Geostatistical inference under preferential sampling: Final presentation

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

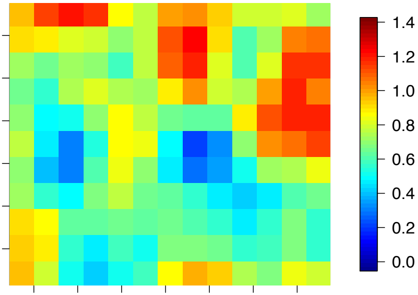
John Paige

Statistics Department
University of Washington

June 2, 2016

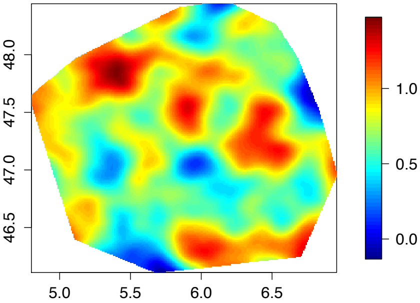
Divisions of Spatial Statistics

- Cressie (1991) and Gelfand (2010) divide spatial statistics into 3 areas:
  - discrete data
  - continuous data
  - point patterns



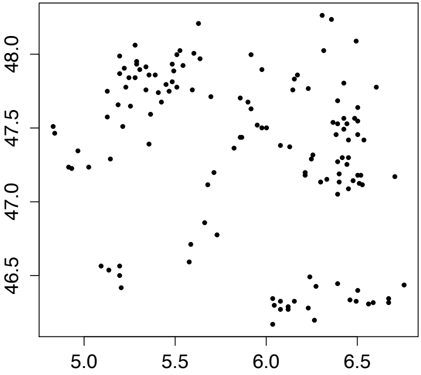
Divisions of Spatial Statistics

- Cressie (1991) and Gelfand (2010) divide spatial statistics into 3 areas:
  - discrete data
  - continuous data
  - point patterns



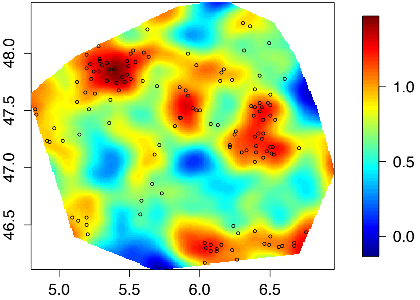
Divisions of Spatial Statistics

- Cressie (1991) and Gelfand (2010) divide spatial statistics into 3 areas:
  - discrete data
  - continuous data
  - point patterns



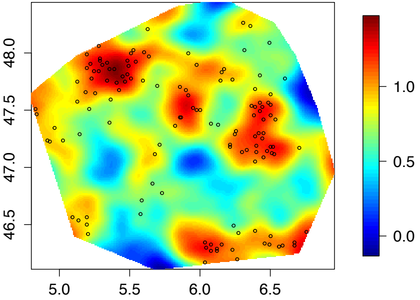
Divisions of Spatial Statistics

- Cressie (1991) and Gelfand (2010) divide spatial statistics into 3 areas:
  - discrete data
  - continuous data
  - point patterns
- Diggle et al. (2013) instead gives 2 subdivisions of spatial statistics:
  - continuous data
  - discrete data
- This emphasizes random nature of sampling locations



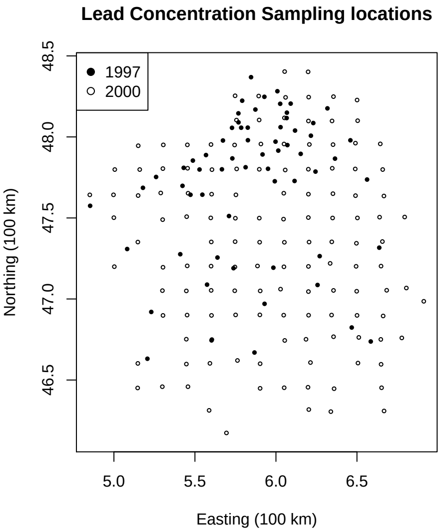
Divisions of Spatial Statistics

- Classically, data locations are assumed to be fixed constants
- What happens when the sample locations depend on the measured process itself?
    - This is called **preferential sampling**



