ISRIC Spring School 2018

Introduction to Geostatistics

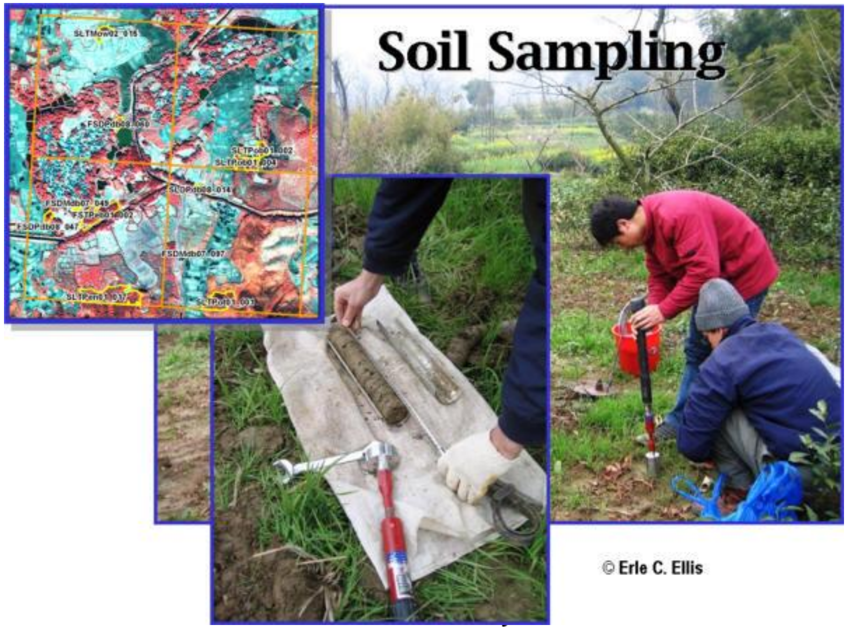
Bas Kempen

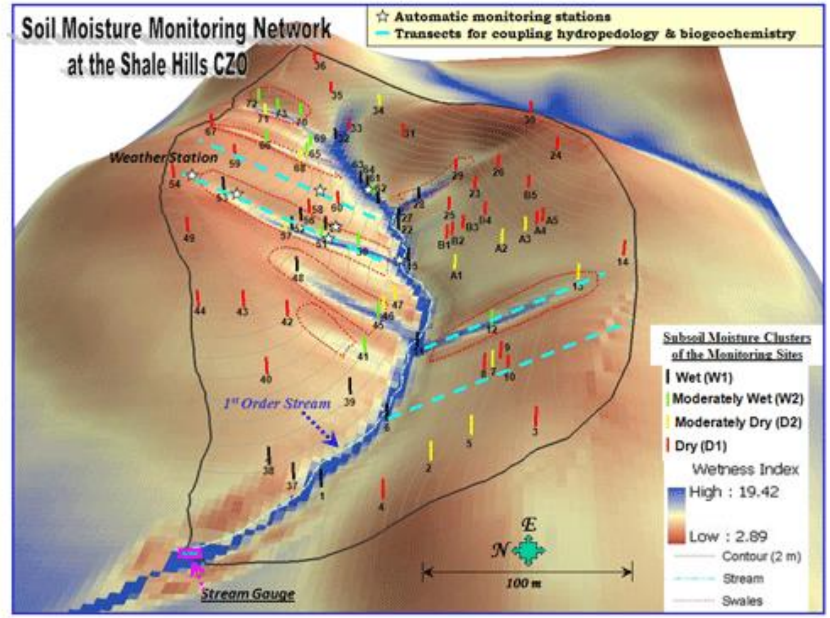
World Soil Information

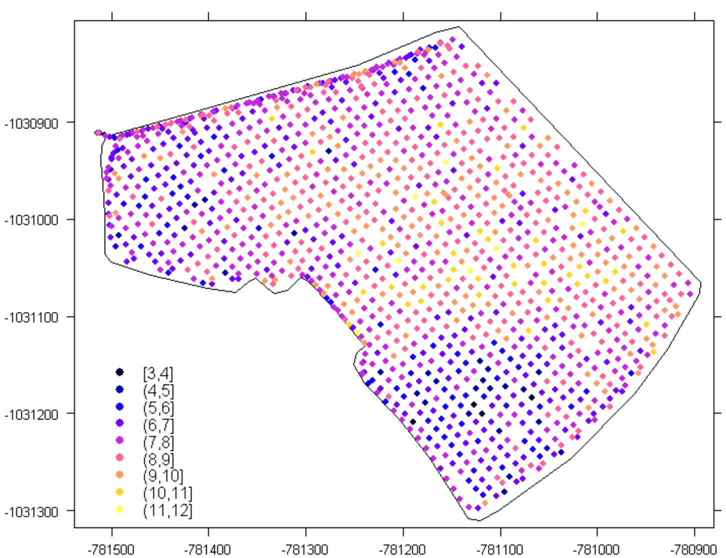
Outline of this lecture

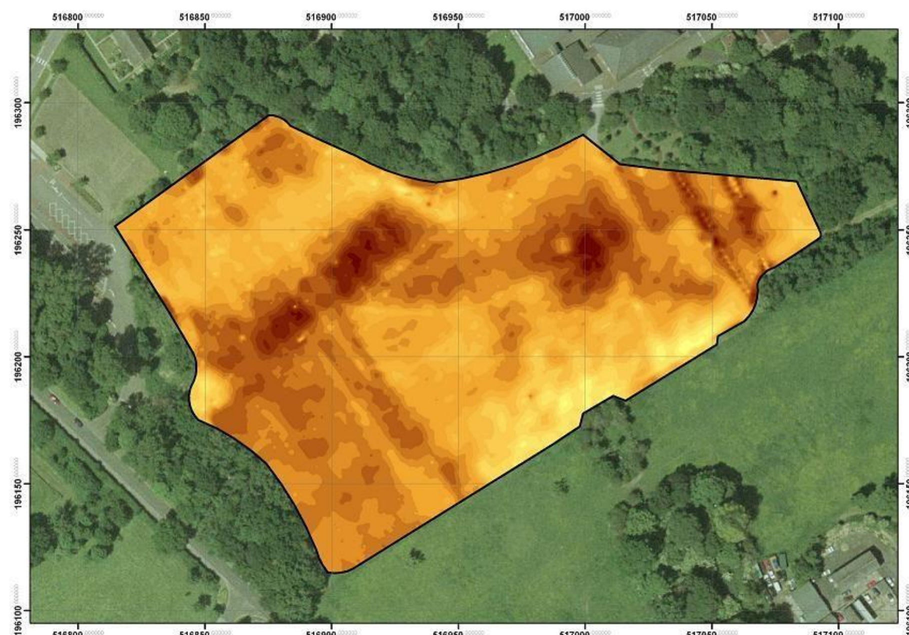
- Explore spatial variation
- Quantify spatial variation with the semivariogram
- Estimate the semivariogram from point observations
- Use the semivariogram for spatial interpolation with ordinary kriging
- Extend ordinary kriging to regression kriging

