Geostatistical inference under preferential sampling: Final presentation

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

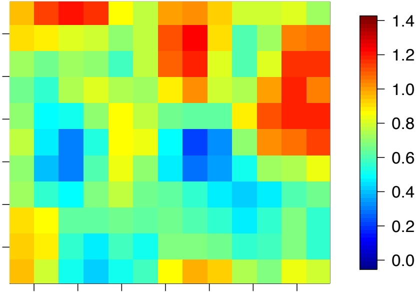
John Paige

Statistics Department
University of Washington

June 9, 2016

Divisions of Spatial Statistics

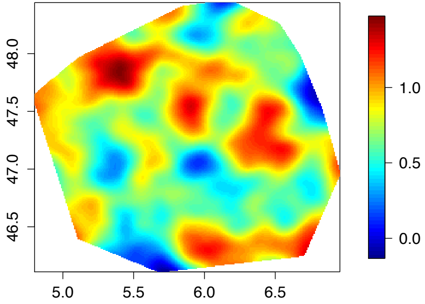
- Cressie (1991) and Gelfand (2010) divide spatial statistics into 3 areas:
  - discrete data
  - continuous data
  - point patterns



Data simulated with `RandomFields` R package (Schlather et al., 2016)

Divisions of Spatial Statistics

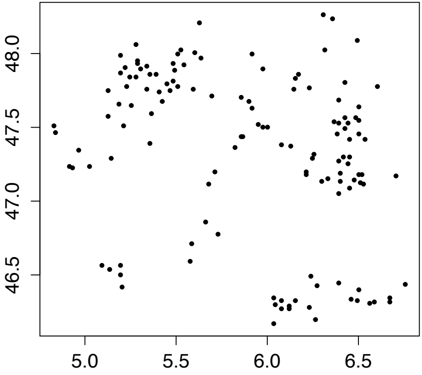
- Cressie (1991) and Gelfand (2010) divide spatial statistics into 3 areas:
  - discrete data
  - continuous data
  - point patterns



Data simulated with `RandomFields` R package (Schlather et al., 2016)

Divisions of Spatial Statistics

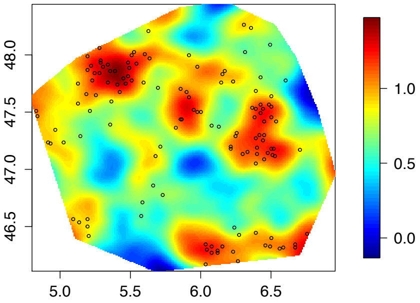
- Cressie (1991) and Gelfand (2010) divide spatial statistics into 3 areas:
    - discrete data
    - continuous data
    - point patterns



Data simulated with `RandomFields` R package (Schlather et al., 2016)

Divisions of Spatial Statistics

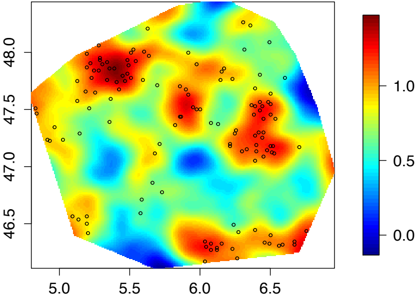
- Cressie (1991) and Gelfand (2010) divide spatial statistics into 3 areas:
  - discrete data
  - continuous data
  - point patterns
- Diggle et al. (2013) instead gives 2 subdivisions of spatial statistics:
  - continuous data
  - discrete data
- This emphasizes random nature of sampling locations



Data simulated with `RandomFields` R package (Schlather et al., 2016)

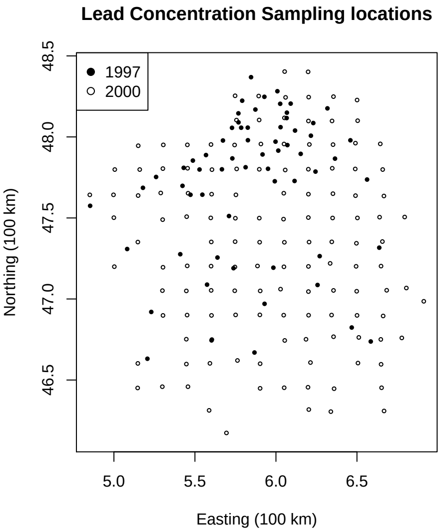
Divisions of Spatial Statistics

- 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

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

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

$\Rightarrow \vec{Y} \sim \mathrm{MVN}(\vec{\mu}, \Sigma_0)$, 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,
  $\text{Var}\left(S(x_i) - S(x_j)\right) = V(x_i - x_j)$
- **Isotropy**: under isotropy (and stationarity),
  $\text{Var}\left(S(x_i) - S(x_j)\right) = V(|x_i - x_j|)$
- **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:

