GlobalSoilMap and Soil Survey Uncertainty Workshop – Lincoln, NE, August 2012

Introductory Short Course

ISRIC World Soil Information

Gerard Heuvelink

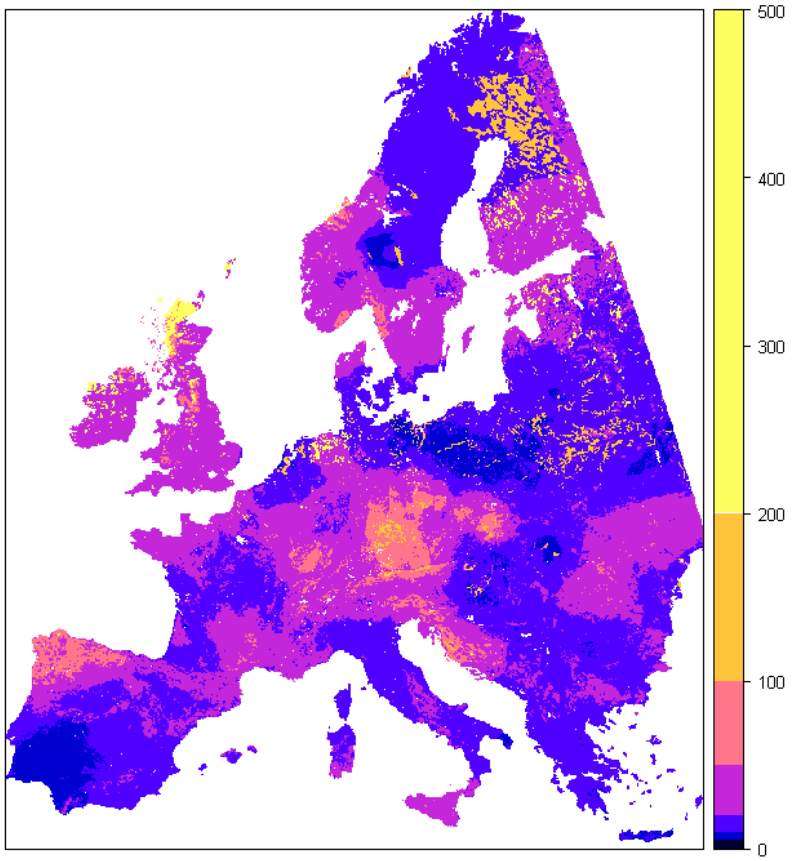
Programme

10.00 – 11.00	Probabilistic modelling of uncertainty
11.00 – 11.30	Coffee break
11.30 – 12.30	Multivariate uncertainty
12.30 – 13.30	Lunch
13.30 – 15.00	Spatial uncertainty
15.00 – 15.30	Coffee break
15.30 – 16.30	Uncertainty quantification for GlobalSoilMap
16.30 – 17.00	Discussion (or Uncertainty Game!?)

PART I: PROBABILISTIC MODELLING OF UNCERTAINTY

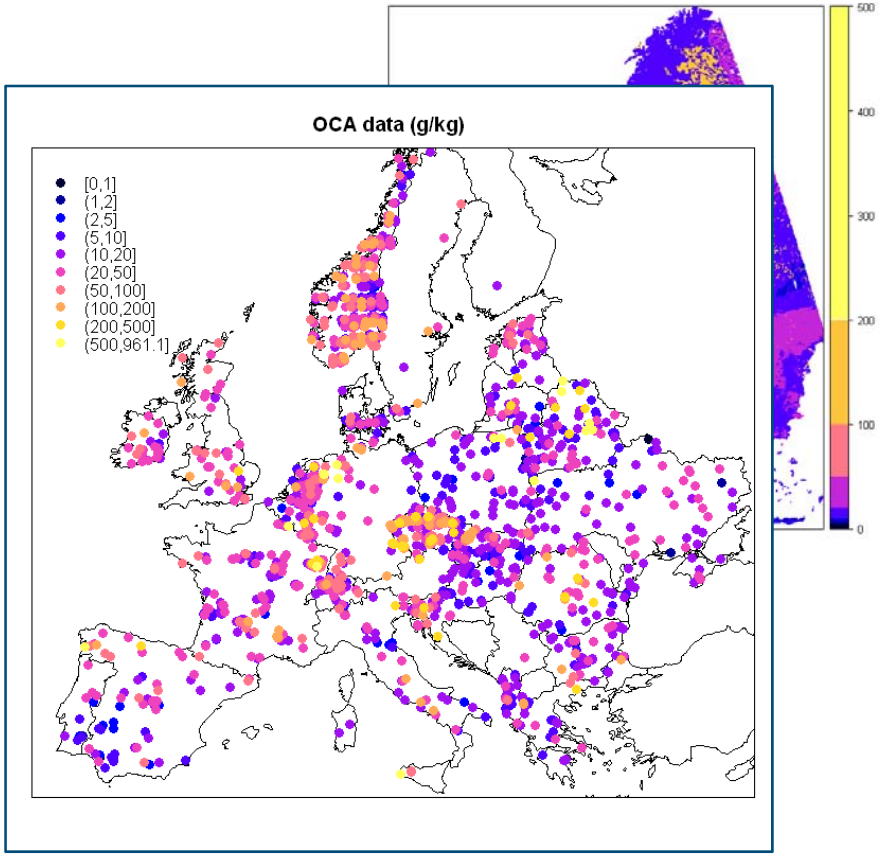
Is this map error-free?

Organic Carbon content
A horizon (g/kg)



Common error sources

- Measurement errors
- Interpretation errors
- Digitization errors
- Classification errors
- Generalisation errors
- Interpolation errors



Error = difference between reality and our representation of reality

- Suppose map says OC=35 g/kg at some location, while true value = 57 g/kg, then there will be no disagreement that in this case the error = 57 – 35 = 22 g/kg
- Other example: number of people in this building
- Or: mean temperature Lincoln, Monday next week (my guess: 77°F = 25°C)
- Problem: we usually do not know the error (because if we did, we would eliminate it!)

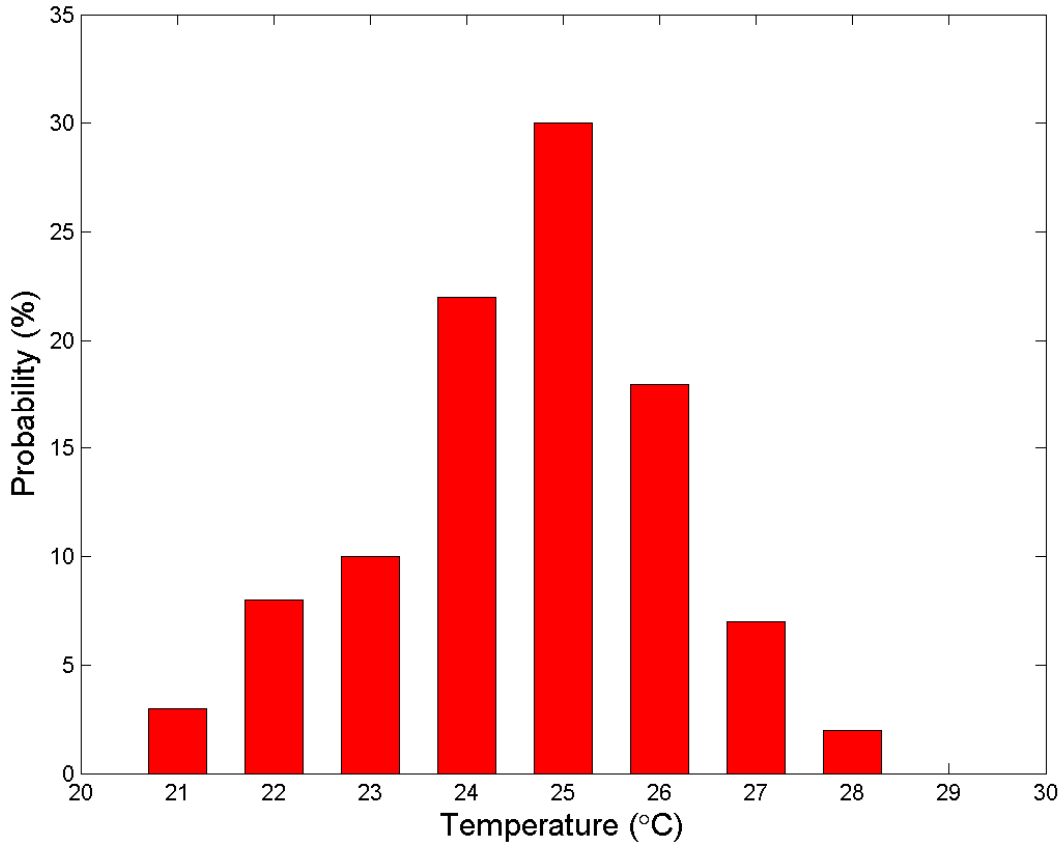
We usually do not know the error

- But in many cases we do know something:
  - We may know that the error has equal chance of being positive or negative, it can go either way
  - We may know that it is unlikely that the absolute value of the error is greater than a given threshold
- In other words, we are uncertain about the true value but not completely ignorant

What is uncertainty?

- Uncertainty arises when we are not sure about the 'true' state of the environment; it is an expression of confidence based on limited knowledge
- Uncertainty is an acknowledgement of error: we are aware that our representation of reality may differ from reality and express this by being uncertain
- Uncertainty is subjective; one person can be more uncertain than another
- In the presence of uncertainty, we cannot identify a true 'reality'. But perhaps we can identify all possible realities and a probability for each one
- Example: temperature Lincoln, Monday next week

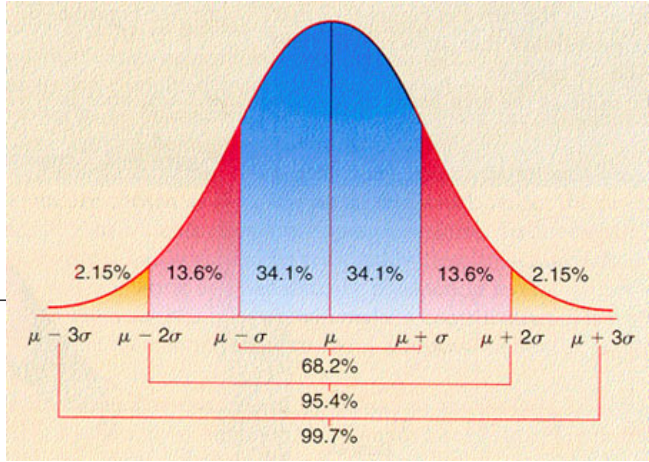
Uncertainty can be described statistically with a probability distribution function (pdf)



