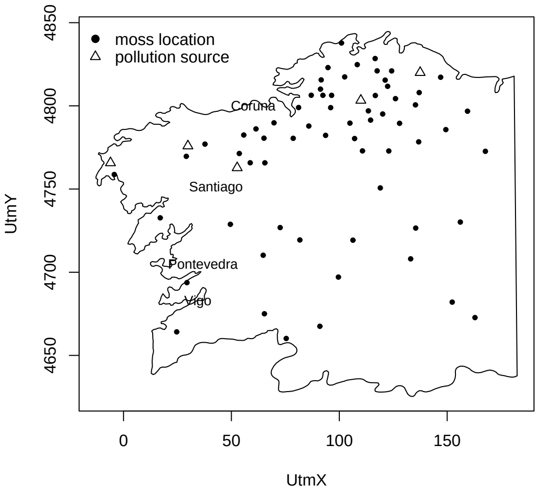
An application to Galicia pollution data and a model for preferential sampling

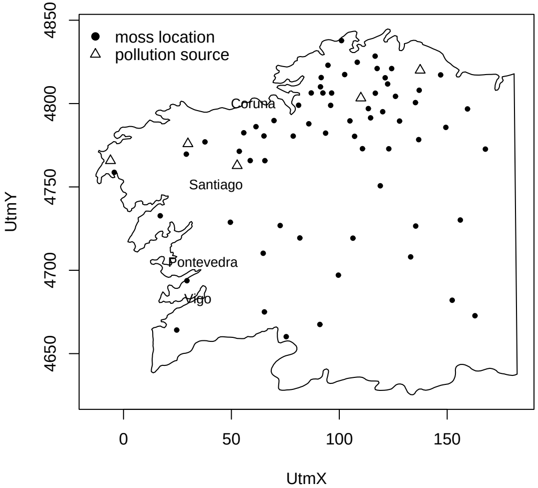
Raquel Menezes, Peter Diggle, M.Febrero-Bande, P.Garcia-Soidán

October 2005

1995 Galicia pollution data - Cr, Ni and Pb



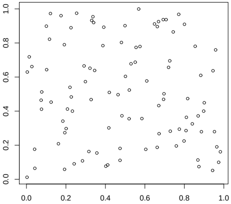
1995 Galicia pollution data - Cr, Ni and Pb



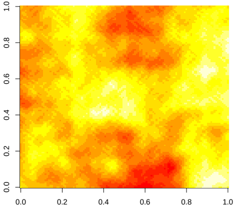
Formal tests of Monte Carlo point to rejection of $H0$: non-PS for Ni and Pb.

The traditional geostatistical model

observed data $(y, \mathbf{x})$



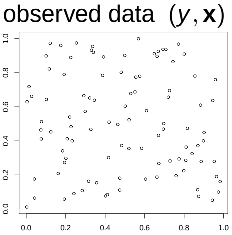
unobserved data s

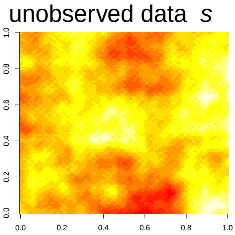


$\mathbf{x_1}, \dots, \mathbf{x_n}$ are locations within an observation region $D \subset R^2$

$y_1, \dots, y_n$ are measurements associated with these locations

The traditional geostatistical model





$\mathbf{x_1}, \ldots, \mathbf{x_n}$ are locations within an observation region $D \subset R^2$

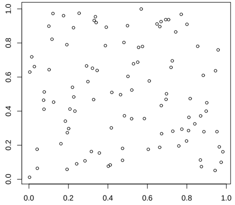
$y_1, \ldots, y_n$ are measurements associated with these locations

So, consider next 3 stochastic processes:

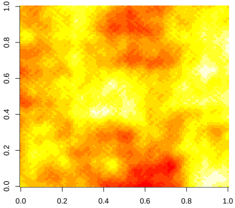
1. Field $S(\mathbf{x}) : \mathbf{x} \in D \subset \boldsymbol{R}^2$ (goal of prediction)

