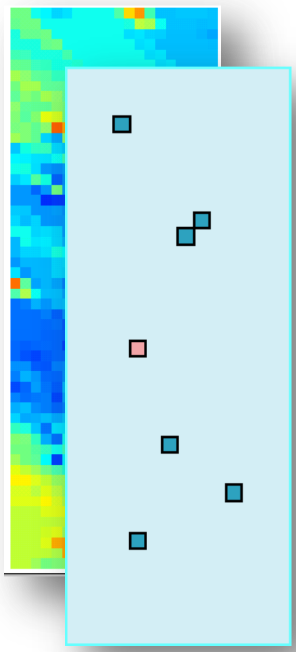
CERENA

GEOSTATISTICS AND APPLICATIONS

Maria João Pereira
maria.pereira@ist.utl.pt

Need for a model of reality



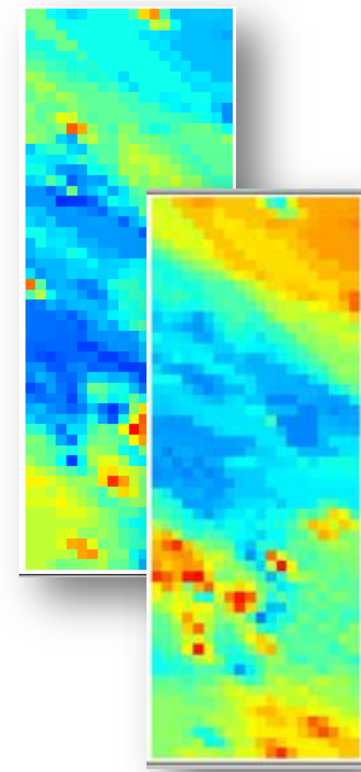
GEOESTATISTICS

Estimation Models

Simulation Models

Sampling Optimization

...



Derteministic model

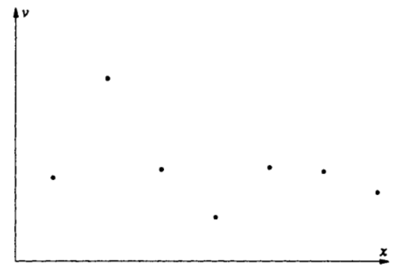
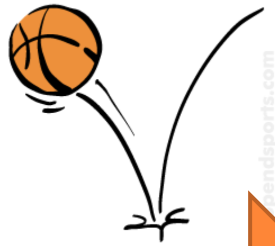


Figure 9.1 An example of an estimation problem. The dots represent seven sample points on a profile to be estimated.



Deterministic model

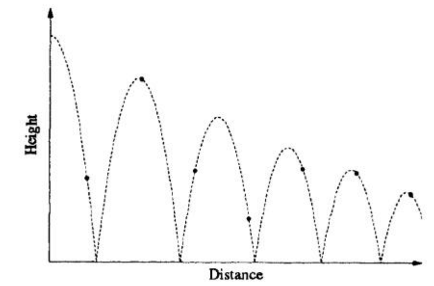
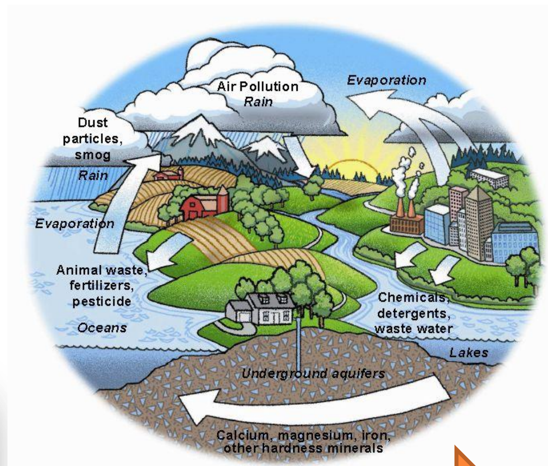


Figure 9.2 With the seven sample points shown in Figure 9.1 viewed as heights of a bouncing ball, the dashed curve shows a deterministic model of the heights at unsampled locations.

Complete knowledge of the model that describes the process

No uncertainty in predictions

Probabilistic model



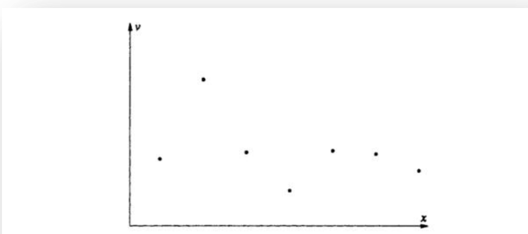


Figure 9.1 An example of an estimation problem. The dots represent seven sample points on a profile to be estimated.

Complex process

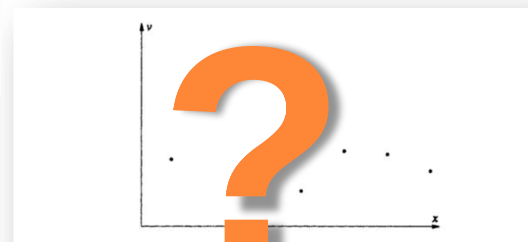
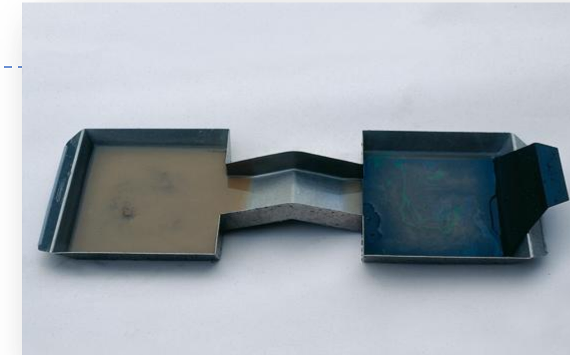


Figure 9.1 An example of an estimation problem. The dots represent seven sample points on a profile to be estimated.

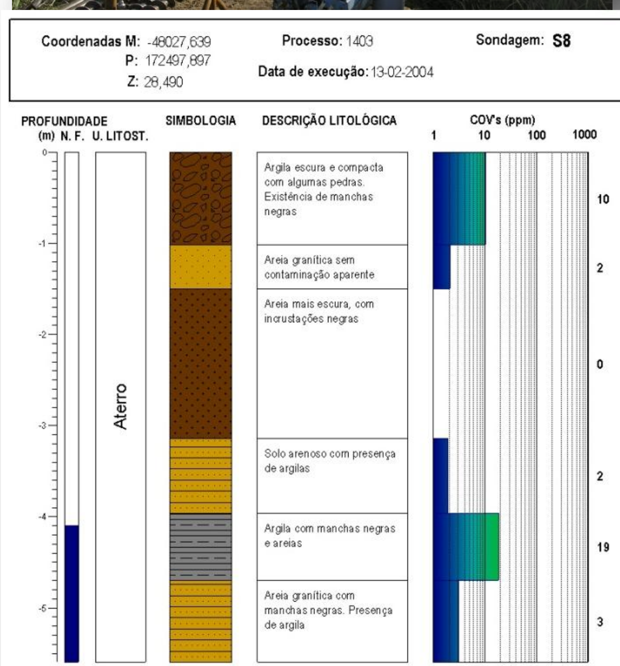
The model that describes the process is imcomplete or unknown

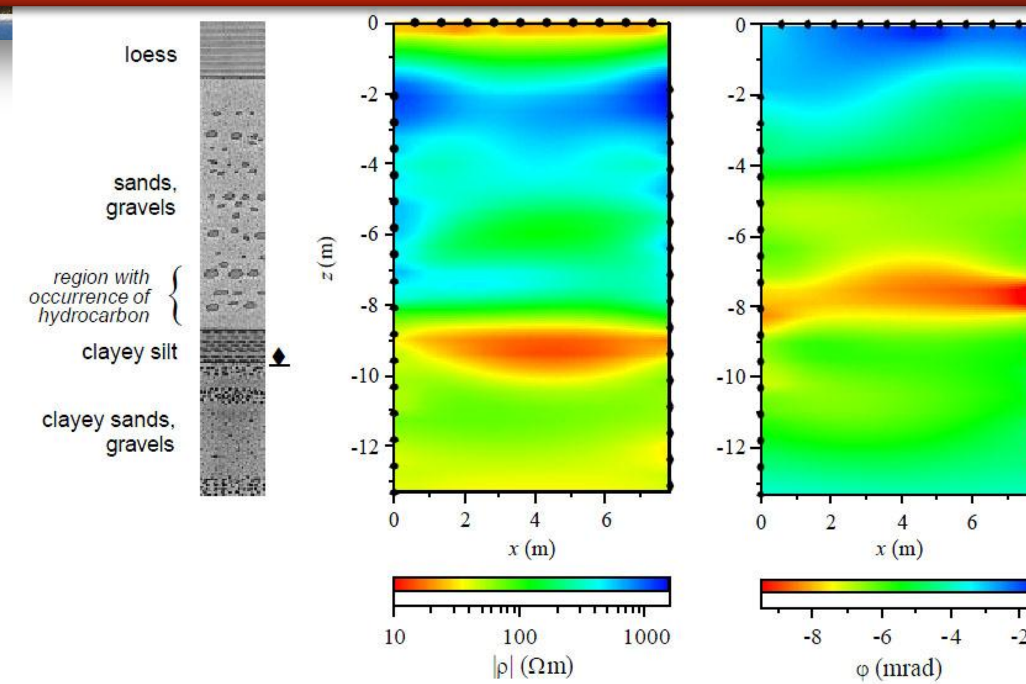
Uncertainty in predictions

Why geostatistics?