What we all know: soils and soil properties vary in space





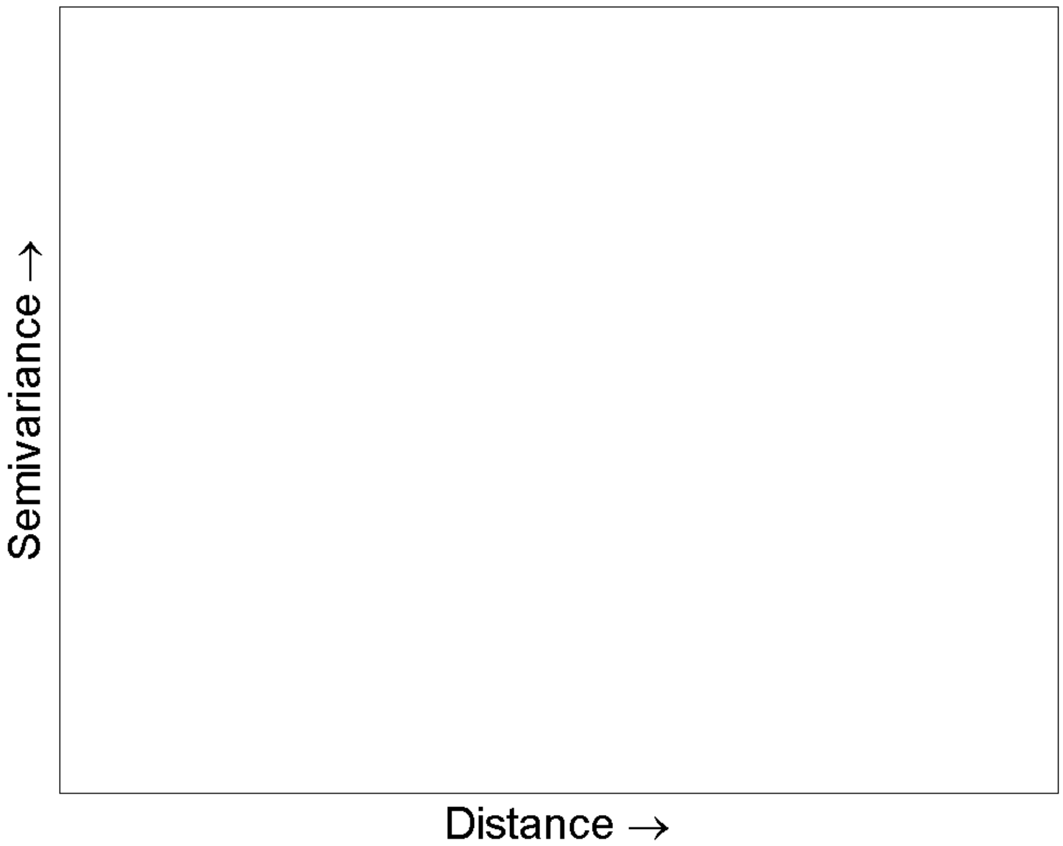




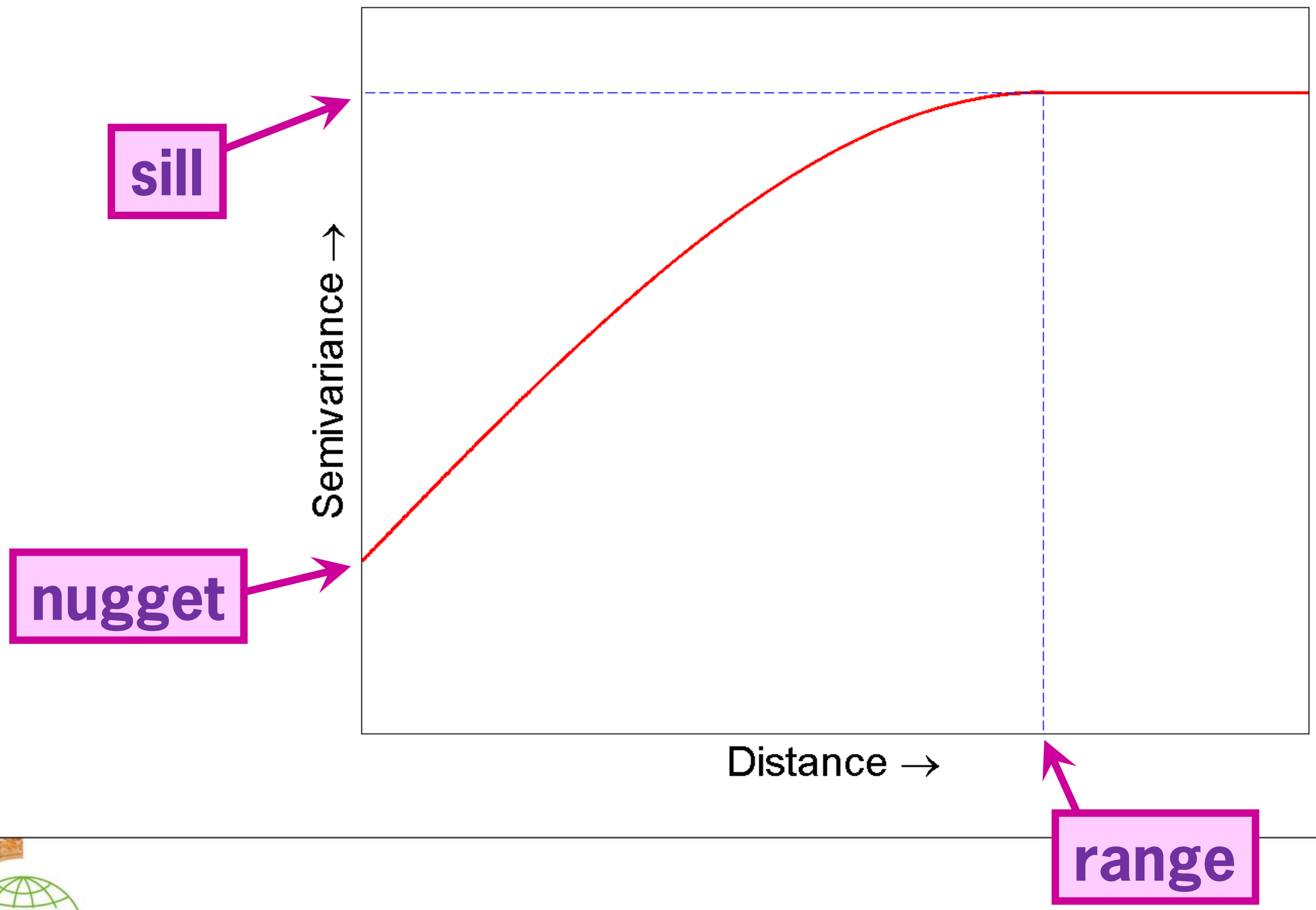
Spatial variation can be quantified using the so-called **semivariance**, this is half the expected squared difference between the values of the variable of interest at two locations

$$\gamma(h) = \frac{1}{2} E\left[\left(Z\underbrace{(x)}_{\text{measurement at location x}} - Z\underbrace{(x+h)}_{\text{measurement at location x+h}}\right)^2\right]$$

Plot of semivariance as a function of the distance is called a **(semi)variogram**



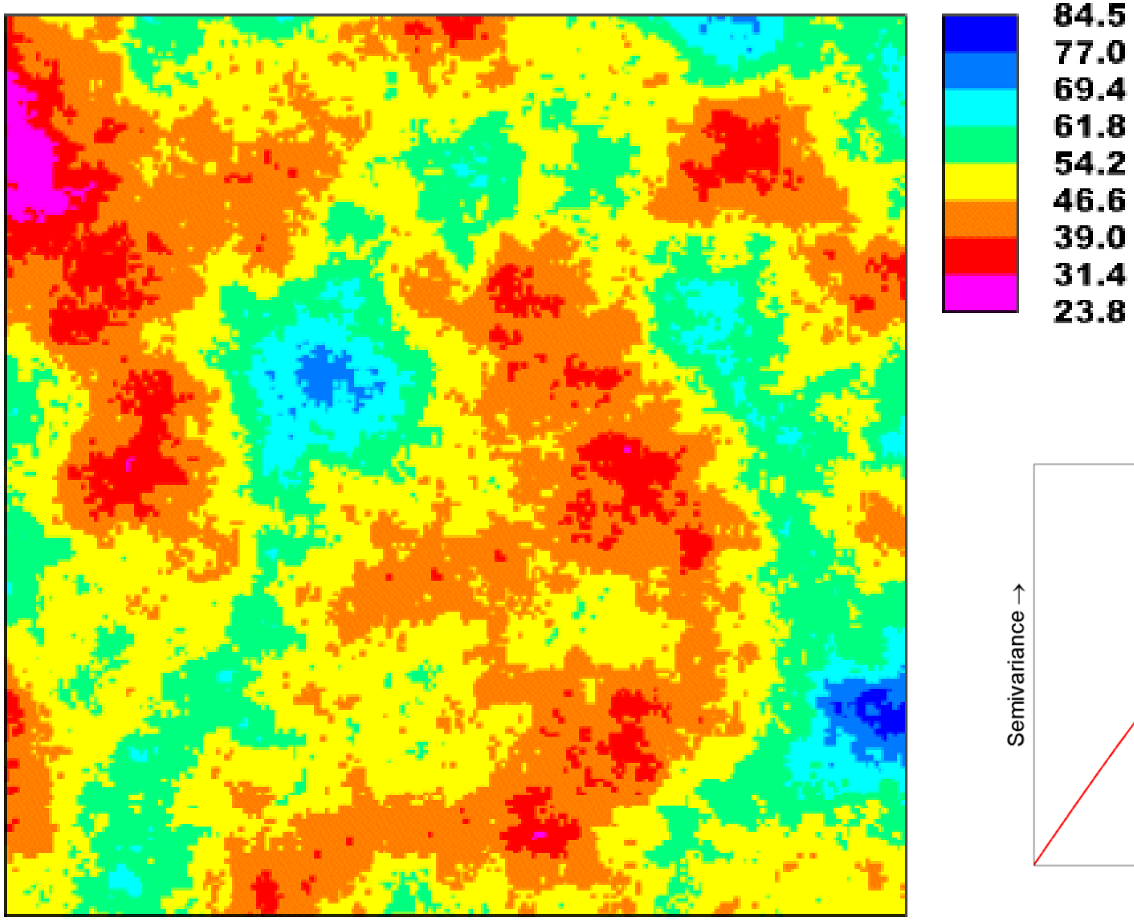
Typical shape of the semivariogram, with parameters

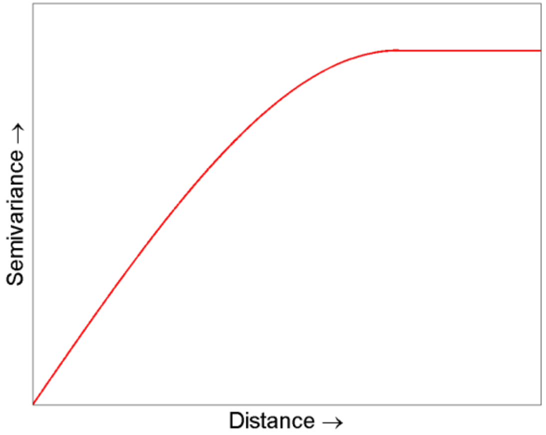


Interpretation of semivariogram parameters

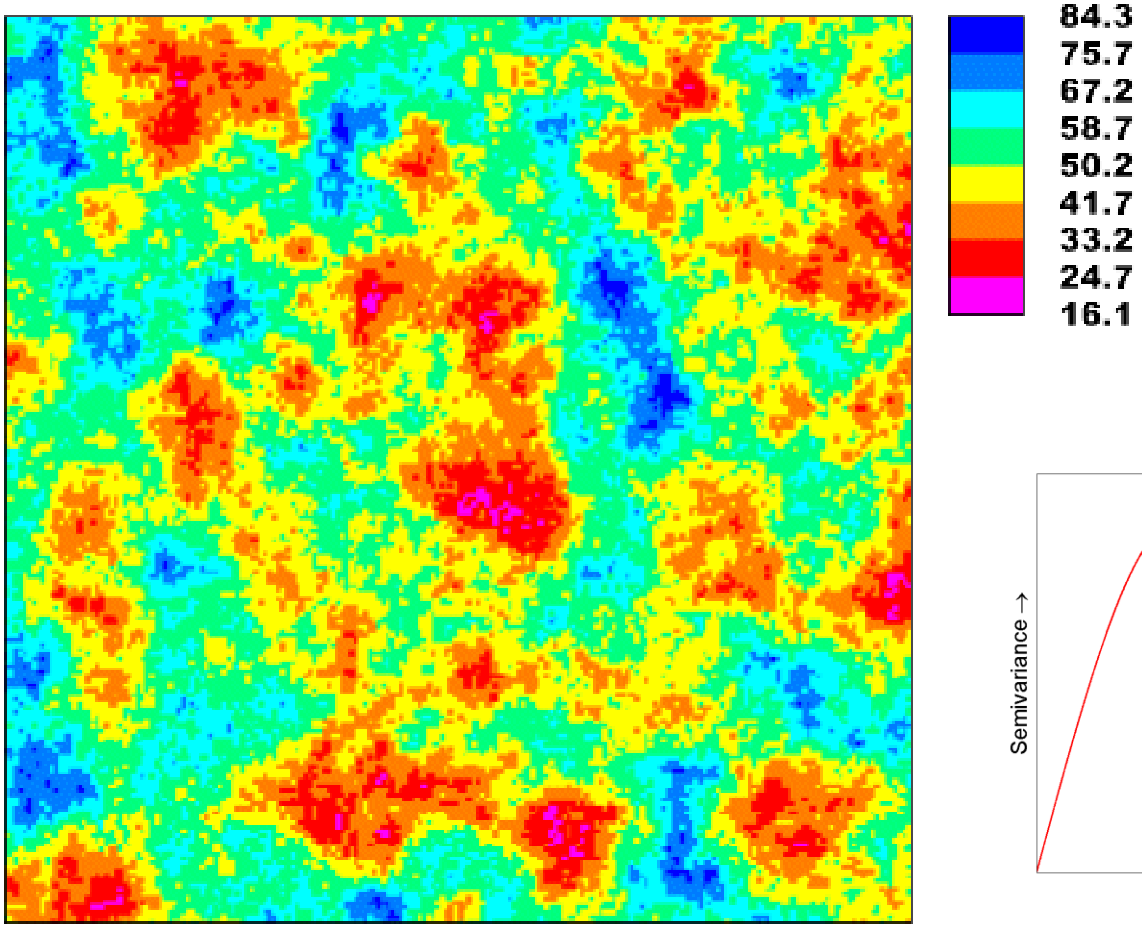
- **Nugget** = measurement errors and short-distance spatial variation
- **Sill** = variance of the variable of interest
- **Range** = distance up to which there is spatial correlation

Example 1: a possible ‘reality’