The pdf characterizes uncertainty completely, it is all we need

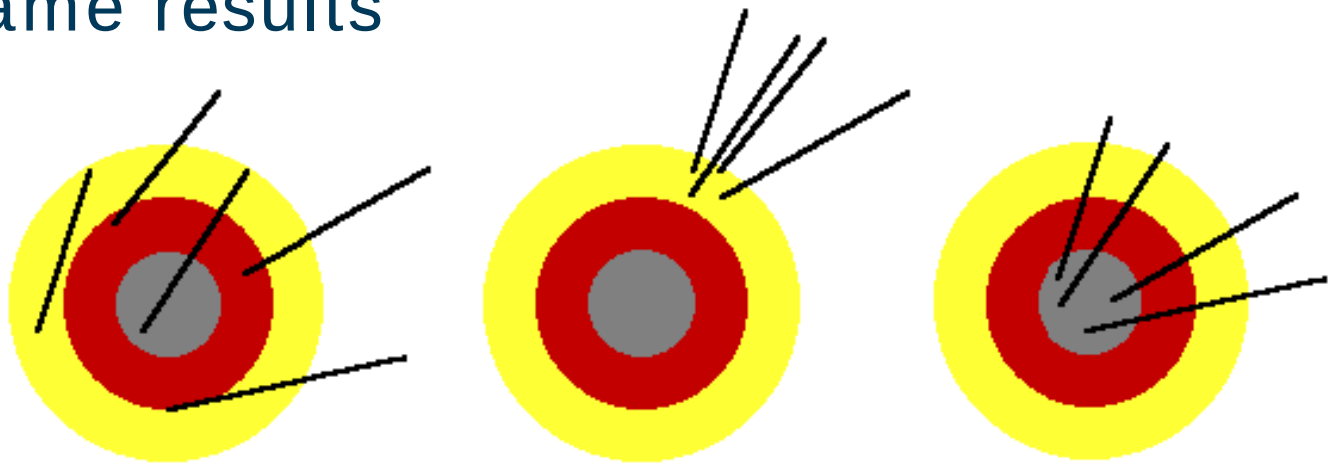
- It is usually parametrized and thus reduced to a few parameters such as the mean and standard deviation
- Common parametrizations: normal, lognormal, exponential, uniform, Poisson, etc.
- The normal distribution is the easiest to deal with and luckily it also follows from the Central Limit Theorem

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\tfrac{1}{2}\left(\tfrac{z-\mu}{\sigma}\right)^2]$$



Accuracy and precision

- Accuracy = degree to which a representation of reality is close to reality (≅ inverse of error)
- Precision = degree to which repeated measurements under unchanged conditions show the same results



good accuracy
poor precision

poor accuracy
good precision

good accuracy
good precision

EXERCISE 1

How many US$ in cash are in this room?

1. Make a guess
2. How good is your guess? Specify lower and upper limit of 90%-confidence interval
3. Even more informative: draw a pdf
4. We'll check at the end how often the true value lies within your confidence intervals (should be about 9 out of 10 times)

PART II: MULTIVARIATE UNCERTAINTY

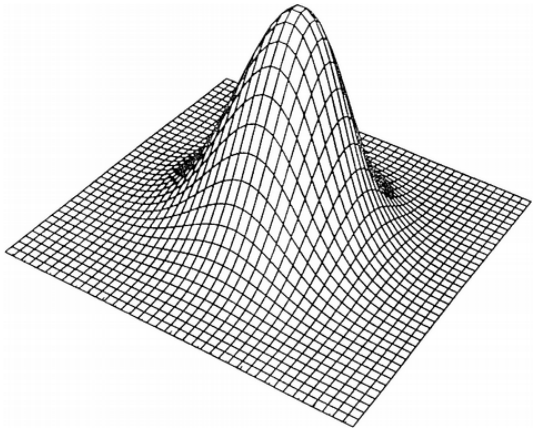
We often need the joint uncertainty of multiple variables:

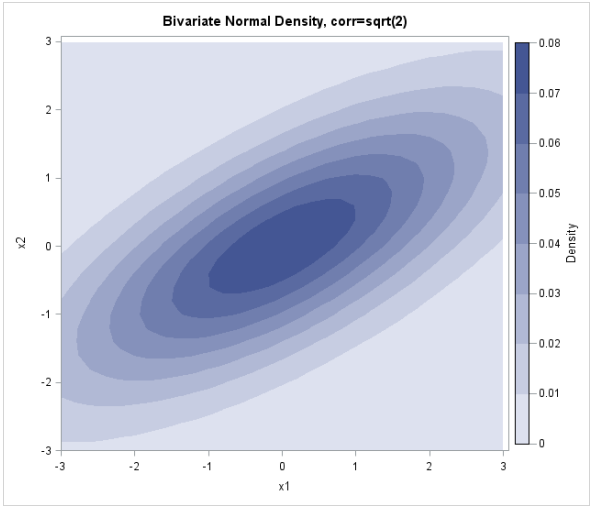
- Temperature and humidity in Lincoln, Monday next week
- Temperature Lincoln on Monday and Tuesday next week
- Soil pH and Organic Carbon at some location
- Soil pH at 0-5 cm and 5-15 cm depths
- Soil pH at location x_1 and location x_2
- Soil pH at x_1 and Organic Carbon at x_2
- Extension to >2 variables also needed

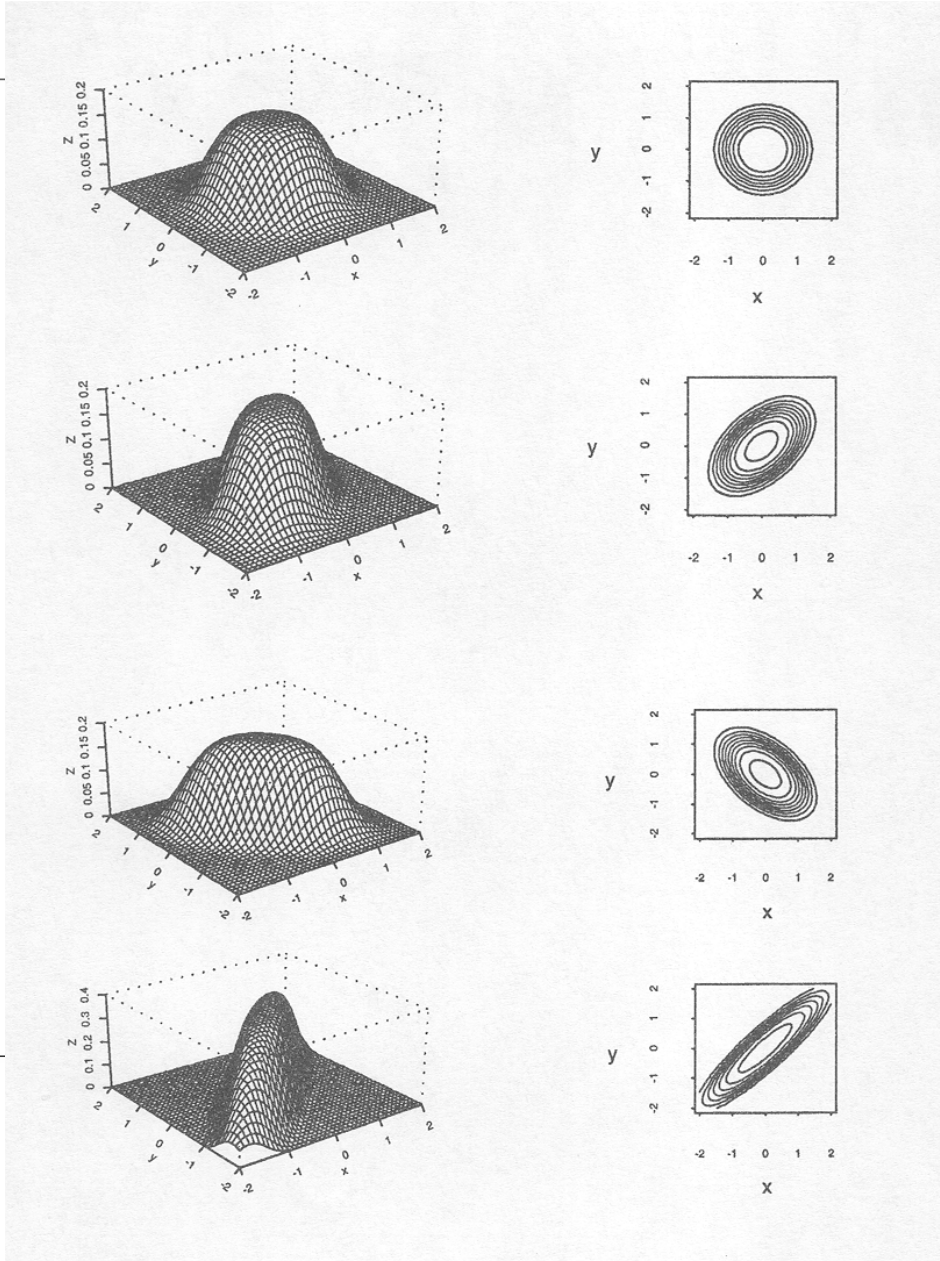
Uncertainty quantification temperature Lincoln, Monday and Tuesday next week

- We need a (marginal) pdf for both variables, but that is not enough!
- The uncertainties are correlated: if temperature on Monday is greater than expected, then on Tuesday it will likely be the same
- We need a bivariate pdf that also specifies the correlation between the uncertain variables
- Let us take a closer look for the bivariate normal distribution

Graphical illustrations







ISRIC World Soil Information

Mathematical formulation

$$f(z) = \frac{1}{(2\pi)^{\frac{m}{2}}|R|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}(z-\mu)^T R^{-1}(z-\mu)]$$

- So, under normality, we need a vector of means μ and a variance-covariance matrix R
- R characterized by the variances of the uncertain variables and their correlation coefficients
- In principle easy to extend from bivariate case to multivariate case (with m uncertain variables)
- However, note that this requires $\frac{1}{2} \cdot m \cdot (m-1)$ correlation coefficients

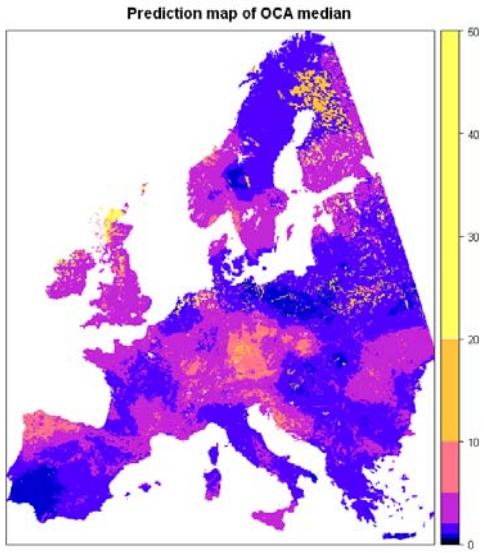
Risky to ignore the correlations

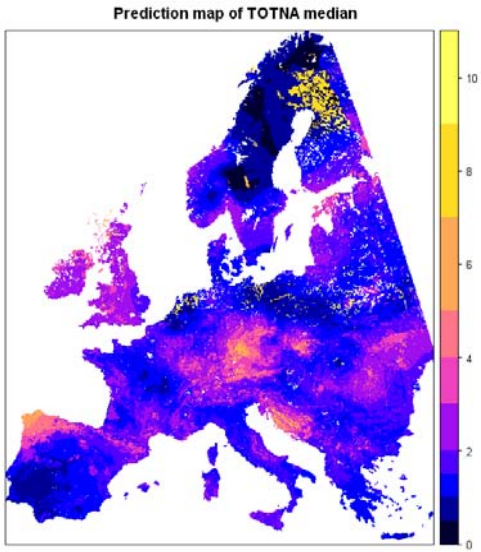
- Uncertainty propagation analysis that uses multiple uncertain inputs may yield incorrect results:
  - Example 1: Degree of pollution = 13 · Cd + Pb. How does the correlation between uncertainty about Cd and Pb influence uncertainty in degree of pollution?
  - Example 2: Plant water availability = water content at field capacity – water content at wilting point. How does the correlation influence uncertainty?