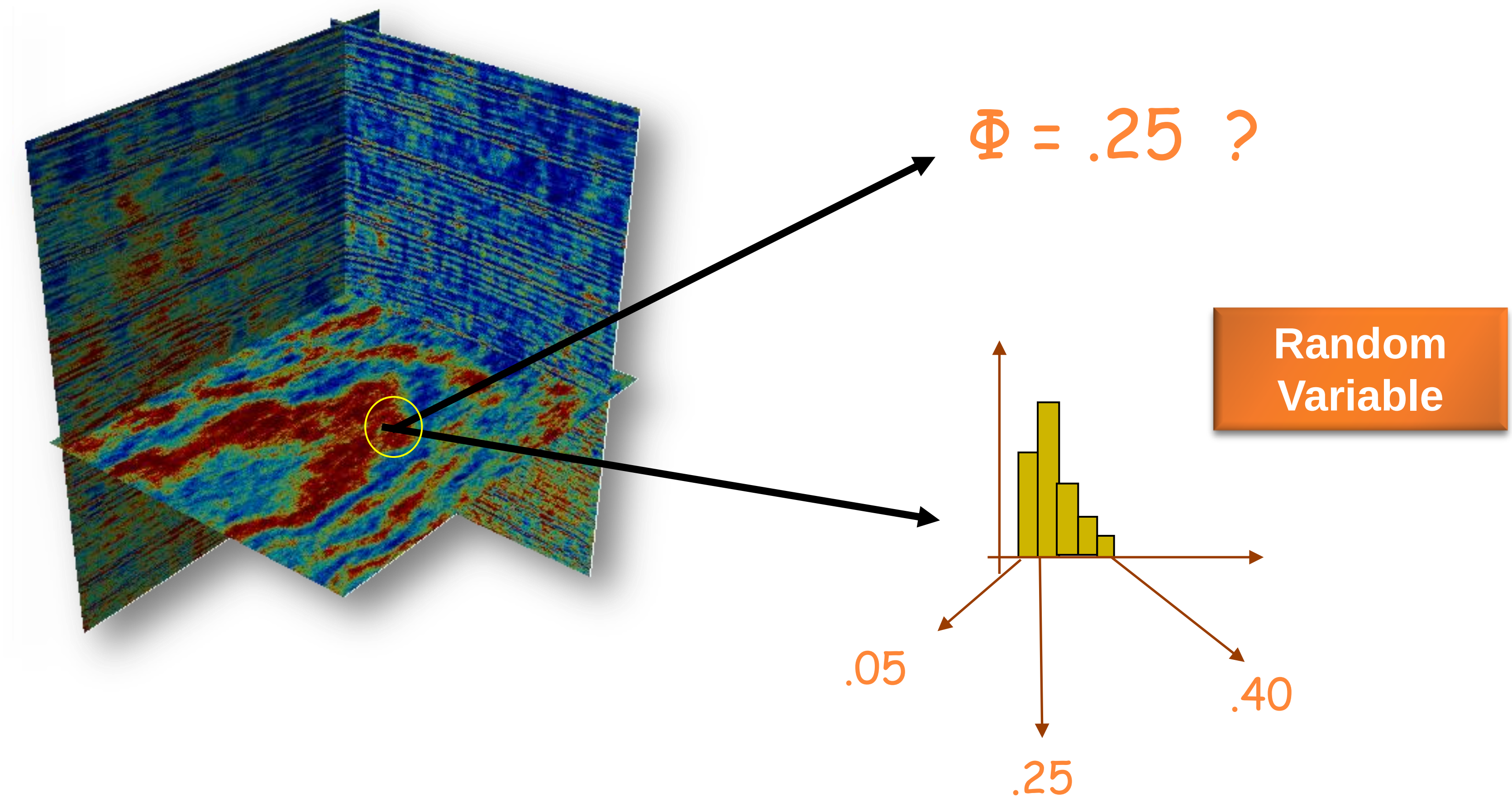


Different levels of knowledge about a certain process

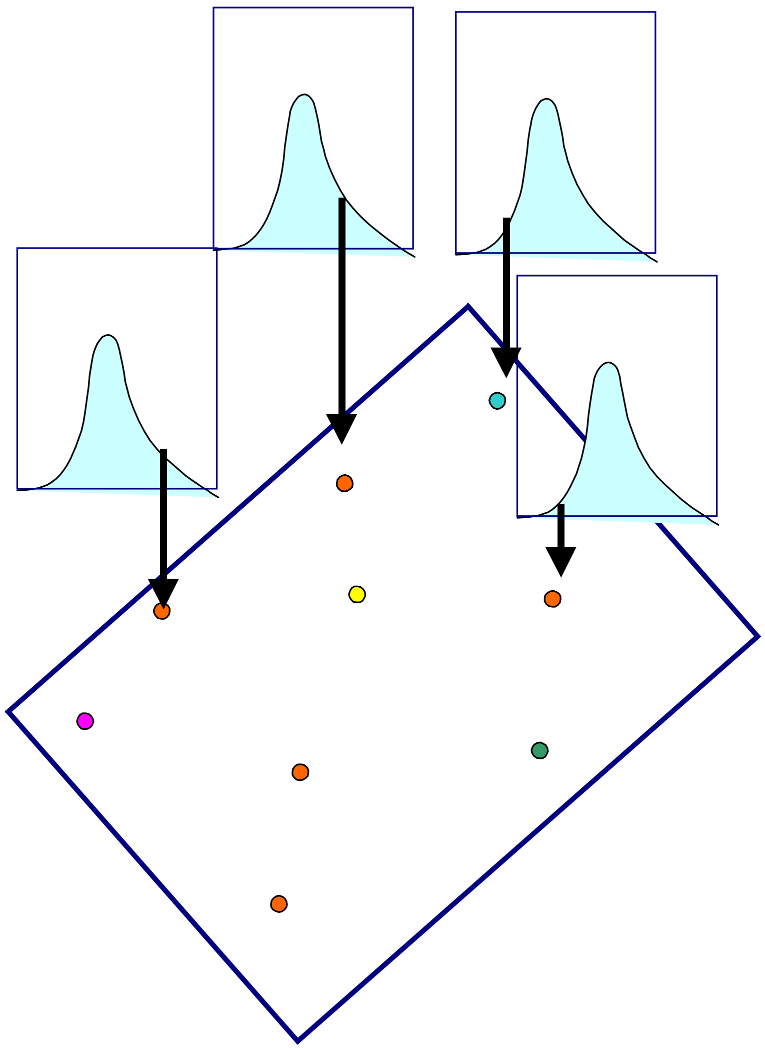




Probabilistic approach



Random Function Model



- Parameters of a random variable:
  - Expected value

$$E\{V\} = \tilde{m} = \sum_{i=1}^{n} p_i v_{(i)}$$

  - Variance

$$Var\{V\} = \tilde{\sigma}^2 = E\{[V - E\{V\}]^2\}$$

- Parameters of joint random functions:
  - Covariance

$$Cov\{UV\} = \tilde{C}_{UV} = E\{(U - E\{U\})(V - E\{V\})\}$$

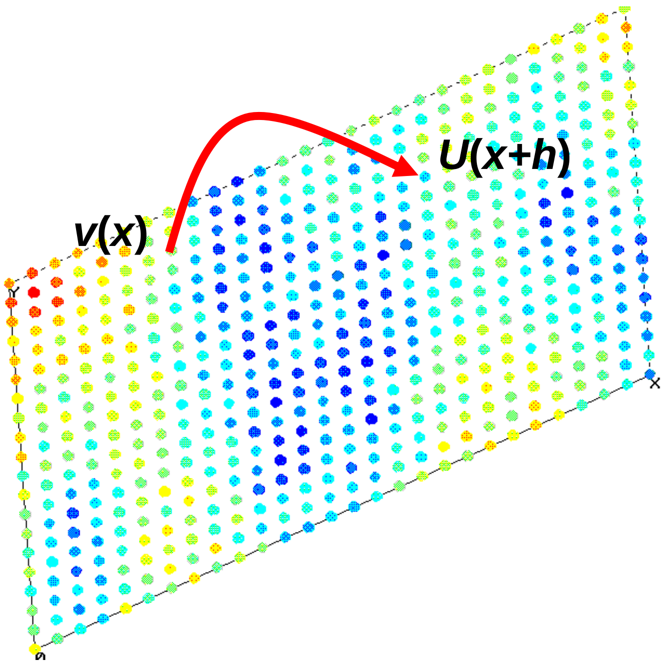
  - Correlation coefficient

$$\tilde{\rho}_{UV} = \frac{\tilde{C}_{UV}}{\sqrt{\tilde{\sigma}_U^2 \tilde{\sigma}_V^2}}$$

Spatial Continuity

Spatial Correlation

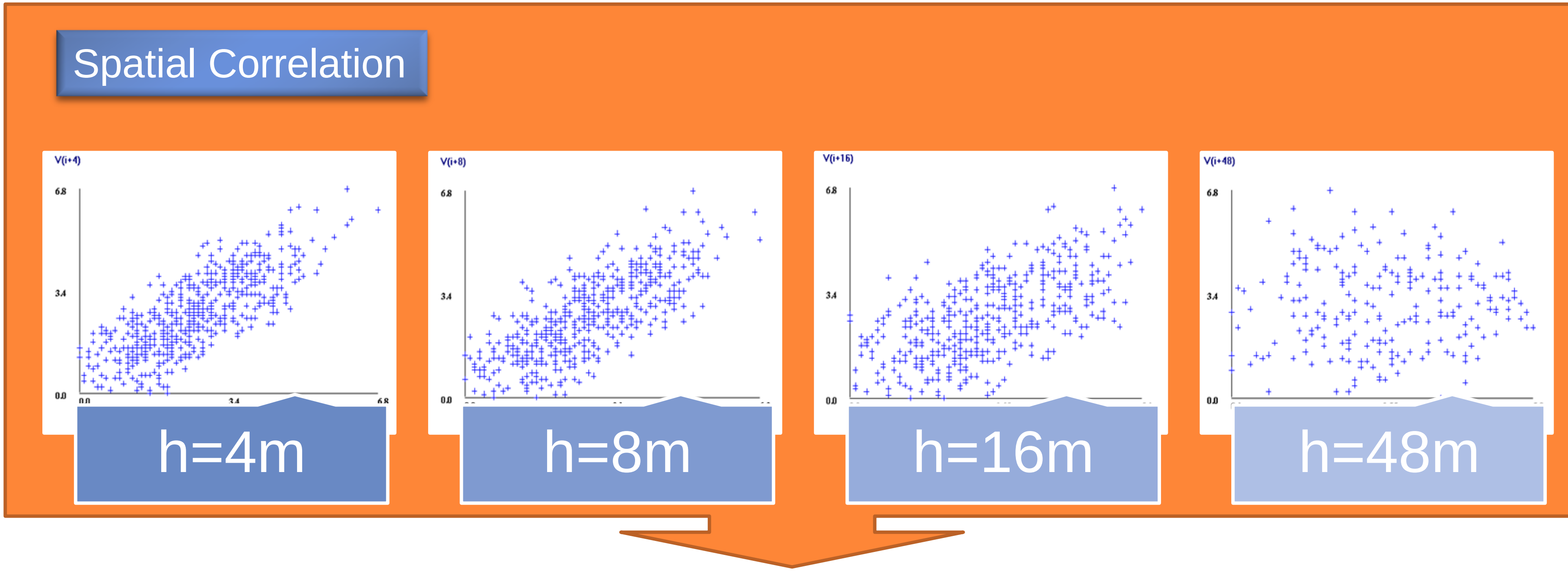
Bi-point statistic

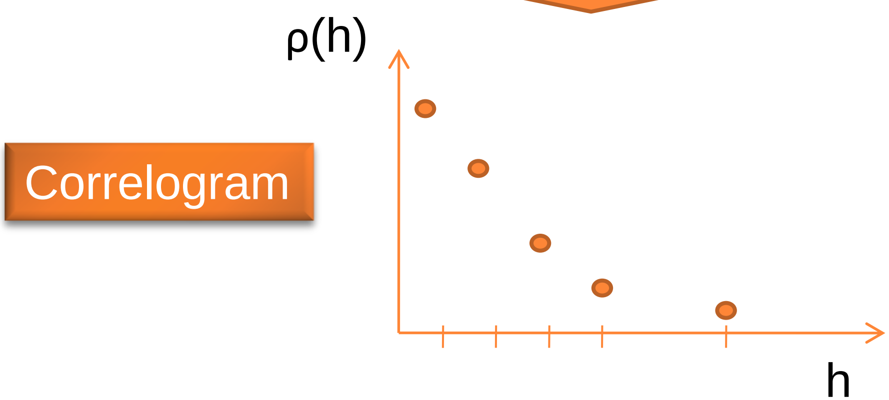


Correlation coefficient

$$\tilde{\rho}_{UV} = \frac{\tilde{C}_{UV}}{\sqrt{\tilde{\sigma}_U^2 \tilde{\sigma}_V^2}}$$

Spatial Continuity

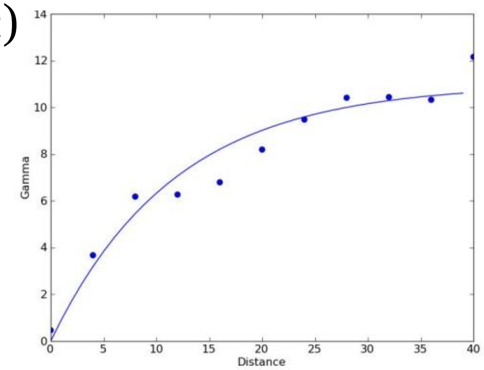




Variogram and Spatial Covariance

- Measures that summarize the dispersion of bi-plots between *z*(*x*) and *z*(*x*+*h*)
- Tools to quantify the spatial continuity of the phenomenon.

