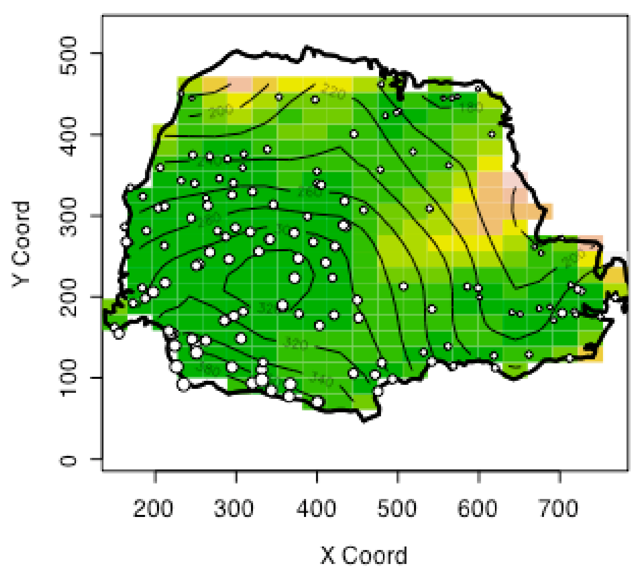
1.2 Kriging



Research goals in air quality research

- Calculate air pollution fields for health effect studies
- Assess deterministic air quality models against data
- Interpret and set air quality standards
- Improved understanding of complicated systems
- Prediction of air quality

The geostatistical model

Gaussian process $Z(s), s \in D \subseteq R^2$

$\mu(s) = EZ(s)$ Var $Z(s) < \infty$

Z is strictly stationary if

$$(Z(s_1),\ldots,Z(s_k)) \stackrel{d}{=} (Z(s_1+h),\ldots,Z(s_k+h))$$

Z is weakly stationary if

$$\mu(s) \equiv \mu \quad \text{Cov}(Z(s_1),Z(s_2)) = C(s_1 - s_2)$$

Z is isotropic if weakly stationary and

$$C(s_1 - s_2) = C_0(\|s_1 - s_2\|)$$

The problem

Given observations at n locations
$Z(s_1),\ldots,Z(s_n)$
estimate

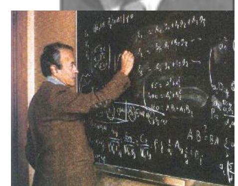
$Z(s_0)$ (the process at an unobserved site)

or $\int_A Z(s)d\nu(s)$ (an average of the process)

In the environmental context often time series of observations at the locations.

Some history

Regression (Galton, Bartlett)



Mining engineers (Krige 1951, Matheron, 60s)

Spatial models (Whittle, 1954)

Forestry (Matérn, 1960)

Objective analysis (Grandin, 1961)

More recent work Cressie (1993), Stein (1999)

A Gaussian formula

If $$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \right)$$

then $$(Y \mid X) \sim N(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X),$$
$$\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})$$

Simple kriging

Let X = $(Z(s_1),\ldots,Z(s_n))^T$, Y = $Z(s_0)$, so that

$\mu_X=\mu 1_n$, $\mu_Y=\mu$,

$\Sigma_{XX}=[C(s_i-s_j)]$, $\Sigma_{YY}=C(0)$, and

$\Sigma_{YX}=[C(s_i-s_0)]$.

Then

$$p(X) \equiv \hat{Z}(s_0) = \mu + \left[C(s_i - s_0)\right]^T \left[C(s_i - s_j)\right]^{-1} \left(X - \mu 1_n\right)$$

This is the best unbiased linear predictor when μ and C are known (simple kriging).