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

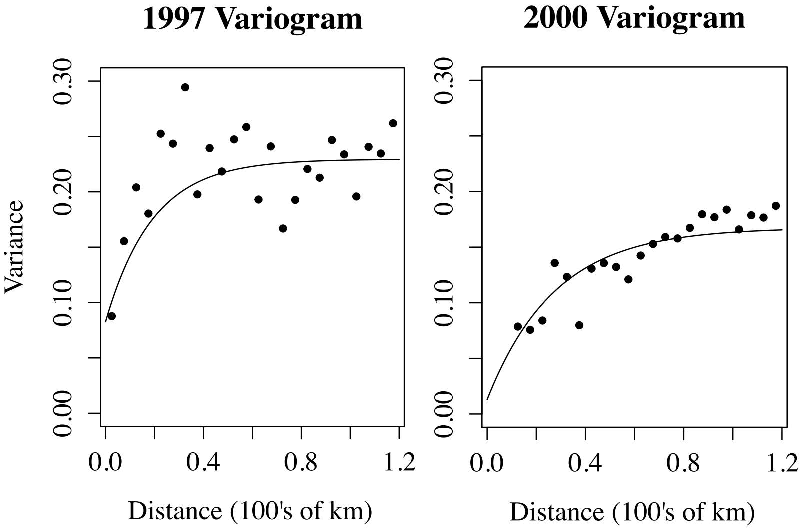
where

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

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, $\vec{X}$ is an inhomogeneous Poisson process wtih random intensity

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

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

$1 + 2 \Leftrightarrow X$ 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:

- $\Lambda(x) \geq 0$ is a random intensity
- $N(B) \sim Pois\left(\int_B \Lambda(x)\, dx\right)$
- $N(B) \perp\!\!\!\perp N(B')$ when $B \cap B' = \emptyset$

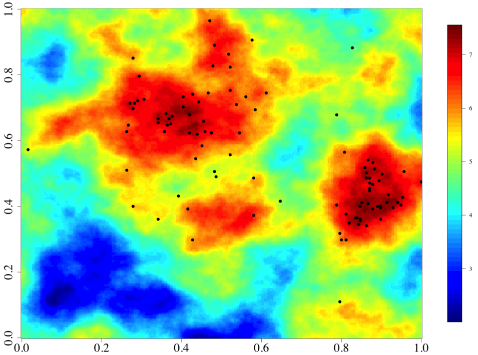
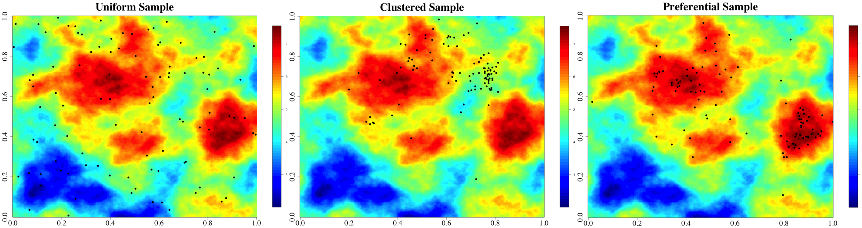


Figure: An example of a LGCP on unit square where $\beta = 2$, and S has Matérn covariance.

Testing Affect of Sample Designs: Samplings Schemes



Sampling processes for Matérn covariance, $\beta = 2$, and $\alpha = 1$.