EXERCISE 2

How does the correlation in uncertain Organic Carbon and Nitrogen influence the C/N ratio?

- Assume $C \sim N(\mu = 80, \sigma^2 = 15^2)$, $N \sim N(\mu = 3, \sigma^2 = 0.6^2)$
- Calculate pdf C/N ratio for ρ=0, 0.5 and 0.8 and compare results
- Use Monte Carlo simulation because analytical solution difficult





ISRIC World Soil Information

Presence of correlation can also be turned to our advantage

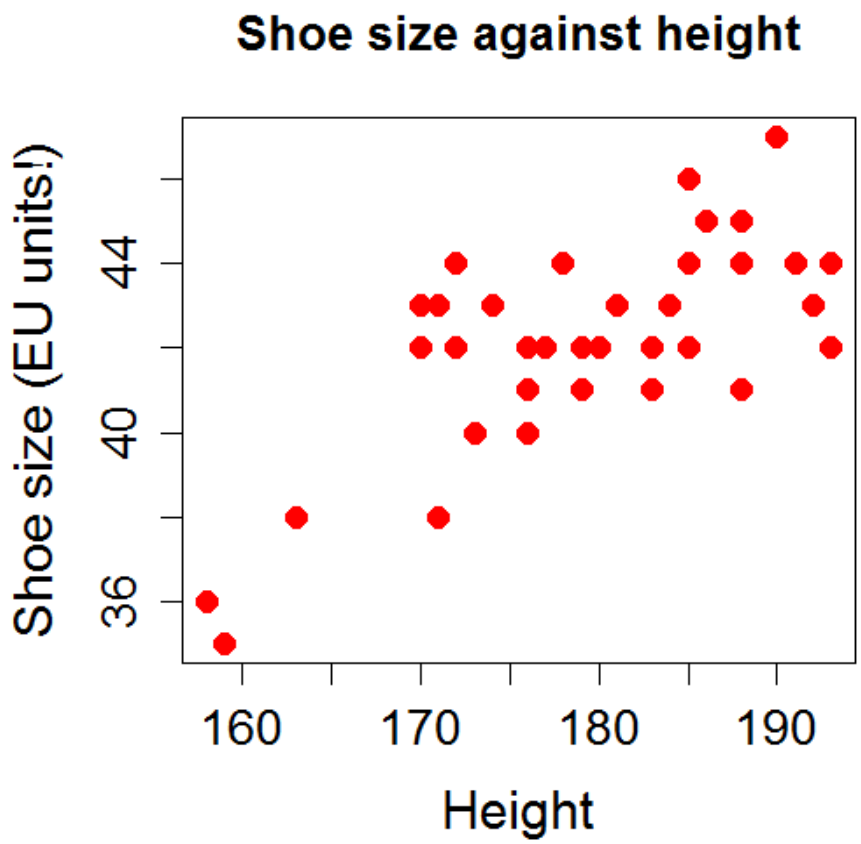
- When variables X and Y are correlated, then knowledge about X helps to know Y
- Conditional (normal) distribution:

$$\{Y|X = x\} \sim N(\mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x - \mu_X), \underbrace{(1 - \rho^2)\sigma_Y{}^2}_{\text{variance reduction!}})$$

- Also exploited in regression:
  - Pedo-transfer functions that predict an expensive variable from cheap ones
  - Digital soil mapping relates soil properties to geomorphometric parameters and remote sensing data
  - Predicting shoe size from height for Wageningen University students

Wageningen students willing to provide their personal data:

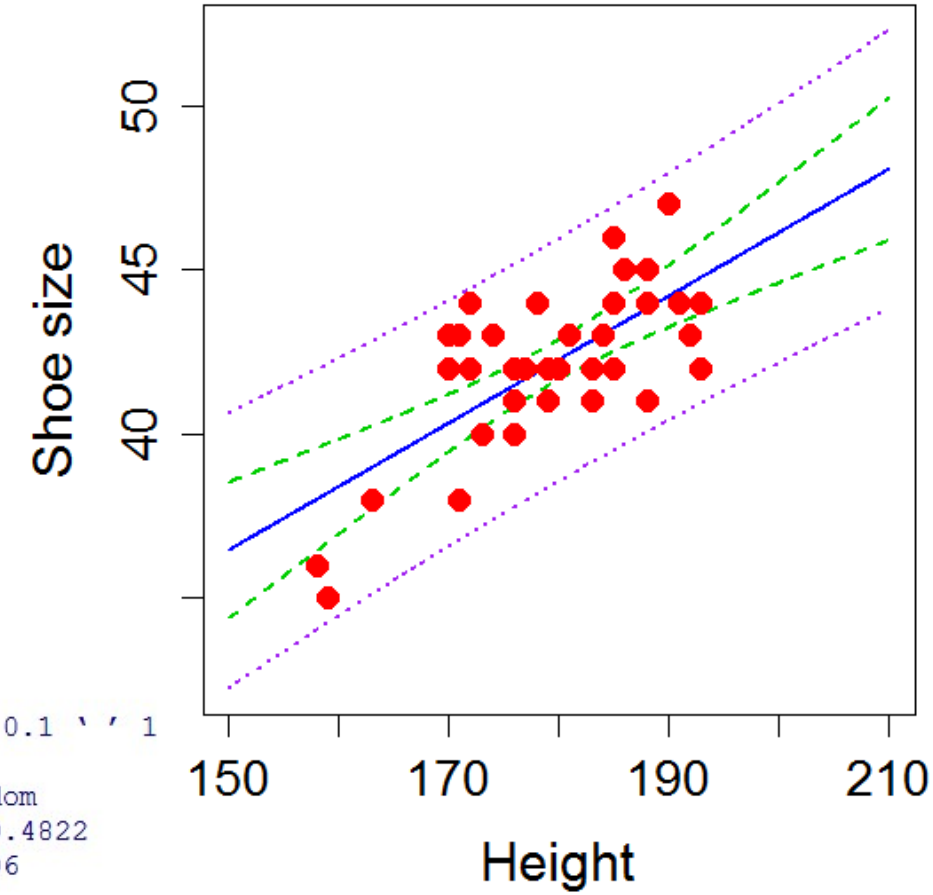
nr	height	shoe size
1	188	44
2	193	42
3	179	42
4	190	47
5	193	44
6	183	42
7	178	44
8	172	44
9	185	46
10	176	42
11	174	43
12	185	44
13	179	41
14	188	41
15	176	40