Problems Addressed

- Determining if data is sampled preferentially
- How preferential sampling affects 'naive' inference
- Effective model for preferential sampling
- Focus on lead levels in Galicia, Spain and simulated experiments



Classical Model

$$Y_i = \mu + S(x_i) + Z_i,$$

Y_i: observation at location x_i

μ: mean

$S(\vec{x}) \sim \text{MVN}\left(\vec{0}, \Sigma(\vec{x})\right)$: spatially correlated portion of process

$Z_i \overset{iid}{\sim} \mathcal{N}\left(0, \tau^2\right)$: measurement noise

Classical Model

$$Y_i = \mu + S(x_i) + Z_i,$$

Y_i: observation at location x_i

μ: mean

$S(\vec{x}) \sim \text{MVN}\Big(\vec{0}, \Sigma(\vec{x})\Big)$: spatially correlated portion of process

$Z_i \overset{iid}{\sim} \mathcal{N}\left(0, \tau^2\right)$: measurement noise

$\Rightarrow \vec{Y} \sim \text{MVN}(\vec{\mu}, \Sigma_0)$, where $\Sigma_0 = \Sigma_0(\vec{\theta}) = \Sigma + \tau^2 I$

Classical Model

$$Y_i = \mu + S(x_i) + Z_i,$$

Y_i: observation at location x_i
μ: mean
$S(\vec{x}) \sim \text{MVN}\Big(\vec{0}, \Sigma(\vec{x})\Big)$: spatially correlated portion of process
$Z_i \overset{iid}{\sim} \mathcal{N}\left(0, \tau^2\right)$: measurement noise
$\Rightarrow \vec{Y} \sim \text{MVN}(\vec{\mu}, \Sigma_0)$, where $\Sigma_0 = \Sigma_0(\vec{\theta}) = \Sigma + \tau^2 I$

Log likelihood:

$$\mathcal{L}(\vec{\theta}) = -\frac{1}{2}\log(|\Sigma_0|) - \frac{1}{2}(\vec{Y} - \vec{\mu})'\Sigma_0^{-1}(\vec{Y} - \vec{\mu}) - \frac{n}{2}\log(2\pi)$$

Classical Model

$$Y_i = \mu + S(x_i) + Z_i,$$

Y_i: observation at location x_i
μ: mean
$S(\vec{x}) \sim \text{MVN}\left(\vec{0}, \Sigma(\vec{x})\right)$: spatially correlated portion of process
$Z_i \overset{iid}{\sim} \mathcal{N}\left(0, \tau^2\right)$: measurement noise

Common assumptions:

- x_i are sampled independently of true process $\mu + S$
- Stationarity

Variograms

- **Stationarity**: under stationarity,
  $\mathrm{Var}\left(S(x_i) - S(x_j)\right) = V(x_i - x_j)$
- **Isotropy**: under isotropy (and stationarity),
  $\mathrm{Var}\left(S(x_i) - S(x_j)\right) = V(|x_i - x_j|)$
- **Variograms** define the spatial structure of the covariance in S
- Empirical estimate given data Y_i at location x_i (under stationarity and isotropy):

$$\widehat{V}(d) = \frac{1}{|N(d)|} \sum_{|x_i - x_j| \in N(d)} (Y_i - Y_j)^2$$

  where $N(d)$ is the set of pairs (x_i, x_j) with $|x_i - x_j| \approx d$
- This estimator assumes non-preferentiality

