Geostatistical Inference under Preferential Sampling

Marie Ozanne and Justin Strait

Diggle, Menezes, and Su, 2010

October 12, 2015

A simple geostatistical model

Notation:

- The underlying spatially continuous phenomenon $S(x), x \in \mathbb{R}^2$ is sampled at a set of locations $x_i, i = 1, \ldots, n$, from the spatial region of interest $A \subset \mathbb{R}^2$
- Y_i is the measurement taken at x_i
- Z_i is the measurement error

The model:

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

- $\{Z_i, i = 1, \ldots, n\}$ are a set of mutually independent random variables with $\mathrm{E}[Z_i] = 0$ and $\mathrm{Var}(Z_i) = \tau^2$ (called the **nugget variance**)
- Assume $\mathrm{E}[S(x)] = 0 \ \forall x$

Thinking hierarchically

Diggle *et al.* (1998) rewrote this simple model hierarchically, assuming Gaussian distributions:

- $S(x)$ follows a latent Gaussian stochastic process
- $Y_i|S(x_i) \sim N(\mu + S(x_i), \tau^2)$ are mutually independent for $i = 1, \ldots, n$

If $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_n)$, and $S(X) = \{S(x_1), \ldots, S(x_n)\}$, this model can be described by:

$$[S, Y] = [S][Y|S(X)] = [S][Y_1|S(x_1)] \ldots [Y_n|S(x_n)]$$

where $[\cdot]$ denotes the distribution of the random variable.
$\rightarrow$ This model treats X as **deterministic**

What is preferential sampling?

- Typically, the sampling locations x_i are treated as stochastically independent of $S(x)$, the spatially continuous process:

$$[S, X] = [S][X]$$

(this is **non-preferential sampling**).

- This means that $[S, X, Y] = [S][X][Y|S(X)]$, and by conditioning on X, standard geostatistical techniques can be used to infer properties about S and Y.
- **Preferential sampling** describes instances when the sampling process depends on the underlying spatial process:

$$[S, X] \neq [S][X]$$

- Preferential sampling complicates inference!

Examples of sampling designs

1. Non-preferential, uniform designs: Sample locations come from an independent random sample from a uniform distribution on the region of interest A (e.g. completely random designs, regular lattice designs).
2. Non-preferential, non-uniform design: Sample locations are determined from an independent random sample from a non-uniform distribution on A.
3. Preferential designs:
   - Sample locations are more concentrated in parts of A that tend to have higher (or lower) values of the underlying process $S(x)$
   - X, Y form a marked point process where the points X and the marks Y are dependent

Schlather *et al.* (2004) developed a couple tests for determining if preferential sampling has occurred.

Why does preferential sampling complicate inference?

Consider the situation where S and X are stochastically dependent, but measurements Y are taken at a different set of locations, independent of X. Then, the joint distribution of S, X, and Y is:

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

We can integrate out X to get:

$$[S, Y] = [S][Y|S]$$

This means inference on S can be done by "ignoring" X (as is convention in geostatistical inference). However, if Y is actually observed at X, then the joint distribution is:

$$[S, X, Y] = [S][X|S][Y|X, S] = [S][X|S][Y|S(X)]$$

Conventional methods which "ignore" X are misleading for preferential sampling!

Shared latent process model for preferential sampling

The joint distribution of S, X, and Y (from previous slide):

$$[S, X, Y] = [S][X|S][Y|X, S] = [S][X|S][Y|S(X)]$$

with the last equality holding for typical geostatistical modeling.

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 x, x' separated by distance u
2. Given S, X is an inhomogeneous Poisson process with intensity

   $$\lambda(x) = \exp(\alpha + \beta S(x))$$

3. Given S and X, $Y = (Y_1, \ldots, Y_n)$ is set of mutually independent random variables such that

   $$Y_i \sim N(\mu + S(x_i), \tau^2)$$

Shared latent process model for preferential sampling

Some notes about this model:

- Unconditionally, X follows a log-Gaussian Cox process (details in Moller *et al.* (1998))
- If we set $\beta = 0$ in $[X|S]$, then unconditionally, Y follows a multivariate Gaussian distribution
- Ho and Stoyan (2008) considered a similar hierarchical model construction for marked point processes