Knowing height reduces uncertainty (variance) about shoe size with 50 per cent

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.48954    5.98082   1.252    0.219
height       0.19312    0.03332   5.796 1.59e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.704 on 34 degrees of freedom
Multiple R-squared: 0.497,      Adjusted R-squared: 0.4822
F-statistic: 33.59 on 1 and 34 DF,  p-value: 1.585e-06
```



PART III: SPATIAL UNCERTAINTY

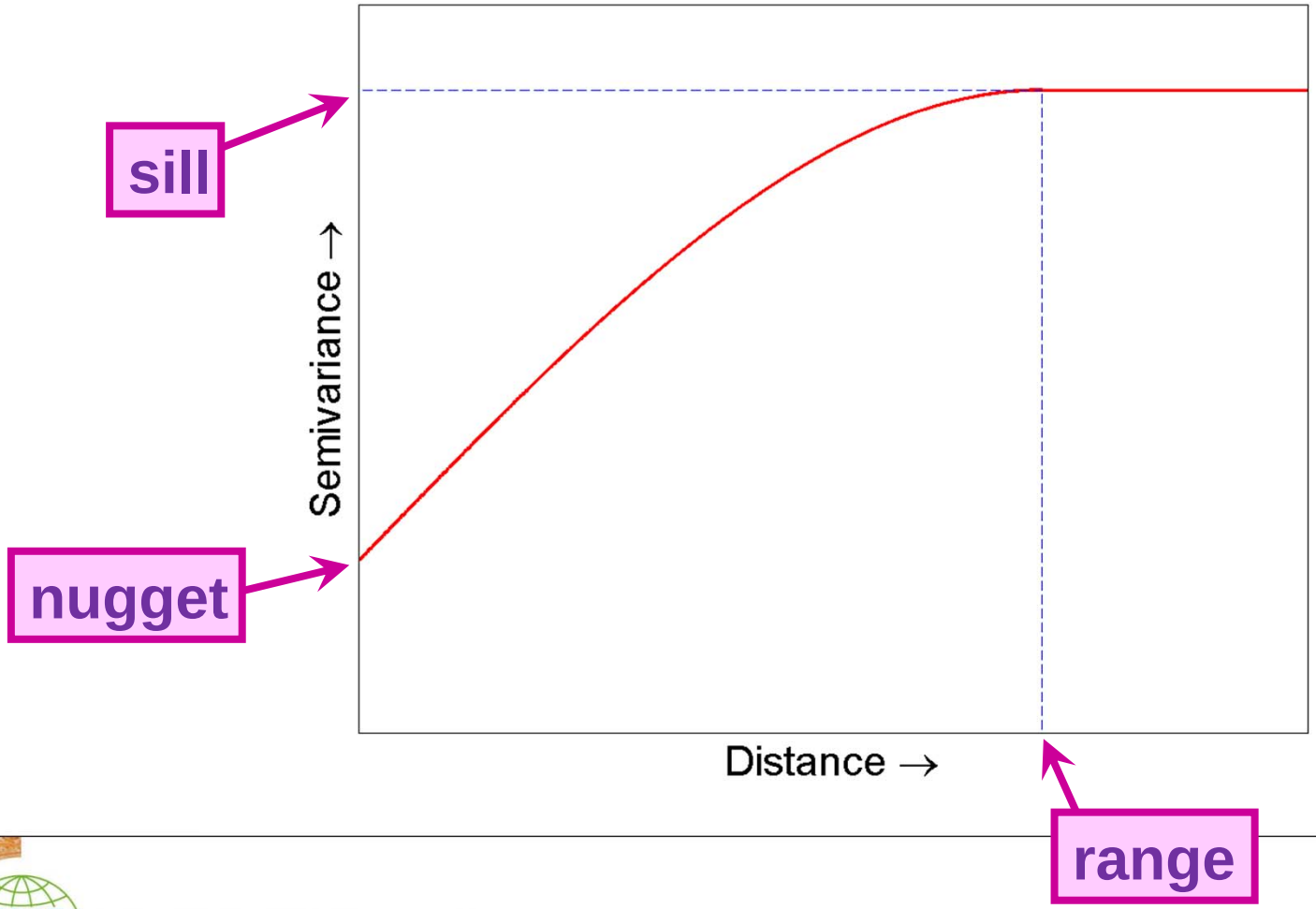
The spatial case may be seen as a multivariate case

- Each location has a variable, the collection of all locations (e.g., the nodes of a grid) yields a (very large) vector of variables
- All we need is a vector of means and a variance-covariance matrix (again assuming normality)
- However, these are too many parameters to be estimated from the limited information that is typically available
- We need an assumption: stationarity

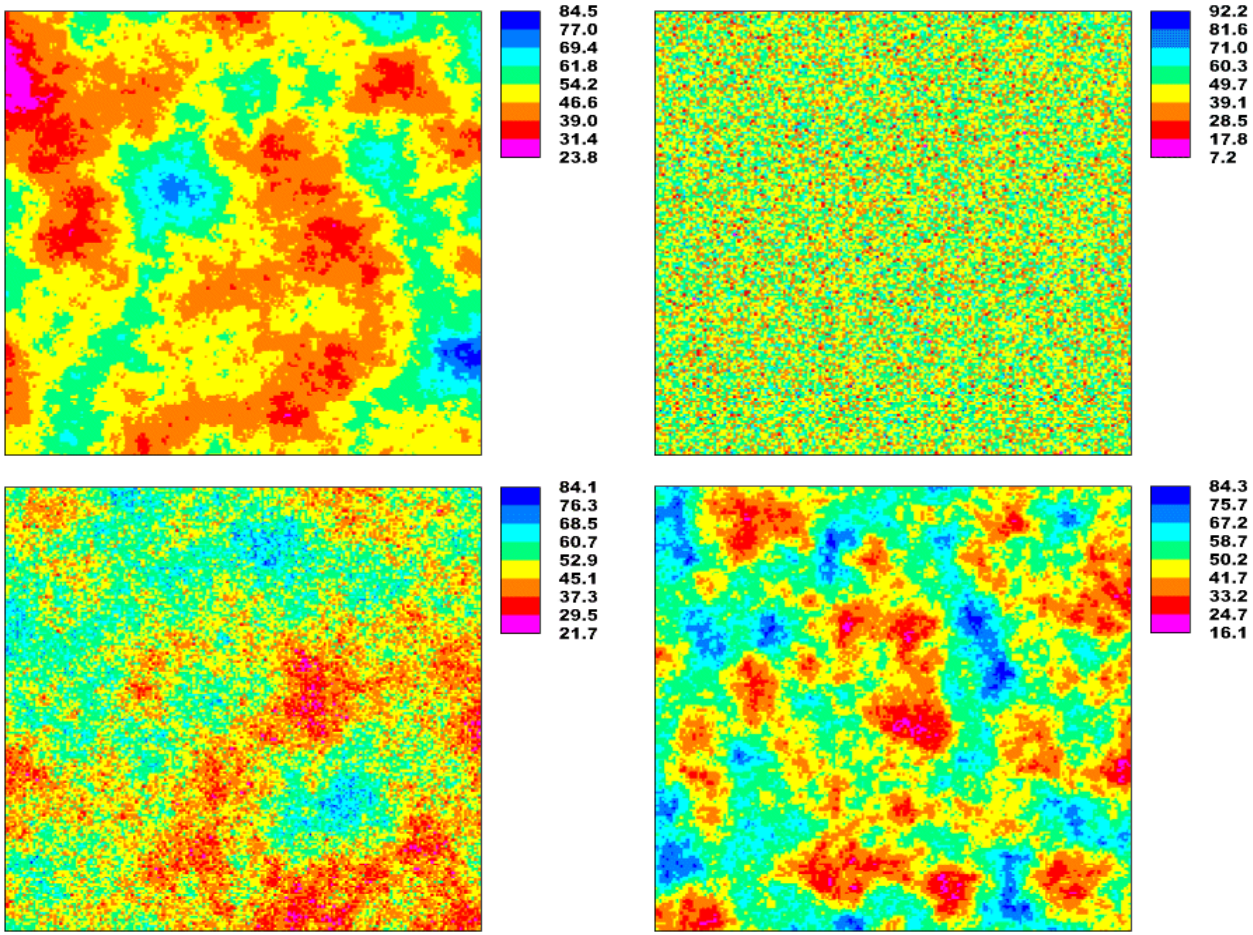
Stationarity assumption

- Assume that the mean and variance are constant, and that spatial correlation only depends on separation distance between locations
- Gross simplification but can be relaxed (later)
- Variance and correlation function are captured by the semivariogram, the key tool of geostatistics
- Mean and semivariogram can reliably be estimated from ≥60 observations

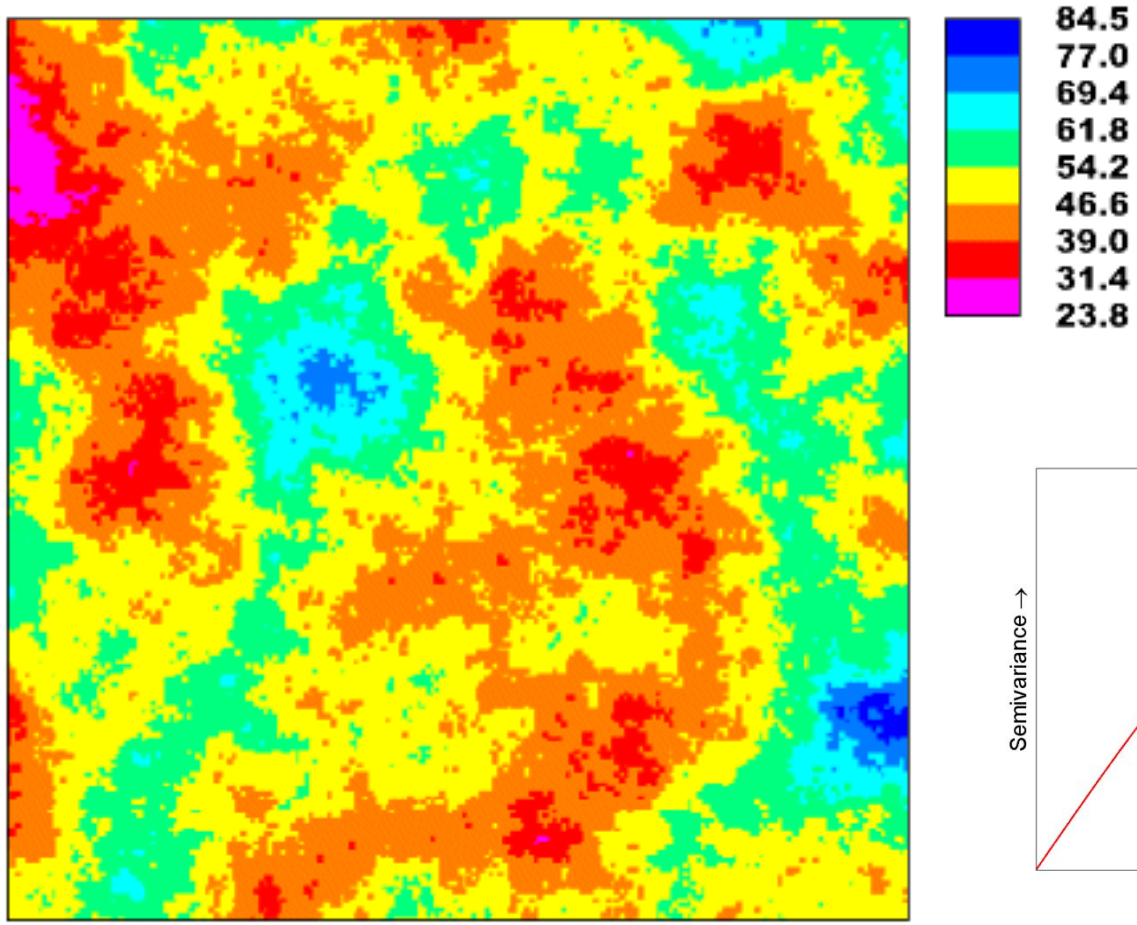
Typical shape of the semivariogram, with parameters



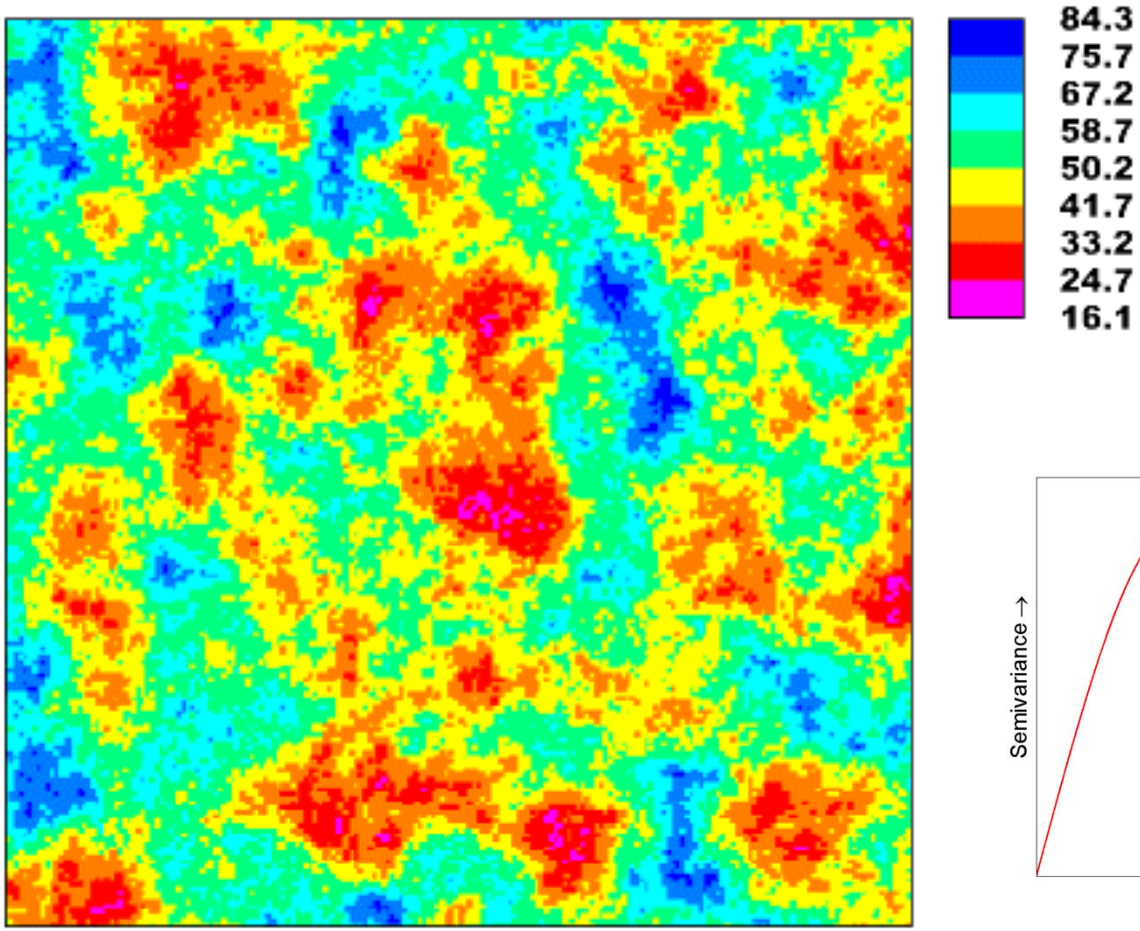
It is good to know which spatial pattern corresponds with which semivariogram

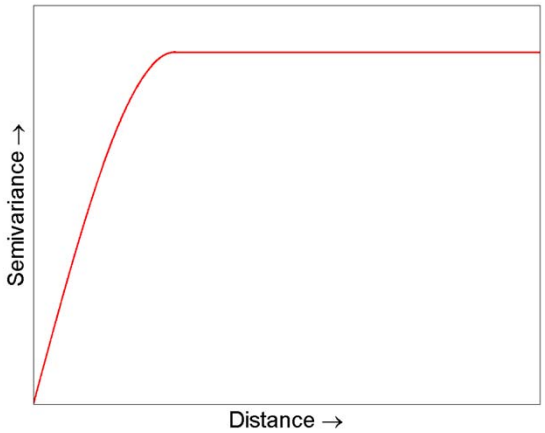


Example 1: a possible ‘reality’

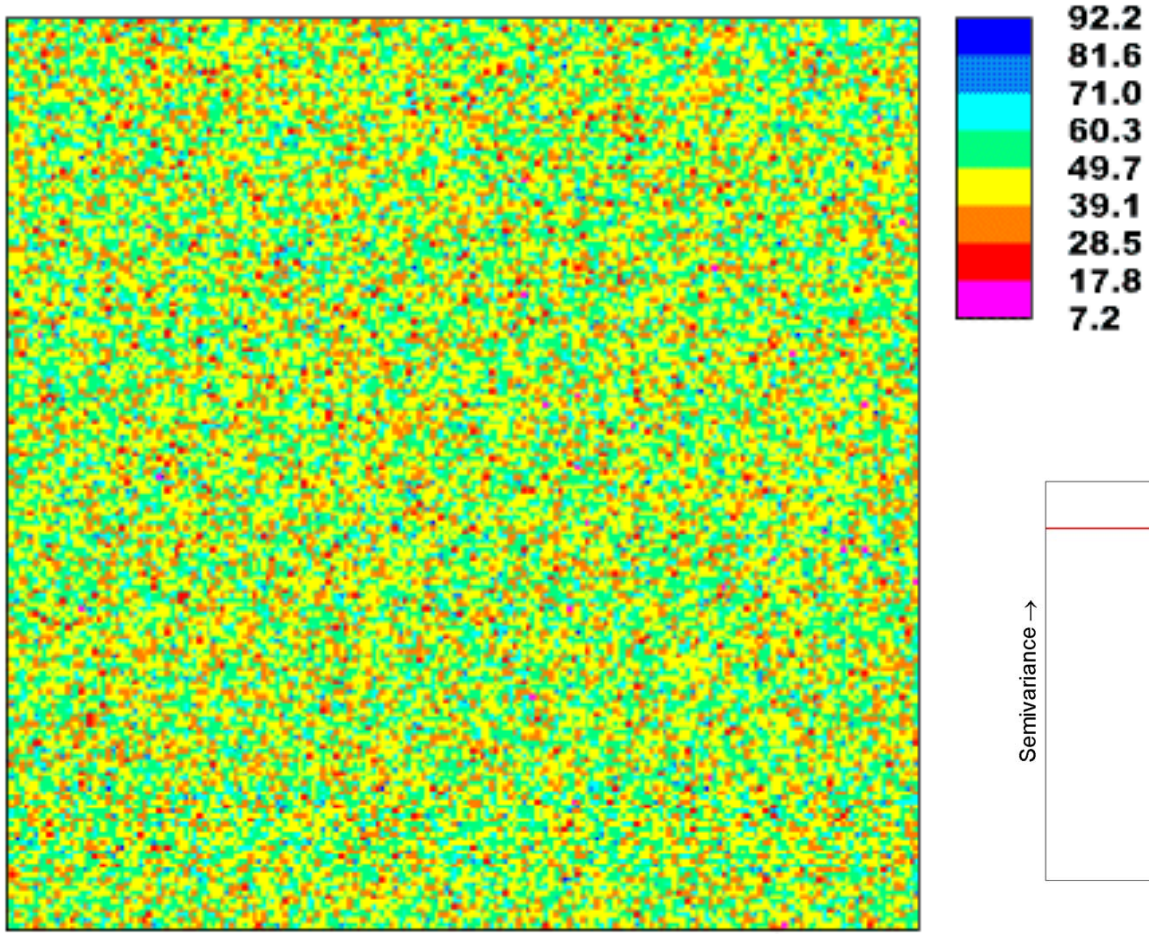


Example 2: second reality

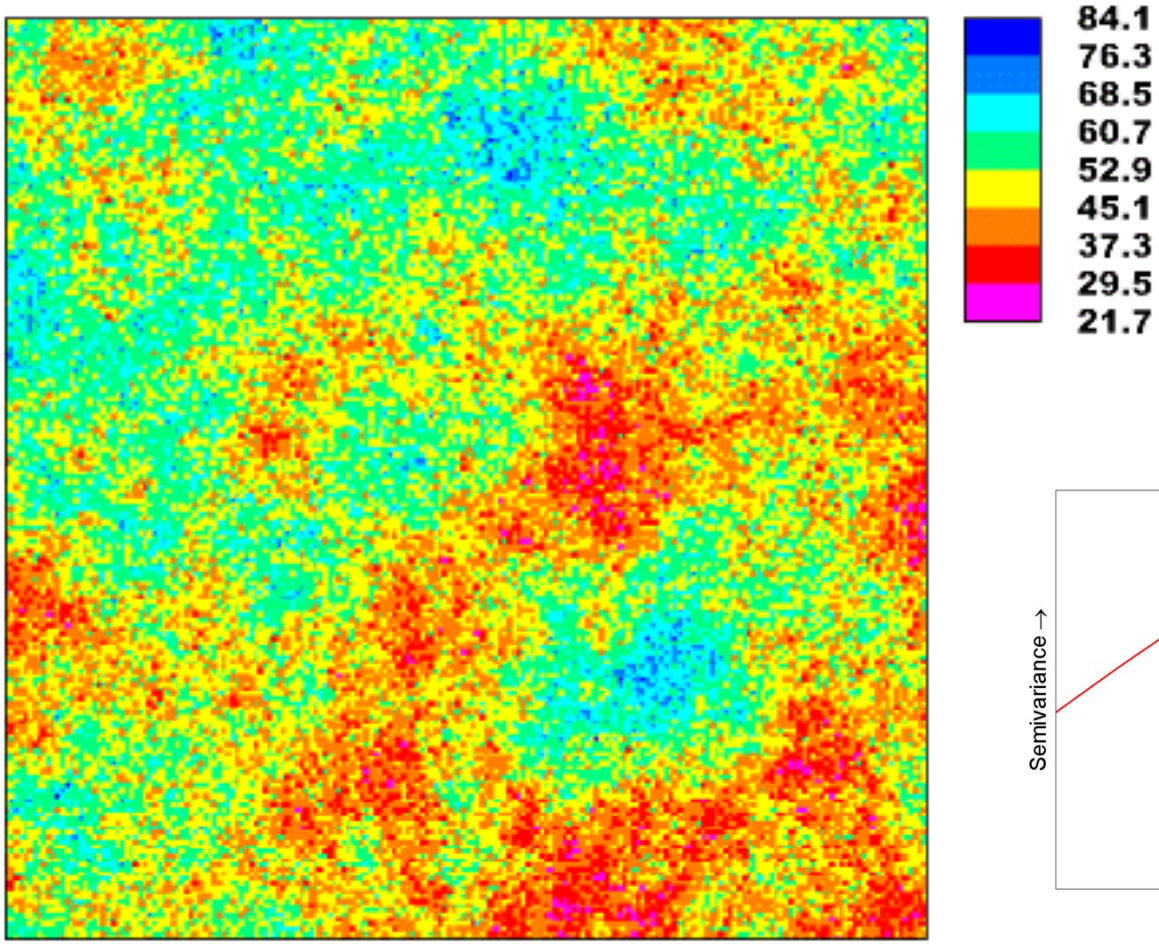


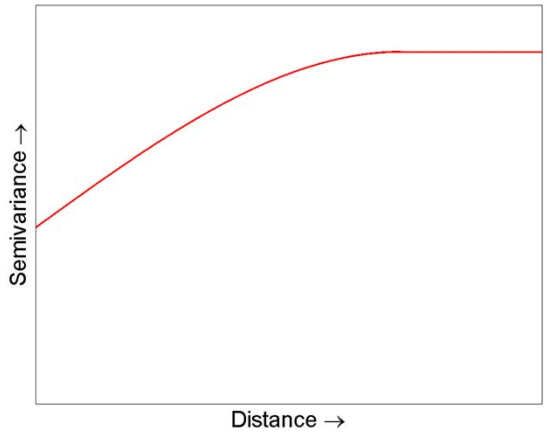


Example 3: third reality

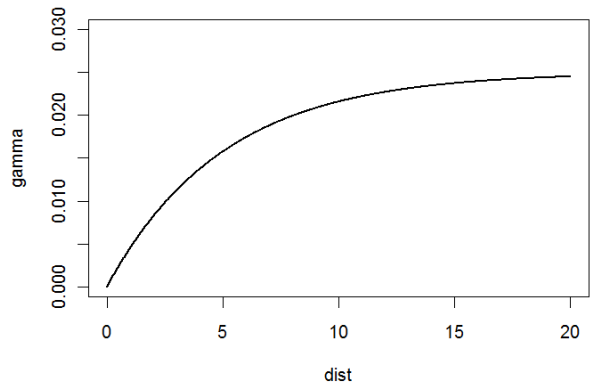


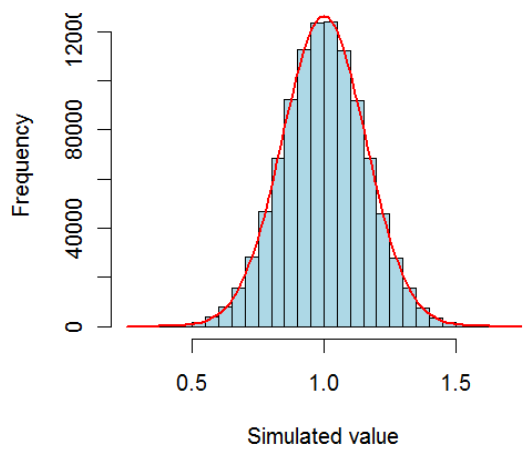
Example 4: fourth reality

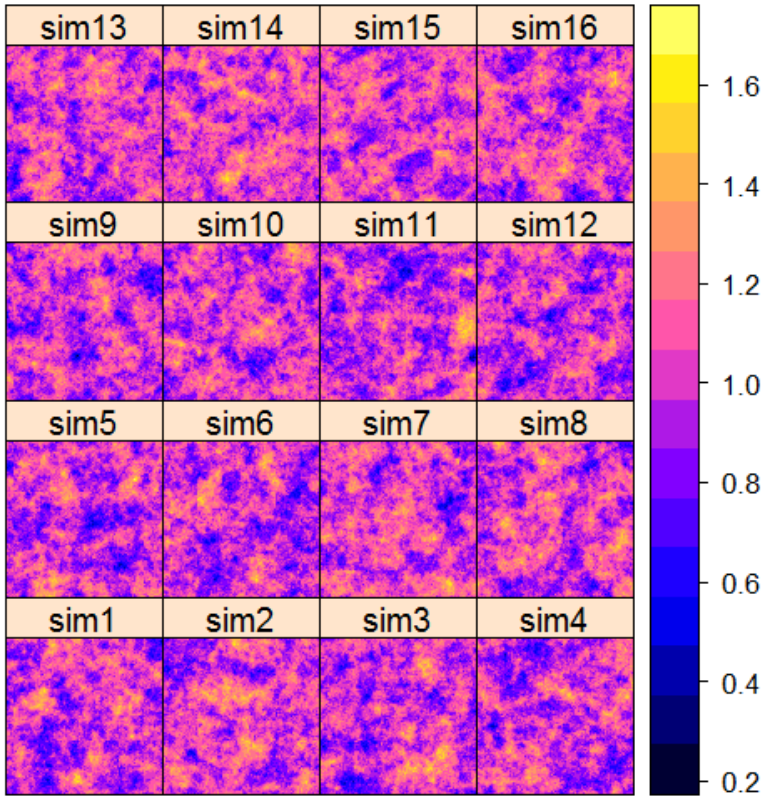


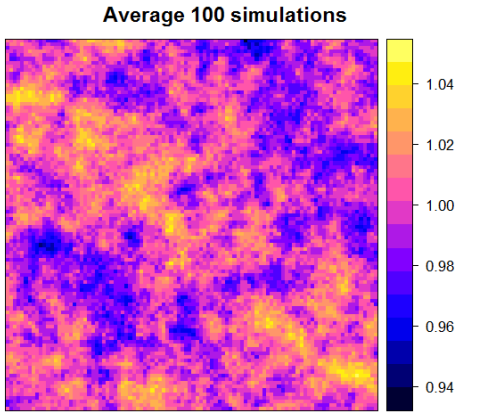


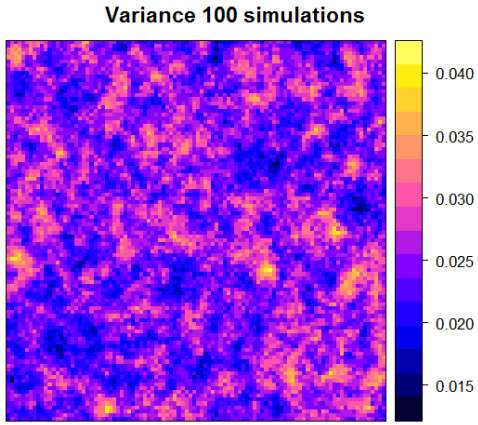
Realisations are a random sample from the imposed pdf: parameters reproduced







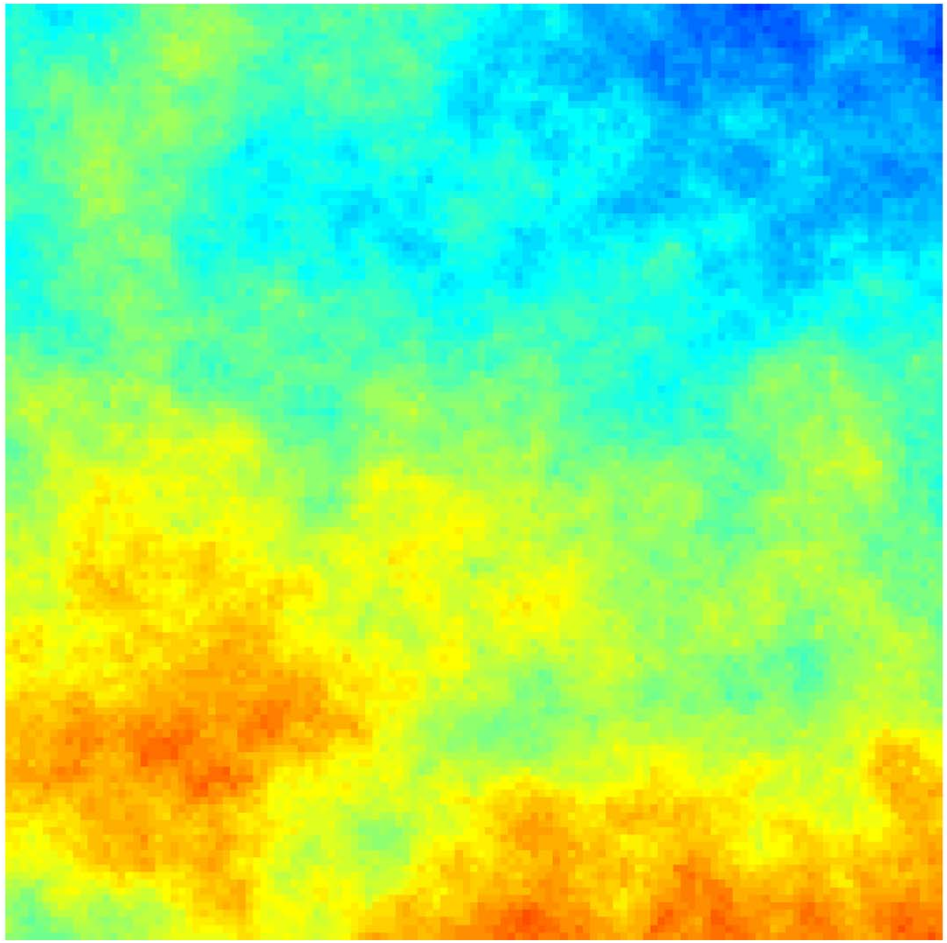




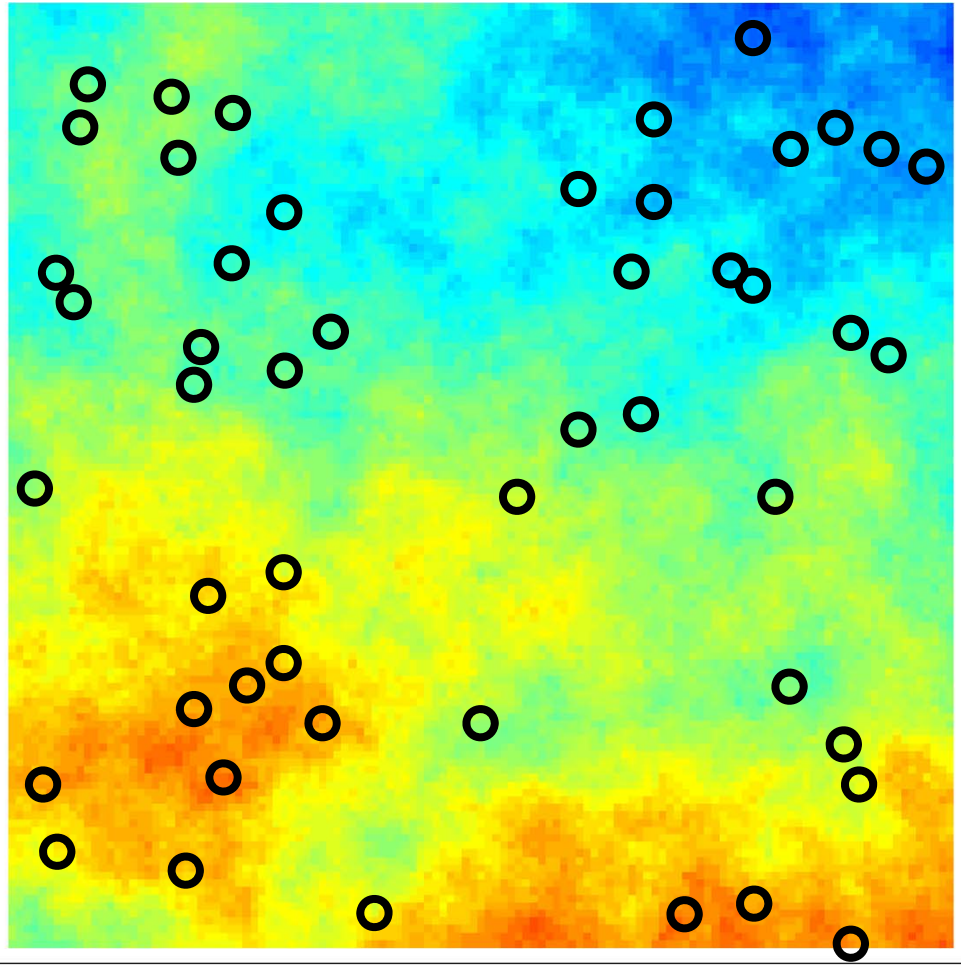
Conditioning to point observations