Variogram estimation tested under 500 simulations from three sampling designs:

a Uniform

b Preferential ($\beta = 2$)

c Clustered

Classical Variogram Bias Under Preferential Sampling

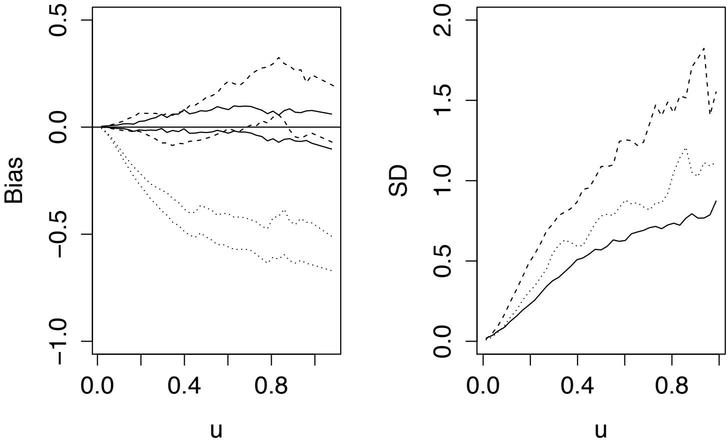


Figure: Variogram bias ± 2 standard errors and standard deviations for uniform (solid), clustered (dashed), and preferential (dotted) sampling schemes.

Classical Prediction Bias Under Preferential Sampling

Mod.	*Param.*	*Confidence intervals*		
		Uniform	*Preferential*	*Clustered*
1	Bias	(-0.03, 0.04)	(0.96, 1.13)	(-0.07, 0.06)
1	RMSE	(0.38, 0.42)	(1.31, 1.52)	(0.69, 0.79)
2	Bias	(-0.03, 0.02)	(-0.11, -0.06)	(-0.02, 0.04)
2	RMSE	(0.30, 0.32)	(0.32, 0.34)	(0.30, 0.32)

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 $\vec{S} = \{\vec{S}_0, \vec{S}_1\}$ on lattice $\vec{X}^* = (x_1^* \ldots x_N^*)'$.

- $\vec{S}_0$ are true values at data locations
- $\vec{S}_1$ are true values at other grid points

$$
\begin{aligned}
L(\vec{\theta}) &= \int \pi(\vec{Y}|\vec{X}, \vec{S})\pi(\vec{X}|\vec{S})\pi(\vec{S})\ d\vec{S} \\
&= \ldots \\
&= E_{\vec{S}|\vec{Y}}\left[\pi(\vec{X}|\vec{S})\frac{\pi(\vec{Y}|\vec{S}_0)}{\pi(\vec{S}_0|\vec{Y})}\pi(\vec{S}_0)\right] \\
&\approx m^{-1}\sum_{j=1}^{m}\pi(\vec{X}|\vec{S}_j)\frac{\pi(\vec{Y}|\vec{S}_{0j})}{\pi(\vec{S}_{0j}|\vec{Y})}\pi(\vec{S}_{0j})
\end{aligned}
$$

where $\vec{S}_j$ is the jth conditional simulation of S conditioned on $\vec{Y}$.

Preferential Model Likelihood

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

Steps for Monte Carlo Simulation:

1. Simulate $\vec{S} \sim \text{MVN}\left(\vec{0}, \Sigma\right)$ using Circulant Embedding (Wood and Chan, 1994)
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 $m^{-1} \sum_{j=1}^{m} \pi(\vec{X}|\vec{S}_j) \frac{\pi(\vec{Y}|\vec{S}_{0j})}{\pi(\vec{S}_{0j}|\vec{Y})} \pi(\vec{S}_{0j})$

Preferential Model Likelihood

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

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

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

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

Goodness of Fit

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

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

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

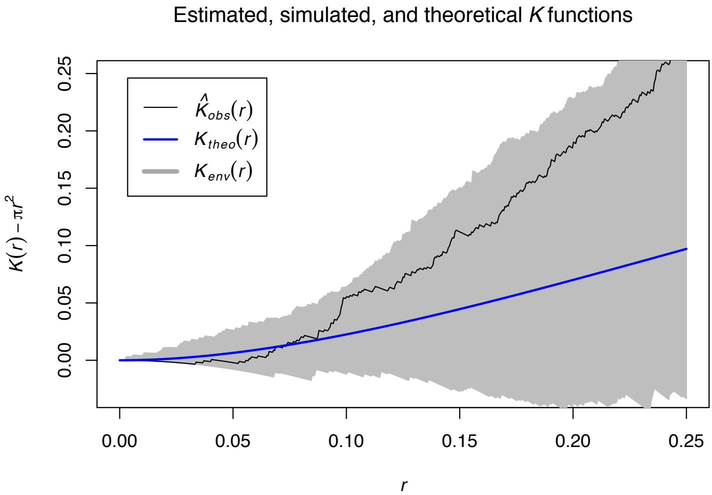
Then our p-value is the rank of T_1 out of $T_1, T_2, ..., T_n$ (*i.e.* if $n = 100$ and T_1 is largest test statistic, $p = .01$).