Variograms

Matérn theoretical variogram:

$$V(d) = \sigma^2(1 - \rho(u \mid \phi, \kappa)) + \tau^2$$

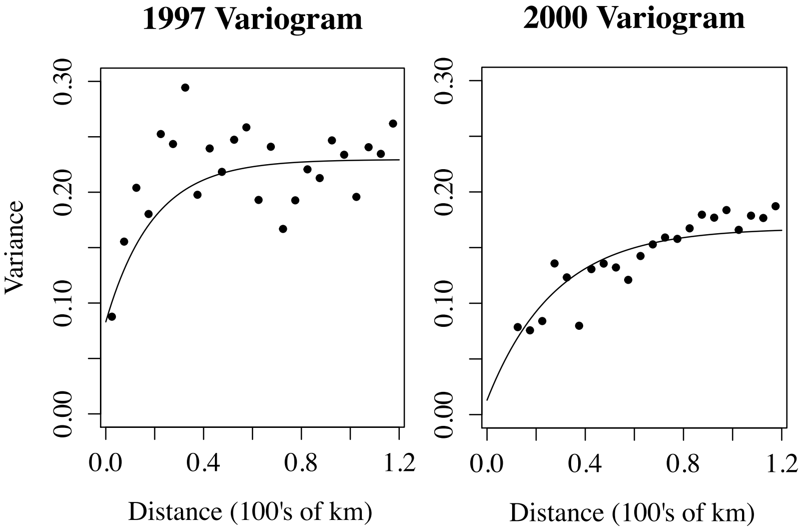
where

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

is the Matérn correlation function

u: distance
ϕ: scale
κ: smoothness
σ^2: is the variance of S
τ^2: measurement variance
$K_\kappa(\cdot)$: Bessel function

Variograms of Log-Lead Data (Classical)



Model for Preferential Sampling

Three assumptions for model:

1. S is still a stationary, mean zero Gaussian process (so $S(\vec{x}) \sim \text{MVN}\left(\vec{0}, \Sigma(\vec{x})\right))$
2. Conditional on S, X is an inhomogeneous Poisson process wtih random intensity

$$\Lambda(x) = \exp\{\alpha + \beta S(x)\}$$

3. $Y_i | S, X \overset{iid}{\sim} \mathcal{N}\left(\mu + S(x_i), \tau^2\right)$

$1 + 2 \Leftrightarrow X$ is a log Gaussian Cox process (LGCP)

Log-Gaussian Cox Processes

A **Cox process** is a stochastic point process that for B, B' bounded Borel sets satisfies:

- $N(B) \sim Pois\left(\int_B \Lambda(x)\, dx\right)$
- $N(B) \perp\!\!\!\perp N(B')$ when $B \cap B' = \emptyset$

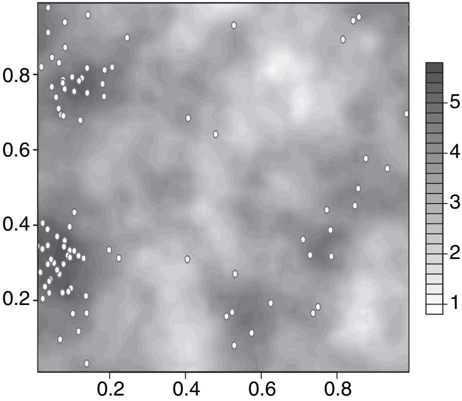


Figure : From Diggle et al. 2010. An example of a LGCP on unit square where $\beta = 2$, $\alpha = 1$, and S has Matérn covariance.

