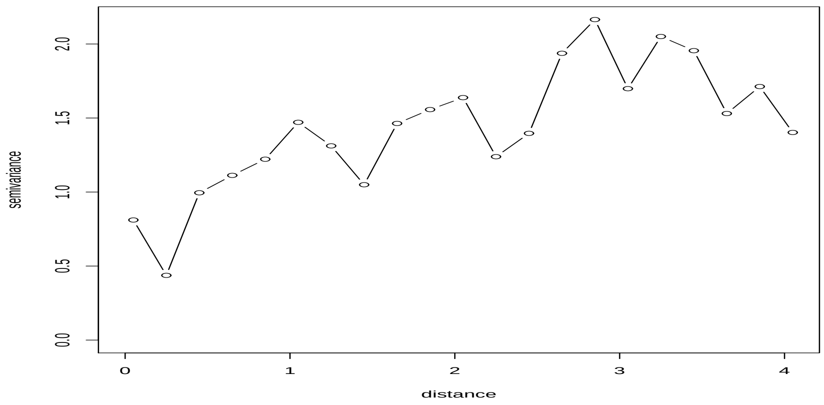

Figure 9: Smoothed empirical variogram of the residual for the PM2.5 data.

7.1.2 The location sub-model

START HERE

The forward selection procedure led to a model with three covariates: Popdensity, College, Income. Table 4 shows parameter estimated and standard errors for the fitted model.

Figure 13 shows the fitted intensity surfaces for the fitted model in Table 4. Left-hand panels are displayed on the untransformed scale, whereas the right-hand panels show log-transformed intensities. The white areas have zero population and have been excluded from the analysis. For each type of monitor, the two models give qualitatively similar fitted intensity surfaces, but the fitted intensity surface shows more spatial variation for O_3 than for PM2.5.

To assess the fit of the Poisson model, we use a standard diagnostic tool of spatial point pattern analysis, the K-function (Ripley ,1976, 1977) as extended to the non-stationary case by Baddelely, Møller and Waagepetersen (2000). The K-function measures, as a function of distance, the extent to which the locations show excess aggregation or regularity, relative to expectation under the fitted Poisson model. For an inhomogeneous Poisson process, $K(s) = \pi s^2$, where s denotes distance. Figure 14 displays estimates of $K(s) - \pi s^2$, together with 95% pointwise Monte Carlo tolerance limits under the fitted inhomogeneous Poisson process model, derived from 100 simulated realisations of the fitted model.

In Figure 14, we consider distances up to 75km, which is about one-quarter of the width of California. Estimates of $K(s)$ become increasingly imprecise as s increases. For the PM2.5 monitors, the K-function analysis suggests some residual spatial aggregation at distances up to about 4.1km, corresponding to groups of two or more monitors being placed close together more often than would be consistent with the fitted Poisson model. Note, however, that with only 84 PM2.5 monitors altogether, this effect could be explained by a few unusually close pairs.

	Maximum likelihood estimates		Bayesian estimates	
effects	Model I	Model II	Model I'	Model II'
Intercept	3.549	3.576	3.484	3.510
	(0.456)	(0.455)	(3.143, 3.824)	(3.153, 3.827)
Popdensity	0.264	0.274	0.273	0.287
	(0.073)	(0.073)	(0.223,0.326)	(0.234,0.338)
College	-1.657	-2.393	-1.729	-2.588
	(0.665)	(1.067)	(-2.212, -1.255)	(-3.366, -1.823)
Income	–	1.097*10^{-5}	–	$1.25 * 10^{-5}$
	–	(1.255*10^{-5})	–	$(3.40 * 10^{-6}, 2.12 * 10^{-5})$
τ^{2*}	0.367	0.373	0.425	0.450
	–	–	(0.325, 0.575)	(0.325, 0.625)
σ^2	1.207	1.189	1.209	1.205
	–	–	(0.976, 1.534)	(0.965, 1.564)
ϕ	0.257	0.259	0.310	0.310
	–	–	(0.259, 0.362)	(0.259, 0.362)

Table 3: Maximum likelihood estimates (with standard error) and Bayesian point estimates (with 95% cridible interval) for the fitted models of the PM2.5 pollutant data. τ^{2*} denotes a relative nugget (τ^2/σ^2).

Effect	Estimate (S.E)
Intercept	-11.698 (0.337)
Intercept*LA	-0.997 (0.375)*
Popdensity	0.829(0.054)*
College	2.571(0.985)*
Income	$-5.024 * 10^{-5}(1.033 * 10^{-5})$*

Table 4: Maximum likelihood estimates and standard errors for the PM2.5 monitor locations. All quantitiative explanatory variables except Popdensity are centered. Asterisks indicate that the corresponding Wald statistic is significant at the 5% level.