Goodness of Fit: Test Statistic

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

$K(s)$: Theoretical K-function under MLEs
$\hat{K}(s)$: Empirical K-function of simulation (or data)
$v(s)$: Variance of $\hat{K}(s)$

Goodness of Fit: Monte Carlo Testing



$p = 0.07$

Likelihood Fitting

- Use Nelder-Mead (Nelder and Mead, 1965) simplex algorithm for optimization with naive initial guesses
  - To simulate the $\vec{S}_j$'s quickly, I use cutoff embedding (Gneiting et al., 2012)
- Standard errors and correlations for MLEs inferred using quadratic fit of likelihood surface

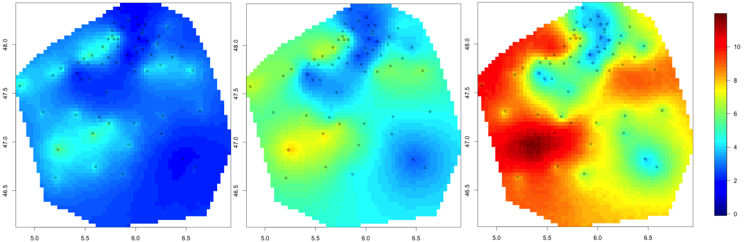
Likelihood Fit: Non-Preferential

	Estimate	*SE*	*Correlation matrix*				
μ_{97}	1.55	0.018	1	0.00	0.02	-0.00	-0.02
μ_{00}	0.73	0.014		1	0.10	0.20	0.07
σ	0.37	0.006			1	0.39	-0.37
ϕ	0.31	0.039				1	0.47
τ	0.23	0.005					1

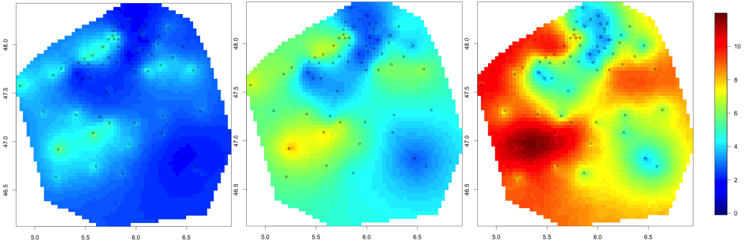
Table: Estimates are under joint 1997 and 2000 model

- Likelihood ratio test performed for joint versus separate models:
  - $p = 0.007$ under 3 degrees of freedom

1997 Predictions



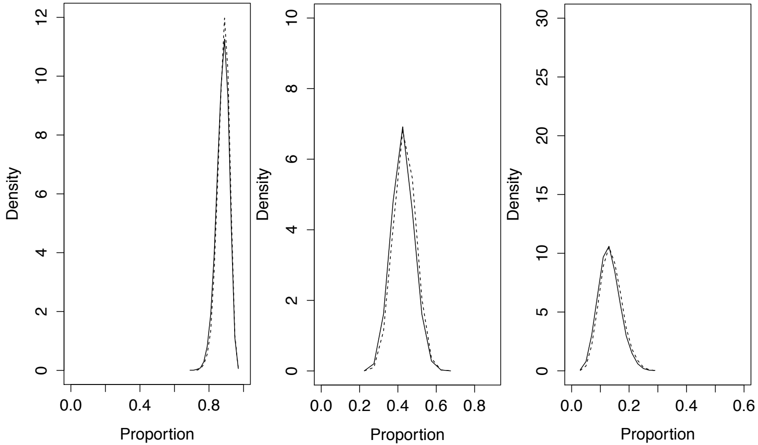
Preferential (MLEs from (Diggle et al., 2010)): ($\hat{\mu}_{97} = 1.515$, $\hat{\mu}_{00} = 0.762$, $\hat{\sigma}^2 = 0.138$, $\hat{\phi} = 0.313$, $\hat{\tau}^2 = 0.059$)



Non-preferential (MLEs fit with naive initial guess):

($\hat{\mu}_{97} = 1.551$, $\hat{\mu}_{00} = 0.727$, $\hat{\sigma}^2 = 0.136$, $\hat{\phi} = 0.305$, $\hat{\tau}^2 = 0.052$)

1997 Predictions



Areal proportion predicted over 3 (left), 5 (middle) and 7 (right) μ g/(g dry weight) for preferential (solid) and non-preferential (dashed) MLEs