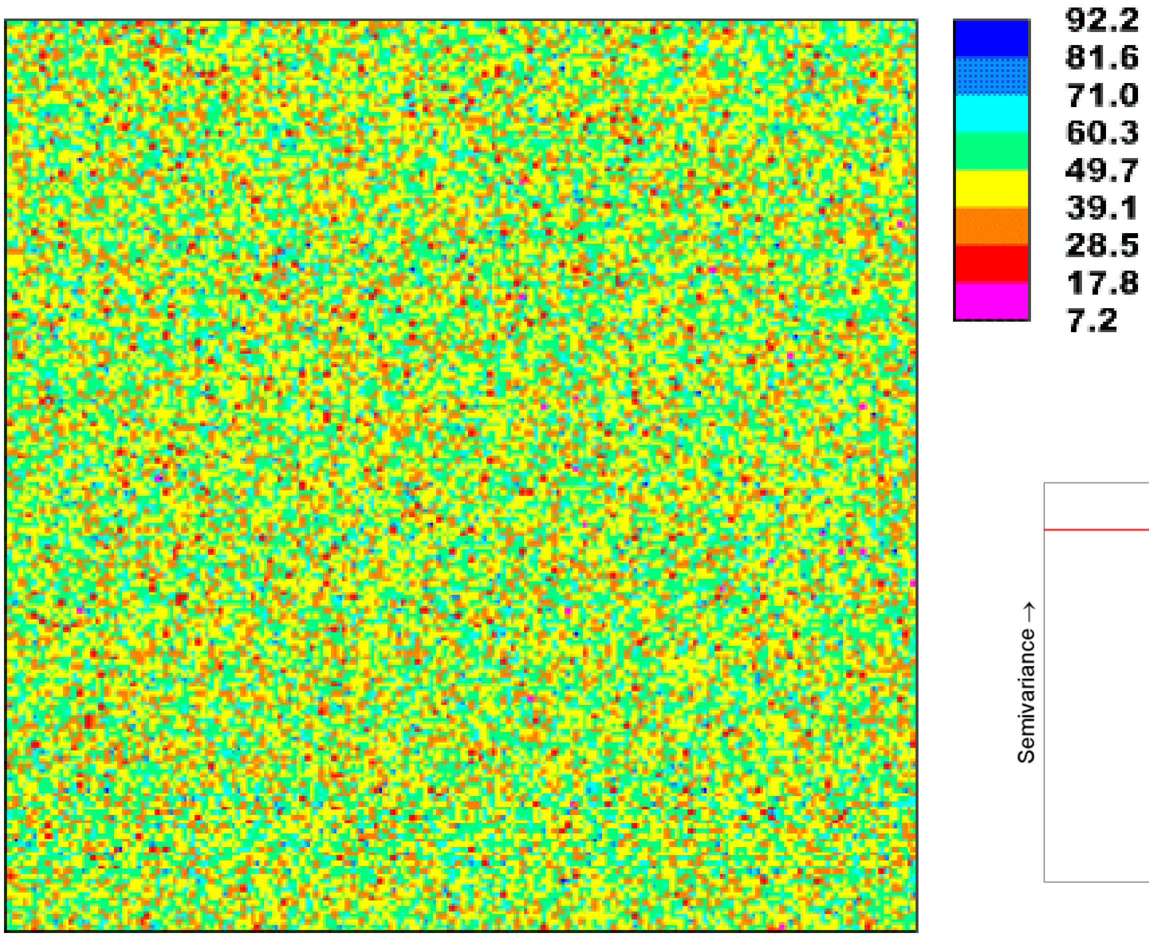


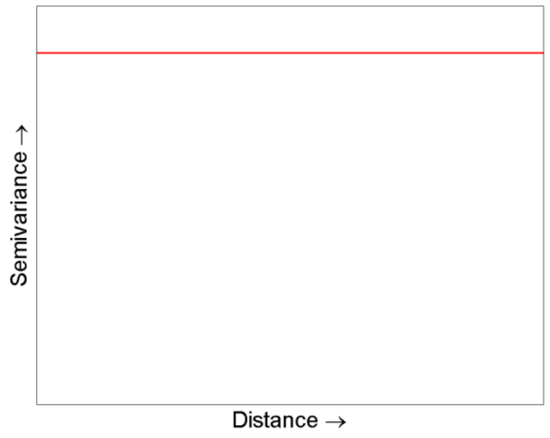


Example 2: another reality

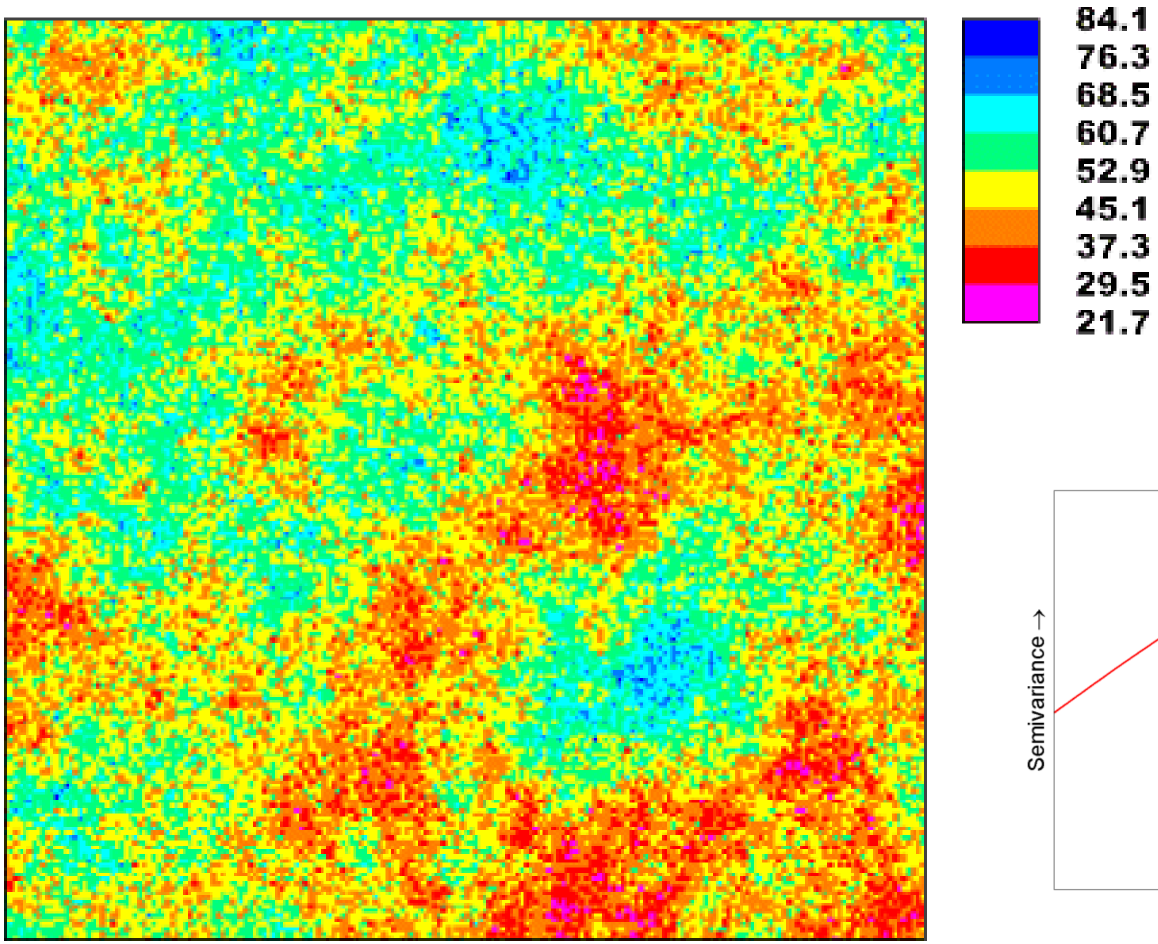


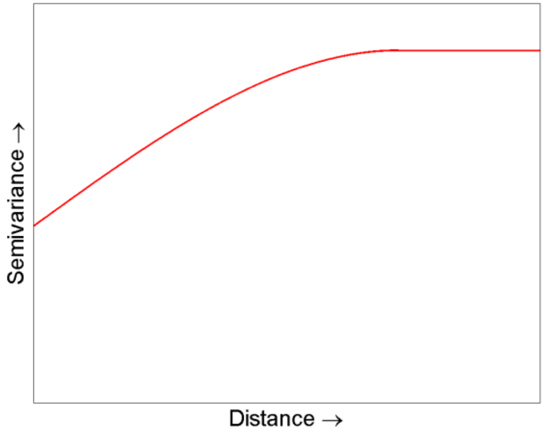
Example 3: third reality

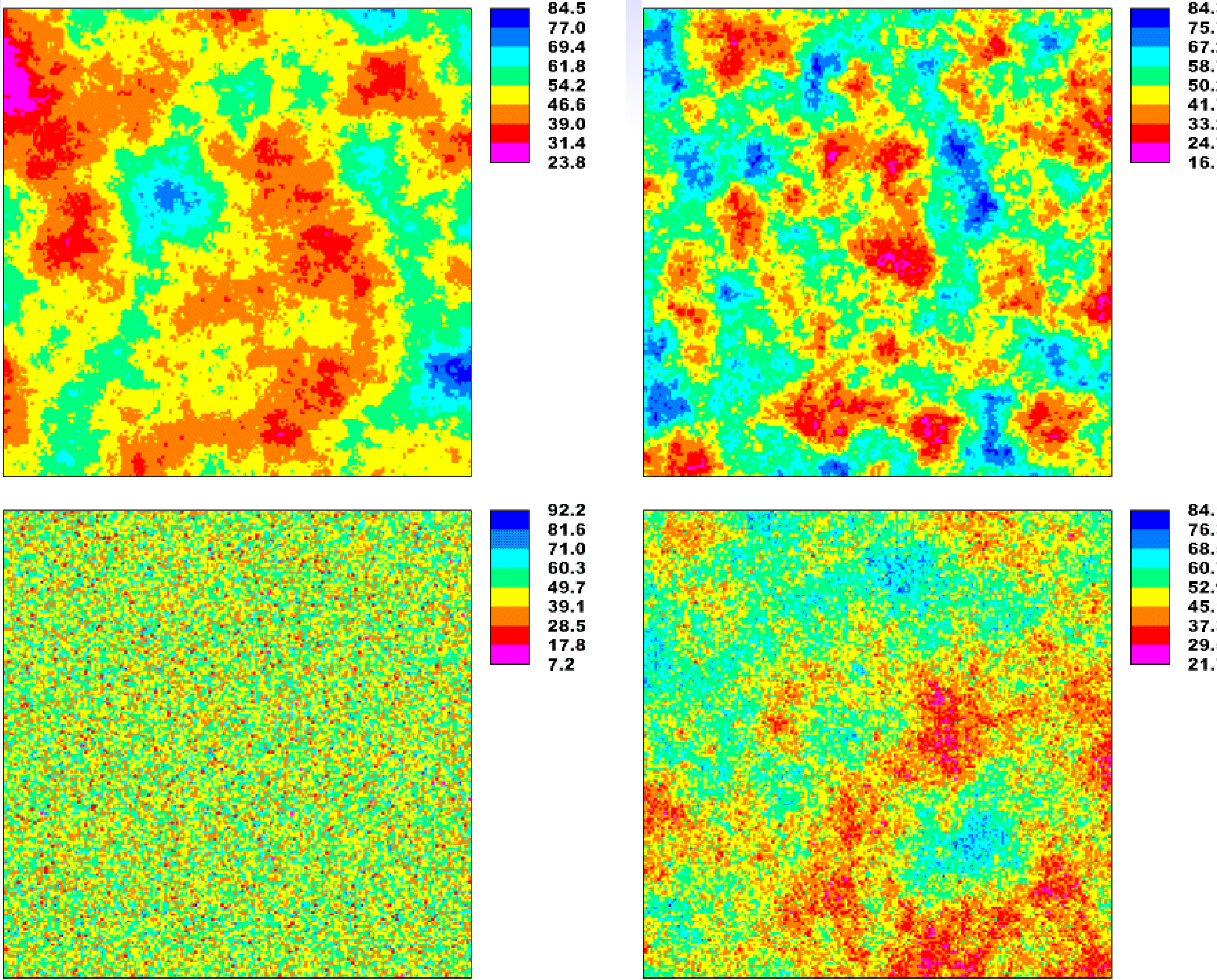




Example 4: fourth reality





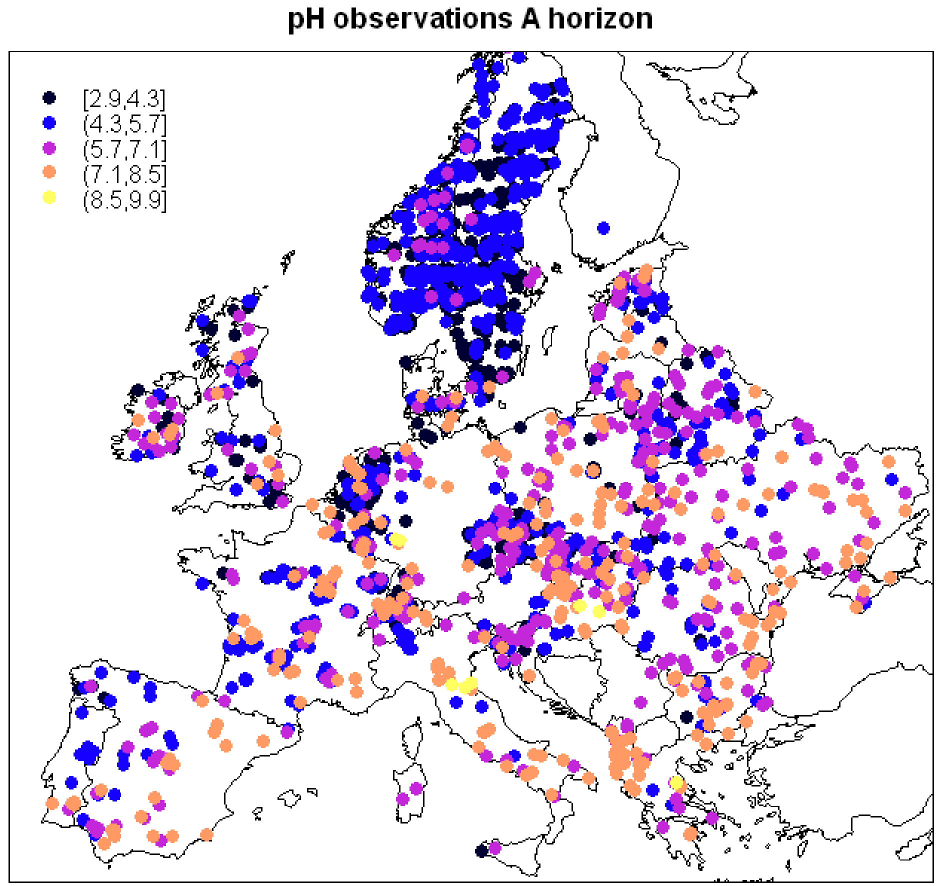

84.5
77.0
69.4
61.8
54.2
46.6
39.0
31.4
23.8
84.3
75.7
67.2
58.7
50.2
41.7
33.2
24.7
16.1
92.2
81.6
71.0
60.3
49.7
39.1
28.5
17.8
7.2
84.1
76.3
68.5
60.7
52.9
45.1
37.3
29.5
21.7

When presenting the semivariogram, two assumptions were implicitly made:

1. The semivariance of $Z(x)$ and $Z(x+h)$ only depends on the distance h and not on the locations x and $x+h$ (**stationarity** assumption)
2. The semivariance is a function of the length of h, not of its direction (**isotropy** assumption)

These assumptions are not always realistic and can be relaxed, but today we won’t go into that

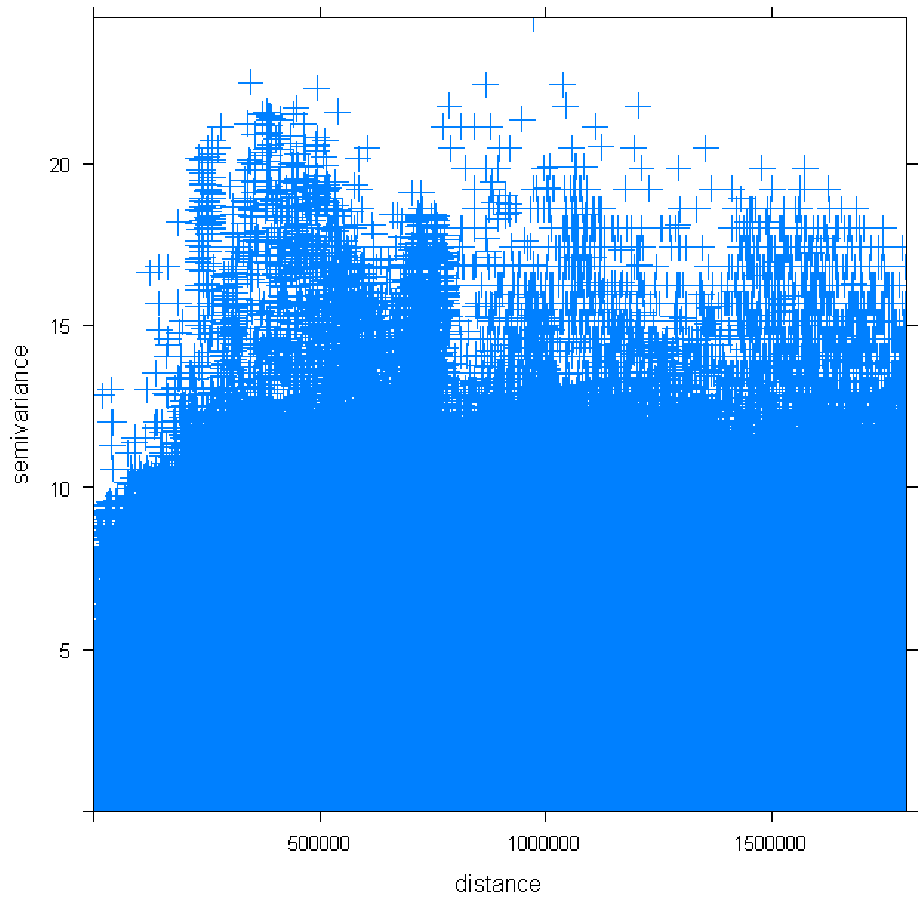
Estimation of the semivariogram from point observations, take pH data Europe as an example



Structural analysis: estimate semivariogram from observations

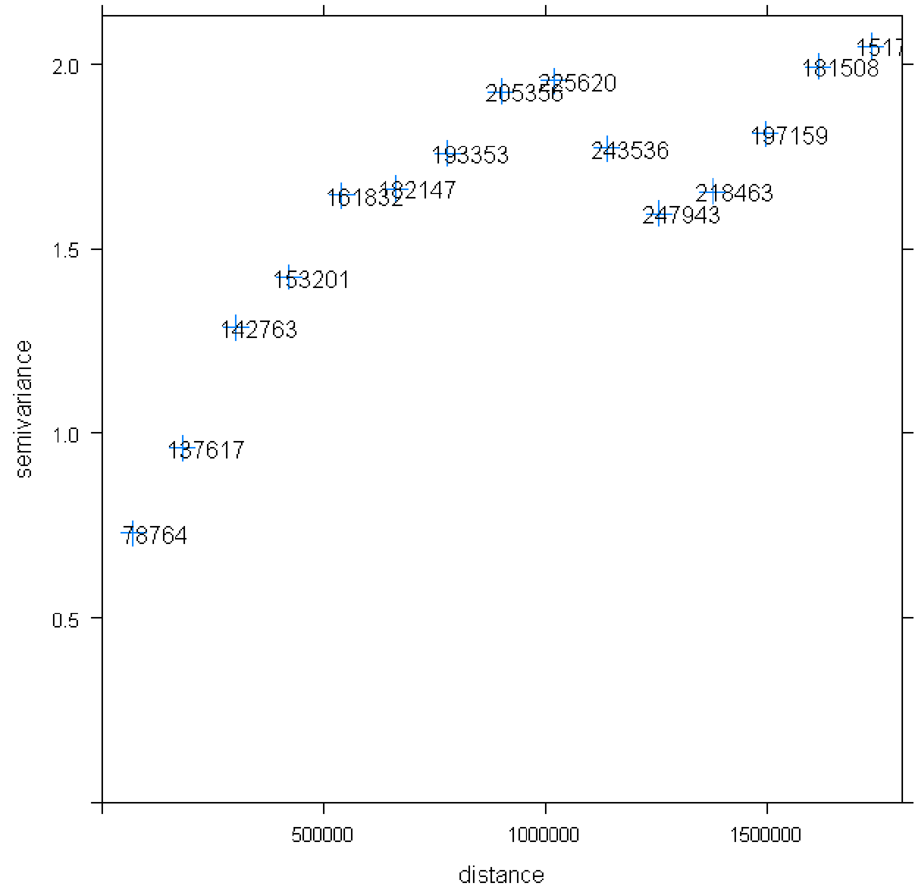
- Suppose there are n observations (in this example n=2582)
- This yields $\frac{1}{2} \times n \times (n-1)$ pairs of observations
