Geostatistical inference under preferential sampling: introduction

By Peter Diggle, Raquel Menezes, and Ting-li Su

John Paige

Statistics Department
University of Washington

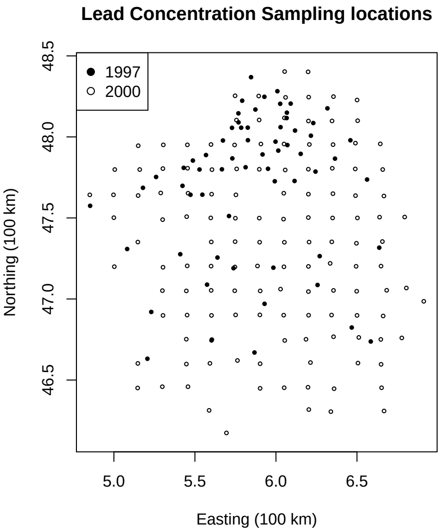
April 7, 2016

Introduction to the Problem

- *Geostatistics* involves modeling a process that is continuous in space measured at discrete locations
- Often, data is assumed to be distributed randomly throughout the domain
- What happens when the chance of sampling the process is tied to the value of the process itself?
  - This is called *preferential sampling*

Motivating Dataset

- Lead concentration ($\mu g/g$ dry weight) in Galicia, Spain
- Samples in 1997 are concentrated to the north, but samples in 2000 are on lattice
- How to tell if data is sampled preferentially (and at what level)?
- If so, how does preferential sampling effect 'naive' inference?
- How can we effectively take preferential sampling into account?



Preferential Sampling

$S = \{S(x) : x \in \mathbb{R}^2\}$: spatially continuous stochastic process
$X = (x_1, ..., x_n)$: set of sample locations

Preferential sampling refers to when $\pi(S, X) \neq \pi(X)\pi(S)$ (S is not independent of X).

Note that non-preferential sampling does not necessitate uniform sampling. Sample locations could still be clustered.

Scientific Context

- *Covariograms* defined the spatial structure of the covariance in S. The traditional estimator is:

$$\widehat{C}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (S(x_i) - \overline{S})(S(x_j) - \overline{S})$$

  where $x_i - x_j \approx \mathbf{h}$ (under isotropic model, $|x_i - x_j| \approx h$)
- Isaaks and Srivastava 1988 and Srivastava and Parker 1989 propose an alternative non-ergodic estimator:

$$\widehat{C}_{ne}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (S(x_i) - \overline{S}(\mathbf{h}_i))(S(x_j) - \overline{S}(\mathbf{h}_j))$$

  where $\overline{S}(\mathbf{h}_i)$ is the sample mean of S at all the points $x_i \in \mathbf{h}_i$
  - They claim this works better under preferential sampling
- Curriero et al. 2002 shows this is "equivalent" yet "worse" than the traditional estimator under isotropy

Scientific Context

- Schlather et al. 2004 proposes tests for prefential sampling assuming stationarity
  - Assumption of stationarity may affect legitimacy of results

Model for Preferential Sampling

Three assumptions for model:

1. S is a stationary, mean zero Gaussian process
2. Conditional on S, X is an inhomogeneous Poisson process wtih intensity
$$\lambda(x) = \exp\{\alpha + \beta S(x)\}$$
3. $Y_i | S, X \overset{iid}{\sim} \mathcal{N}\left(\mu + S(x_i), \tau^2\right)$

$1 + 2 \Rightarrow X$ is a log Gaussian Cox process

Log-Gaussian Cox Processes

A *Cox process* is a stochastic point process satisfying:

- $\Lambda(x)$ is a random rate process
- Conditioned on $\Lambda(x) = \lambda(x)$, a Cox process is an inhomogeneous Poisson process with rate $\lambda(x)$

A *log Gaussian* Cox process also satisfies $\Lambda(x) = \exp\{Z(x)\}$, where $Z(x)$ is a Gaussian random field.