Testing Affect of Sample Designs: Samplings Schemes

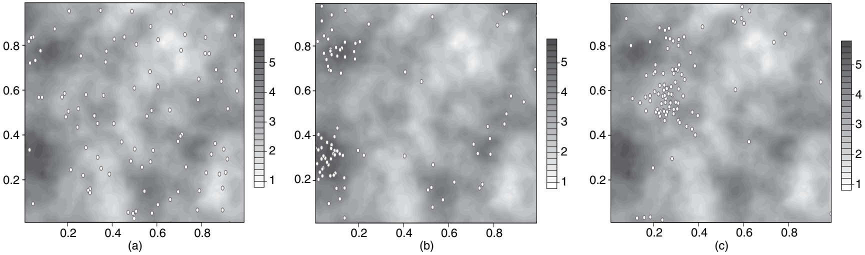


Figure : From Diggle et al. 2010

Variogram estimation tested under 500 simulations from three sampling designs:

- a Uniform
- b Preferential ($\beta = 2$)
- c Clustered

Classical Variogram Bias Under Preferential Sampling

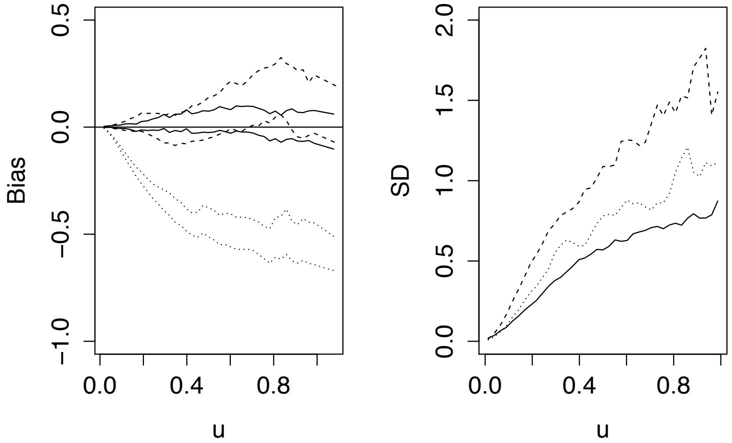


Figure : Variogram bias ±2 standard errors and standard deviations for uniform (solid), clustered (dashed), and preferential (dotted) sampling schemes.

Classical Prediction Bias Under Preferential Sampling

Mod.	*Param.*	*Confidence intervals*		
		Uniform	*Preferential*	*Clustered*
1	Bias	(-0.029, 0.038)	(0.956, 1.123)	(-0.074, 0.064)
1	RMSE	(0.354, 0.410)	(1.318, 1.501)	(0.717, 0.851)
2	Bias	(-0.040, 0.030)	(-0.265, -0.195)	(-0.040, 0.032)
2	RMSE	(0.375, 0.425)	(0.434, 0.491)	(0.382, 0.432)

Table : Classical predictions using 95% Confidence intervals for the given parameters under the given models and sampling schemes

Models:

1. $(\mu = 4, \sigma^2 = 1.5, \phi = .15, \kappa = 1, \beta = 2)$
2. $(\mu = 1.51, \sigma^2 = .14, \phi = .31, \kappa = .5, \beta = -2.20, \tau^2 = .059)$

Preferential Model Likelihood

Make gridded approximation of $S = \{S_0, S_1\}$ on lattice $X^* = \{x_1^*, ..., x_N^*\}$.

- 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.

Preferential Model Likelihood

Define C as a $n \times N$ matrix with a single 1 in each row and all else 0 s.t. $X = CX^*$.

Steps for Monte Carlo Simulation:

1. Simulate $\vec{S} \sim \text{MVN}\left(\vec{0}, \Sigma\right)$ using Circulant Embedding (**?**)
2. Compute jth simulation of $\vec{S}|\vec{Y}$:

$$\vec{S}_j \equiv \vec{S} + \Sigma C' \Sigma_0^{-1}(\vec{Y} - \vec{\mu} + \vec{Z} - C\vec{S})$$

   Where $Z_i \overset{iid}{\sim} \mathcal{N}\left(0, \tau^2\right)$

3. Calculate $m^{-1} \sum_{j=1}^{m} \pi(X|\vec{S}_j) \frac{\pi(\vec{Y}|\vec{S}_{0j})}{\pi(\vec{S}_{0j}|\vec{Y})} \pi(\vec{S}_{0j})$