- Each pair of observations $\{z(x_i), z(x_j)\}$ provides information about the semivariance over distance $|x_i - x_j|$ (by computing $\frac{1}{2} \times (z(x_i) - z(x_j))^2$)
- Presented in a graph: semivariogram cloud

Semivariogram cloud pH data



- How many blue crosses are in this figure?

Averaging over 'lags' (intervals) gives experimental semivariogram



Last step of structural analysis: fit a function through the experimental semivariogram

1. Choose a function shape, common choices:

spherical:

$$\gamma(h) = 0 \quad \text{for } h = 0$$

$$\gamma(h) = c_0 + c \cdot \left\{\frac{3}{2} \cdot \frac{h}{a} - \frac{1}{2} \cdot \left(\frac{h}{a}\right)^3\right\} \quad \text{for } 0 < h \leq a$$

$$\gamma(h) = c_0 + c \quad \text{for } h > a$$

linear:

$$\gamma(h) = 0 \quad \text{for } h = 0$$

$$\gamma(h) = c_0 + b \cdot h \quad \text{for } h > 0$$

exponential:

$$\gamma(h) = 0 \quad \text{for } h = 0$$

$$\gamma(h) = c_0 + c \cdot (1 - e^{-\frac{h}{a}}) \quad \text{for } h > 0$$

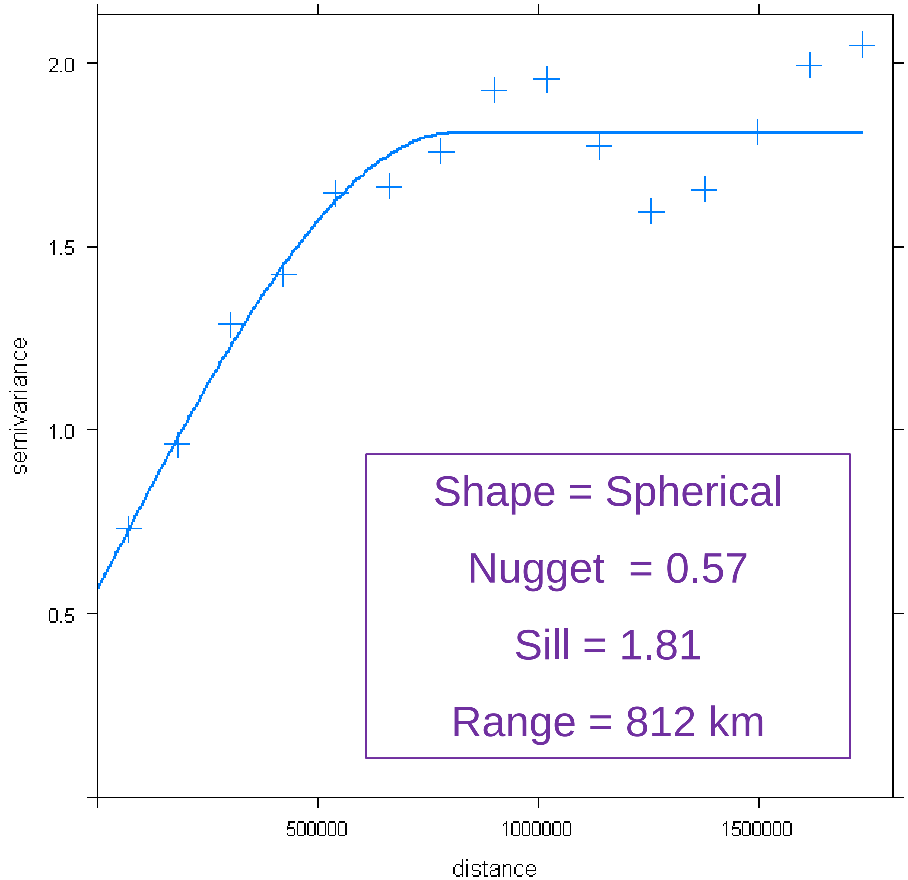
Gaussian:

$$\gamma(h) = 0 \quad \text{for } h = 0$$

$$\gamma(h) = c_0 + c \cdot (1 - e^{-\left(\frac{h}{a}\right)^2}) \quad \text{for } h > 0$$

2. Estimate parameters of the chosen shape (e.g. using weighted least squares fitting)

Resulting variogram model for pH example:



- Does it agree with our hypothesis?