For the O_3 monitors, the estimated K-function falls well outside the simulation envelope over most of the plotted range of distance. This is caused by a pair of monitors situated close together in a zip-code with only 64 people. We labelled this pair of monitors with crosses in Figure **??**. If we treat one of the two monitors as an outlier and delete it from the analysis, the fitted model can be shown as model (IV) on Table **??**. Removing one of these two monitors does not greatly change the estimated covariate effects. However, the estimated K-function as shown on the bottom right plot in Figure 14 now falls within the simulation envelope, or nearly so, throughout the plotted range of distances.

Bearing in mind the implicit multiple testing in the pointwise comparisons between data and simulation envelopes, and the idealised nature of the Poisson model, the overall fit with the "outlier" removed seems acceptable.

7.2 Heavy-metal bio-monitoring in Galicia

Empirical evidence for preferential sampling?

How to exploit the two-stage sampling design, in which one of the two stages is potentially preferential, the other unambiguously non-preferential.

References

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

Baddeley A, Møller J. and Waagepetersen R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* **54**, 329-350.

Chilès, J-P and Delfiner, P. (1999). *Geostatistics.* New York : Wiley.

Cox, D.R.(1972). The statistical analysis of dependencies in point processes. In *Stochastic Point Processes*, ed P.A.W. Lewis, 55-66. New York : Wiley.

Cressie, N.A.C. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association of Mathematical Geology*, **17**, 563–86.

Curriero, F.C., Hohn, M.E., Liebhold, A.M. and Lele, S.R. (2002). A statistical evaluation of non-ergodic variogram estimators. *Environmental and Ecological Statistics*, **9**, 89–110.

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

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

Fernández, J.A., Rey, A. and Carballeira, A. (2000). An extended study of heavy metal deposition in Galicia (NW Spain) based on moss analysis. *Science of the Total Environment*,

254, 31–44.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of measurements and event time data. *Biostatistics*, **1**, 465–480.

Lin, H, Scharfstein, D.O. and Rosenheck, R.A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society*, B 66, 791–813.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (second edition). London : Chapman and Hall.

Møller, J., Syversveen, A. and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, **25**, 451–82.

Rathbun, S.L. (1996). Estimation of Poisson intensity using partially observed concomitant variables. *Biometrics*, **52**, 226–42.

Ripley, B.D. (1976). The second order analysis of stationary point processes. *Journal of Applied Probability* **13**, 255-266.

Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society* B **39**, 172–212.

Schlather, M., Ribeiro, P. J. and Diggle, P. J. (2004). Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society*, B **66**, 79–93.

Wood, A. T. A. and Chan, G. (1994). Simulation of stationary Gaussian processes in $[0,1]^d$. *Journal of Computational and Graphical Statistics*, **3**, 409–432.

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

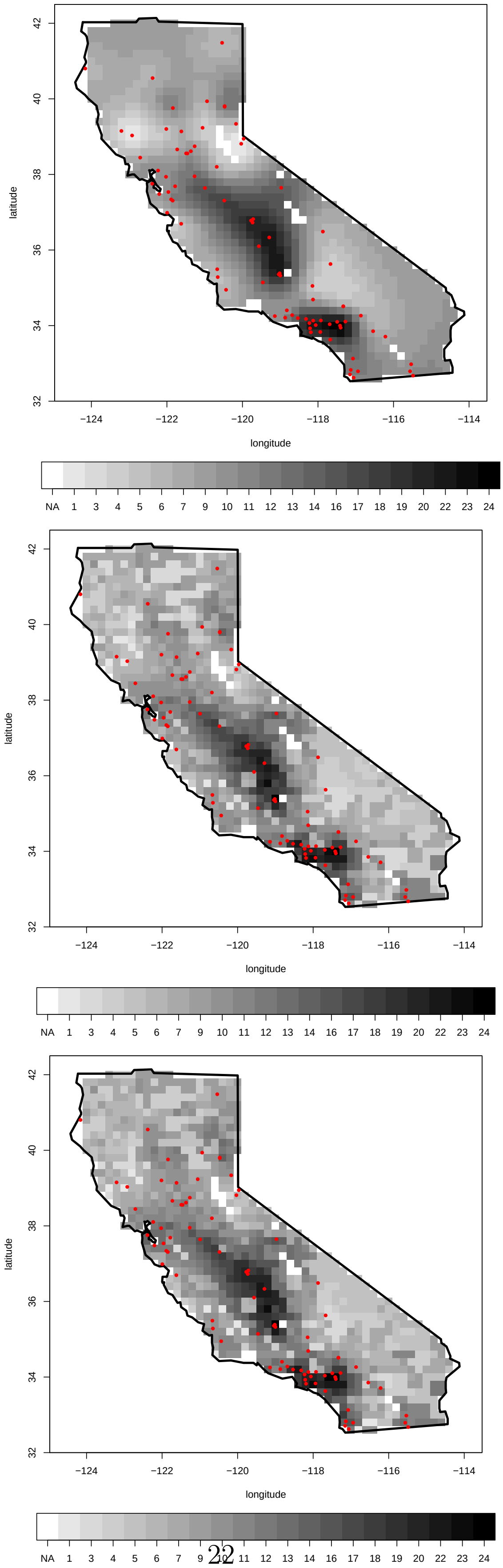
42
40
38
36
34
32
latitude
−124
−122
−120
−118
−116
−114
longitude
NA 1 3 4 5 6 7 9 10 11 12 13 14 16 17 18 19 20 22 23 24
42
40
38
36
34
32
latitude
−124
−122
−120
−118
−116
−114
longitude
NA 1 3 4 5 6 7 9 10 11 12 13 14 16 17 18 19 20 22 23 24
42
40
38
36
34
32
latitude
−124
−122
−120
−118
−116
−114
longitude
NA 1 3 4 5 6 7 9 10 11 12 13 14 16 17 18 19 20 22 23 24