The prediction variance is

$$m_1 = C(0) - \left[C(s_i - s_0)\right]^T \left[C(s_i - s_j)\right]^{-1} \left[C(s_i - s_0)\right]$$

Some variants

Ordinary kriging (unknown μ)

$$p(X) \equiv \hat{Z}(s_0) = \hat{\mu} + \left[C(s_i - s_0)\right]^T \left[C(s_i - s_j)\right]^{-1} \left(X - \hat{\mu} 1_n\right)$$

where

$$\hat{\mu} = \left(1_n^T \left[C(s_i - s_j)\right]^{-1} 1_n\right)^{-1} 1_n^T \left[C(s_i - s_j)\right]^{-1} X$$

Universal kriging ($\mu(s)$=A(s)β for some spatial variable A)

$$\hat{\beta} = (\left[A(s_i)\right]^T \left[C(s_i - s_j)\right]^{-1} \left[A(s_i)\right])^{-1}$$

$$\left[A(s_i)\right]^T \left[C(s_i - s_j)\right]^{-1} X$$

Still optimal for known C.

Universal kriging variance

$$E\left(\hat{Z}(s_0) - Z(s_0)\right)^2 = \boxed{m_1} +$$

simple kriging variance

$$\left(A(s_0) - [A(s_i)^T[C(s_i - s_j)]^{-1}[C(s_i - s_0)]\right)^T$$
$$\times\left(\left[A(s_i)\right]^T\left[C(s_i - s_j)\right]^{-1}\left[A(s_i)\right]\right)^{-1}$$
$$\times\left(A(s_0) - [A(s_i)^T[C(s_i - s_j)]^{-1}[C(s_i - s_0)]\right)$$

variability due to estimating β

The (semi)variogram

$$\gamma(\|h\|) = \frac{1}{2}\text{Var}(Z(s+h) - Z(s)) = C(0) - C(\|h\|)$$

Intrinsic stationarity

Weaker assumption (C(0) needs not exist)

Kriging predictions can be expressed in terms of the variogram instead of the covariance.

Ordinary kriging

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

where

$$\lambda^T = \left(\gamma + 1 \frac{1 - 1^T \Gamma^{-1} \gamma}{1^T \Gamma^{-1} 1} \right)^T \Gamma^{-1}$$