Preferential Model Likelihood

$$\pi(X|\vec{S}_j) = \left(\prod_{i=1}^{n} \Lambda(x_i)\right) \left(\int \Lambda(x)\ dx\right)^{-n}$$

$$\vec{Y}|\vec{S}_{0j} \sim \text{MVN}\left(\vec{S}_{0j}, \tau^2 I\right)$$

$$\vec{S}_{0j}|\vec{Y} \sim \text{MVN}\left(\Sigma C' \Sigma_0^{-1}(\vec{Y} - \vec{\mu}), \Sigma - \Sigma C' \Sigma_0^{-1} C \Sigma\right)$$

$$\vec{S}_{0j} \sim \text{MVN}\left(\vec{0}, C \Sigma C'\right)$$

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: distance
$\rho(u;\phi,\kappa)$: Matérn correlation function

K-functions commonly used in goodness of fit tests

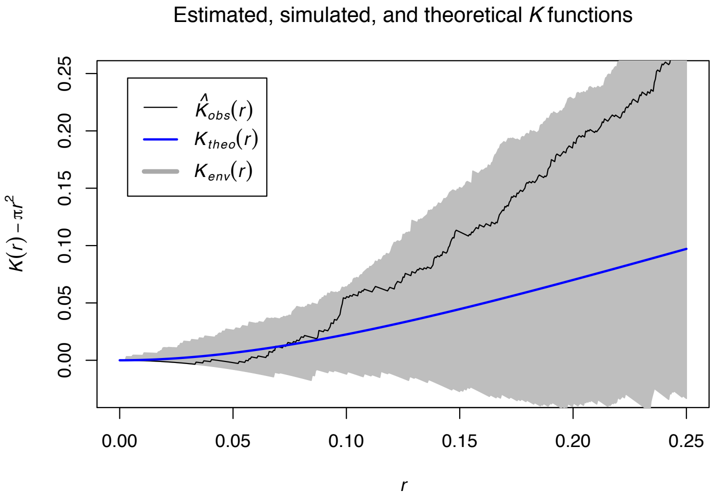
Goodness of Fit: Monte Carlo Testing

For any Monte Carlo test statistic, T, where higher T casts doubt on H_0, assume:

- T_1 is from data
- $T_2, ..., T_n$ are simulated under H_0

Then our p-value is the rank of T_1 out of $T_1, T_2, ..., T_n$ (*i.e.* if $n = 100$ and T_1 is largest test statistic, $p = .01$).

Goodness of Fit: Monte Carlo Testing



$p = 0.07$

Problems with the Methodology

- No cross-validation performed (effectiveness of K-function goodness of fit tests unclear)
- Predictive distribution assumes non-preferentiality with plug-in parameters from preferential model
- Predictive distribution has unreasonable certainty in locations far from data
- Joint non-preferential model gives parameters similar to preferential model parameters:
  - My non-preferential model: $(\hat{\mu}_{97} = 1.551, \hat{\mu}_{00} = 0.727, \hat{\sigma}^2 = 0.136, \hat{\phi} = 0.305, \hat{\tau}^2 = 0.052)$
  - Their preferential model: $(\hat{\mu}_{97} = 1.515, \hat{\mu}_{00} = 0.762, \hat{\sigma}^2 = 0.138, \hat{\phi} = 0.313, \hat{\tau}^2 = 0.059)$

Conclusions

- For preferential simulations, variograms estimated naively were biased
- 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
- LGCP isn't necessarily the best fit for the log-lead data
- LGCP model gives tractable Monte Carlo likelihood
- LGCP model has easy goodness of fit tests

References

Cressie, Noel (1991), *Statistics for spatial data*. New York: Wiley.

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.

Diggle, Peter J, Paula Moraga, Barry Rowlingson, Benjamin M Taylor, et al. (2013), "Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm." *Statistical Science*, 28, 542–563.

Gelfand, A. E. (2010), "Misaligned spatial data: The change of support problem." In Handbook of Spatial Statistics.

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.