The traditional geostatistical model

observed data $(y, \mathbf{x})$



unobserved data s



$\mathbf{x_1}, \ldots, \mathbf{x_n}$ are locations within an observation region $D \subset R^2$

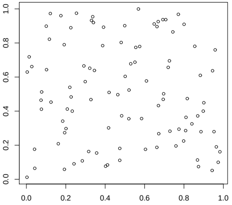
$y_1, \ldots, y_n$ are measurements associated with these locations

So, consider next 3 stochastic processes:

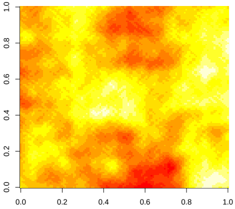
1. Field $S(\mathbf{x}) : \mathbf{x} \in D \subset \boldsymbol{R}^2$ (goal of prediction)
2. Point process P of sample locations $\mathbf{x}$

The traditional geostatistical model

observed data $(y, \mathbf{x})$



unobserved data s



$\mathbf{x_1}, \ldots, \mathbf{x_n}$ are locations within an observation region $D \subset R^2$

$y_1, \ldots, y_n$ are measurements associated with these locations

So, consider next 3 stochastic processes:

1. Field $S(\mathbf{x}) : \mathbf{x} \in D \subset \boldsymbol{R}^2$ (goal of prediction)
2. Point process P of sample locations $\mathbf{x}$
3. Measurement process $Y \equiv Y(\mathbf{x_i})$ (noisy version of S)

Motivation

Geostatistical methods rely on the fundamental assumption that the sampling points have been chosen independently of the values of the spatial variable. (Diggle et al, 2003)

Motivation

Geostatistical methods rely on the fundamental assumption that the sampling points have been chosen independently of the values of the spatial variable. (Diggle et al, 2003)

Additionally, most methods consider the sample locations as being uniformly spread over the observed region.

Motivation

> *Geostatistical methods rely on the fundamental assumption that the sampling points have been chosen independently of the values of the spatial variable.* (Diggle et al, 2003)

Additionally, most methods consider the sample locations as being uniformly spread over the observed region.

What happens if these assumptions fail?

- preferability issue
- clustering issue

A variogram robust to clusters

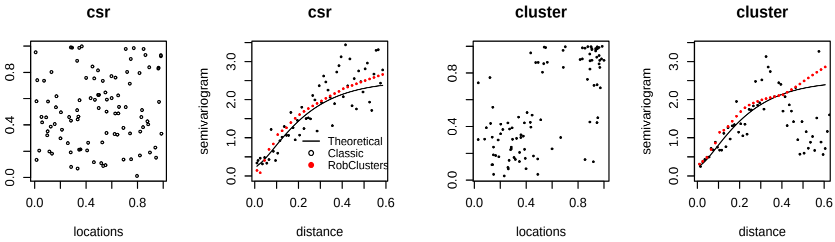
$$2\widehat{\gamma}(u) = \frac{\sum_i \sum_j w_{ij}(u)[Y(\mathbf{x_i}) - Y(\mathbf{x_j})]^2}{\sum_i \sum_j w_{ij}(u)}$$

- $w_{ij}(u) = \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right)$, where $n_i = \sum_k I_{\{\|\mathbf{x_i} - \mathbf{x_k}\| \leq \delta\}}$
- $w_{ij}(u) = 1$ (classic estimator)

A variogram robust to clusters

$$2\widehat{\gamma}(u) = \frac{\sum_i \sum_j w_{ij}(u)[Y(\mathbf{x_i}) - Y(\mathbf{x_j})]^2}{\sum_i \sum_j w_{ij}(u)}$$