$$\gamma = (\gamma(s_0 - s_1), \ldots, \gamma(s_0 - s_n))^T$$

$$\Gamma_{ij} = \gamma(s_i - s_j)$$

and kriging variance

$$m_1(s_0) = 2\sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j)$$

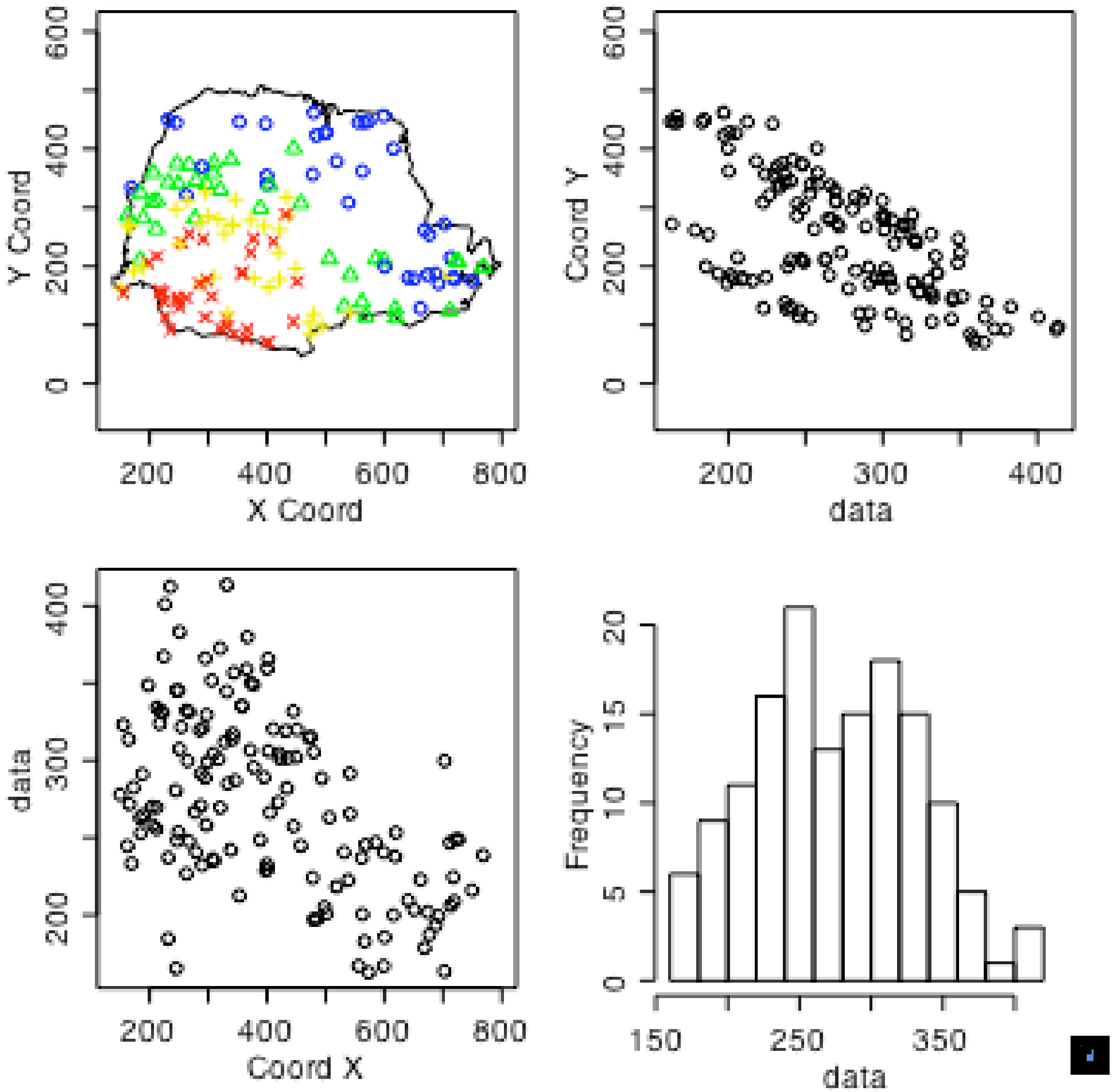
Parana data

Built-in geoR data set

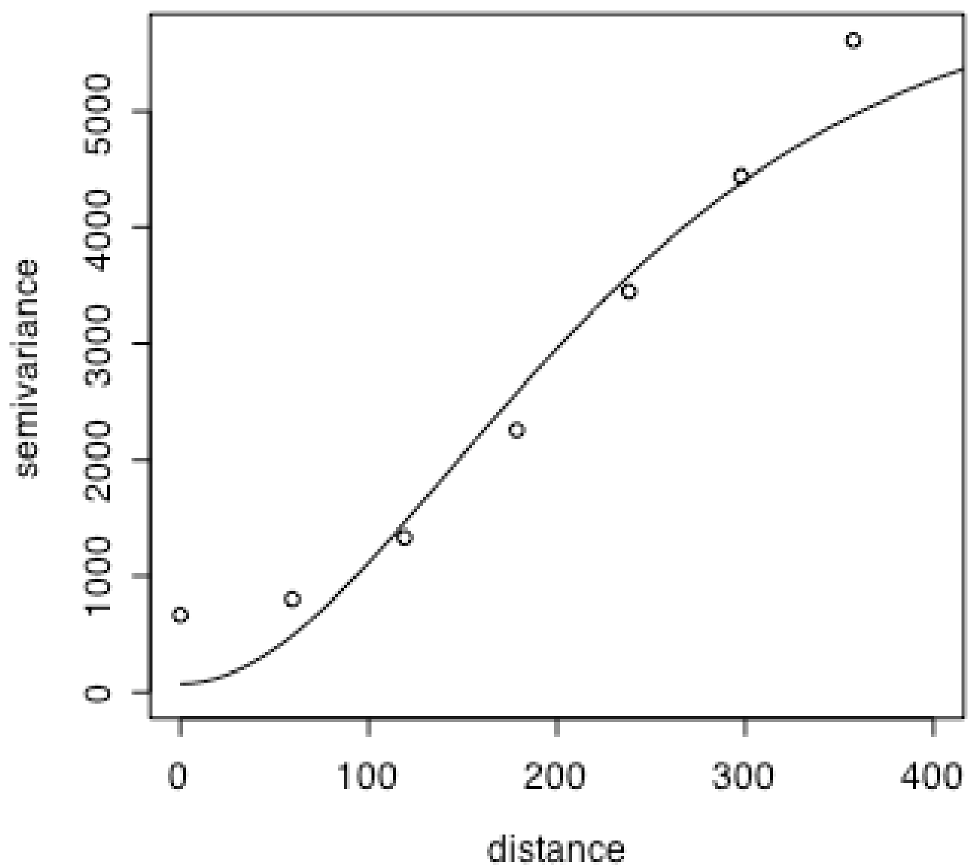
Average rainfall over different years for May-June (dry-season)