Geostatistical interpolation: Kriging

- Introduced in the 1950s by Daniel Krige: mining engineer from South-Africa
- Kriging comes in many forms, we focus on Ordinary Kriging but will also look at Regression Kriging
- Principle: prediction at a location is a linear combination of observations nearby
- The weight that is given to each observation depends on the degree of (spatial) correlation: the semivariogram plays an important role

Ordinary Kriging

Predict $Z(x_0)$ at unobserved location s_0 using observations $Z(x_i)$, $i=1,\ldots,n$ as follows:

$$\hat{Z}_{OK}(x_0) = \sum_{i=1}^{n} \lambda_i \cdot Z(x_i)$$

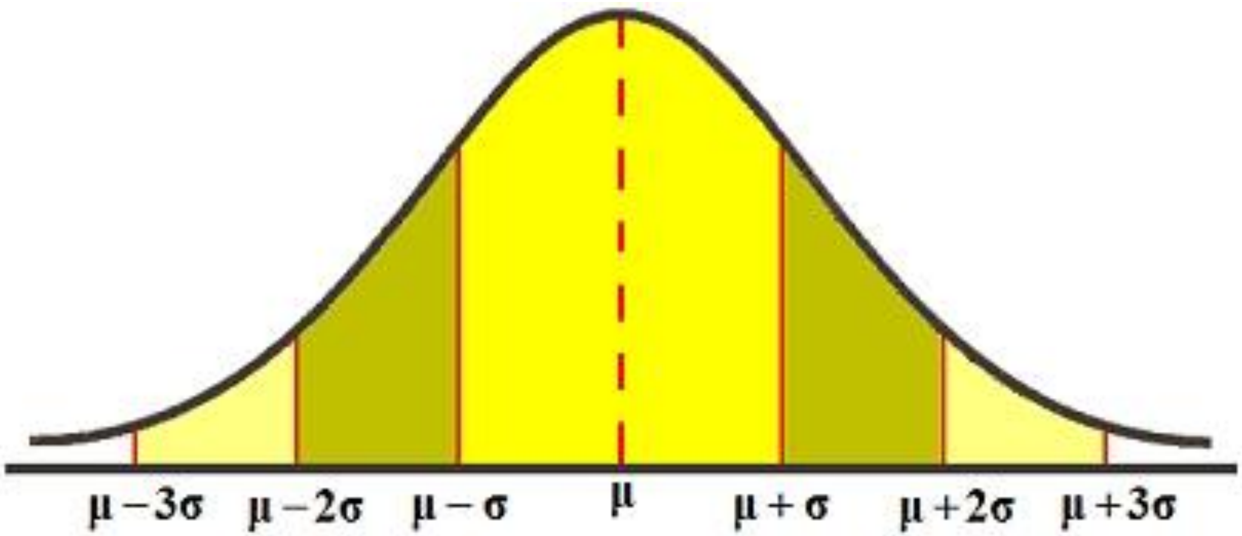
Kriging weight

How to choose the kriging weights?

- What criterion would you recommend?

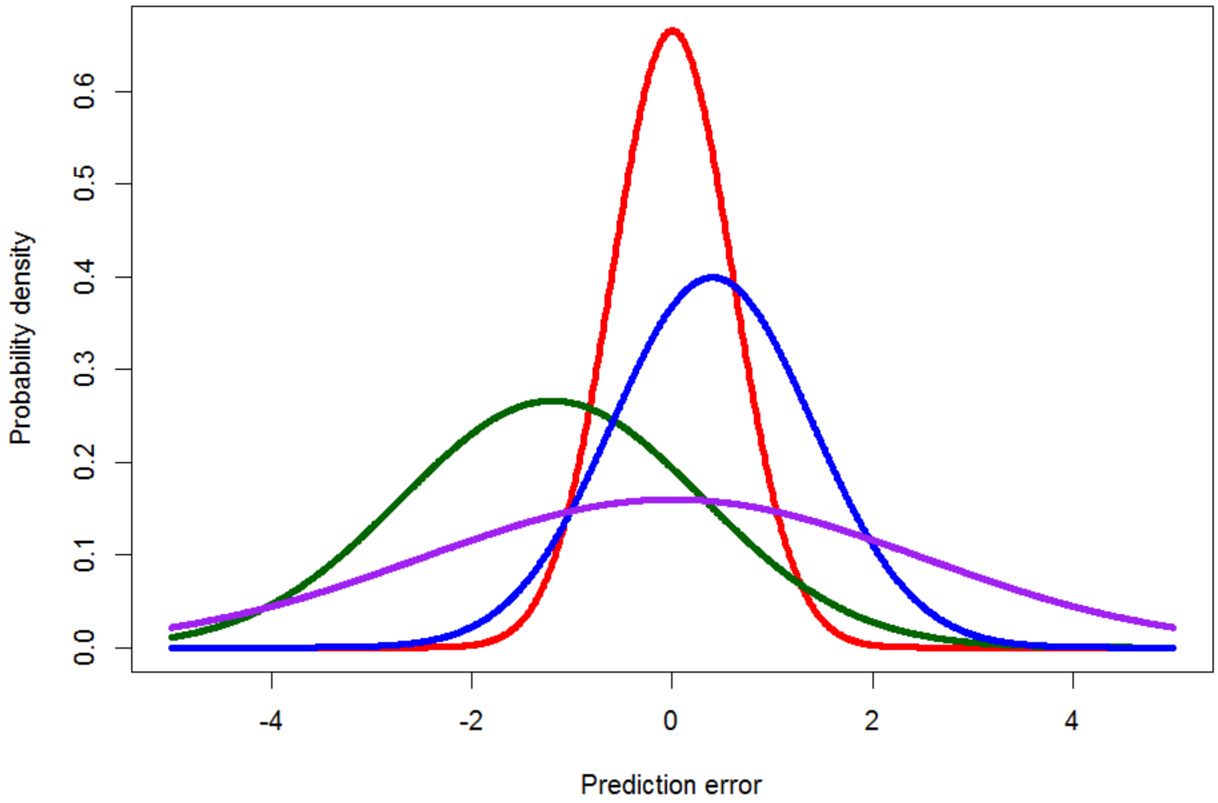
How to choose the kriging weights?

- What criterion would you recommend?



- What would be nice properties of the kriging prediction error $\hat{Z}(x_0) - Z(x_0)$?

Which probability distribution of $\hat{Z}(x_0) - Z(x_0)$ do you prefer?



Computation of kriging weights

Minimise the expected squared prediction error, in other words choose the coefficients λ_i such that:

$$E[(\hat{Z}(x_0) - Z(x_0))^2]$$

is as small as possible, under the unbiasedness condition:

$$\sum_{1=1}^{n} \lambda_i = 1$$

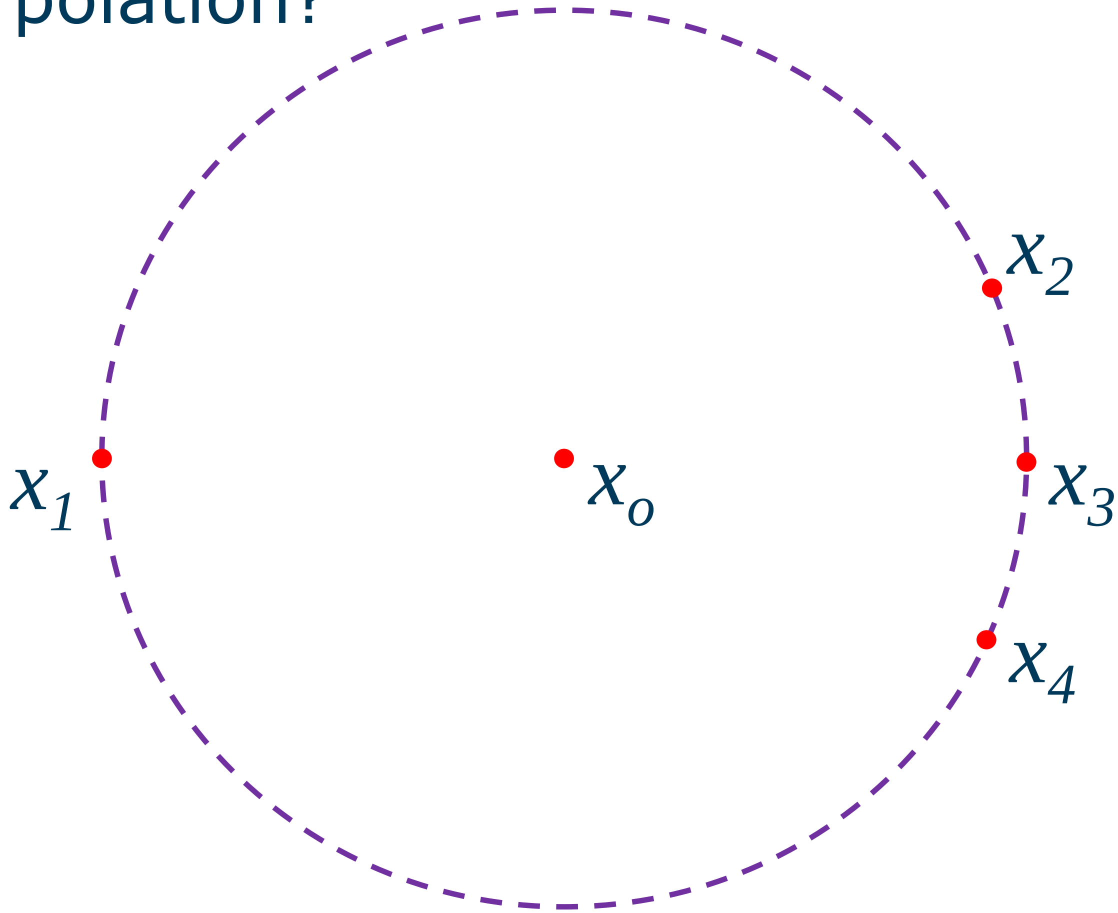
Solution in case of Ordinary Kriging:

Lagrange parameter

$$\sum_{j=1}^{n} \lambda_j \cdot \gamma(|x_i - x_j|) + \varphi = \gamma(|x_i - x_0|) \quad \text{for all } i = 1, \ldots, n$$

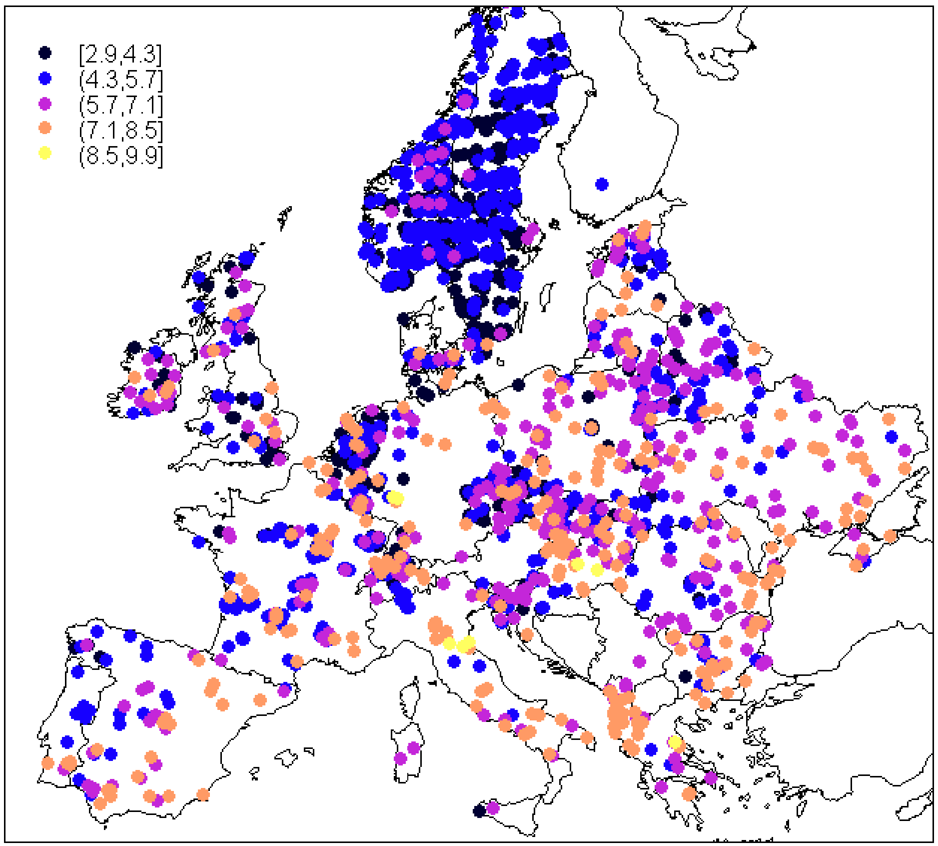
$$\text{In addition} \quad \sum_{i=1}^{n} \lambda_i = 1$$

Which observation gets the largest weight? Which the smallest? What would we get in case of inverse distance interpolation?