$\gamma(h)$

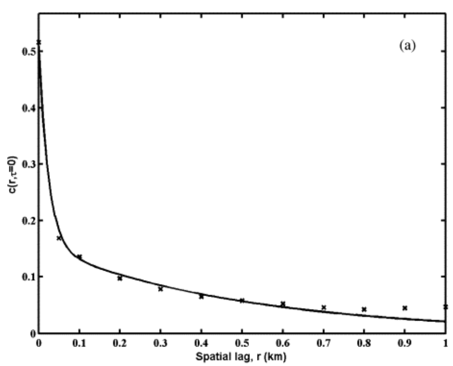


h

variogram

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[Z(x) - Z(x+h)\right]^2$$

$C(h)$



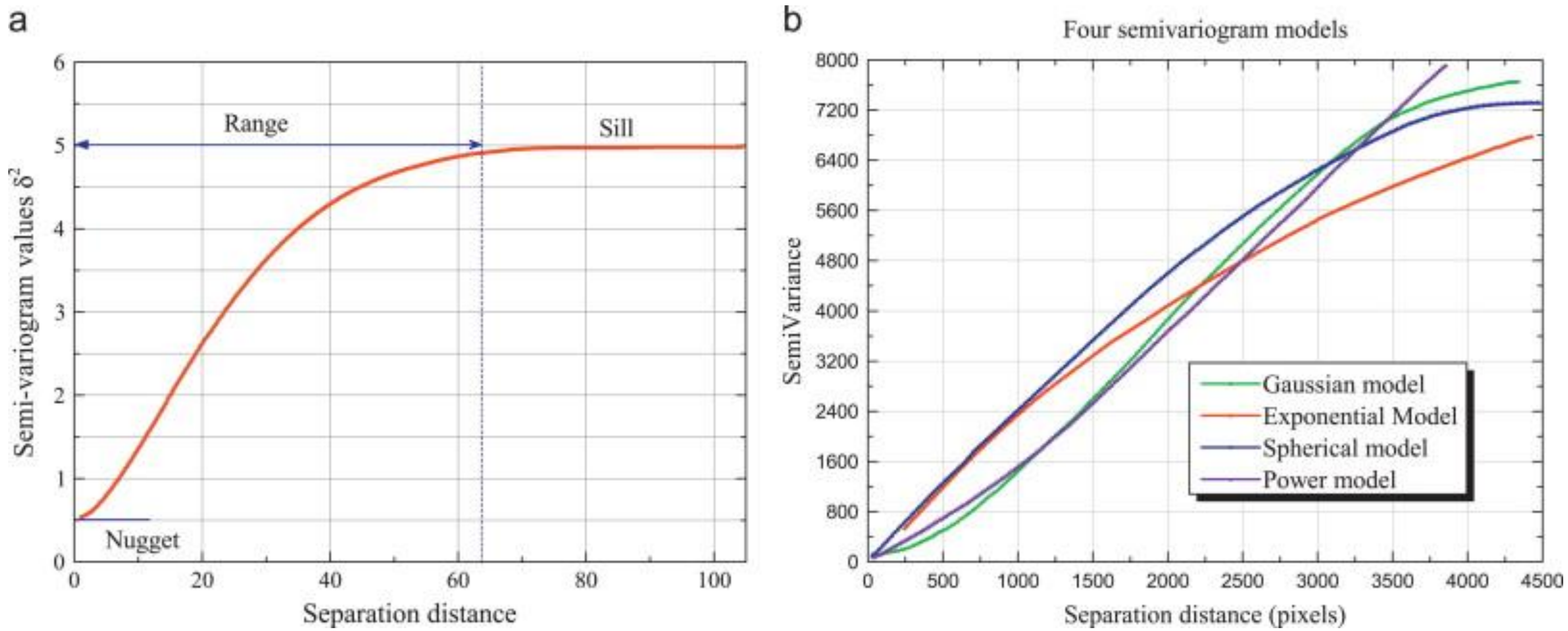
h

Spatial Covariance

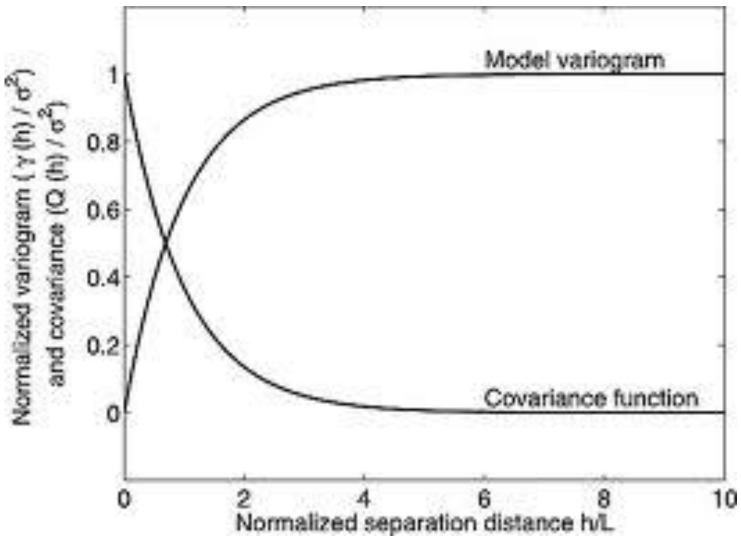
$$C(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} \left[Z(x).Z(x+h)\right] - m(x).m(x+h)$$

$$\gamma(h) = C(0) - C(h)$$

Variogram models

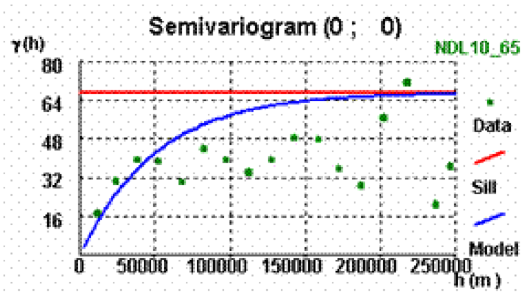


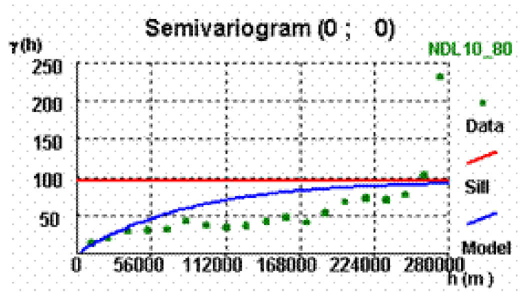
Variogram vs Covariance

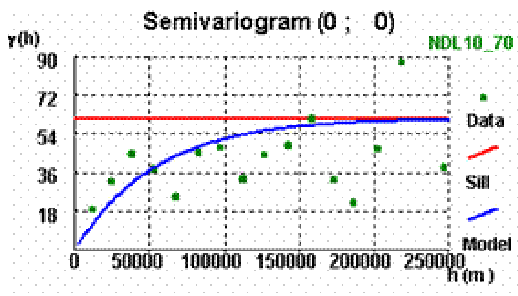


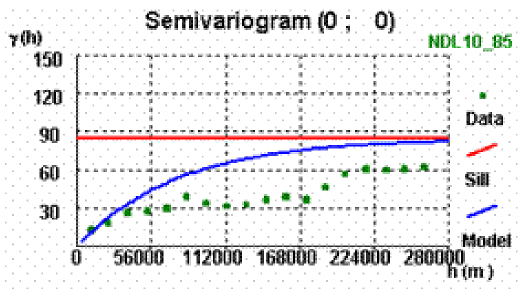
$$\gamma(h) = C(0) - C(h)$$

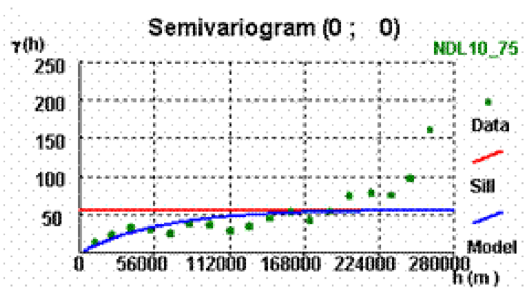
Spatial Continuity

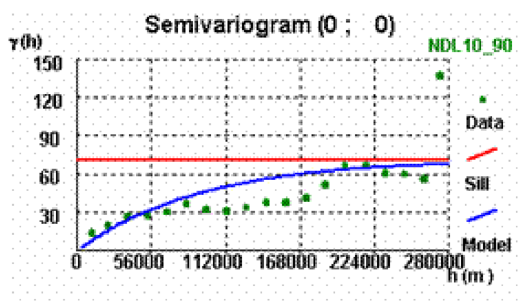






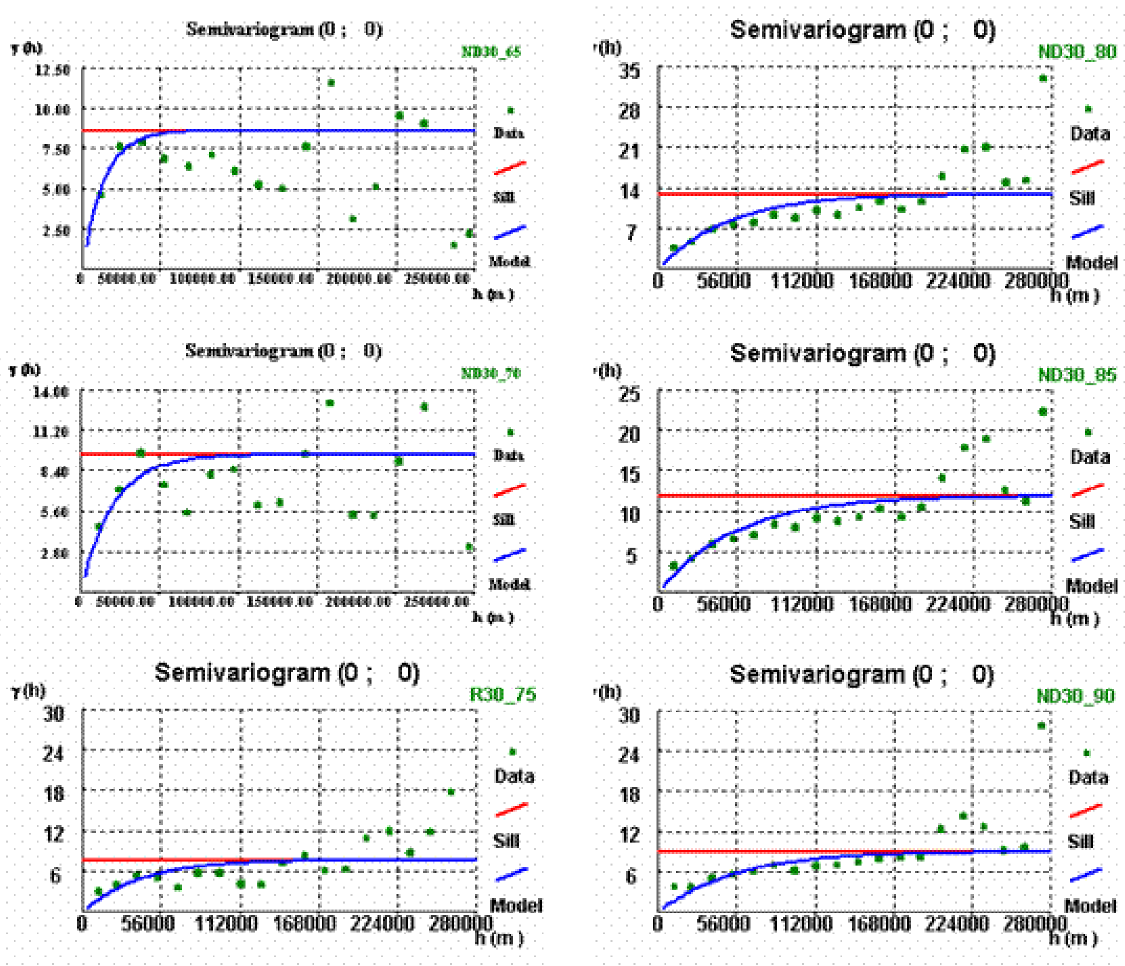






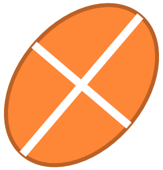
RL10 – Draught index: number of days per year with precipitation lower than 10 mm