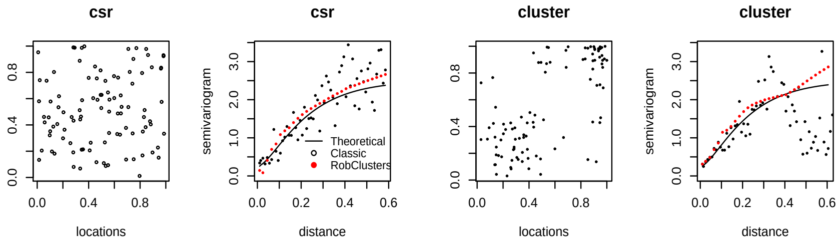
- $w_{ij}(u) = \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right)$, where $n_i = \sum_k I_{\{\|\mathbf{x_i} - \mathbf{x_k}\| \leq \delta\}}$
- $w_{ij}(u) = 1$ (classic estimator)



A variogram robust to clusters

$$2\widehat{\gamma}(u) = \frac{\sum_i \sum_j w_{ij}(u)[Y(\mathbf{x_i}) - Y(\mathbf{x_j})]^2}{\sum_i \sum_j w_{ij}(u)}$$

- $w_{ij}(u) = \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right)$, where $n_i = \sum_k I_{\{\|\mathbf{x_i} - \mathbf{x_k}\| \leq \delta\}}$
- $w_{ij}(u) = 1$ (classic estimator)



This new variogram is proven to enjoy good properties, such as asymptotic unbiasedness and consistency.

Model-based approach for preferential sampling

Stochastic dependence between S and $P \Longrightarrow [S, P] \neq [S][P]$

Log-Gaussian Cox processes – $[P \mid S]$

Model-based approach for preferential sampling

Stochastic dependence between S and $P \implies [S, P] \neq [S][P]$

Log-Gaussian Cox processes – $[P \mid S]$

1. Cox process – *Diggle(2003)*:
   Useful to model aggregated spatial point patterns where the aggregation is due to a stochastic environmental heterogeneity.
2. Log-Gaussian process – *Møller et al.(1998)*:
   Takes $\log \Lambda(\mathbf{x}) = Z(\mathbf{x})$, where $Z(\mathbf{x})$ is some Gaussian process.

Model-based approach for preferential sampling

Stochastic dependence between S and $P \Longrightarrow [S, P] \neq [S][P]$

Log-Gaussian Cox processes – $[P \mid S]$

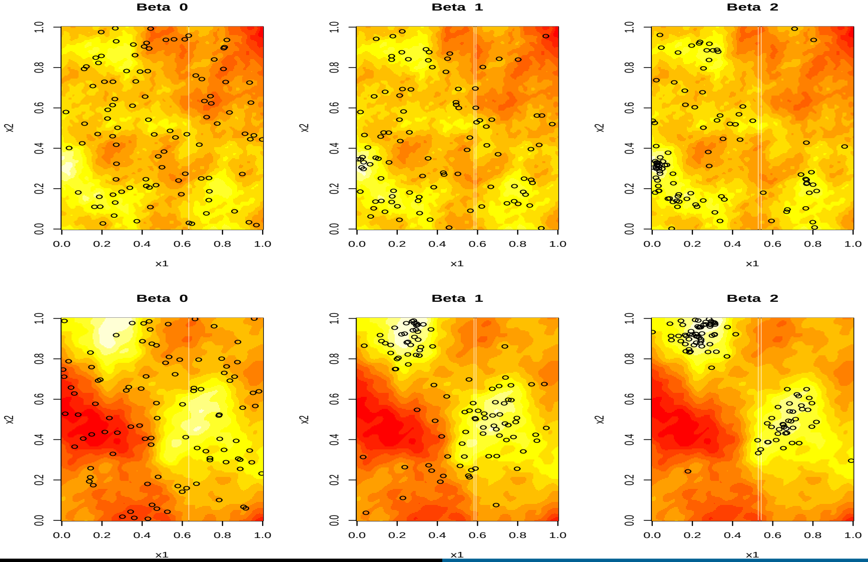
1. Cox process – *Diggle(2003)*:
   Useful to model aggregated spatial point patterns where the aggregation is due to a stochastic environmental heterogeneity.
2. Log-Gaussian process – *Møller et al.(1998)*:
   Takes $\log \Lambda(\mathbf{x}) = Z(\mathbf{x})$, where $Z(\mathbf{x})$ is some Gaussian process.