Simulation experiment

- Approximately simulate the stationary Gaussian process S on the unit square by simulating on a finely spaced grid, and then treating S as constant within each cell.
- Then, sample values of Y according to one of 3 sampling designs:
  1. Completely random (non-preferential): Use sample locations x_i that are determined from an independent random sample from a uniform distribution on A.
  2. Preferential: Generate a realization of X by using $[X|S]$, with $\beta = 2$, and then generate Y using $[Y|S(X)]$.
  3. Clustered: Generate a realization of X by using $[X|S]$, but then generate Y on locations X using a separate independent realization of S.
     - This is non-preferential, but marginally X and Y share the same properties as the preferential design.

Specifying the model for simulation

- S is stationary Gaussian with mean $\mu = 4$, variance $\sigma^2 = 1.5$ and correlation function defined by the Matérn class of correlation functions:

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

  where K_κ is the modified Bessel function of the second kind. For this simulation, $\phi = 0.15$ and $\kappa = 1$.
- Set the nugget variance $\tau^2 = 0$ so that y_i is the realized value of $S(x_i)$.

Simulation sampling location plots

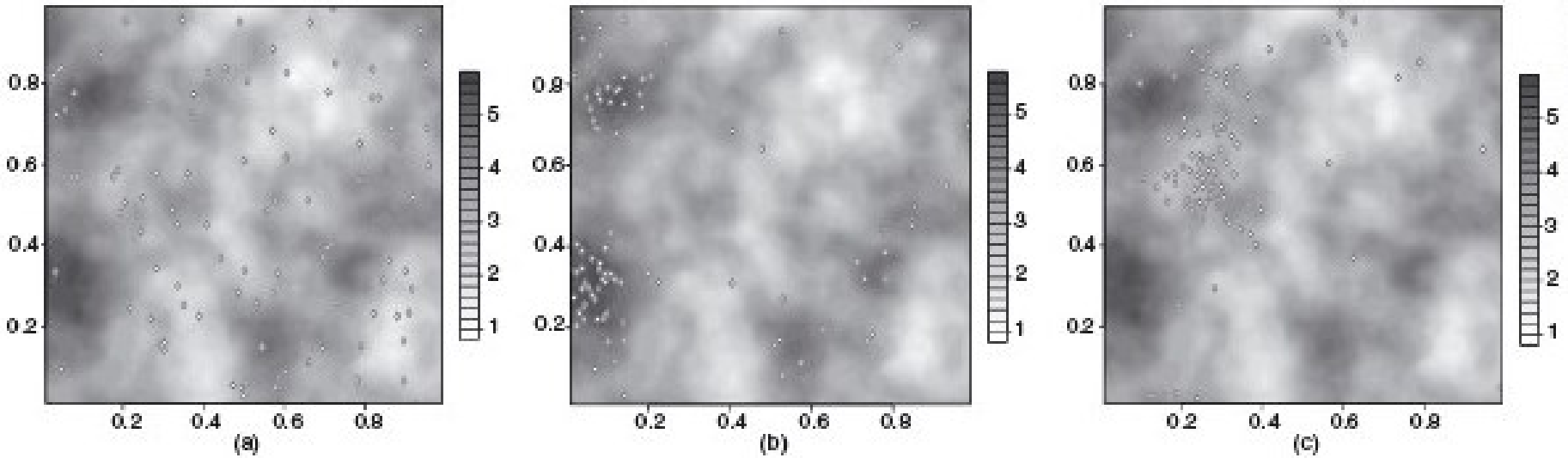

Figure: Underlying process realization and sampling locations from the simulation for (a) completely random sampling, (b) preferential sampling, and (c) clustered sampling

Estimating the variogram

Theoretical variogram of spatial process $Y(x)$:

$$V(u) = \frac{1}{2}\mathrm{Var}(Y(x) - Y(x'))$$

where x and x' are distance u apart

Empirical variogram ordinates: For (x_i, y_i), $i = 1, \ldots, n$ where x_i is the location and y_i is the measured value at that location:

$$v_{ij} = \frac{1}{2}(y_i - y_j)^2$$

- Under non-preferential sampling, v_{ij} is an unbiased estimate of $V(u_{ij})$, where u_{ij} is the distance between x_i and x_j
- A variogram cloud plots v_{ij} against u_{ij}; these can be used to find an appropriate correlation function. For this simulation, simple binned estimators are used.