Figure 10: Bayesian Kriging maps for PM2.5. Top row is produced by Model (0), middle row is by Model (I'), and bottom row is produced by Model (II') on Table 3.

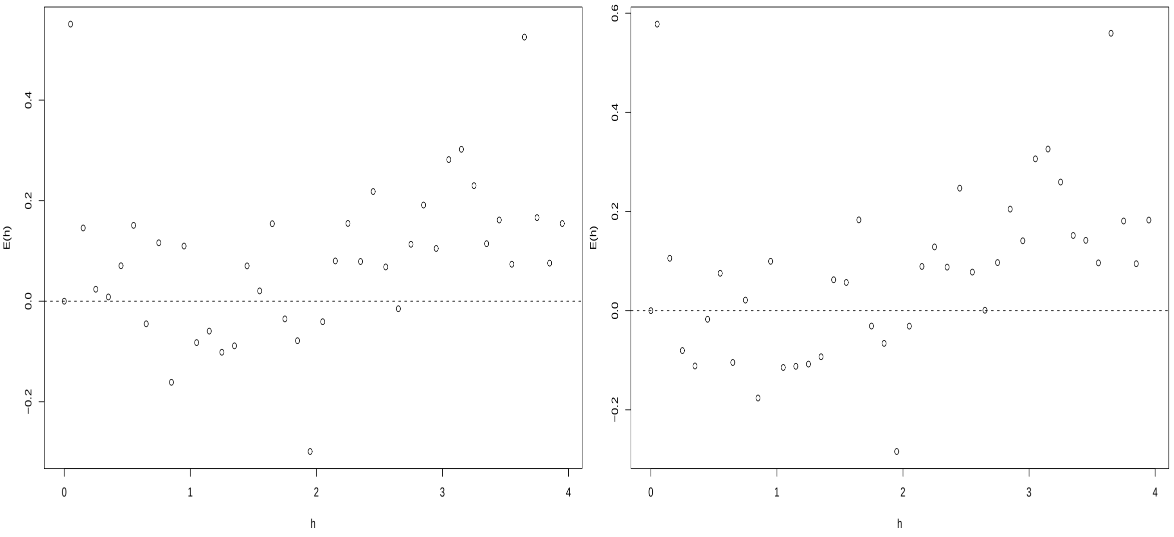

Figure 11: Test for independence between PM2.5 monitor locations and pollutant levels. The empirical function $\hat{E}(h)$ was calculated after applying a square-root transformation to PM2.5 levels and removing the trend surfaces (left panel: Model I, right-panel: Model II on Table 3). See text for details.