$$P \mid S \sim \mathit{Poisson}\,(\exp\{\alpha + \beta\, S(\mathbf{x})\})$$

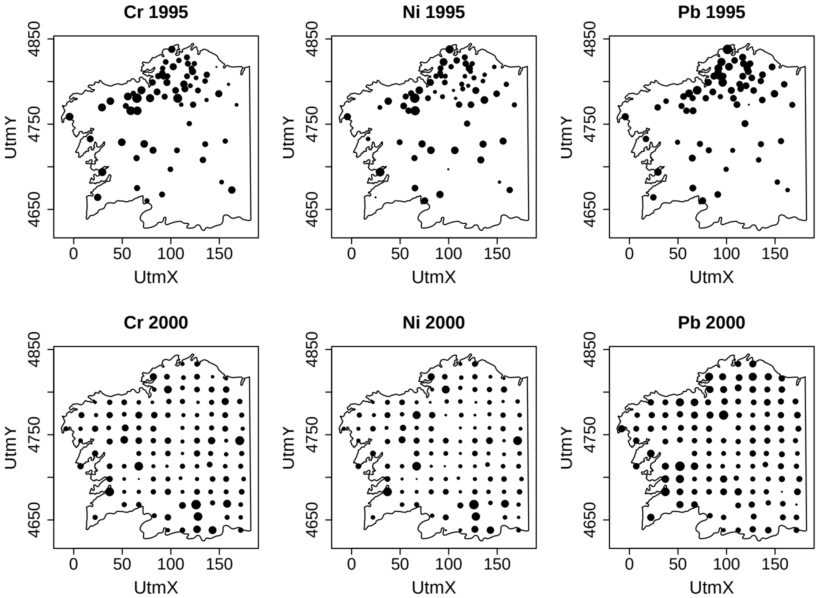
where $|\beta|$ identifies the *degree of "preferability"*, i.e., it measures how much the "intensity" depends on S

Influence of β on sample locations ($\alpha = 0$)

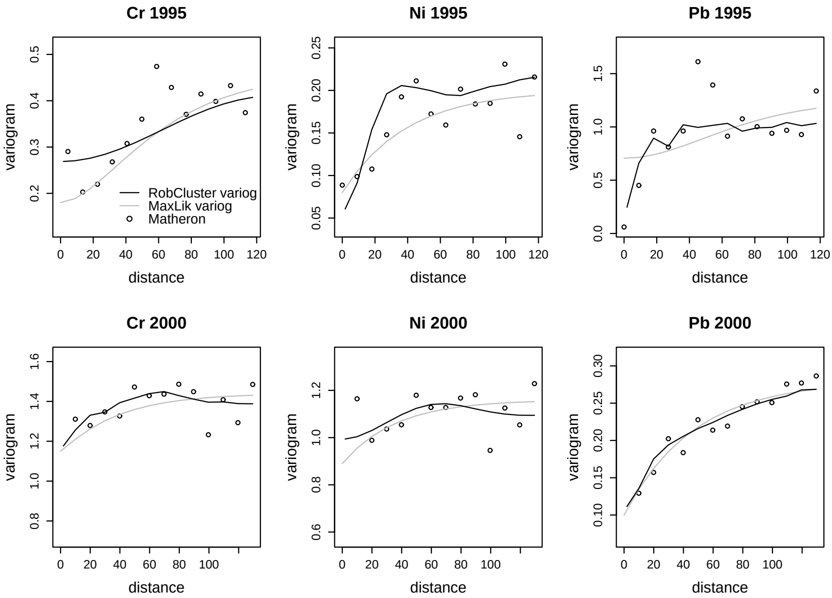
Given two distinct Gaussian fields $S(.)$