Empirical variograms under different sampling regimes

Looking at 500 replicated simulations, the pointwise bias and standard deviation of the smoothed empirical variograms are plotted:

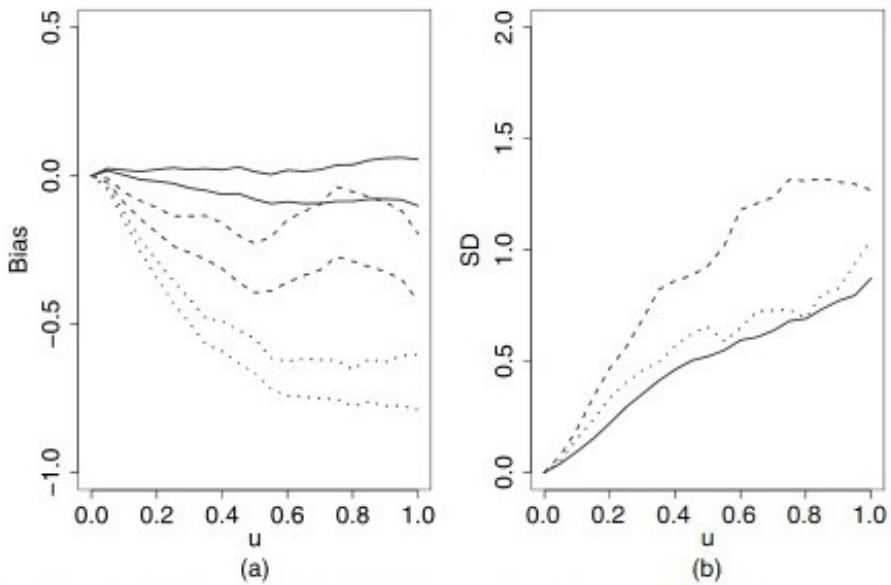

Fig. 2. Estimated bias and standard deviation of the sample variogram under random (———), preferential (· · · · · · ·) and clustered (- - - - - - -) sampling (see the text for a detailed description of the simulation model): (a) pointwise means plus and minus two pointwise standard errors; (b) pointwise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

- Under preferential sampling, the empirical variogram is biased and less efficient!
- The bias comes from sample locations covering a much smaller range of $S(x)$ values

Spatial prediction

Goal: Predict the value of the underlying process S at a location x_0, given the sample (x_i, y_i), $i = 1, \ldots, n$.

- Typically, ordinary kriging is used to estimate the unconditional expectation of $S(x_0)$, with plug-in estimates for covariance parameters.
- The bias and MSE of the kriging predictor at the point $x_0 = (0.49, 0.49)$ are calculated for each of the 500 simulations, and used to form 95% confidence intervals:

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	RMSE	(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	RMSE	(0.202,0.228)	(0.247,0.292)	(0.214,0.247)

Kriging issues under preferential sampling

- For both models, the completely random and clustered sampling designs lead to approximately unbiased predictions (as expected).
- Under the Model 1 simulations, there is large, positive bias and high MSE for preferential sampling (here, $\beta = 2$) - this is because locations with high values of S are oversampled.
- Under the Model 2 simulations, there is some negative bias (and slightly higher MSE) due to preferential sampling (here, $\beta = -2$) ; however, the bias and MSE are not as drastic because:
  - the variance of the underlying process is much smaller; the **degree of preferentiality** $\beta\sigma$ is lower here than for Model 1.
  - the nugget variance is non-zero for Model 2.

Fitting the shared latent process model

Data: X, Y

Likelihood for the data:

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

where θ consists of all parameters in the model

- To evaluate $[X|S]$, the realization of S at all possible locations $x \in A$ is needed; however, we can approximate S (which is spatially continuous) by a set of values on a finely spaced grid, and replace exact locations X by their closest grid point.
- Let $S = \{S_0, S_1\}$, where S_0 represents values of S at the n observed locations $x_i \in X$ and S_1 denotes values of S at the other $N - n$ grid points.
- Unfortunately, estimating the likelihood with a sample average over simulations S_j fails when the nugget variance is 0 because simulations of S_j usually will not match up with the observed Y.