143 recording stations throughout Parana State, Brazil

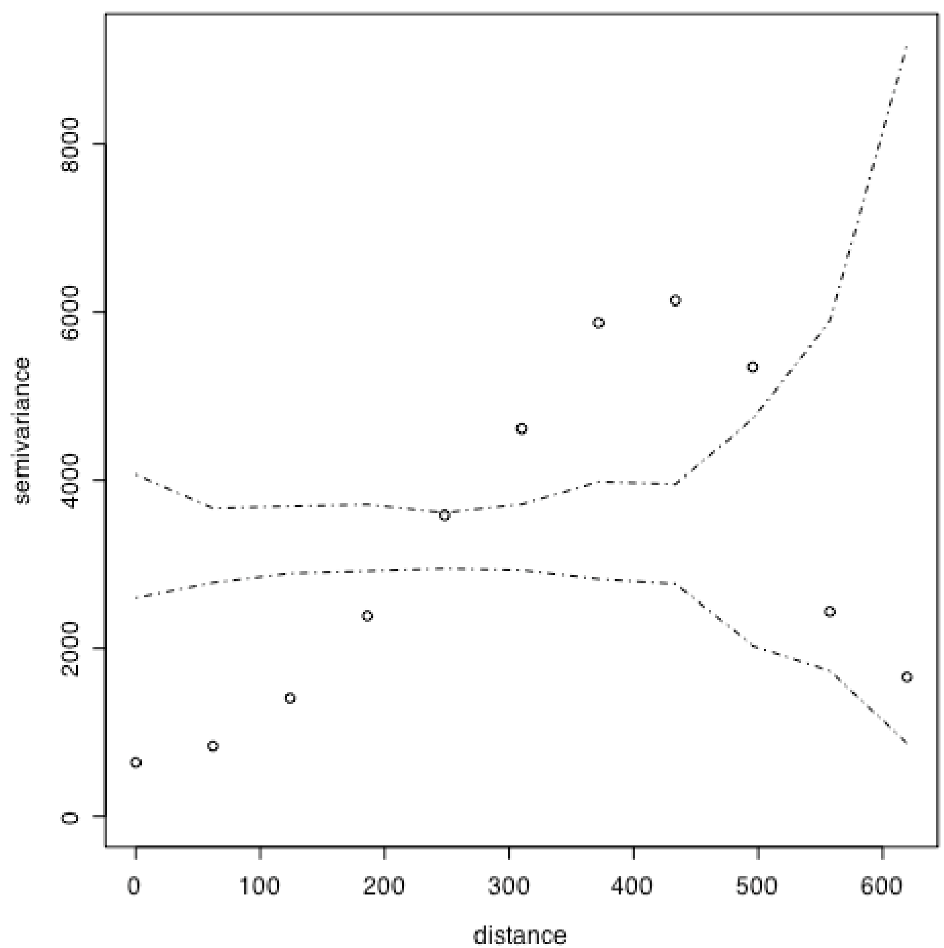
Parana precipitation