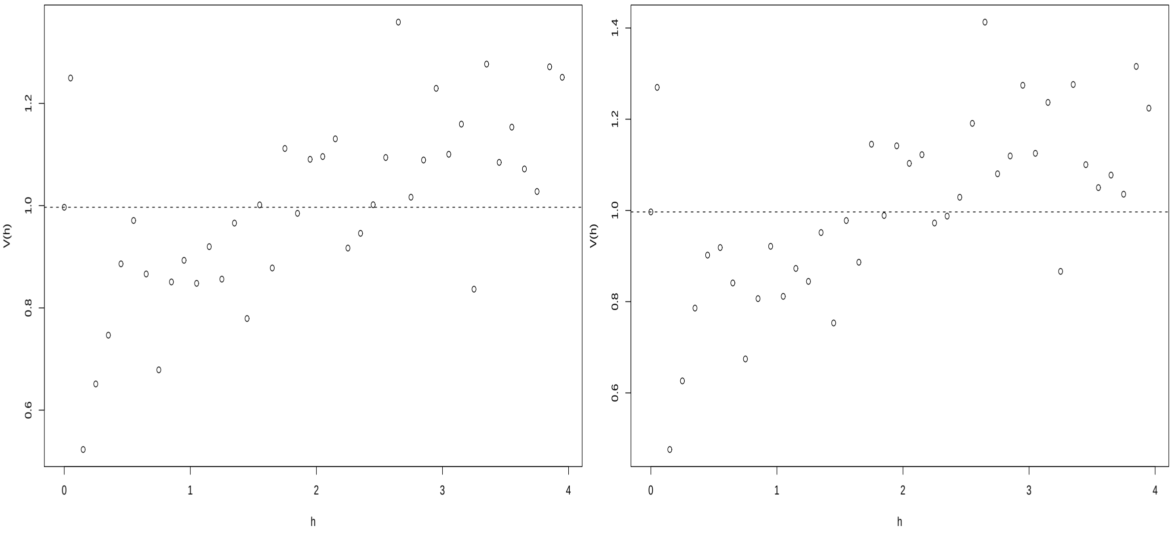

Figure 12: Test for independence between PM2.5 monitor locations and pollutant levels. The empirical function $\hat{V}(h)$ was calculated after applying a square-root transformation to PM2.5 levels and removing the trend surfaces (left panel: Model I, right-panel: Model II on Table 3). See text for details.