Evaluating the likelihood

$$
\begin{aligned}
L(\theta) &= \int [X|S][Y|X,S][S]dS \\
&= \int [X|S][Y|X,S]\frac{[S|Y]}{[S|Y]}[S]dS \\
&= \int [X|S][Y|S_0]\frac{[S|Y]}{[S_0|Y][S_1|S_0,Y]}[S_0][S_1|S_0]dS \\
&= \int [X|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0][S|Y]dS
\end{aligned}
\tag{1}
$$

The third equality uses $[S] = [S_0][S_1|S_0]$, $[S|Y] = [S_0|Y][S_1|S_0, Y]$, and $[Y|X,S] = [Y|S_0]$. The last equality uses $[S_1|S_0, Y] = [S_1|S_0]$. Hence:

$$
L(\theta) = \mathrm{E}_{S|Y}\left[[X|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0]\right] \tag{2}
$$

Approximating the likelihood

A Monte Carlo approximation can be used to approximate the likelihood:

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

where S_j are simulations of $S|Y$.

- Antithetic pairs of realizations are used to reduce Monte Carlo variance
- To simulate from $[S|Y]$, we can simulate from several other unconditional distributions, and then notice that:

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

has the distribution of $S|Y = y$, where:

- $S \sim MVN(0, \Sigma), Y \sim MVN(\mu, \Sigma_0)$, $Z \sim N(0, \tau^2)$
- C is an n x N matrix which identifies the position of the data locations within all possible prediction locations

Goodness of fit

- We can use K-functions to assess how well the shared latent process model under preferential sampling fits the data.
- The K-function $K(s)$ is defined by $\lambda K(s) = \mathrm{E}[N_0(s)]$, where $N_0(s)$ is the number of points in the process within distance s of a chosen origin and λ is the expected number of points in the process per unit area.
- Under our preferential sampling model, X marginally follows a log-Gaussian Cox process with intensity $\Lambda(x) = \exp(\alpha + \beta S(x))$. The corresponding K-function is:

$$K(s) = \pi s^2 + 2\pi \int_0^s \gamma(u) u du$$

  where $\gamma(u)$ is the covariance function of $\Lambda(x)$ (Diggle (2003))
- By comparing the estimated K-function from the data to an envelope of estimates obtained from simulated realizations of the fitted model, goodness of fit can be determined.

Lead biomonitoring in Galicia, Spain

Background

- Uses lead concentration, [Pb] (μg/g dry weight), in moss samples as measured variable
- Initial survey conducted in Spring 1995 to 'select the most suitable moss species and collection sites' (Fernandez et al., 2000)
- Two further surveys of [Pb] in samples of *Scleropodium purum*
  - October 1997: sampling conducted more intensively in subregions where large gradiants in [Pb] expected
  - July 2000: used approximately regular lattice design; gaps arise where different moss species collected