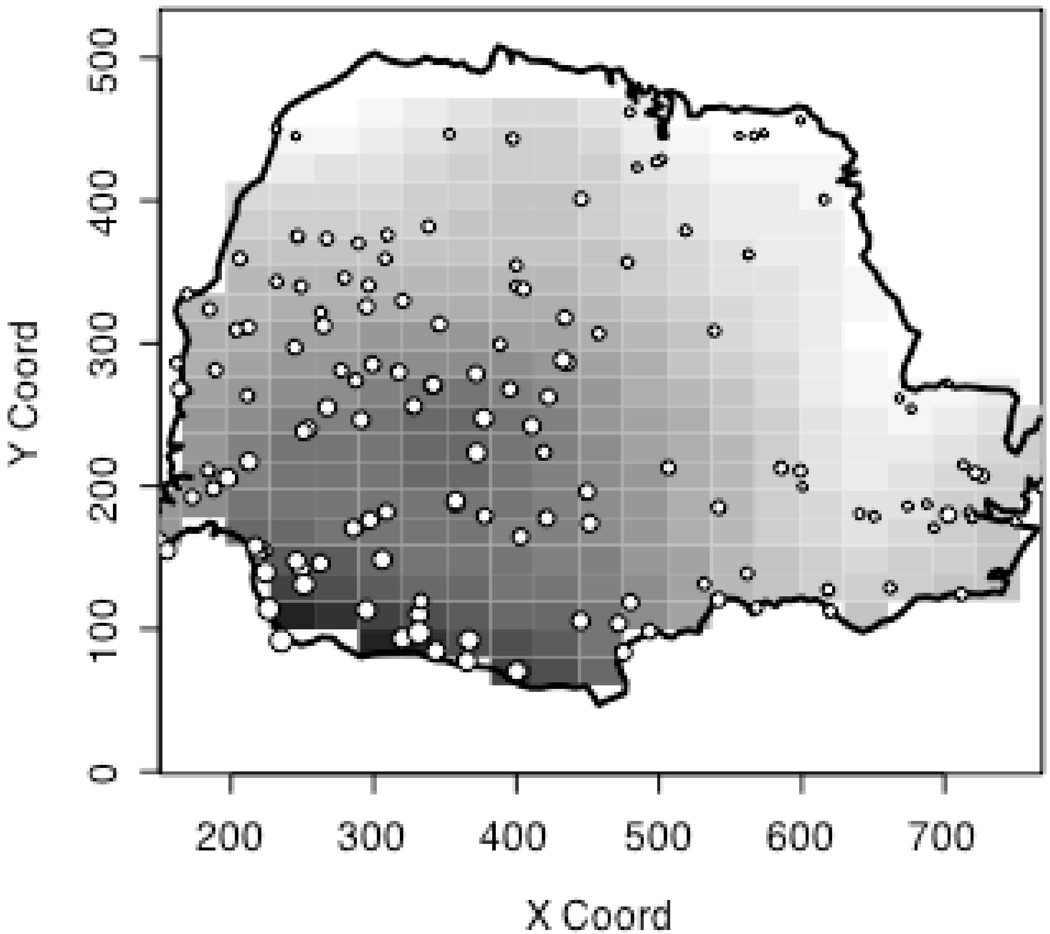
Fitted variogram



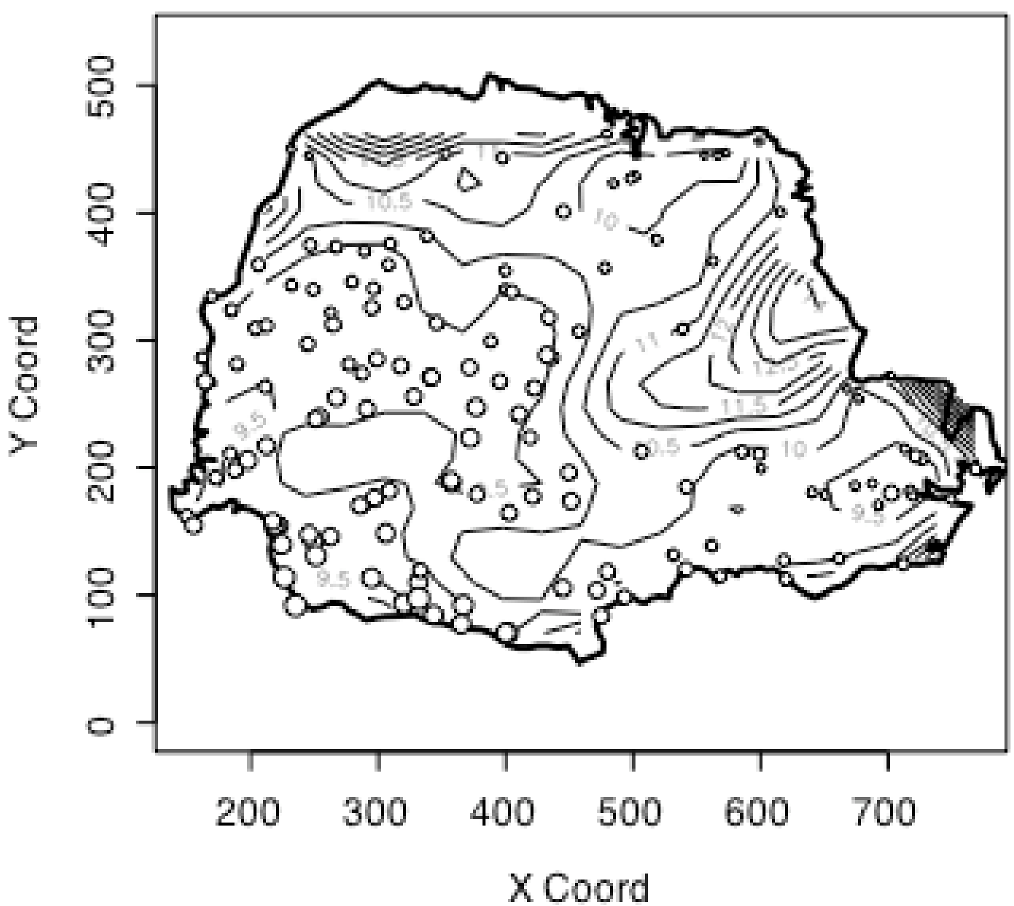
Is it significant?