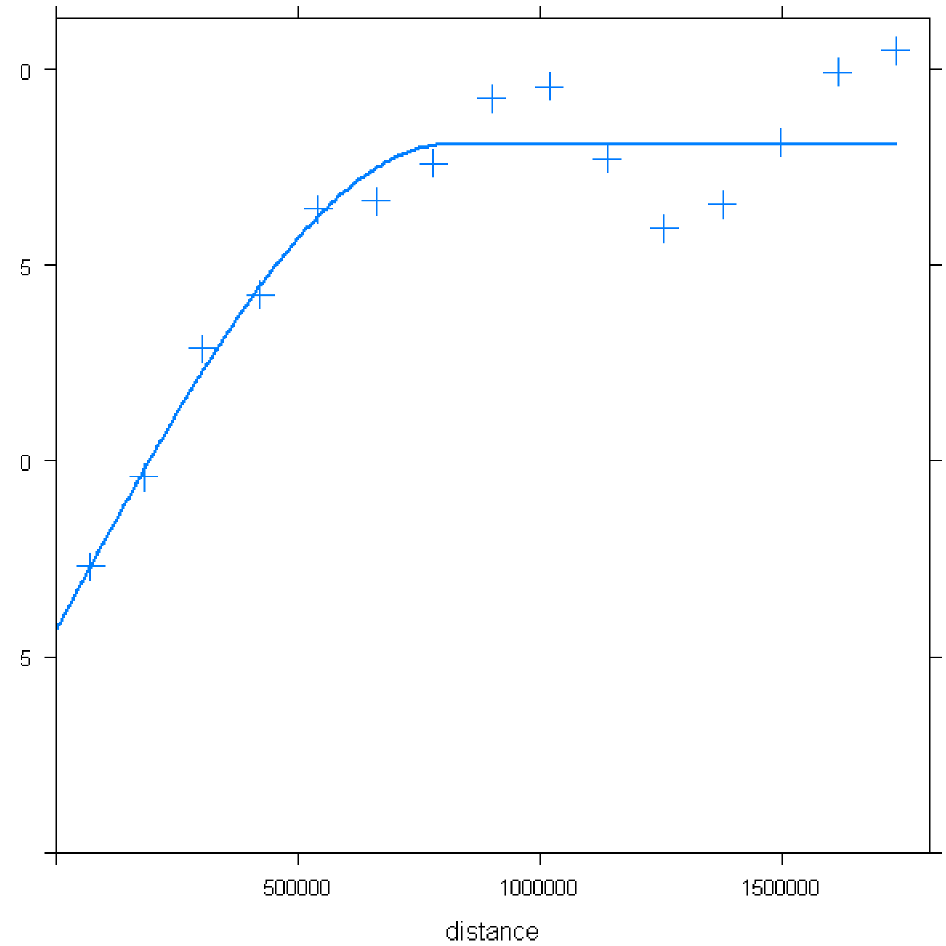


Application to European pH data

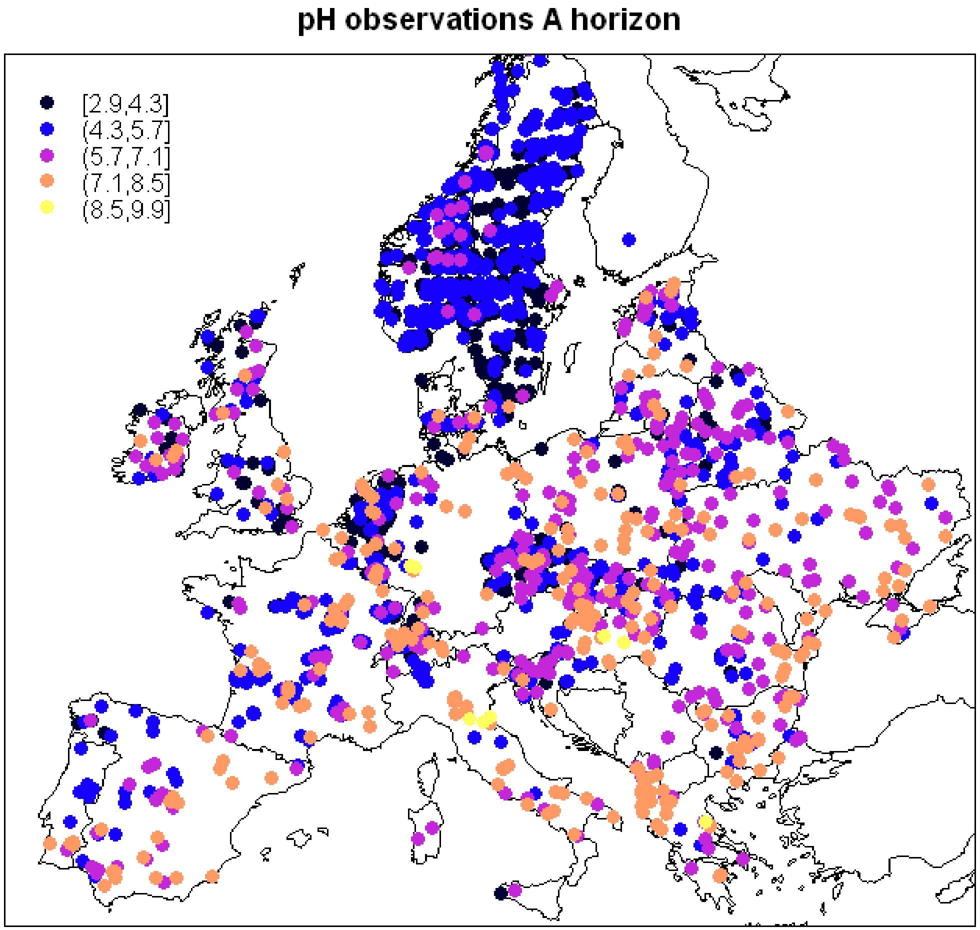
pH observations A horizon

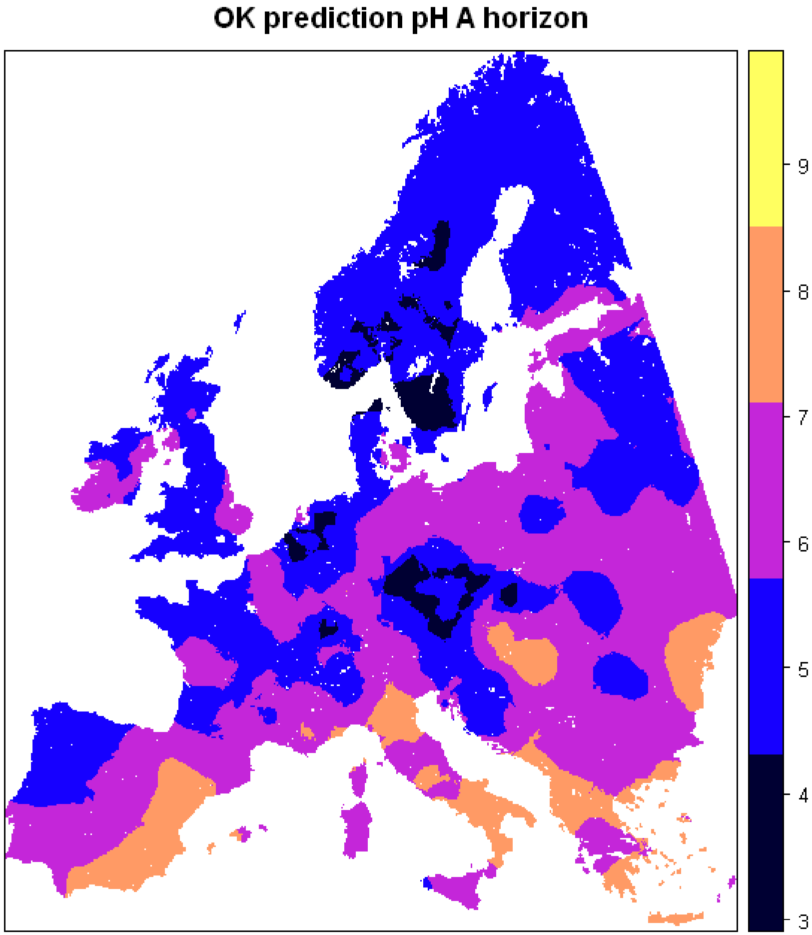


Semivariogam pH A horizon



OK prediction yields a smoothed representation of reality





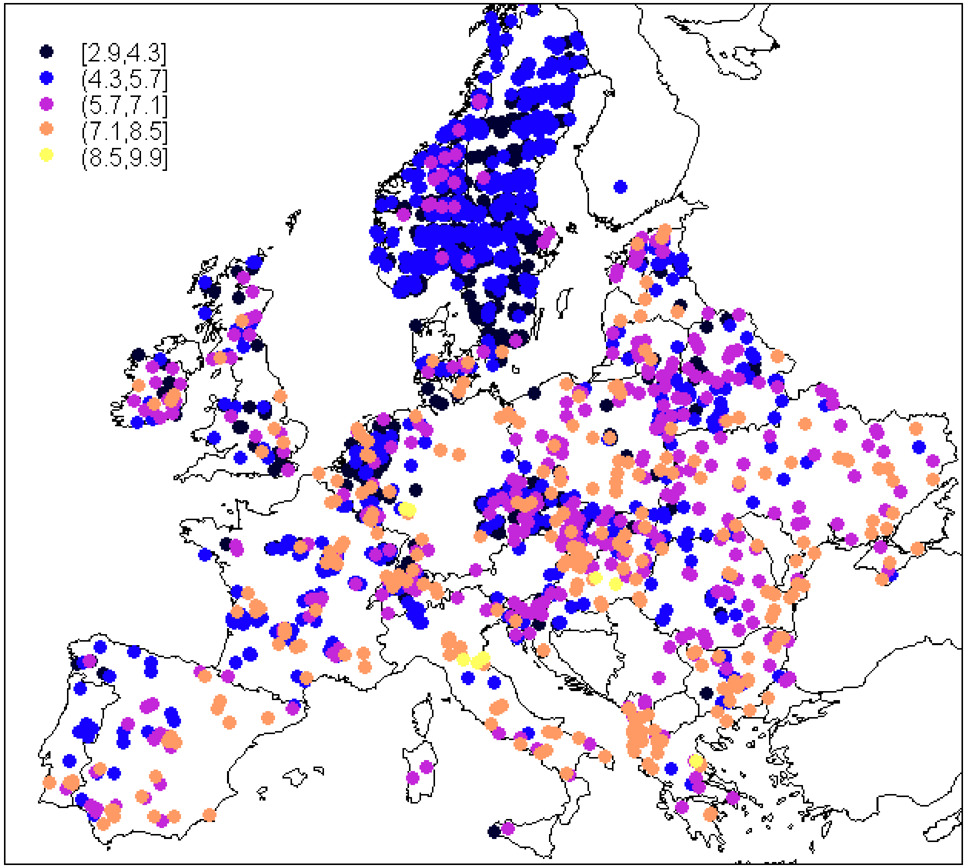
The kriging interpolation error is quantified by the Ordinary Kriging variance:

$$\sigma^2_{OK}(x_0) = E\left[(Z(x_0) - \hat{Z}(x_0))^2\right] = \sum_{i=1}^{n} \lambda_i \cdot \overbrace{\gamma(x_i - x_0)}^{\text{semivariance at distance } x_i - x_0} + \varphi$$