Spatial Continuity

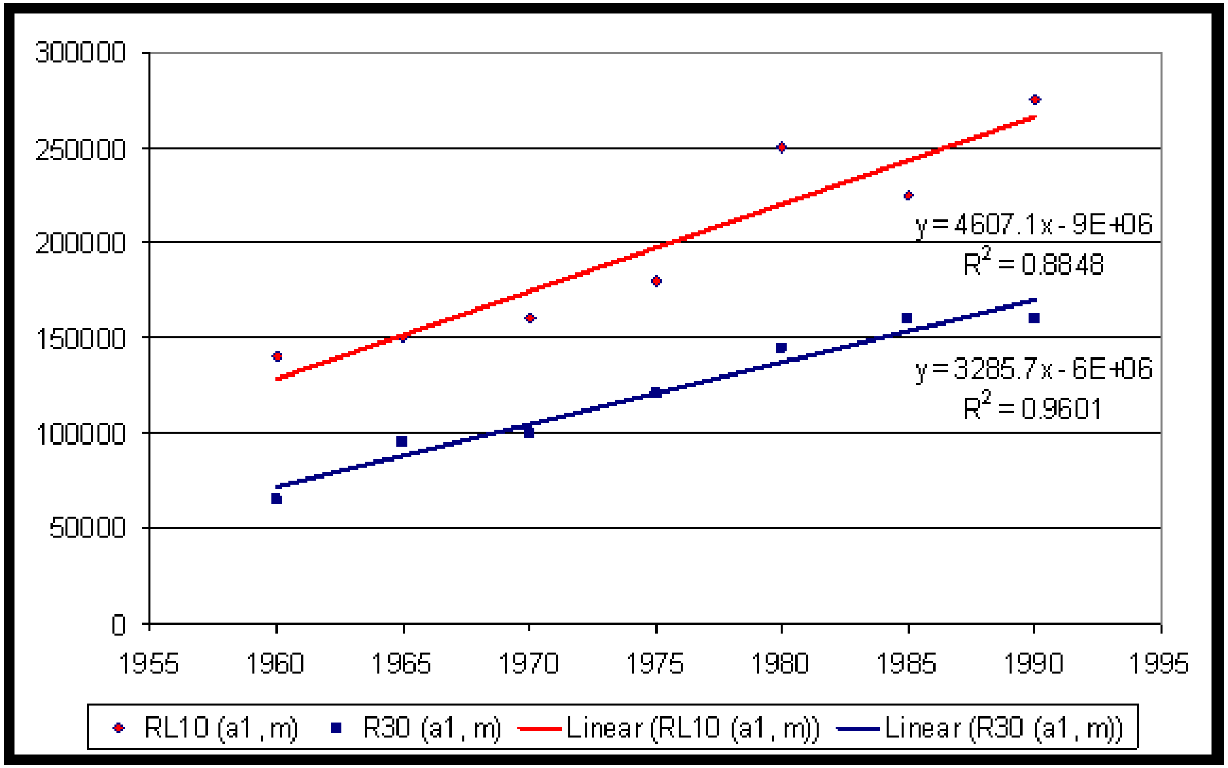


R30 – number of days per year with precipitation higher than 30 mm

Anisotropy



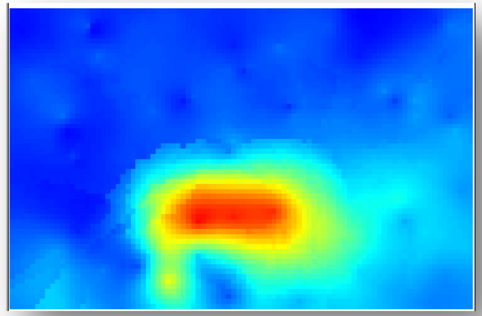
Spatial Continuity



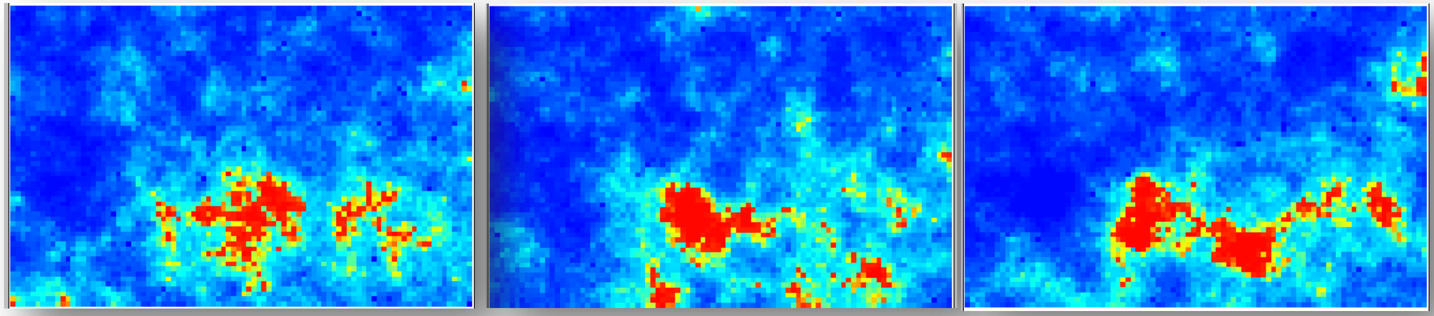
RM Durao, MJ Pereira, AC Costa, J Delgado, G Del Barrio, A Soares
International Journal of Climatology 30 (10), 1526-1537

Estimation or simulation?

The mean image is sufficient to represent the knowledge of a quantity?



Estimation



Simulation

... depends upon the variability/variance of the feature
... depends on our ignorance
… depends on the objective of the study

Geostatical Methods

Estimation

- Ordinary kriging
- Simple kriging
- Universal kriging
- Kriging with external drift
- Morphological kriging
- Cokriging
- Collocated Cokriging

Simulation

- Sequential Gaussian Simulation
- Sequetial Indicator Simulation
- Probability Field Simulation
- Sequential Direct Simulation
  - SDS with local anisotropies
  - Co-simulation
  - Block simulation

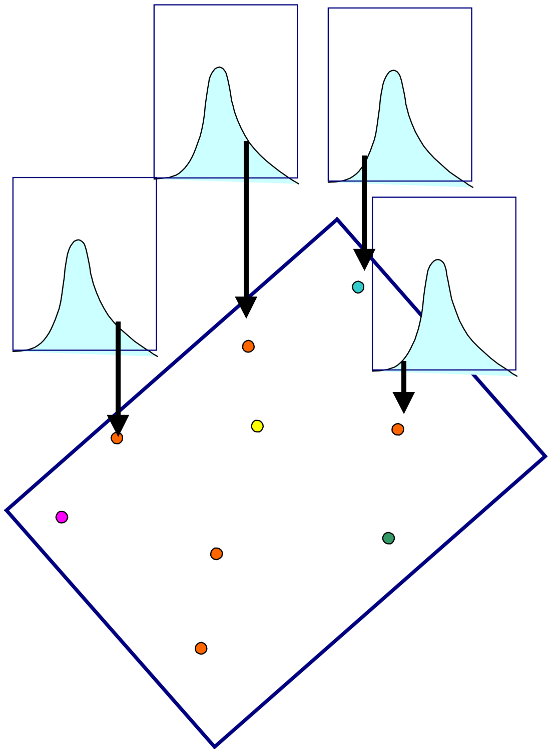
Spatial Inference

Probabilistic Framework of the Geostatistical Estimator

Problem: We do not know the real values

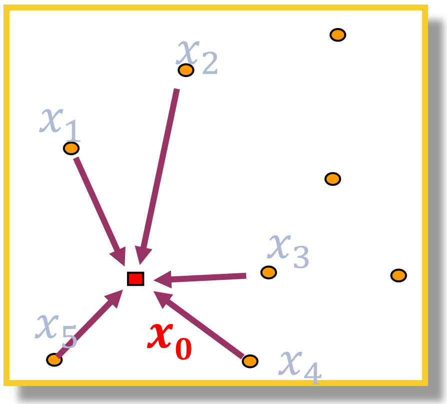
Solution: Build a Model that appropriately accommodates our data set and the physical phenomena.

$z(x_0)$ and the neighbouring samples $z(x_\alpha), \alpha = 1, N$ are considered to be outcomes of a set of random variables located in x_0 and $x_\alpha, \alpha = 1, N$.



Estimation

- Estimating the value of an attribute $Z(x_0)$ at a point or a local area x_0, based on a set of n samples $Z(xi)$.

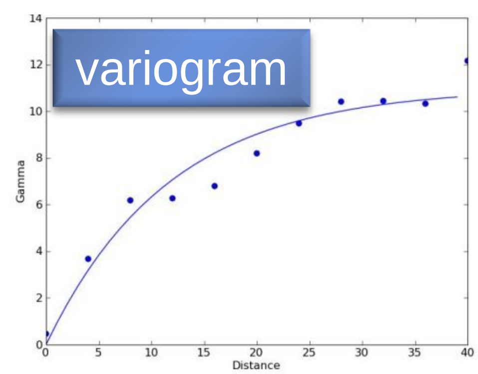


$$[z(x_0)]^* = \sum_{\alpha=1}^{N} \lambda_\alpha z(x_\alpha)$$

The weights λ_α
- Reflect the structural proximity of samples $Z(x_\alpha)$ to the point $Z(x_0)$
- Should have a disaggregating effect of preferred groups of samples

Ordinary Kriging System

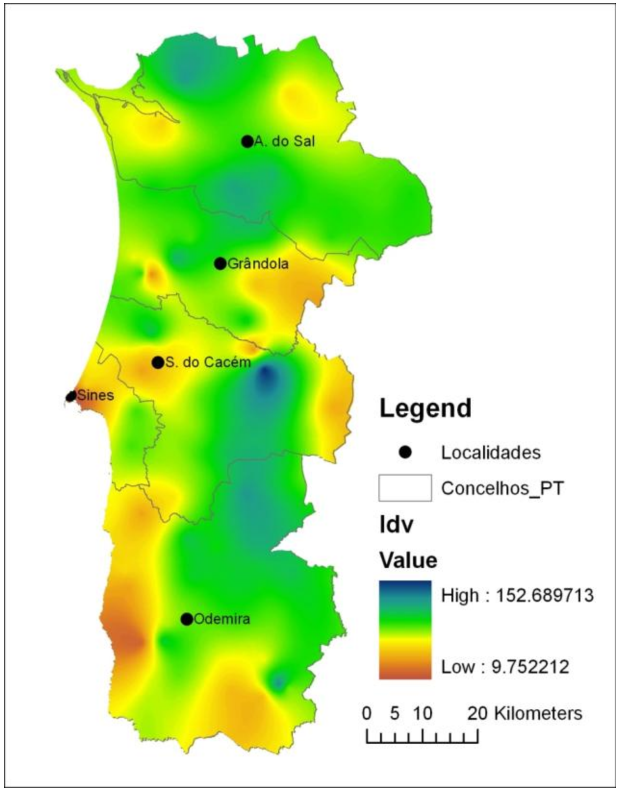
$$\begin{cases} \sum_{\beta}^{N} \lambda_\beta C(x_\alpha, x_\beta) + \mu = C(x_\alpha, x_0) \\ \sum_{\beta} \lambda_\beta = 1 \end{cases} \qquad \alpha = 1, \dots, N$$