Lead biomonitoring in Galicia, Spain

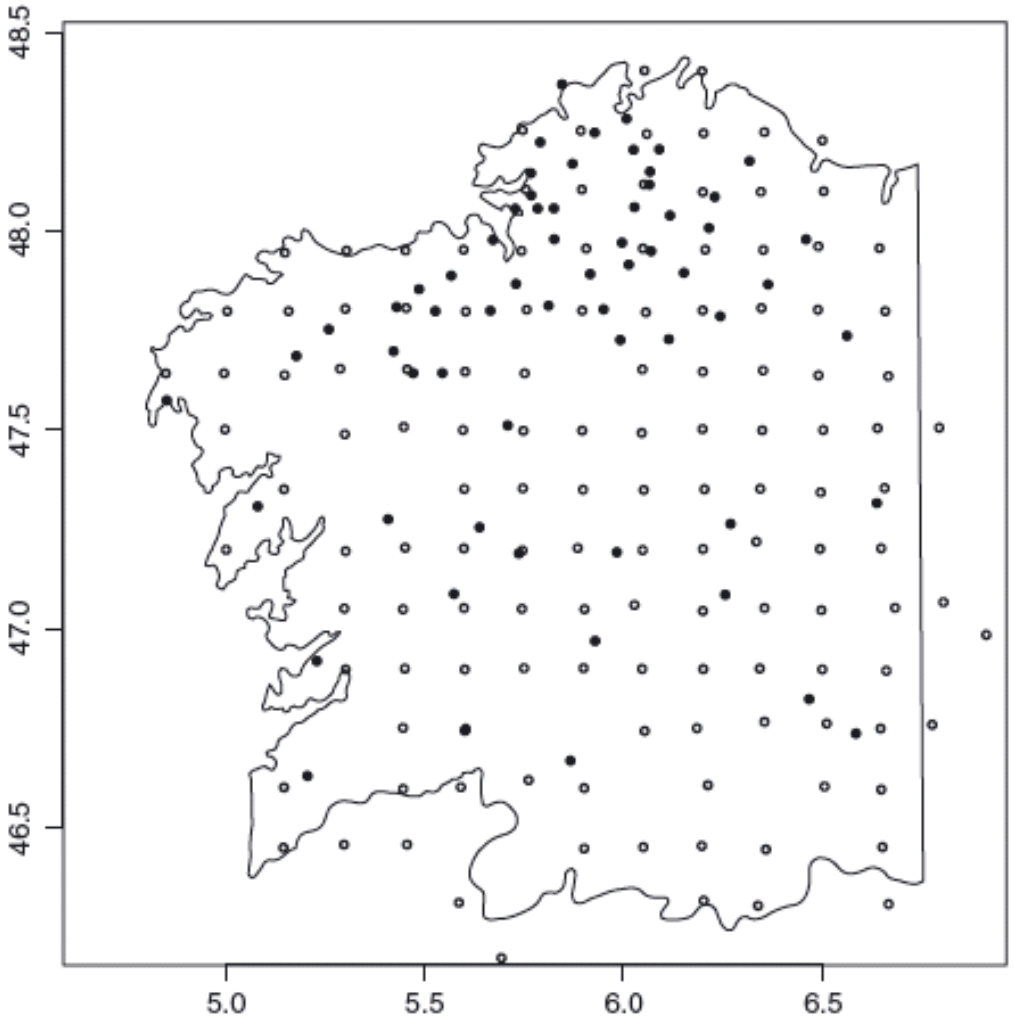

Fig. 3. Sampling locations for 1997 (•) and 2000 (∘): the unit of distance is 100 km; two outliers in the 1997 data were at locations (6.50,46.90) and (6.65,46.75)

Lead biomonitoring in Galicia, Spain

Summary statistics:

	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

Lead biomonitoring in Galicia, Spain

Standard geostatistical analysis

Assumptions:

- standard Gaussian model with underlying signal $S(x)$
- $S(x)$ is a zero-mean stationary Gaussian process with:
  - variance σ^2
  - Matern correlation function $\rho(u; \phi, \kappa)$
  - Gaussian measurement errors, $Z_i \sim N(0, \tau^2)$