Problems with the Methodology

- Using the joint model for 1997 predictions
- Assuming independence of 1997 and 2000 data in fitting
- No cross-validation performed
- No evaluation of preferential model in simulation study
- True preferential predictive distribution not used nor given
  - True predictive distribution may be Gaussian, although its parameters are unclear
- Non-preferential MLEs similar to preferential MLEs:
  - 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

- Accounting for preferentiality is important
- For empirical variogram estimation and classical prediction:
  - Uniform sampling performed best, then clustered, then preferential
  - For preferential data, naive variogram estimates and predictions were biased
- Proposed class of models is flexible and tractable for geostatistics with preferential data

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.

Gneiting, Tilmann, Hana Ševčíková, Donald B Percival, Martin Schlather, and Yindeng Jiang (2012), "Fast and exact simulation of large Gaussian lattice systems in 2: Exploring the limits." *Journal of Computational and Graphical Statistics*.

Nelder, John A and Roger Mead (1965), "A simplex method for function minimization." *The computer journal*, 7, 308–313.

Schlather, Martin, Alexander Malinowski, Marco Oesting, Daphne Boecker, Kirstin Strokorb, Sebastian Engelke, Johannes Marktini, Felix Ballani, Olga Moreva, Christoph Berreth, Peter Menck, Debastian Gross, Ulrike Ober, Katharina Burmeister, Juliane Manitz, Paulo Ribeiro, RIchard Singleton, Ben Pfaff, and R Core Team (2016), *RandomFields: Simulations and Analysis of Random Fields*. URL `https://cran.r-project.org/web/packages/RandomFields/RandomFields.pdf`. R package version 3.1.12.

Wood, Andrew TA and Grace Chan (1994), "Simulation of stationary Gaussian processes in [0, 1] d." *Journal of computational and graphical statistics*, 3, 409–432.

Appendix: Understanding the Preferential Likelihood

$$
\begin{aligned}
L(\theta) &= \int_S [\vec{X}|\vec{S}][\vec{Y}|\vec{S},\vec{X}]\frac{[\vec{S}|\vec{X},\vec{Y}]_{np}}{[\vec{S}|\vec{X},\vec{Y}]_{np}}[\vec{S}]\ d\vec{S} \\
&= \int_S [\vec{X}|\vec{S}][\vec{Y}|\vec{S},\vec{X}]\frac{[\vec{S}|\vec{X},\vec{Y}]_{np}}{[\vec{S}_1|\vec{S}_0,\vec{X},\vec{Y}]_{np}[\vec{S}_0|\vec{X},\vec{Y}]_{np}}[\vec{S}_1|\vec{S}_0][\vec{S}_0]\ d\vec{S} \\
&= \int_S [\vec{X}|\vec{S}]\frac{[\vec{Y}|\vec{S},\vec{X}]}{[\vec{S}_1|\vec{S}_0,\vec{X},\vec{Y}]_{np}[\vec{S}_0|\vec{X},\vec{Y}]_{np}}[\vec{S}_1|\vec{S}_0][\vec{S}_0][\vec{S}|\vec{X},\vec{Y}]_{np}\ d\vec{S} \\
&= \int_S [\vec{X}|\vec{S}]\frac{[\vec{Y}|\vec{S}_0,\vec{X}]}{[\vec{S}_1|\vec{S}_0,\vec{X}]_{np}[\vec{S}_0|\vec{X},\vec{Y}]_{np}}[\vec{S}_1|\vec{S}_0][\vec{S}_0][\vec{S}|\vec{X},\vec{Y}]_{np}\ d\vec{S} \\
&= \int_S [\vec{X}|\vec{S}]\frac{[\vec{Y}|\vec{S}_0,\vec{X}]}{[\vec{S}_1|\vec{S}_0,\vec{X}]_{np}[\vec{S}_0|\vec{X},\vec{Y}]_{np}}[\vec{S}_1|\vec{S}_0][\vec{S}_0][\vec{S}|\vec{X},\vec{Y}]_{np}\ d\vec{S} \\
&= E_{[S|\vec{X},\vec{Y}]_{np}}\left[[\vec{X}|\vec{S}]\frac{[\vec{Y}|\vec{S}_0,\vec{X}]}{[\vec{S}_0|\vec{X},\vec{Y}]_{np}}[\vec{S}_0]\right]
\end{aligned}
$$