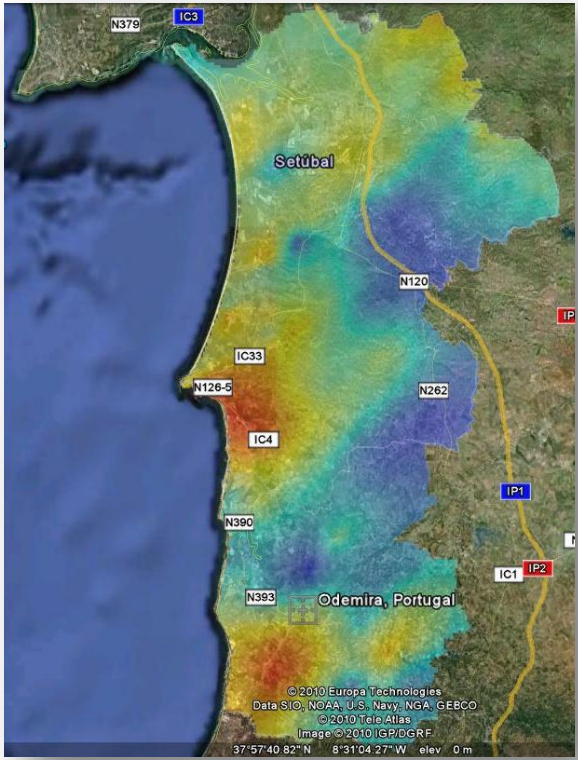
Applications

Air pollution

LDV

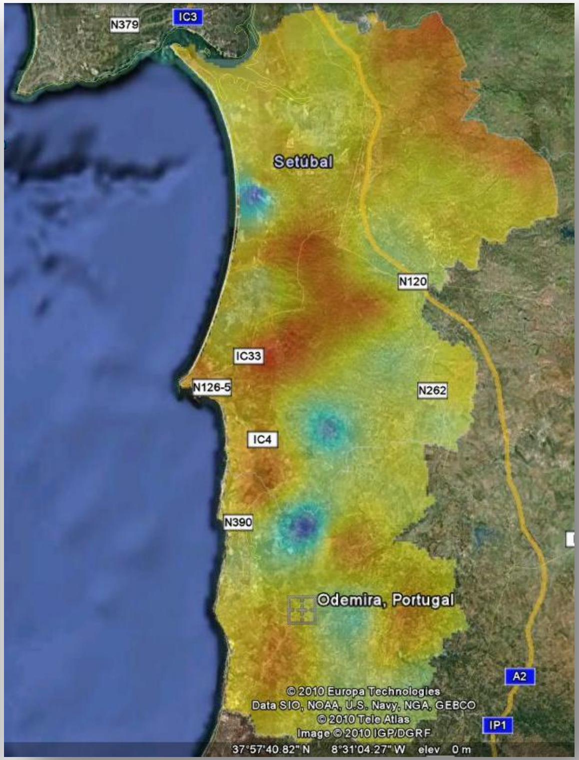


O_3 – June 2009



MÁX: **88** µg/m^3; **MIN: 48** µg/m^3

O_3 – October 2009

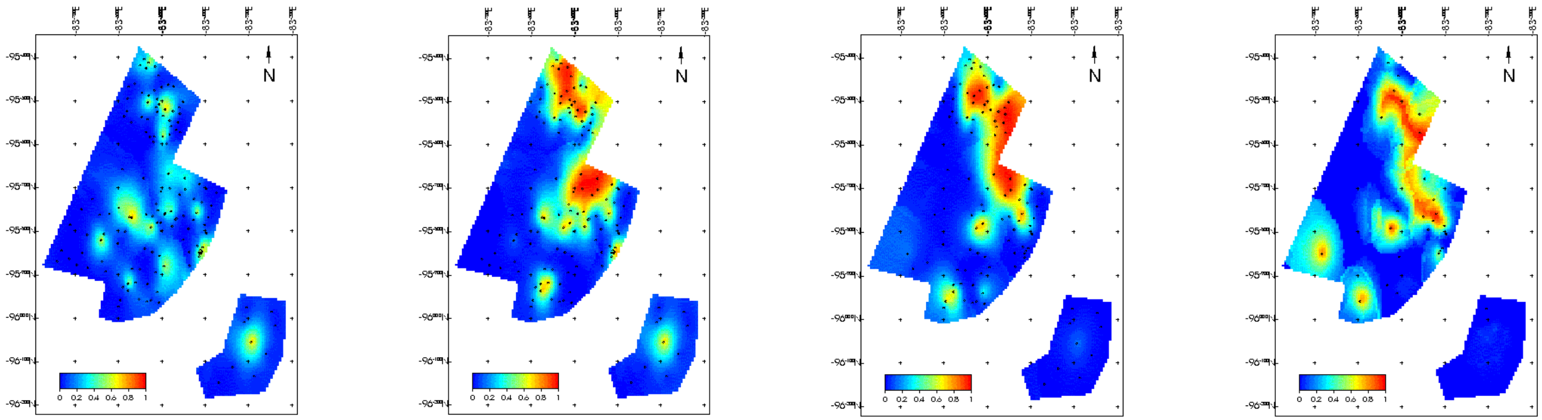


MÁX: **76** µg/m^3; **MIN: 21** µg/m^3

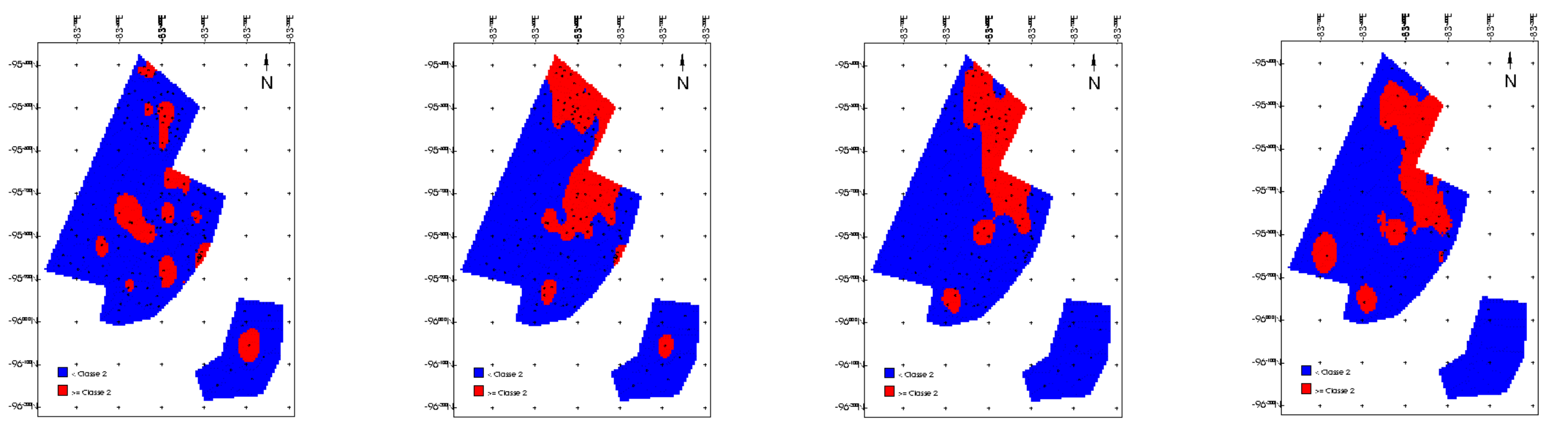
Applications

Soil pollution

- Indicator kriging of the soft variable – *Prob*[oil contamination class ≥ 2]



- Morphological classification

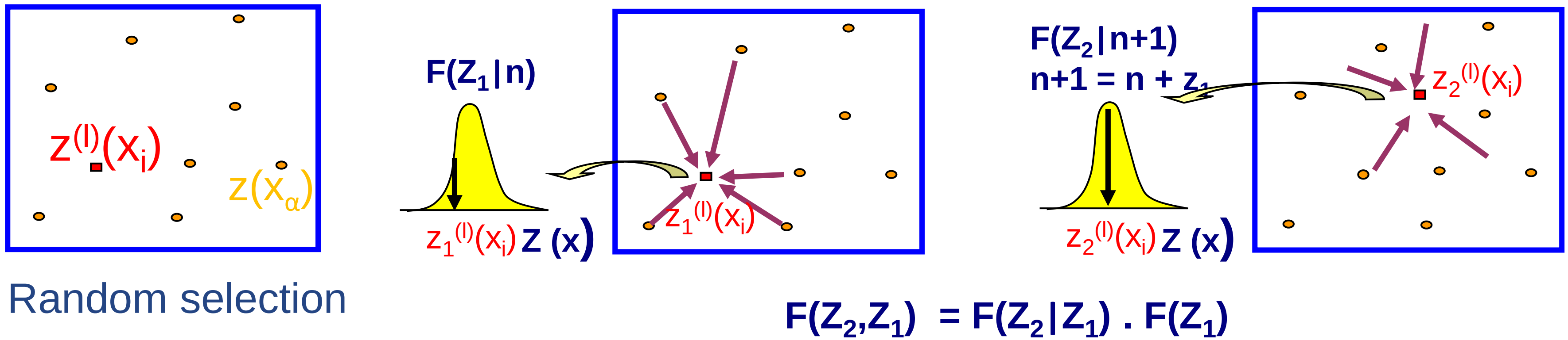


Sequential Simulation

- Stochastic Simulation: Reproduction of the variability of the phenomenon through the distribution function F(x) = Prob {Z (x) <z} and the variogram γ(h)

- Any simulated image reproduces the distribution function, and the variogram of the experimental values.

- Sequential simulation is based on the successive application of Bayes theorem in sequential steps



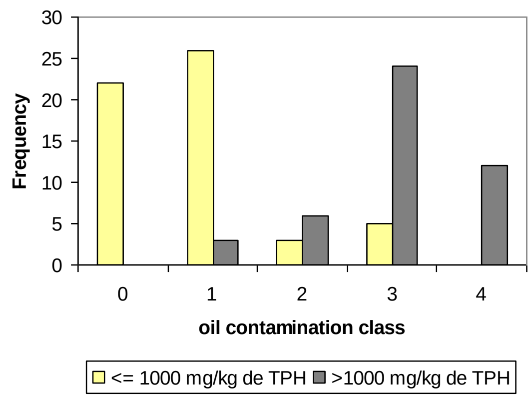
Applications

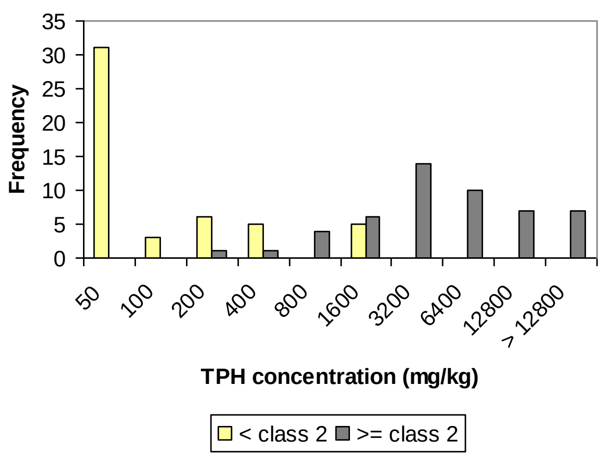
Soil pollution

Data with different levels of uncertainty

TPH concentration > 1000 mg kg^{-1}

"contaminated" samples





misclassification of soft data:

- 6.7% of samples with oil contamination class < 2 are "contaminated"
- 14.3% of samples with oil contamination class ≥ 2 are "clean".

Applications

Soil pollution

- Determination of the misclassified areas inside both "contaminated" and "clean" spots.

TPH dataset

TPH$_1$

$z_1(u_{\alpha 1}) = z(u_{\alpha 1}) \quad if \;\; s(u_{\alpha 1}) = 0{,}1$

TPH$_2$

$z_2(u_{\alpha 1}) = z(u_{\alpha 1}) \quad if \;\; s(u_{\alpha 1}) = 2{,}3{,}4$

$$prob\{z(x) > z_c | z_1(x_\alpha), \alpha = 1, N_{c1}\}$$

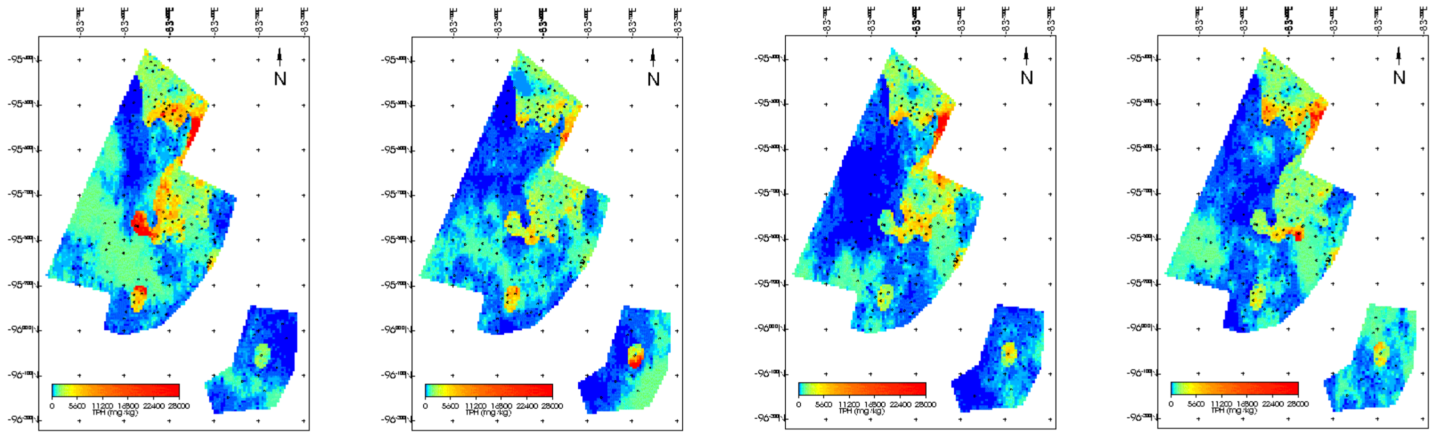
$$prob\{z(x) > z_c | z_2(x_\alpha), \alpha = 1, N_{c2}\}$$