Models fitted separately for 1997 and 2000 data

Lead biomonitoring in Galicia, Spain

Standard geostatistical analysis

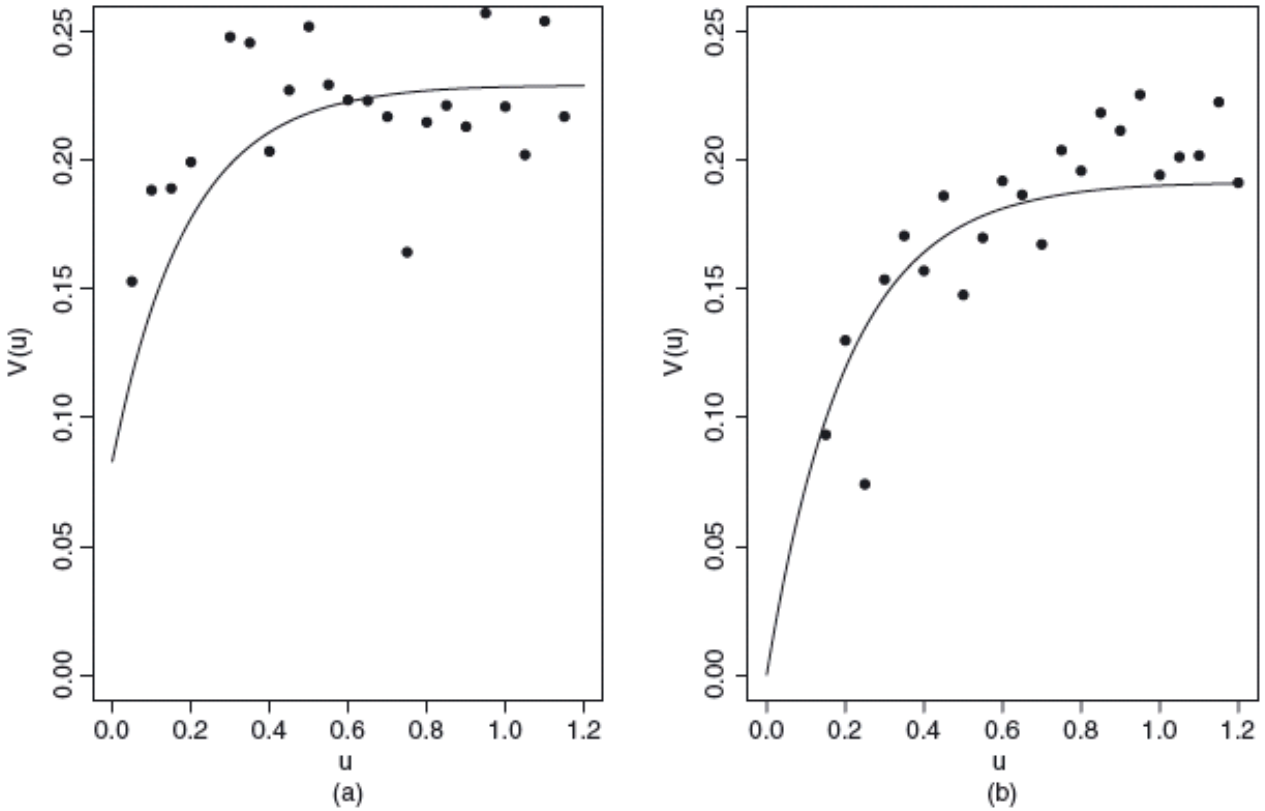

Fig. 5. Smoothed empirical (•) and fitted theoretical (———) variograms for (a) 1997 and (b) 2000 log-transformed lead concentration data

Analysis under preferential sampling

Parameter estimation

Goal: To investigate whether the 1997 sampling is preferential

- Use Nelder-Mead simplex algorithm (Nelder and Mead, 1965) to estimate model parameters
- $m = 100,000$ Monte Carlo samples reduced standard error to approximately 0.3 and approximate generalized likelihood ratio test statistic to test $\beta = 0$ was 27.7 on 1 degree of freedom ($p < 0.001$)

Lead biomonitoring in Galicia, Spain

Analysis under preferential sampling

Parameter estimation

Goal: To test the hypothesis of shared values of σ, ϕ, and τ

- Fit joint model to 1997 and 2000 data sets, treated as preferential and nonpreferential, respectively
- Fit model with and without constaints on σ, ϕ, and τ to get generalized likelihood ratio test statistic of 6.2 on 3 degrees of freedom ($p = 0.102$)

Using shared parameter values (when justified) improves estimation efficiency and results in a better identified model (Altham, 1984)

Lead biomonitoring in Galicia, Spain

Analysis under preferential sampling

Parameter estimation

- Monte Carlo maximum likelihood estimates obtained for the model with shared σ, ϕ, and τ
- Preferential sampling parameter estimate is negative, $\hat{\beta} = -2.198$; dependent on allowing two separate means