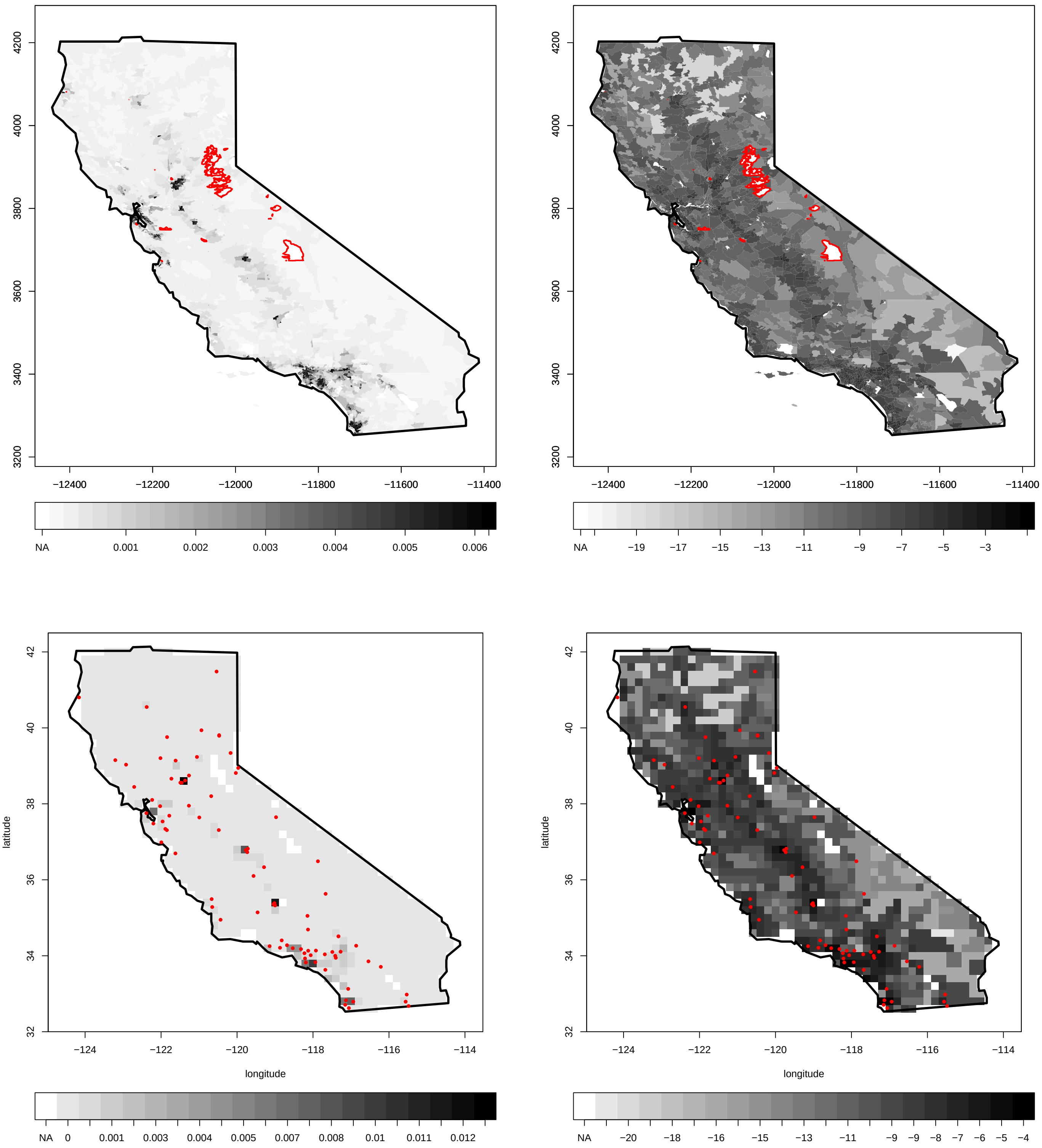

Figure 13: Estimated PM2.5 monitor intensity using both original scales (left-hand panels) and log scale (right-hand panels) for PM2.5 monitors in California. The fitted model is specified in Table 4. There is no population living in the pure white areas. The top row uses postcode entities and the bottom row uses pixels as units to draw.

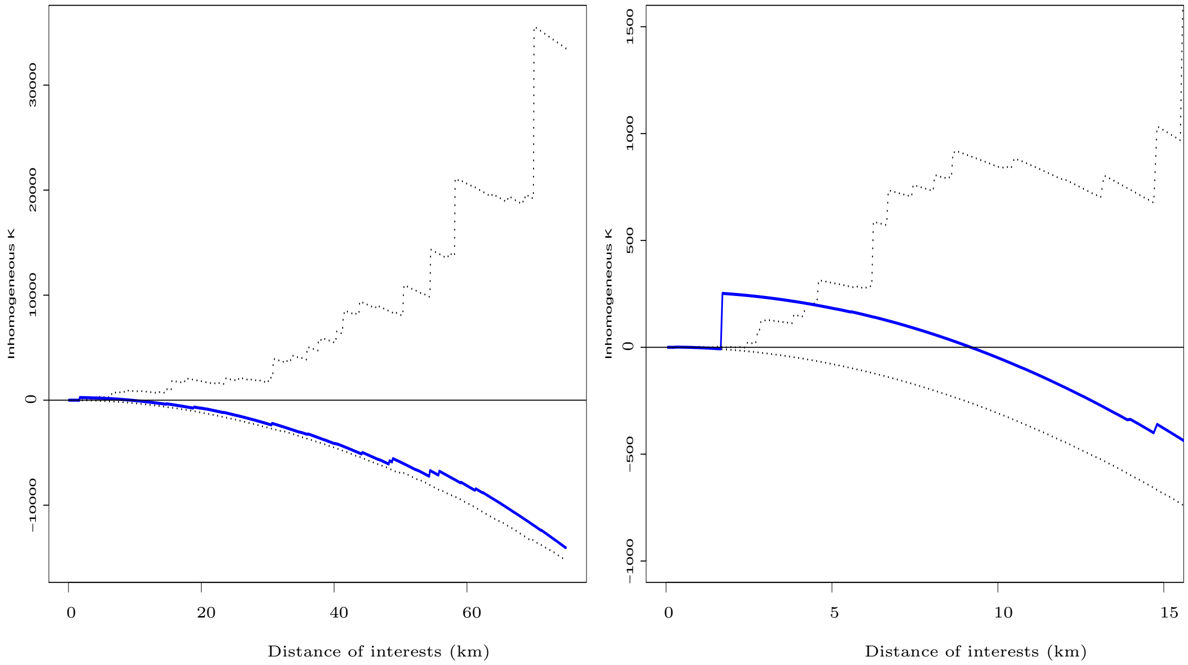

Figure 14: Estimates of $K(s) - \pi s^2$ (solid line) and 95% pointwise Monte Carlo tolerance limits based on 100 simulations of the fitted model (dashed lines) for PM2.5 monitors. Left-hand panel covers distances from 0 to 75km, right-hand panel distances from 0 to 15km.