- Just like we can condition shoe size to height we can also condition the value of a spatial variable at any location to observations at measurement locations
- This is done with kriging: no need to explain
- It will reduce uncertainty in the neighbourhood of the measurement locations, provided there is spatial correlation
- The conditional pdf is complicated (no longer stationary), but we don't need it in explicit form

Conditioning reduces uncertainty



Conditioning reduces uncertainty



Uncertainty can also be reduced by including covariate information

- Many variables, including soil properties, are correlated with exhaustively available covariates (DEM and derivatives, remote sensing imagery)
- We can use these to drop the constant-mean assumption and include a trend
- Simplest with multiple linear regression, but alternatives are CART, ANN, RF or even mechanistic models
- The trend will explain part of the spatial variability and hence reduce the interpolation error

Regression kriging

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

dependent, target variable

trend, explanatory part

stochastic residual, unexplanatory part, can be spatially correlated!

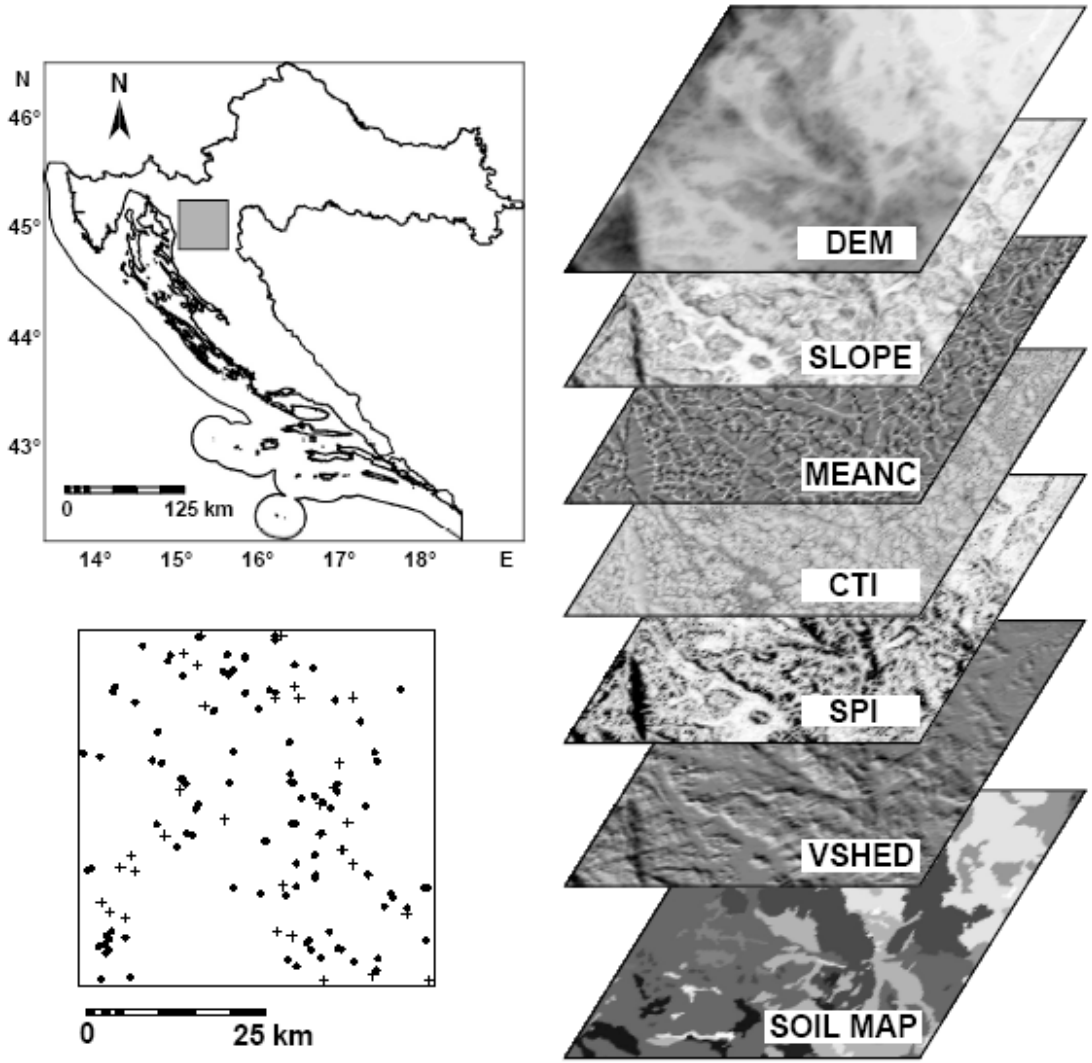
Unlike ordinary kriging, in regression kriging the trend is no longer constant but a function of 'explanatory' variables, for example:

$$\text{Soil depth}(s) = \beta_0 + \beta_1 \cdot \text{elevation}(s) + \beta_2 \cdot \text{slope angle}(s) + \beta_3 \cdot \text{vegetation density}(s) + \beta_4 \cdot \text{upstream area}(s) + \text{residual}(s)$$

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

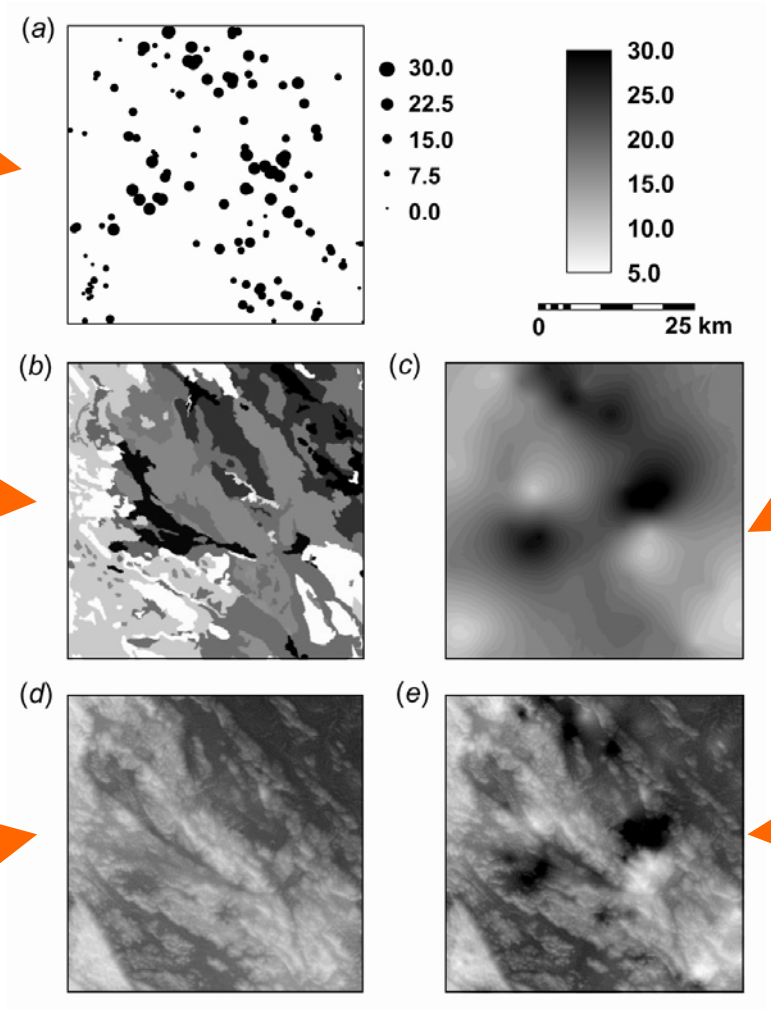


Results using various interpolation methods

observations

soil map only predictor

regression only



ordinary kriging

regression kriging

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

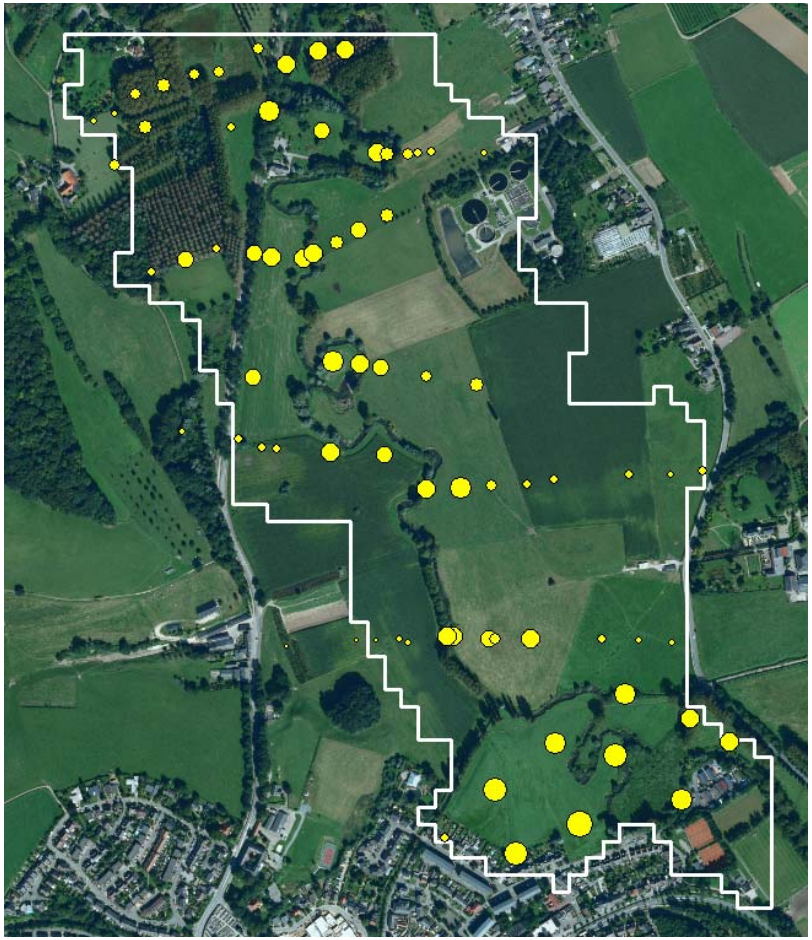
Summary spatial uncertainty quantification

- As before, all we need is a probability distribution function (pdf), nothing new there
- The pdf is complex, even if we assume multivariate normality we still need a large vector of means and a huge variance-covariance matrix
- Solution is to assume stationarity, this brings parametrization down to a scalar mean (or known trend in covariates) and a semivariogram