stochastic simulation

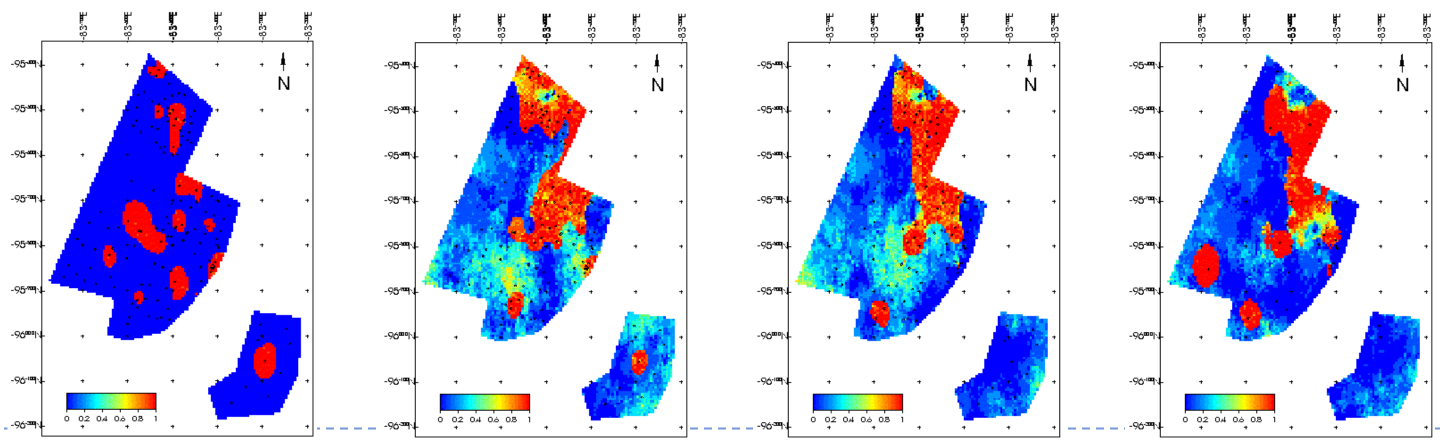
Applications

Soil pollution

- Simulation of TPH concentration values – horizon 2



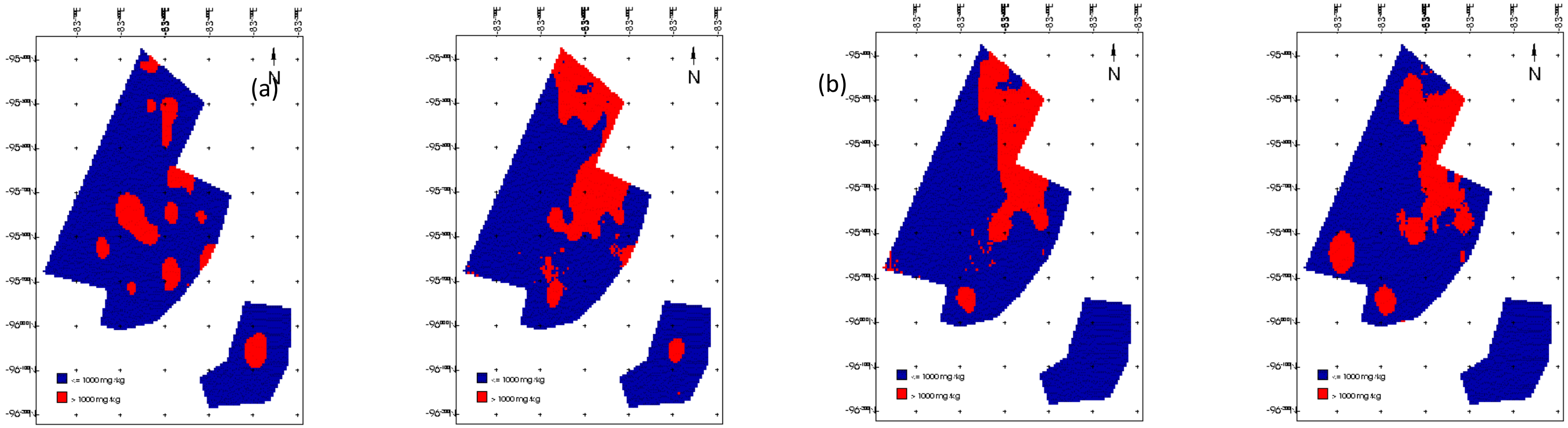
- Probability of exceeding 1000 mg/kg



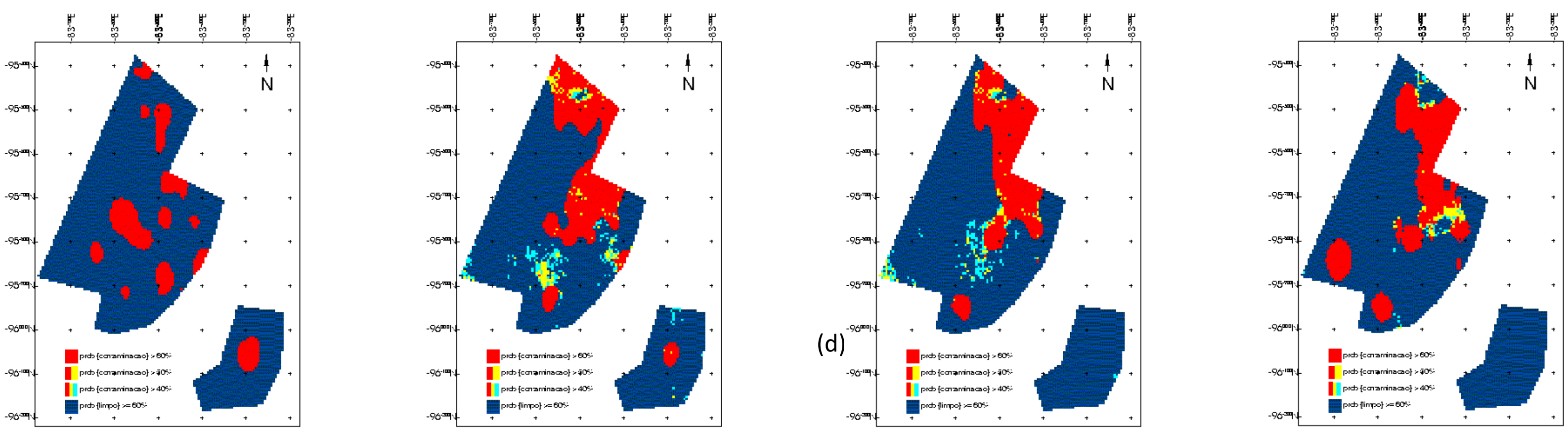
Applications

Soil pollution

- Mean “contaminated” and “clean” spots



- “contaminated” and “clean” spots for different levels of uncertainty



Applications

Soil pollution

- Volumes

Table 1. *Contaminated soil volumes for three different uncertainty levels: 40% probability, 50% probability and 60% probability; the values in brackets refer to the relative difference to the 50% probability case.*

Layer	Contaminated soil volume (m^3)		
(m a.t.s.)	40%	50%	60%
[0,1]	23 900	23 900	23 900
	(0%)		(0%)
[1,2]	46 150	41 250	38 975
	(+11.9%)		(-5.5%)
[2,3]	41 175	36 950	35 175
	(+11.4%)		(-4.8%)
[3,4]	38 850	37 575	36 025
	(+3.4%)		(-4.1%)
[0,4]	150 075	139 675	134 075
	(+7.4%)		(-4.0%)

- Methodology allowed the delineation of areas targeted for remediation with different levels of uncertainty.

- Accounting for soft information permitted a less uncertain quantification of the contaminated soils without increasing the quantity of hard data required, which would result in a considerable increase in study costs.

M Pereira, J Almeida, C Costa, A Soares
geoENV III—Geostatistics for Environmental Applications, 475-485

Uncertainty and Support

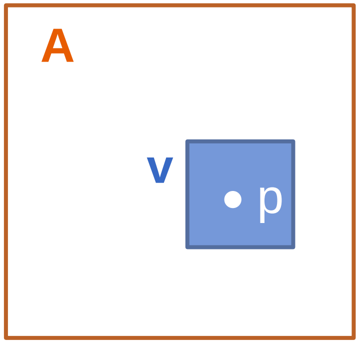
- Relation

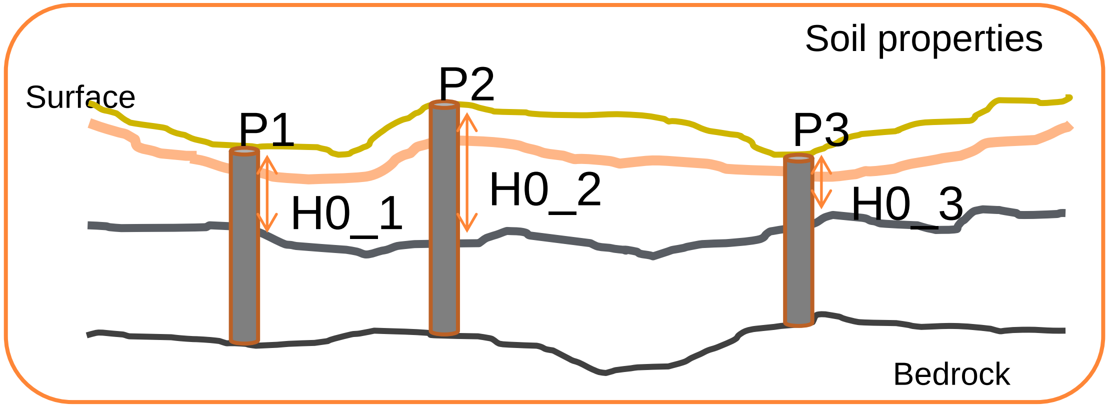
$$\sigma^2_{p/A} = \sigma^2_{p/v} + \sigma^2_{v/A}$$ (Journel and Huijbreghts, 1978)

$\sigma^2_{p/A}$ is the variance of a punctual value p in the domain A

$\sigma^2_{p/v}$ is the variance of p in a volume v inside A

$\sigma^2_{v/A}$ is the variance of v inside A





Applications

Air pollution

Data with different levels of uncertainty and different supports

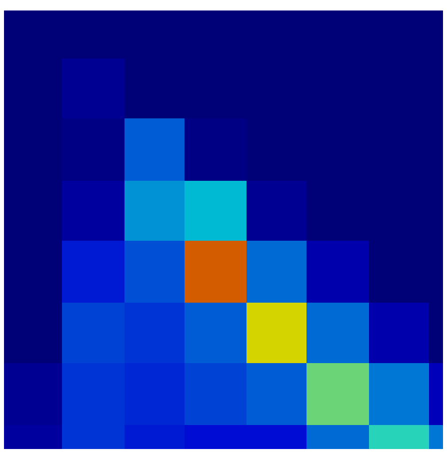
Input data:
•Meteorological variables
•Emissions

Air quality dispersion model

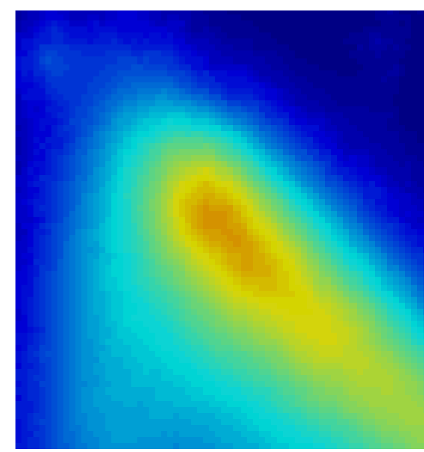
Point data
• pollutant atmospheric concentration measurements