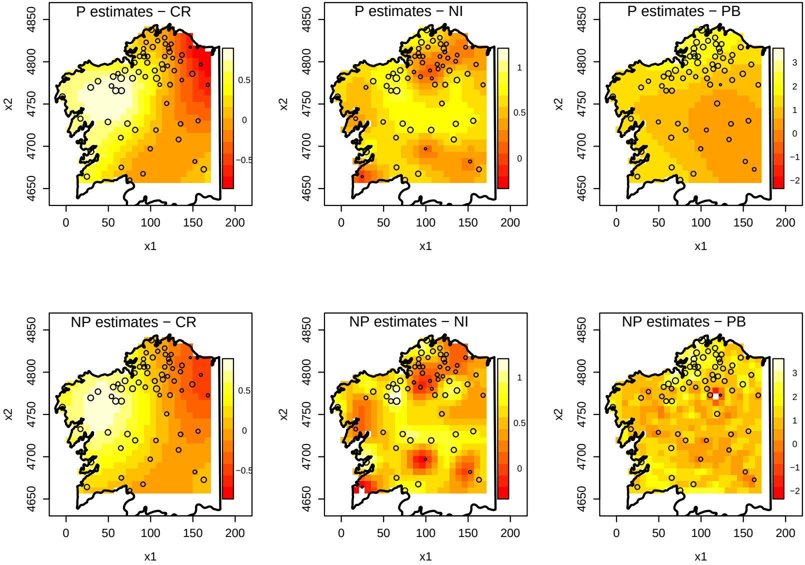
1995 versus 2000 Galicia pollution data



Variogram estimation for Cr, Ni and Pb data



1995 kriging — Parametric versus Non-Parametric



Cross-validation or *leave-one-out* method

Aims of CV:

- compare the P and the NP variogram models;
- inspect the effect of sampling procedure (1995 vs 2000).

Cross-validation or *leave-one-out* method

Aims of CV:

- compare the P and the NP variogram models;
- inspect the effect of sampling procedure (1995 vs 2000).

Fundamental idea behind CV:

- estimate the concentration measurement $Y(\mathbf{x})$ at each sample point $\mathbf{x_i}$ from neighbouring data $Y_j = Y(\mathbf{x_j}),\ j \neq i,$ as if $Y_i = Y(\mathbf{x_i})$ were unknown.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{-i})^2 \quad \text{and} \quad MSSE = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{Y}_{-i})^2}{\widehat{\sigma^2_{-i}}}$$

Main conclusions

		MSE			MSSE		
		Cr	Ni	Pb	Cr	Ni	Pb
1995	P	0.265	0.187	0.798	1.167	1.292	0.996
	NP	0.247	0.172	0.801	0.837	1.305	1.032
2000	P	1.738	1.331	0.217	1.036	1.028	1.012
	NP	1.672	1.276	0.217	0.974	0.978	0.986

With respect to MSE, smaller values are normally associated with the NP approach, reinforcing the advantages of the NP variogram.

Note that MSE is sensible to data poorly *explained* by their neighbors, which possibly explains higher values for Cr and Ni in 2000.

Main conclusions

		MSE			MSSE		
		Cr	Ni	Pb	Cr	Ni	Pb
1995	P	0.265	0.187	0.798	1.167	1.292	0.996
	NP	0.247	0.172	0.801	0.837	1.305	1.032
2000	P	1.738	1.331	0.217	1.036	1.028	1.012
	NP	1.672	1.276	0.217	0.974	0.978	0.986

With respect to MSE, smaller values are normally associated with the NP approach, reinforcing the advantages of the NP variogram.

Note that MSE is sensible to data poorly *explained* by their neighbors, which possibly explains higher values for Cr and Ni in 2000.

With respect to MSSE, note that these values are closer to 1 in 2000. We think such results, mainly those for Ni, support the need for a model-based approach to preferential sampling.