- Information in observations will narrow down the pdf, this is achieved by kriging
- Resulting pdf will again be complex but we do not need it in explicit form

EXERCISE 3 (can skip if running out of time)

Topsoil lead concentration of river Geul valley, Netherlands

- Build geostatistical model
- Use kriging to condition to observations
- Show uncertainty with kriging standard deviation and conditional simulations
- Calculate area for which P(Pb>250 mg/kg)<0.05



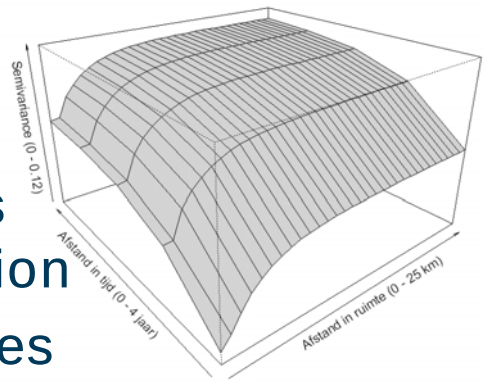
PART IV: UNCERTAINTY QUANTIFICATION FOR GLOBALSOILMAP

Most of it covered in Part III (spatial uncertainty quantification)

- However, we need a 3D extension
- We must also take cross-correlations between multiple soil properties into account
- And then there are three more issues that deserve attention too: unambiguous definition of soil property, support and expert elicitation
- Conclude with three approaches to uncertainty quantification for the GlobalSoilMap project

3D modelling, two basic approaches:

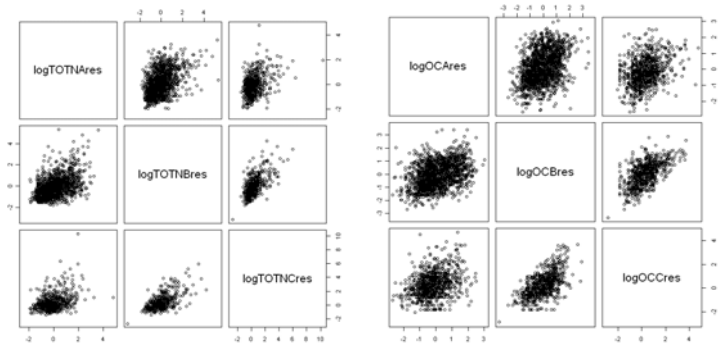
- True-3D: include depth as a third dimension
  - Requires a semivariogram that characterizes spatial variation in lateral and vertical direction
  - There will be zonal and geometric anisotropies



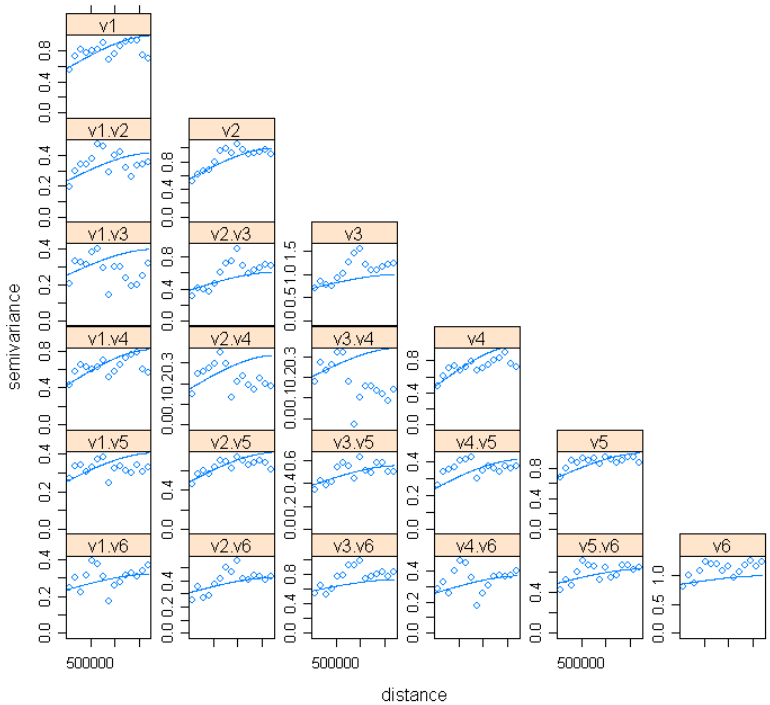
- Pseudo-3D: treat 3rd dimension as a stack of layers (in GlobalSoilMap we have six)
  - Model the layers one by one or better jointly (see cross-correlation approach hereafter)
  - Requires observations for each layer: in deeper layers often fewer observations

Cross-correlation between multiple soil properties

- Standard geostatistical solution: cokriging
- Doable as long as the number of correlated variables is not too large (say up to 15?)
- Kriging prediction and standard deviation maps perhaps do not change that much, but joint simulations do: important for uncertainty propagation analysis (recall C/N ratio exercise!)



logOCTOTN residual variograms



Three more issues (1/3)

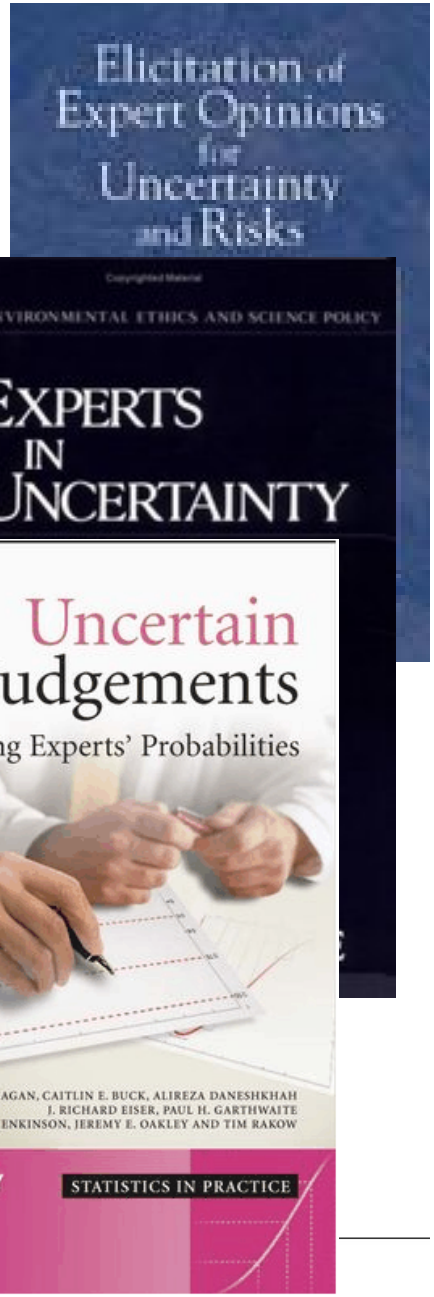
1. We need unambiguous definitions of soil properties
   - pH(H_2O)≠pH(KCl)≠pH(CaCl), and that is one of the easy ones, Organic Carbon can be measured in so many ways
   - GlobalSoilMap is very aware of the problem and has decided on a standard for all soil properties considered