Recall:
Given S, X is an inhomogeneous Poisson process with intensity

$$\lambda(x) = \exp(\alpha + \beta S(x))$$

Analysis under preferential sampling

Goodness of Fit

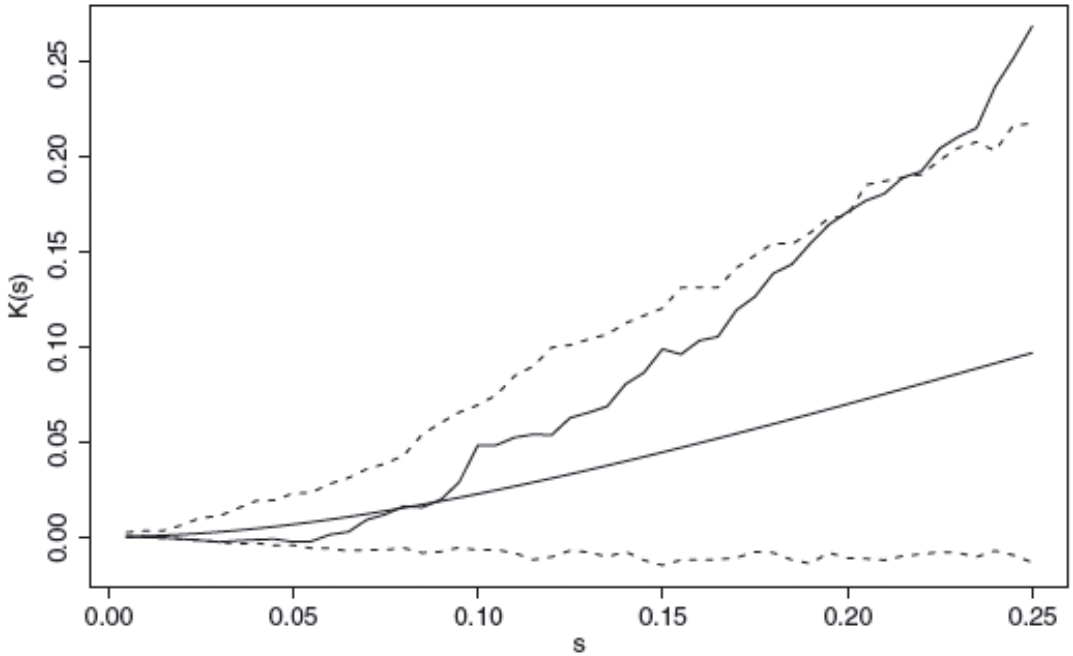

Fig. 6. Estimated K-function of the 1997 sample locations (——) and envelope from 99 simulations of the fitted log-Gaussian Cox process (-------)

Goodness of fit assessed using statistic T; the resultant p-value $= 0.03$

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

Lead biomonitoring in Galicia, Spain

Analysis under preferential sampling

Prediction

- Figures in paper show predicted surfaces $\hat{T}(x) = \mathrm{E}[T(x)|X, Y]$, where $T(x) = \exp\{S(x)\}$ denotes the [Pb] on the untransformed scale
- Predictions based on the preferential sampling have much wider range over lattice of prediction locations compared to those that assume non-preferential sampling (1.310-7.654 and 1.286-5.976 respectively)
- **Takeaway:** Recognition of the preferential sampling results in a pronounced shift in the predictive distribution

Discussion

- Conventional geostatistical models and associated statistical methods can lead to misleading inferences if the underlying data have been preferentially sampled
- This paper proposes a simple model to take into account preferential sampling and develops associated Monte Carlo methods to enable maximum likleihood estimation and likelihood testing within the class of models proposed
- This method is computationally intensive - each model takes several hours to run

References

- Diggle, P.J., Menezes, R., Su, T.-l., 2010. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 59, 191-232.
- Menezes, R., 2005. Assessing spatial dependency under non-standard sampling. Ph.D. Dissertation, Universidad de Santiago de Compostela.
- Pati, D., Reich, B.J., Dunson, D.B., 2011. Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98, 35-48.
- Gelfand, A.E., Sahu, S.K., Holland, D.M., 2012. On the effect of preferential sampling in spatial prediction. *Environmetric*, 23, 565-578.
- Lee, A., Szpiro, A., Kim, S.Y., Sheppard, L., 2015. Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*.