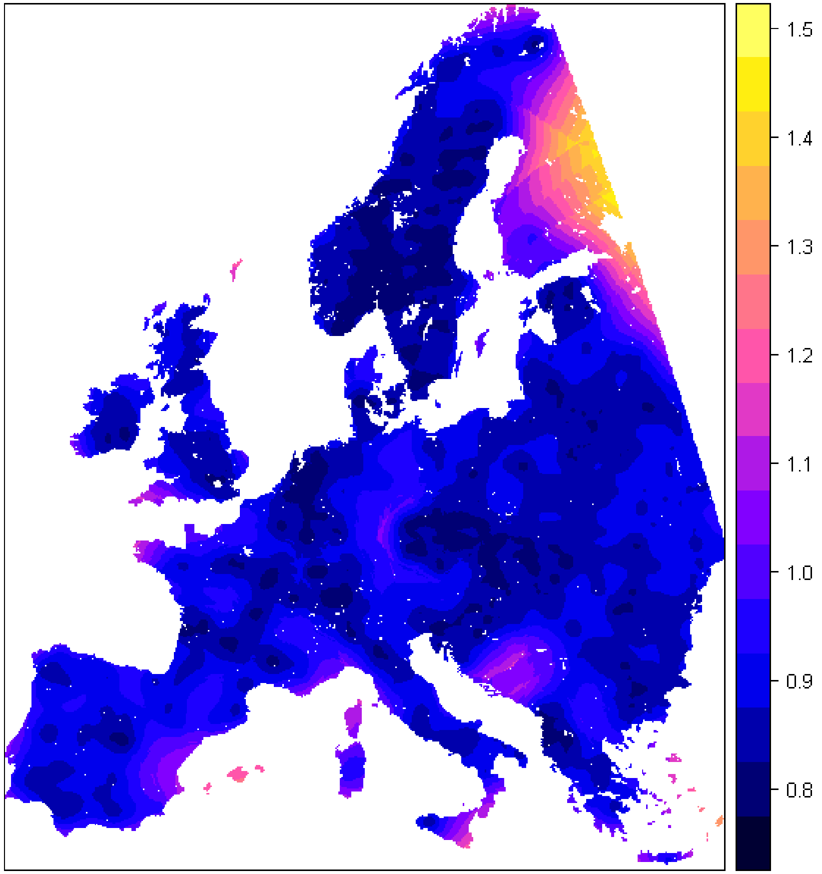
This shows that the interpolation error is small when there is strong spatial correlation and/or when there are many observations in the local neighbourhood

Map of OK standard deviations ($=\sqrt{\sigma_{OK}{}^2}$) shows observation density

pH observations A horizon



OK standard deviation pH A horizon



ISRIC World Soil Information

Regression kriging

$$Z(x) = m(x) + \varepsilon(x)$$

dependent, target variable

trend, explanatory part

stochastic residual, unexplanatory part, can be spatially correlated!

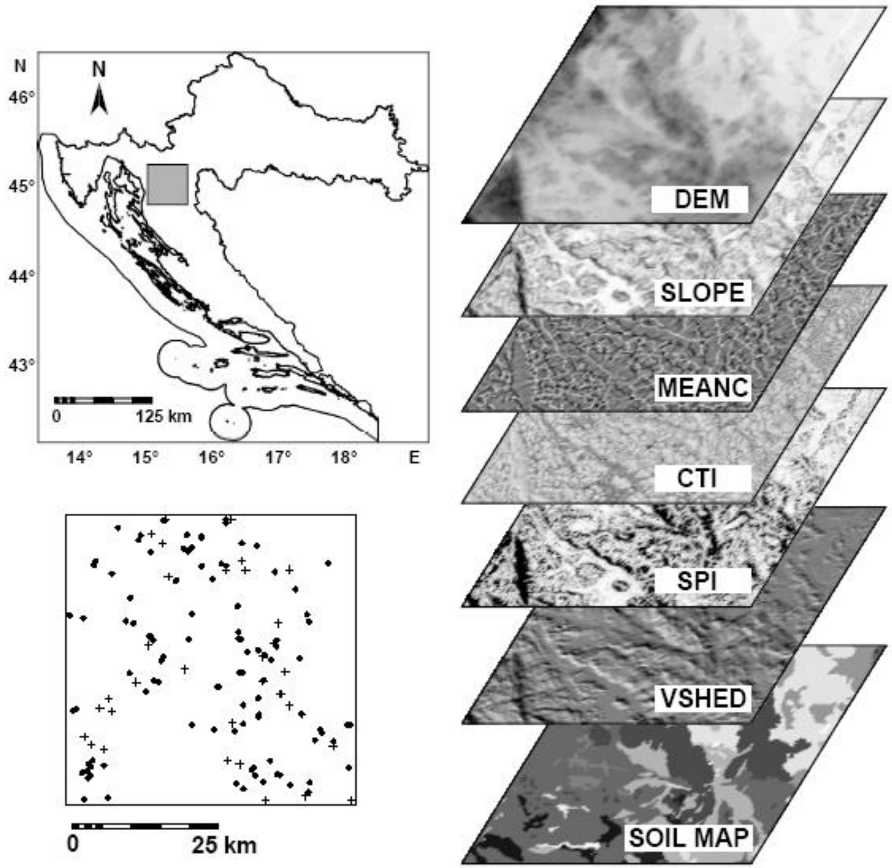
Unlike ordinary kriging, in regression kriging the trend is no longer constant but a function of ’explanatory’ variables, for example:

$$soil\ depth(x) = \begin{matrix} \beta_0 + \beta_1 \cdot elevation(x) + \beta_2 \cdot slope\ angle(x) \\ + \beta_3 \cdot vegetation\ density(x) \\ + \beta_4 \cdot upstream\ area(x) \end{matrix} + residual(x)$$

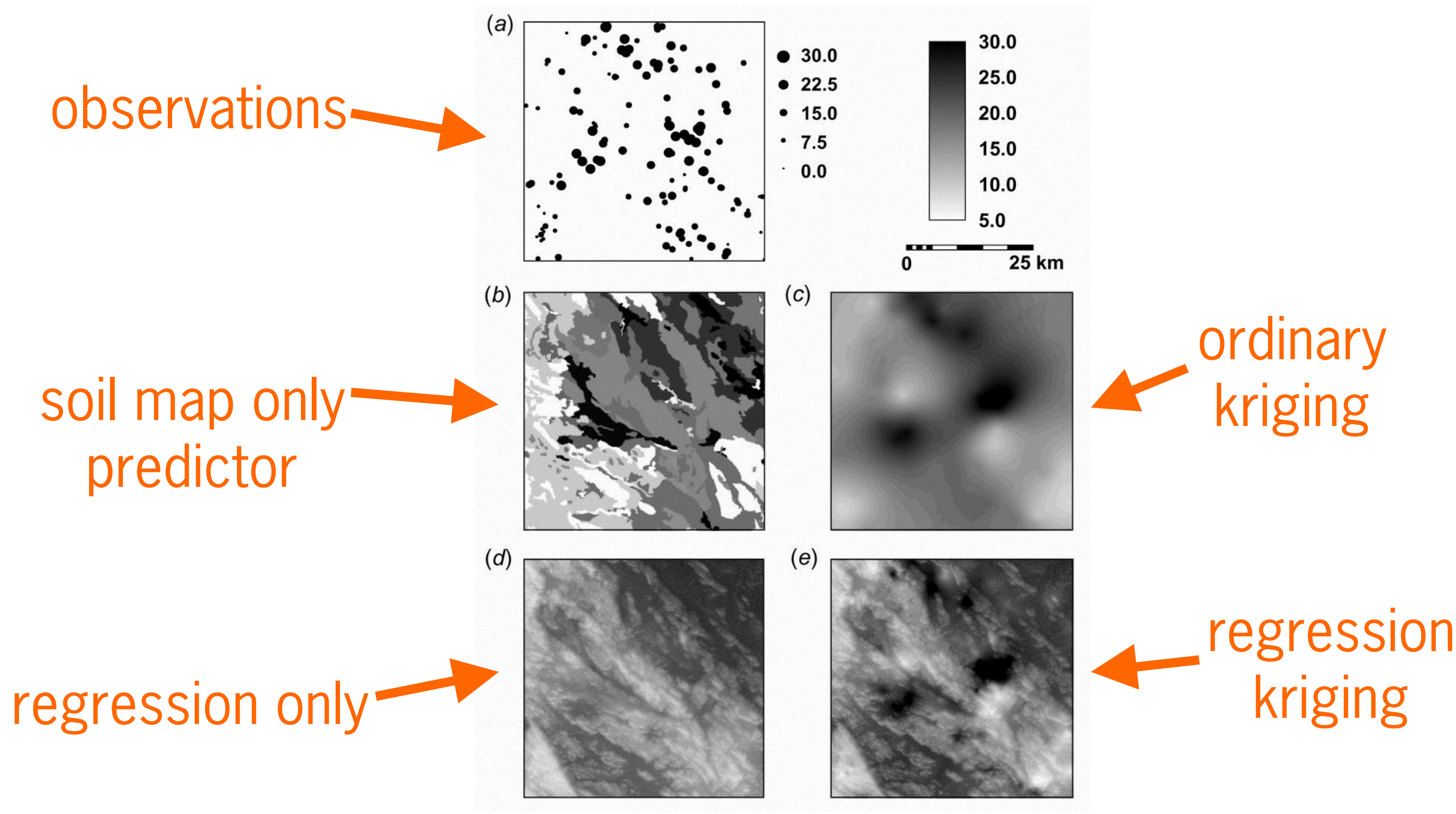
Regression kriging algorithm

1. select explanatory variables and fit regression model (estimate regression coefficients)
2. compute residuals (by subtracting the fitted trend from the observations) at observation locations and compute from them a semivariogram
3. apply the regression model to all unobserved locations (usually a grid)
4. krige the residuals
5. add up the results of steps 3 and 4

Example from Hengl et al. (Geoderma, 2004): predicting soil depth for a 50 × 50 km area in Croatia