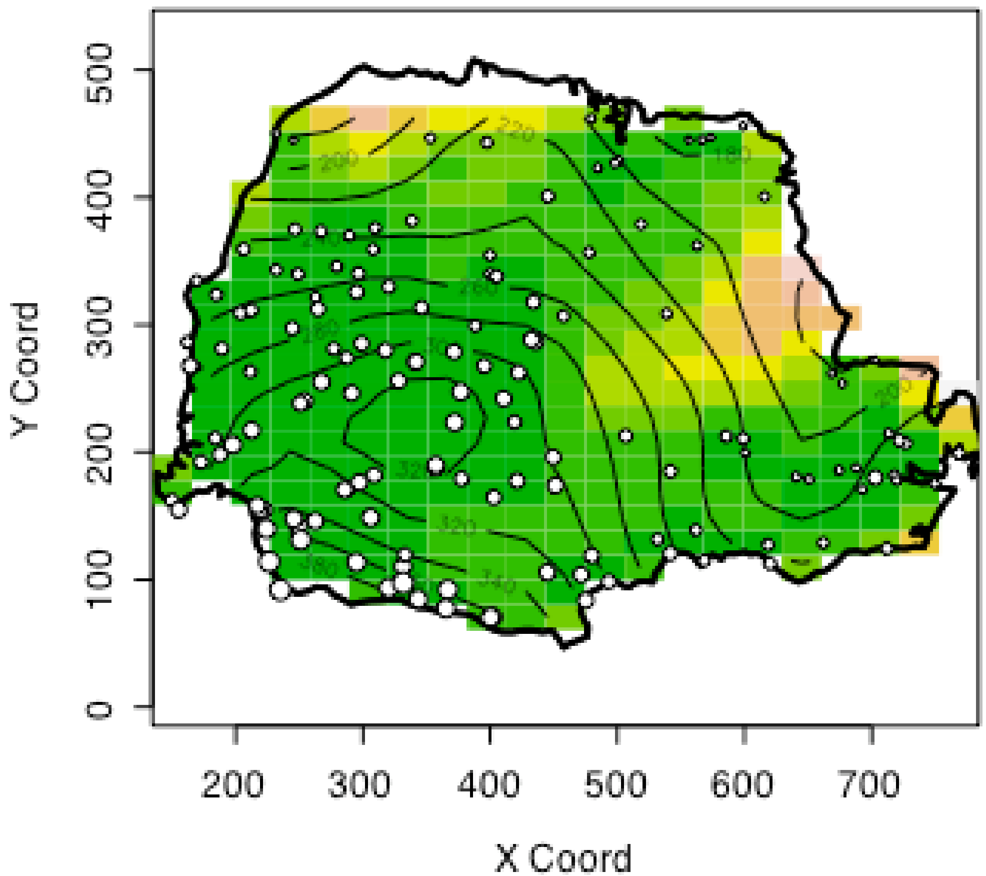
Kriging surface



Kriging standard error



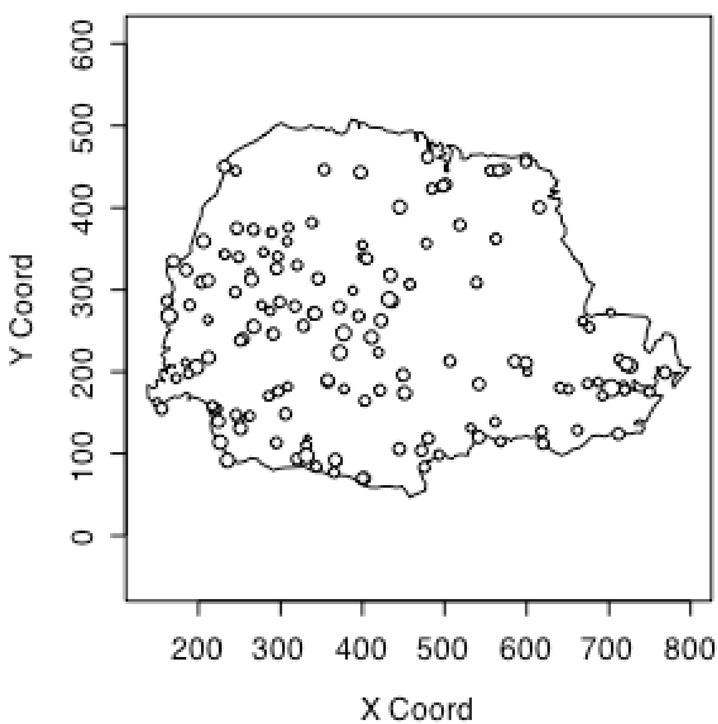
A better combination



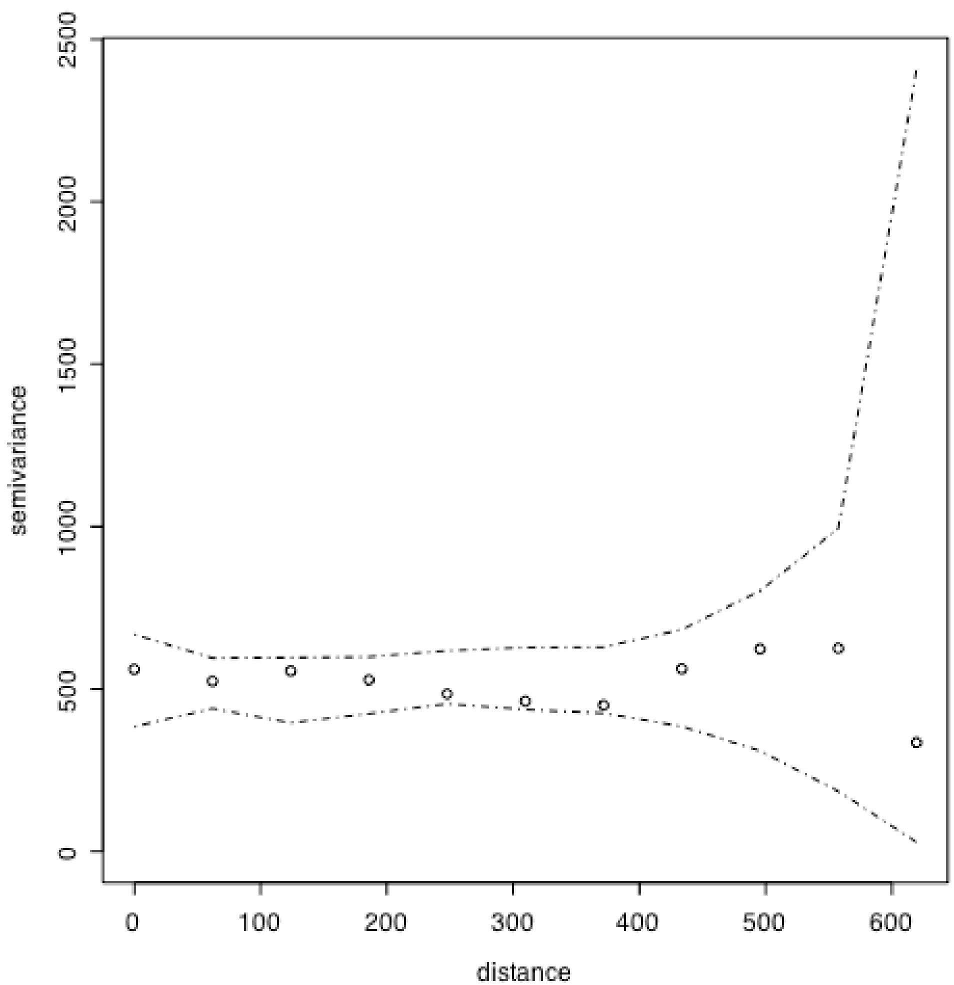
Spatial trend

Indication of spatial trend

Fit quadratic in coordinates



Residual variogram



Effect of estimated covariance structure

The usual geostatistical method is to consider the covariance known. When it is estimated

- the predictor $p_2(X) = p(X;\hat{\theta}(X))$ is not linear
- nor is it optimal
- the "plug-in" estimate $m_1(\hat{\theta}(X))$ of the variability often has too low mean

Let $m_2(\theta) = E_\theta(p_2(X) - \mu)^2$. Is $m_1(\hat{\theta})$ a good estimate of $m_2(\theta)$?

Some results

1. Under Gaussianity, $m_2(\theta) \geq m_1(\theta)$ with equality iff $p_2(X)=p(X;\theta)$ a.s.
2. Under Gaussianity, if $\hat{\theta}$ is sufficient, and if the covariance is linear in θ, then

$$E_\theta m_1(\hat{\theta}) = m_2(\theta) - 2(m_2(\theta) - m_1(\theta))$$

3. An unbiased estimator of $m_2(\theta)$ is

$$2\hat{m} - m_1(\hat{\theta})$$

where $\hat{m}$ is an unbiased estimator of $m_1(\theta)$.

Better prediction variance estimator

(Zimmerman and Cressie, 1992)

$$\mathrm{Var}(\hat{Z}(s_0;\hat{\theta})) \approx m_1(\hat{\theta}) + 2\,\mathrm{tr}\left[\mathrm{cov}(\hat{\theta}) \cdot \mathrm{cov}(\nabla\hat{Z}(s_0;\hat{\theta})\right]$$

(Taylor expansion; often approx. unbiased)

A Bayesian prediction analysis takes account of all sources of variability (Le and Zidek, 1992; 2006)

Some references

N. Cressie (1993) *Statistics for Spatial Data.* Rev. ed. Wiley. Pp. 105-112,119-123, 151-157.

Zimmerman, D. L. and Cressie, N (1992): Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics* 44: 27-43.

Le N. D. and Zidek J. V. (1992) Interpolation With Uncertain Spatial Covariances–A Bayesian Alternative To Kriging. 43 (2): 351-374.

N. D. Le and J. V. Zidek (2006) *Statistical Analysis of Environmental Space-Time Processes.* Springer-Verlag.