Geostatistical stochastic simulation

Block data



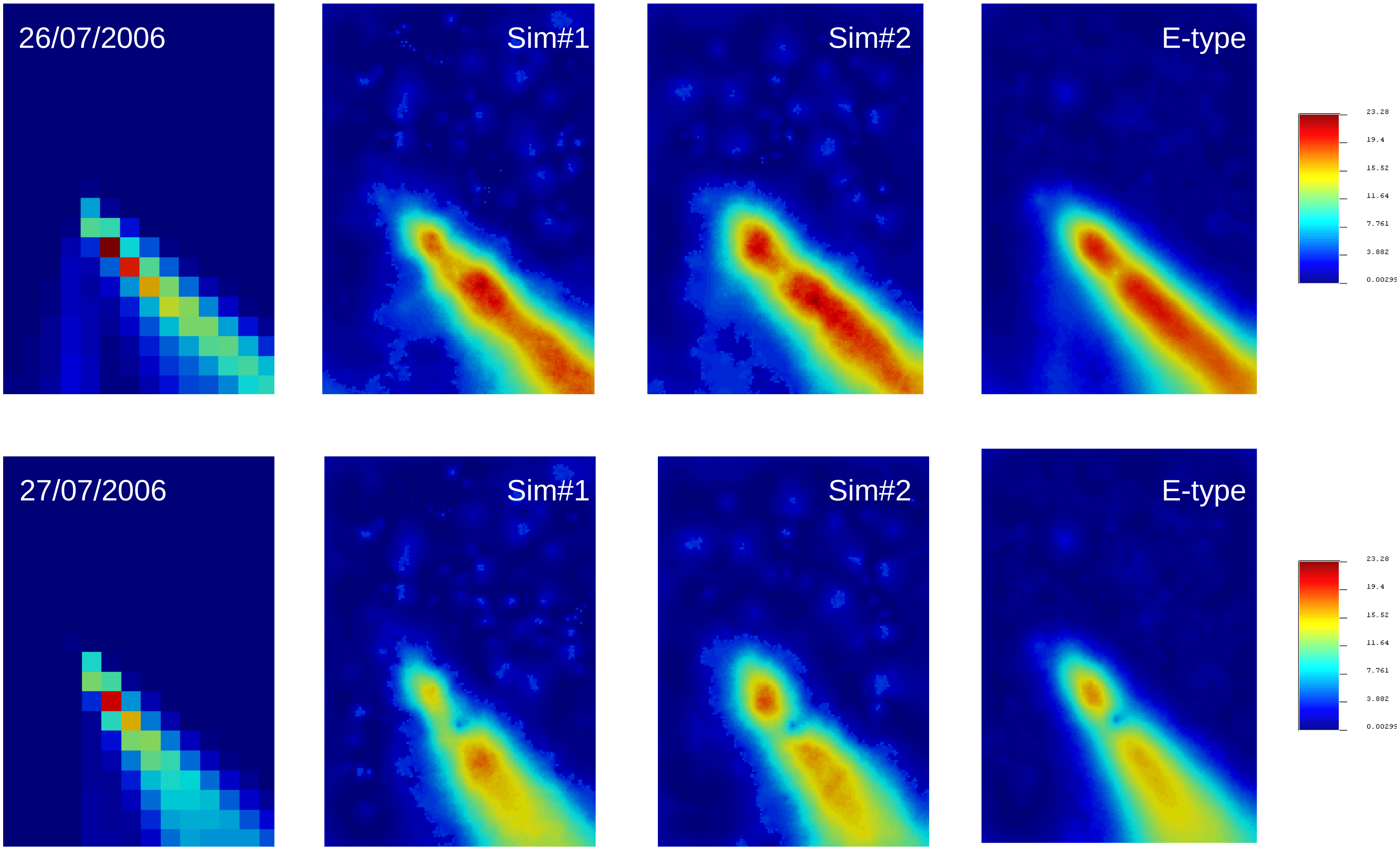
Block Sequential Simulation



Maria João Pereira, Rita Durão, Amílcar Soares.2009.
IAMG'09, Stanford University, CA, USA

Applications

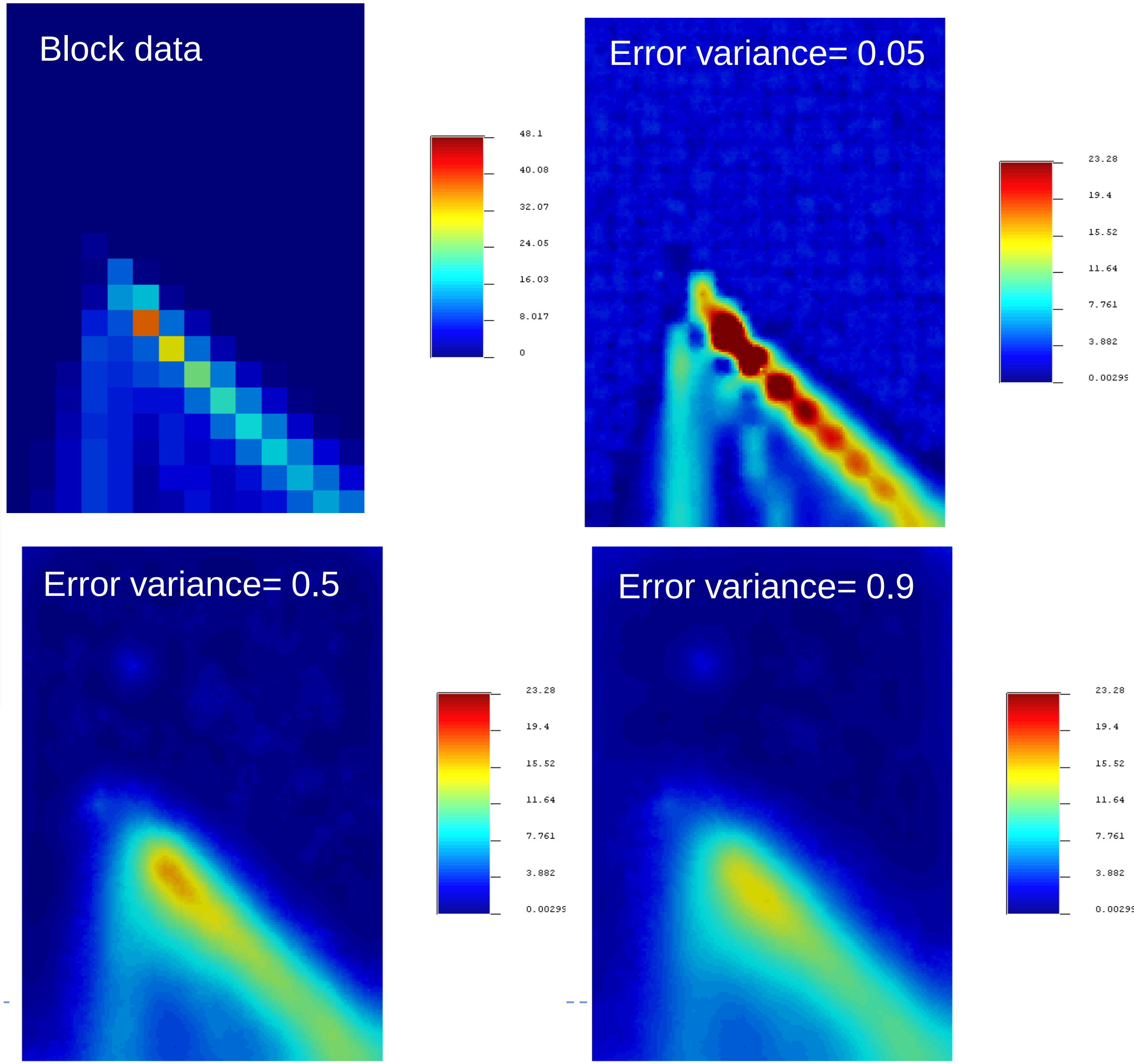
Air pollution



Applications

Air pollution

Block error effect

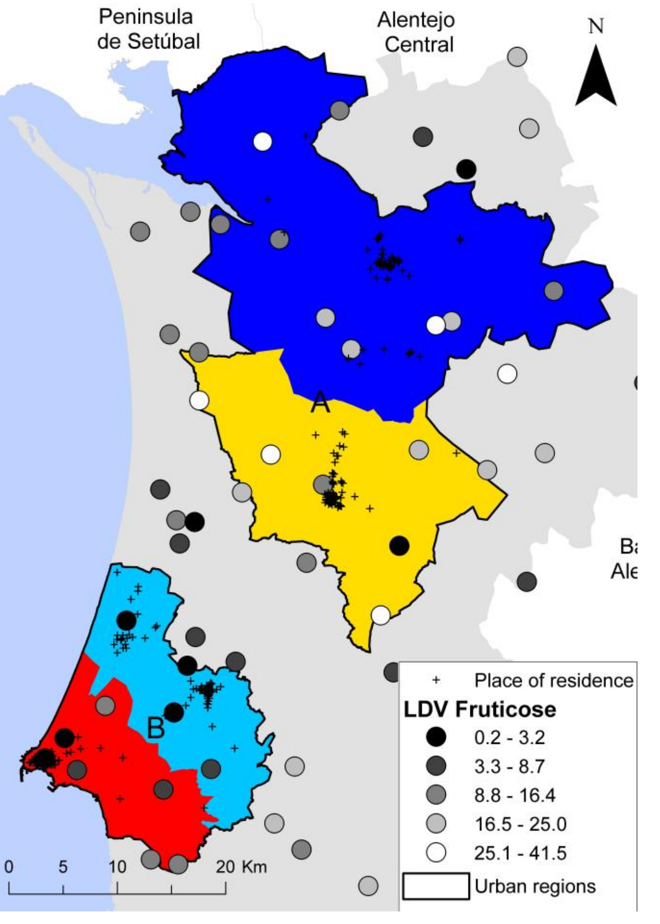


- This hybrid multi-scale approach is a valuable methodology for predicting air quality
- Downscaling and calibration (with experimental point data) of maps obtained by deterministic dispersion models is achieved by the proposed method
- The possibility of incorporating the block-data error into the kriging system is an important feature of the method

Applications

Environmental health

Is there an association between air quality and birth weight in regions A and B ?



<u>Health data</u>

Predominantly urban areas (PUA[d])

- Region A: PUA of A. do Sal and Grândola
- Region B: PUA of Sines, S. André, S. Cacém

<u>Air quality data</u>

- Lichen diversity biomonitoring program
- Lichen Diversity Value (LDV) for Fruticose species

Applications

Environmental health

Individual exposure model based on stochastic simulation

$$E_j = \sum_{s=1}^{S} t_{sj} * q_{sj}$$

E_j (weighted) average exposure of jth mother during pregnancy

t_{sj} time (as a proportion of overall time of gestation) spent by jth mother during pregnancy, at location s

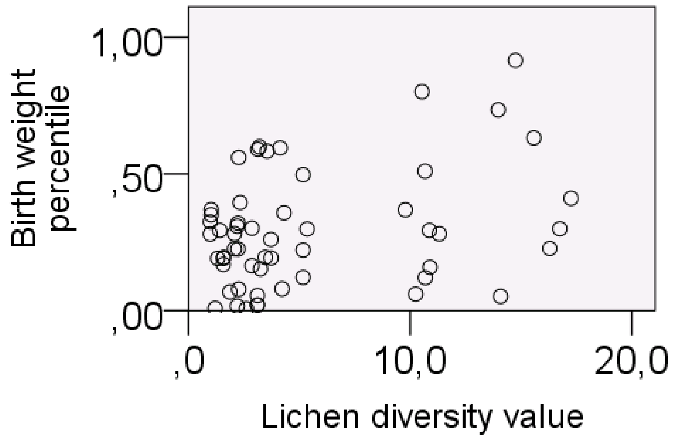
q_{sj} Air quality index for jth mother during pregnancy, at location s

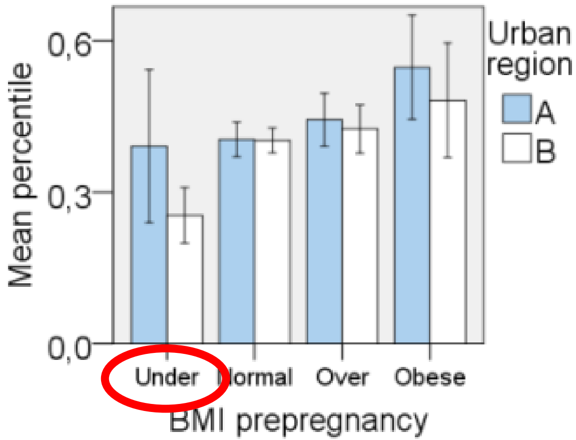
We geocoded mother's residential history during gestational period (place of residence, place of work).

Applications

Environmental health

Underweight category subset (n=55) where we found significant associations between average individual exposure measures and birth weight percentile (ρ=0,326; p-value=0,015).