Results using four interpolation methods

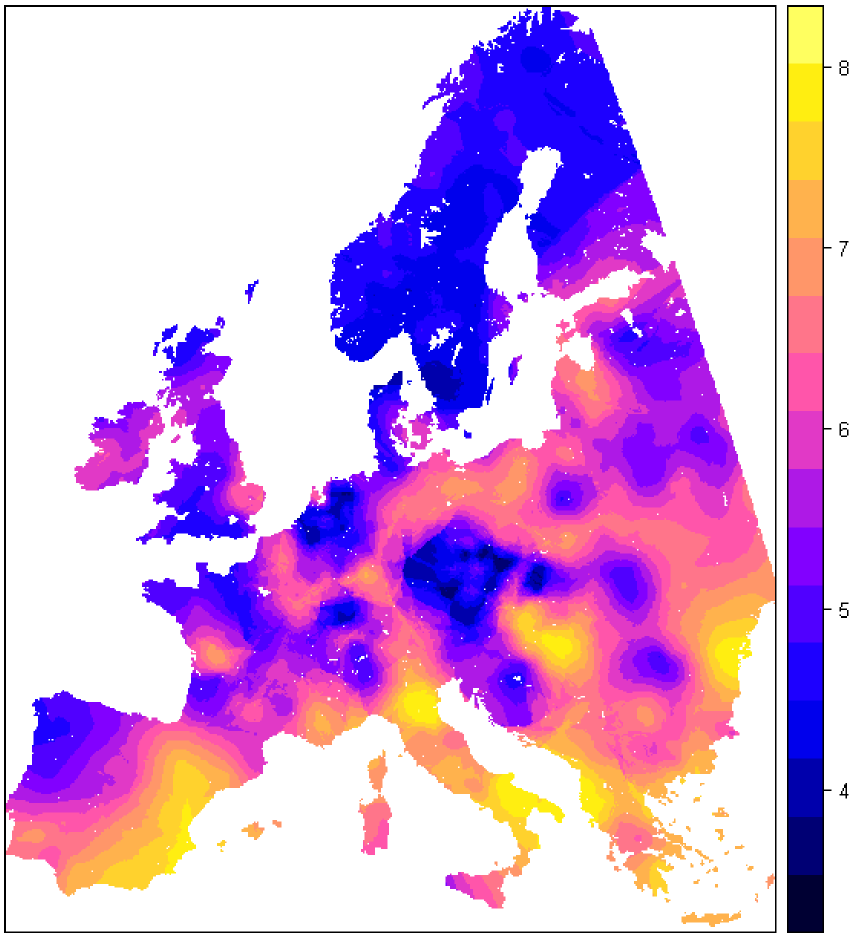


Validation on 35 independent observations

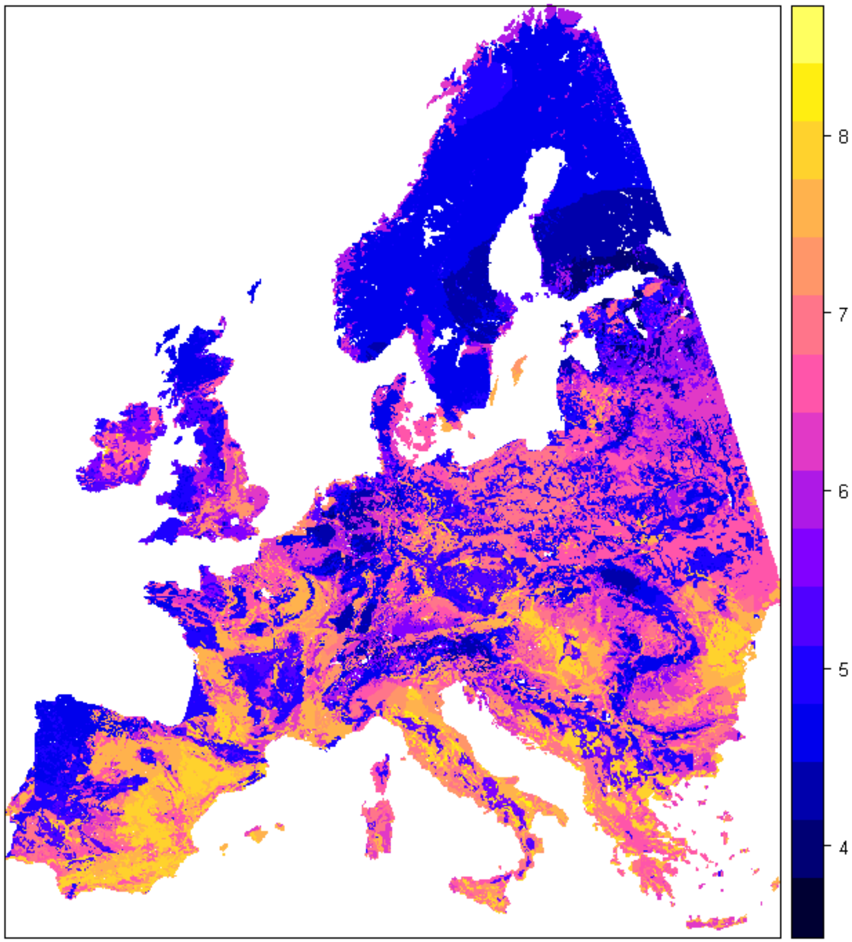
	Mean error [cm]	Root mean squared error [cm]
Soil map	1.42	9.1
Ordinary kriging	0.69	8.5
Multiple regression	1.69	8.8
Regression kriging	0.15	6.8

Regression kriging also yields more realistic results for pH in Europe

Ordinary Kriging

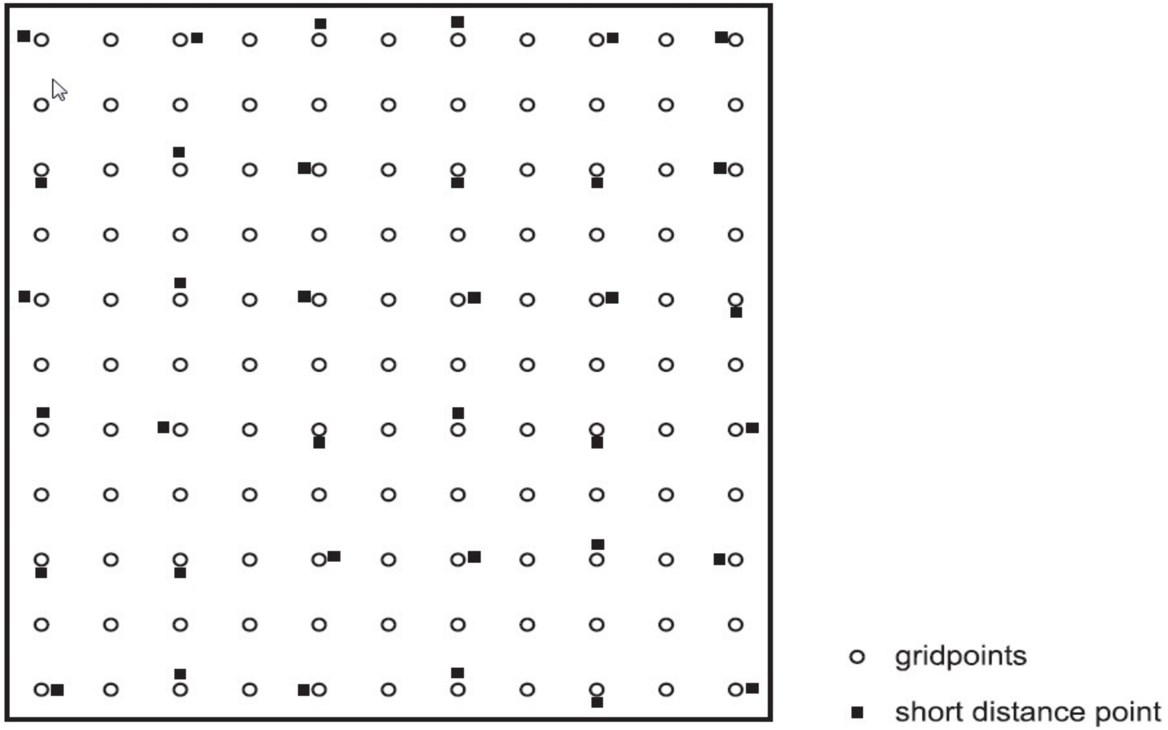


Regression Kriging

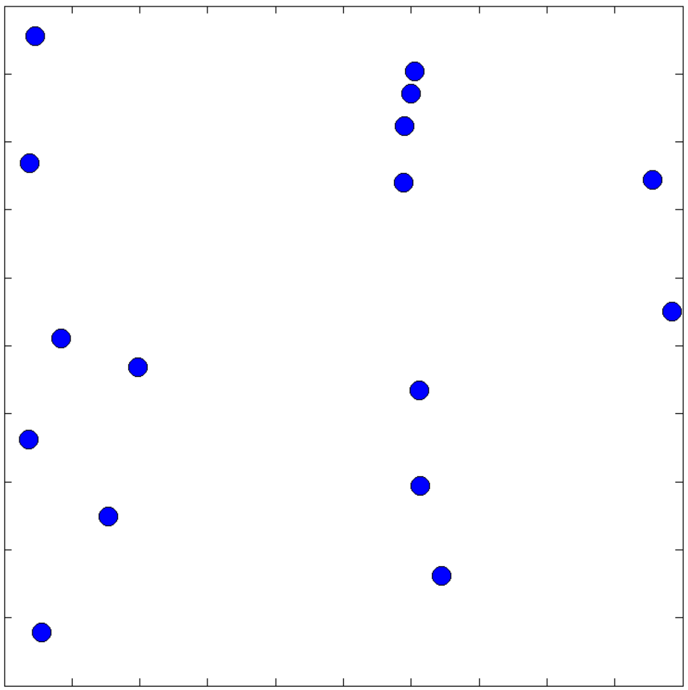


Which sampling design is best?

- Variogram estimation has different requirements than kriging: quantification of short-distance spatial variation versus uniform spread, pragmatic solution:



If you really want the optimum design: given the semivariogram, it can be obtained with numerical techniques such as spatial simulated annealing



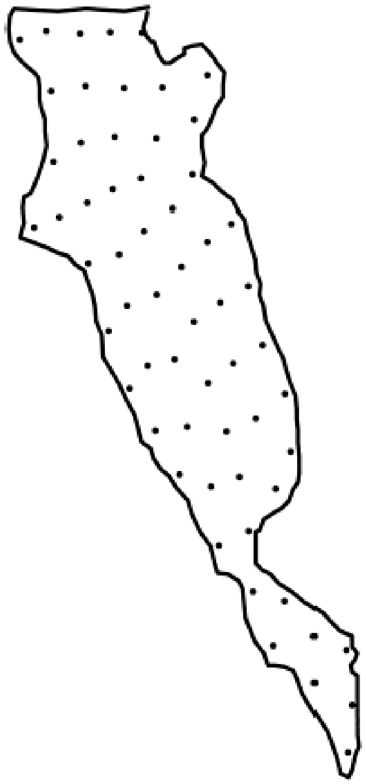


Fig. 12. An a priori optimised sampling scheme for anisotropic sand percentage.

Optimization in feature space

$\sigma_{RK}{}^2(x) =$

spatial

ELSEVIER

Computers & Geosciences 32 (2006) 1378–1388

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

A conditioned Latin hypercube method for sampling in the presence of ancillary information ☆

Budiman Minasny*, Alex B. McBratney

Australian Centre for Precision Agriculture, Faculty of Agriculture, Food and Natural Resources, The University of Sydney, Australia

Received 9 September 2005; received in revised form 3 December 2005; accepted 16 December 2005

Abstract

This paper presents the conditioned Latin hypercube as a sampling strategy of an area with prior information represented as exhaustive ancillary data. Latin hypercube sampling (LHS) is a stratified random procedure that provides an efficient way of sampling variables from their multivariate distributions. It provides a full coverage of the range of each variable by maximally stratifying the marginal distribution. For conditioned Latin hypercube sampling (cLHS) the problem is: given N sites with ancillary variables (X), select x a sub-sample of size n ($n \ll N$) in order that x forms a Latin hypercube, or the multivariate distribution of X is maximally stratified. This paper presents the cLHS method with a search algorithm based on heuristic rules combined with an annealing schedule. The method is illustrated with a simple 3-D example and an application in digital soil mapping of part of the Hunter Valley of New South Wales, Australia. Comparison is made with other methods: random sampling, and equal spatial strata. The results show that the cLHS is the most effective way to replicate the distribution of the variables.