   - But metadata do not always specify the measurement method and conversion from one method to another not always easy
   - This all augments to the uncertainty, perhaps it can be addressed as follows: assign larger 'measurement' errors to these type of observations?

Three more issues (2/3)

2. We need to match the 'support' of soil observations
   - Bulk sampling yields observations that have less variation than point observations: observations at different supports cannot be merged into one dataset unless the change of support is taken into account
   - Equally holds for the vertical: we cannot merge observations from 0-5cm with observations from 0-10cm
   - Spline fitting in the allows for change of support in the vertical and can hence solve the problem, but we ought not ignore the spline interpolation error (a spline interpolation cannot replace a real measurement)
   - Perhaps regularization theory provides a better solution

Three more issues (3/3)

3. We may need to resort to expert elicitation of uncertainties
   - Ideally we quantify uncertainty using real data (such as we do in kriging), but sometimes there are no data or there are not enough data
   - Expert elicitation is a mature research field which provides a sound scientific basis to reliably and consistently extract knowledge from experts
   - It is cumbersome and cannot replace inference using real data but it deserves more attention



Three approaches to uncertainty quantification of GlobalSoilMap products

1. Geostatistical modelling of soil properties
   - This is what we did in this short course: use the point data and covariates to build a geostatistical model that renders a spatial multivariate pdf that automatically captures the uncertainty ('free' by-product)
2. Geostatistical modelling of errors in existing (legacy) soil property maps
   - Similar as above with an important difference that an existing map is taken as is and uncertainty modelling focuses on the error in that map
3. Validation with independent data
   - Not yet discussed, it does not yield a full-statistical model of uncertainty but has some attractive properties

Validation with independent data

- Useful if (geo)statistical modelling is infeasible or unwanted because of the many assumptions