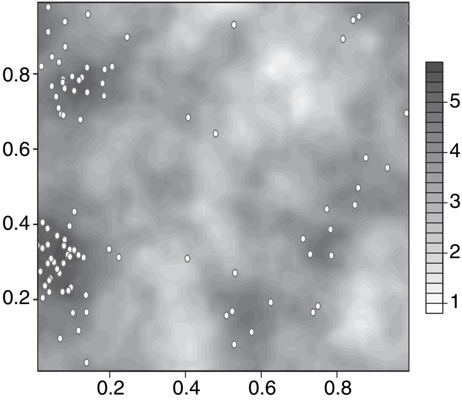


Figure : From Diggle et al. 2010. An example of a log Gaussian Cox process on unit square where $\beta = 2$, $\alpha = 1$, and S has Matérn covariance.

Spatial Covariance Model

The Matérn correlation model, as a function of distance, u, is:

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

κ: shape parameter
ϕ: scale parameter
K_{κ}: modified Bessel function of the second kind, of order κ

- ρ is called the *correlogram* when viewed as a function of distance
- The covariance as a function of distance is the *covariogram*
- The variance as a function of distance is the *variogram*

Testing Affect of Sample Designs

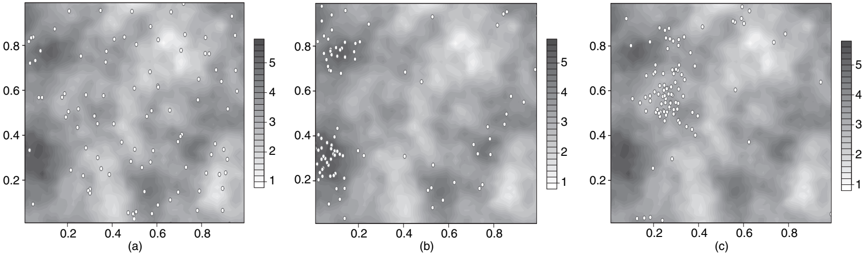


Figure : From Diggle et al. 2010

Variogram estimation tested under 500 simulations from three sampling designs:

a Uniform

b Preferential ($\beta = 2$)

c Clustered

Fitting the Model

Make gridded approximation of $S = \{S_0, S_1\}$

- S_0 are data
- S_1 are the values at other grid points

$$
\begin{aligned}
L(\vec{\theta}) &= \int \pi(Y|X,S)\pi(X|S)\pi(S)\ dS \\
&= \dots \\
&= E_{S|Y}\left[\pi(X|S)\frac{\pi(Y|S_0)}{\pi(S_0|Y)}\pi(S_0)\right] \\
&\approx m^{-1}\sum_{j=1}^{m}\pi(X|S_j)\frac{\pi(Y|S_{0j})}{\pi(S_{0j}|Y)}\pi(S_{0j})
\end{aligned}
$$

where S_j is the jth conditional simulation of S conditioned on Y.

Goodness of Fit

Reduced second moment measure (or K-function) for defined model is given by:

$$K(s) = \pi s^2 + 2\pi \int_0^s \left(\exp\left\{\beta^2\sigma^2\rho(u;\kappa,\phi)\right\} - 1\right) u \, du$$

- s represents the maximum distance apart points can be
- $\rho(u;\phi) := \mathrm{corr}\left(S(x), S(x')|\phi, \kappa, |x - x'| = u\right)$
- ϕ: Matérn scale parameter
- κ: Matérn smoothness parameter

Goodness of Fit

Define test statistic:

$$T = \int_0^{0.25} \frac{(\widehat{K}(s) - K(s))^2}{\nu(s)}\, ds$$

- $\widehat{K}(s)$: empirical K-function
- $\nu(s) := \mathrm{Var}\left(\widehat{K}\right)$

Conclusions

- Taking into account preferential sampling is important!
  - In the simulations (albiet with high β), variograms estimated naively were estimated poorly
- Uniform sampling performed best, then clustered, then preferential
- Proposed class of models is flexible and values for β can be tested directly with likelihood ratio test

References

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

Diggle, Peter, Raquel Menezes, and Ting-li Su (2010), "Geostatistical inference under preferential sampling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59, 191–232.

Isaaks, EH and R Mohan Srivastava (1988), "Spatial continuity measures for probabilistic and deterministic geostatistics." *Mathematical geology*, 20, 313–341.

Schlather, Martin, Paulo J Ribeiro, and Peter J Diggle (2004), "Detecting dependence between marks and locations of marked point processes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 79–93.

Srivastava, R Mohan and Harry M Parker (1989), "Robust measures of spatial continuity." In *Geostatistics*, 295–308, Springer.