Applications

Environmental health

After checking assumptions of linear model for this subset, we estimated univariate linear models for this subset. Significant covariates found:

- Previous low birth weight
- Gestational BMI gain
- Gestational diabetes
- Air quality

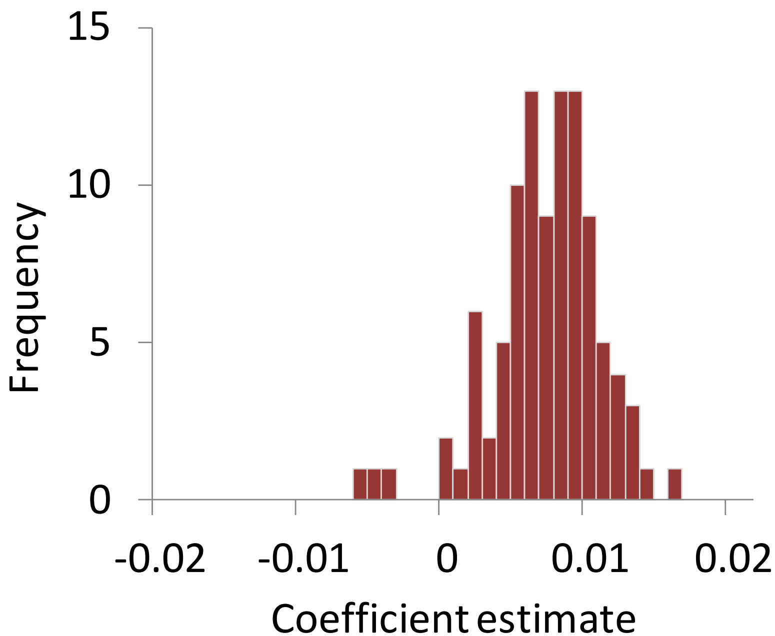
…then we performed 100 multivariate analysis using simulation data for air quality:

$$g(\mu_j) = \beta_0 + \beta_1 X_j + \boldsymbol{\beta}_z \mathbf{Z}_j, \ \mathrm{j} = 1,2,\ldots,\mathrm{n}$$

Applications

Environmental health

Distribution of 100 estimated $\hat{\beta}_1$ (air quality exposure coefficient).



- Mean of distribution estimates coefficient air quality=0,0065
- Standard deviation=0,00376
- Empirical distribution varies between (0,0005; 0,011) for CI 90%.
- For CI 90%, air quality index is significantly associated with birth weight increase

An increase of 1 unit of LDV index is associated with an 0,0065 birth weight percentile units increase.

Applications

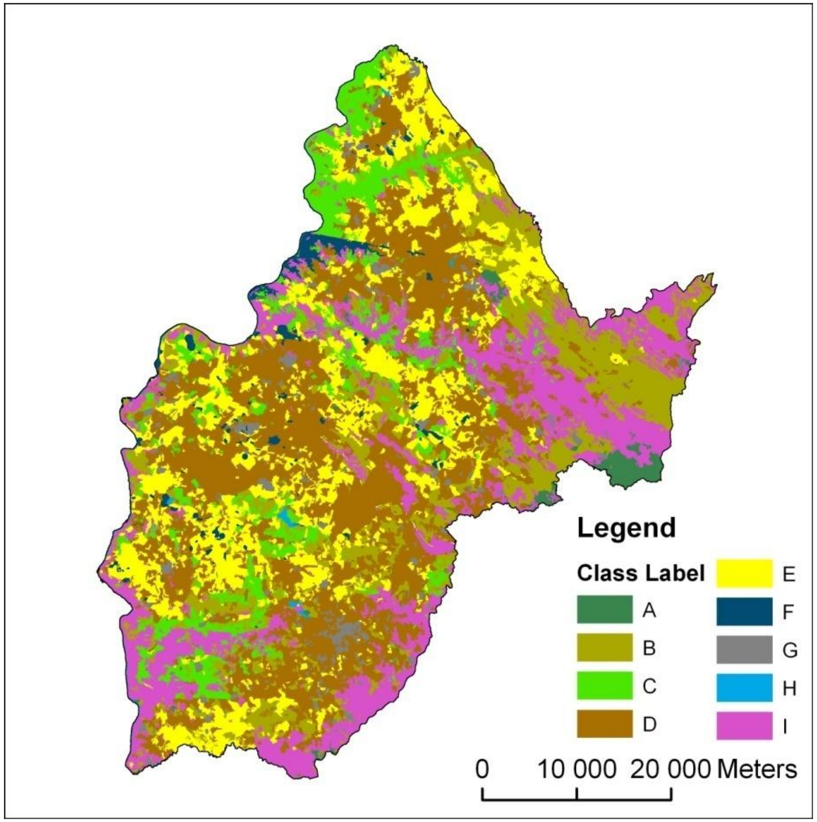
Uncertainty assessment

Hipothesis

Classification algorithms of RS images tend to produce more errors in given classes than in others and for each thematic class different errors occur depending on sensors and ground conditions

Solution

1. calculate the trend of the errors m_i by each image derived thematic class i.
2. local errors are calculated conditioned to the mean error of the predicted class for that location and to the neighboring error values



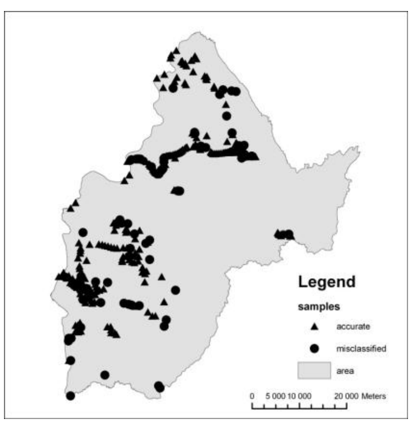
Geostatistical stochastic simulation
SIS with local varying means

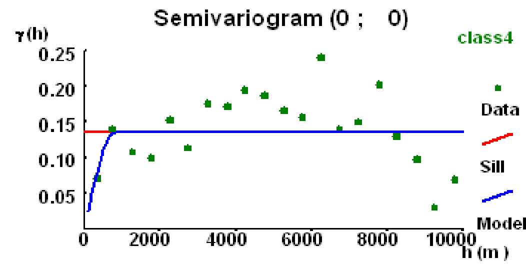
Applications

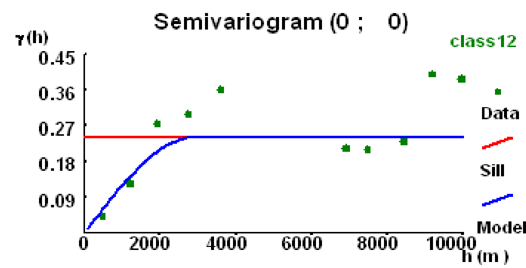
Uncertainty assessment

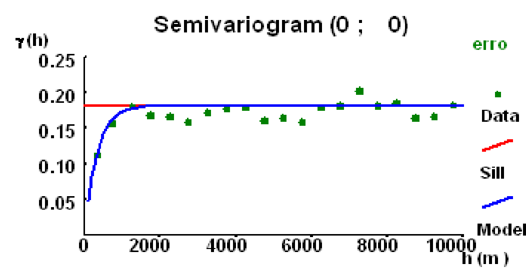
	Ground-based									User's accuracy
Class labels	A	B	C	D	E	F	G	H	I	
Image-derived										
A	6	3	0	2	0	0	0	0	6	0.35
B	0	39	1	0	1	0	0	0	0	0.95
C	0	7	14	0	0	0	0	0	3	0.58
D	0	5	4	68	4	0	0	0	0	0.84
E	1	1	1	3	42	1	0	0	0	0.86
F	0	0	0	1	2	17	0	0	0	0.85
G	0	1	0	5	3	1	10	0	0	0.50
H	0	0	0	0	0	0	0	20	0	1.00
I	3	11	1	3	1	0	0	0	26	0.58
Producer's accuracy	0.60	0.58	0.67	0.83	0.79	0.89	1.00	1.00	0.74	

Table 1. Confusion matrix. Class labels: A – coniferous forest; B – deciduous forest; C – grassland; D – permanent tree crops; E– non-irrigated land; F – irrigated land; G – artificial areas; H – water; I – maquis and mixed forest.



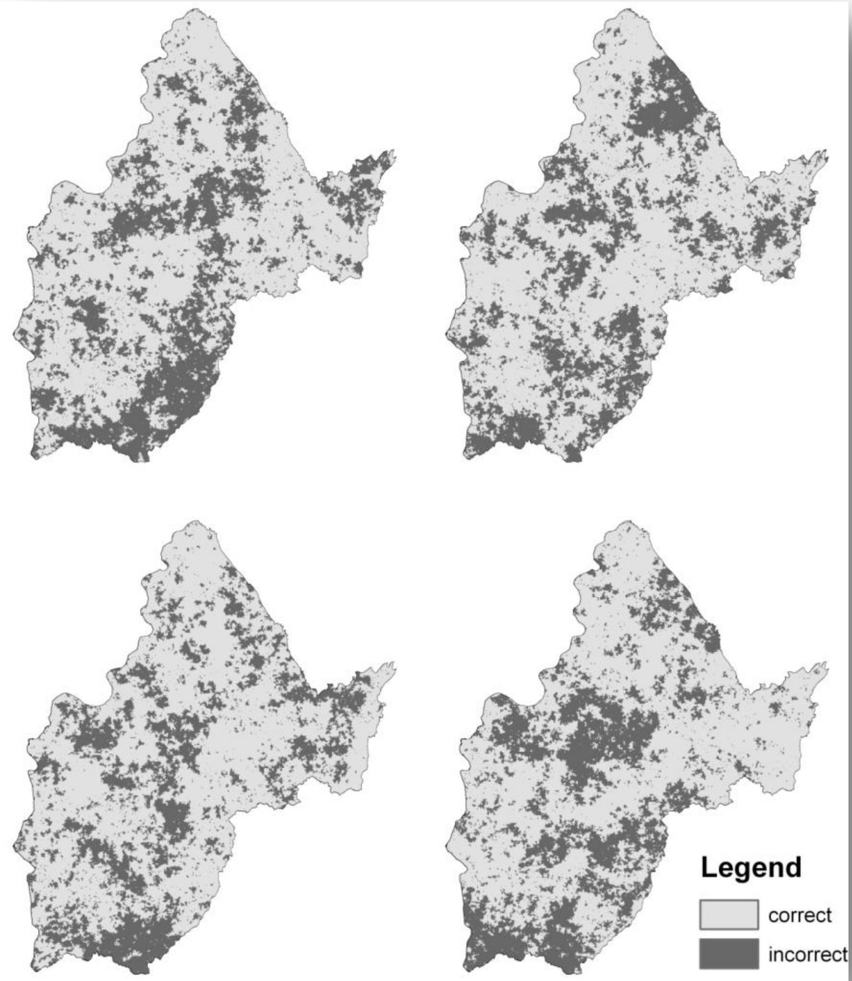


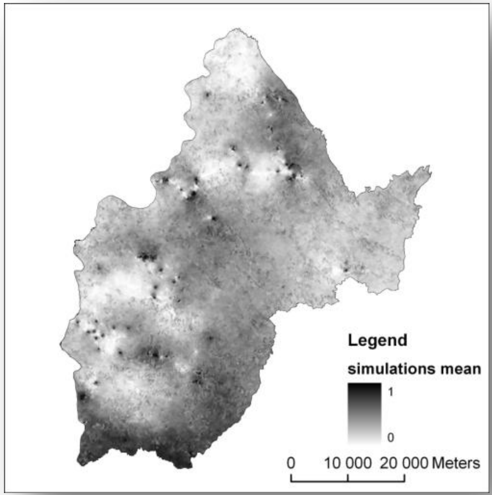


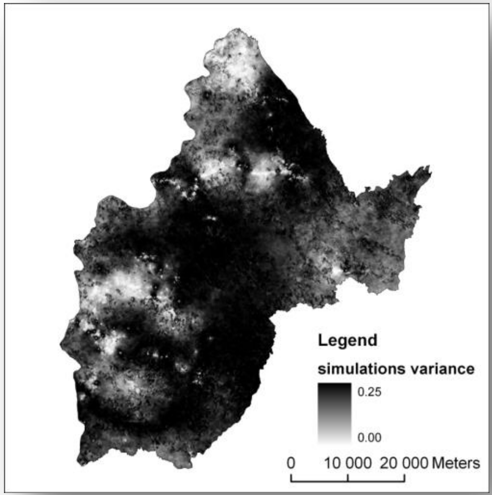


Applications

Uncertainty assessment







Geostatistics goes far beyond estimation…

- Taking into account varying data supports
- Combining different types of data with distinct levels of uncertainty
- Combining physical models with geostatistical models
- Spatial uncertainty assessment

CERENA

Muito Obrigada!

maria.pereira@ist.utl.pt

Projecto PTDC/CTE-SPA/103872/2008
FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR Portugal

2007 CIÊNCIA