- Provides 'model-free' estimates of summary measures of the accuracy, such as the Mean Error and Root Mean Squared Error
- These measures can only be estimated because we have only a (small) sample from the whole area
- Estir
  a pr
  sam

European Journal of Soil Science

European Journal of Soil Science, June 2011, **62**, 394–407 doi: 10.1111/j.1365-2389.2011.01364.x

Sampling for validation of digital soil maps

D.J. Brus, B. Kempen & G.B.M. Heuvelink
Soil Science Centre, Wageningen University and Research Centre, PO Box 47, 6700 AA Wageningen, The Netherlands

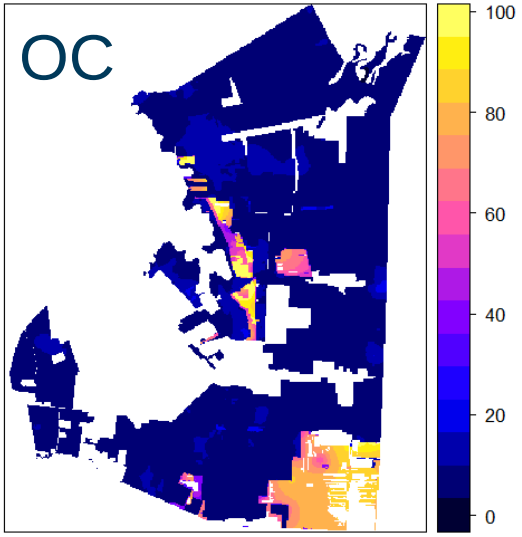
Validation with independent data

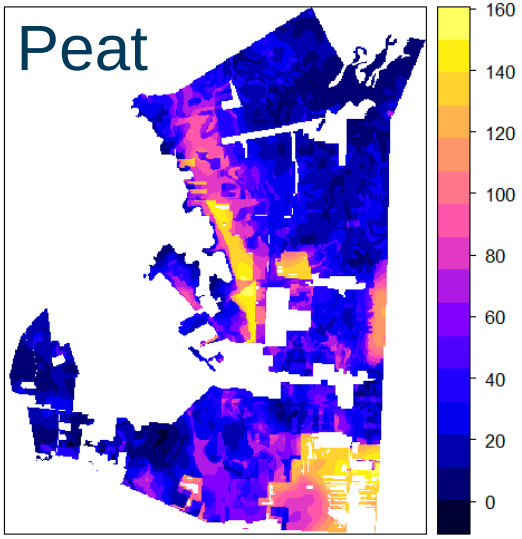
- Validition with non-probability samples is more common but does not allow proper quantification of estimation errors
- Cross-validation also falls under this category, as does validation with legacy data
- In any case, the data must be truly independent, otherwise we might get a false sense of security
- Validation data should also be sufficiently accurate: measurement errors should be small compared to the errors in the map that is validated, or else must be taken into account

EXERCISE 4

Validation of OC and peat thickness maps for part of the Drenthe province, Netherlands

- Overlay validation data with soil prediction map
- Estimate ME and RMSE
- Calculate confidence intervals





Pros and cons of the three approaches

approach	advantages	disadvantages
geostatistical modelling of soil properties	full characterization of uncertainty with spatial multivariate probability distribution; reproducible; flexible with respect to new data or different output formats	only valid under the assumptions made; cannot accomodate all soil mapping approaches; can be computationally demanding
geostatistical modelling of errors in existing soil property maps	attractive when soil property maps have already been prepared; no restrictions on the way that soil property maps are made	requires independent data that cannot be used for soil mapping; takes maps as they are and does not try to improve them
independent validation	makes no model assumptions; fairly easily applied	provides only a summary measure of uncertainty; requires that soil property maps have been prepared; requires independent validation data

That’s it

THANK YOU

ISRIC World Soil Information