ISRIC Spring School – Hand’s on Global Soil Information Facilities, 28 May – 1 June 2018

Uncertainty quantification and propagation

World Soil Information

Gerard Heuvelink

Why pay attention to uncertainty?

GlobalSoilMap.net

Log in | Register

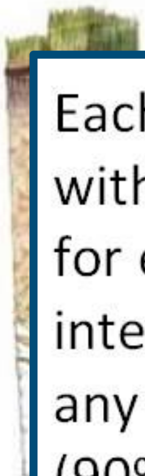
About | Science Committee | Publications | Pictures

Specifications

You are here: Home » About » About This Project

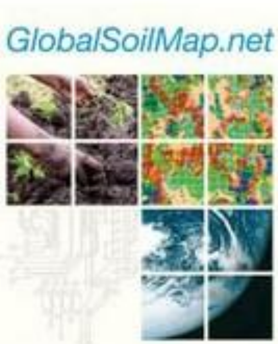
About this project

Submitted by admin on February 1, 2011 - 10:40 ::



Each soil property will have an estimate of the uncertainty associated with the prediction for each depth (for properties reported by depth) for each grid location. Uncertainty here is defined as the 90% prediction interval (PI), which is the range in values within which the true value at any point prediction location is expected to be found 9 times out of 10 (90%).

The project was officially launched on **17th February 2009, New York, USA.**



- Who is who in this project?
- Download the press release (223)
- Download the Science article (399)
- Download the brochure (309)

Donor consultation meeting at the World Bank (click logo):

Why pay attention to uncertainty?

- Any self-respecting researcher should want to check the quality of his/her results before these are made public (do not publish bad maps!)
- Quantified uncertainty allows to compare the performance of methods, evaluate which is best, help to improve methods
- Clients and end users must know the quality of maps to judge their usability for specific purposes
- Uncertainty quantification of soil maps is required to analyse how uncertainty in these maps propagate through environmental models; important because model output may be decisive for environmental policy and decision making
- Note: independent validation addresses many of the issues above, but it only provides summary measures and in many cases we may need spatially explicit uncertainties

Programme of this module

Lecture:

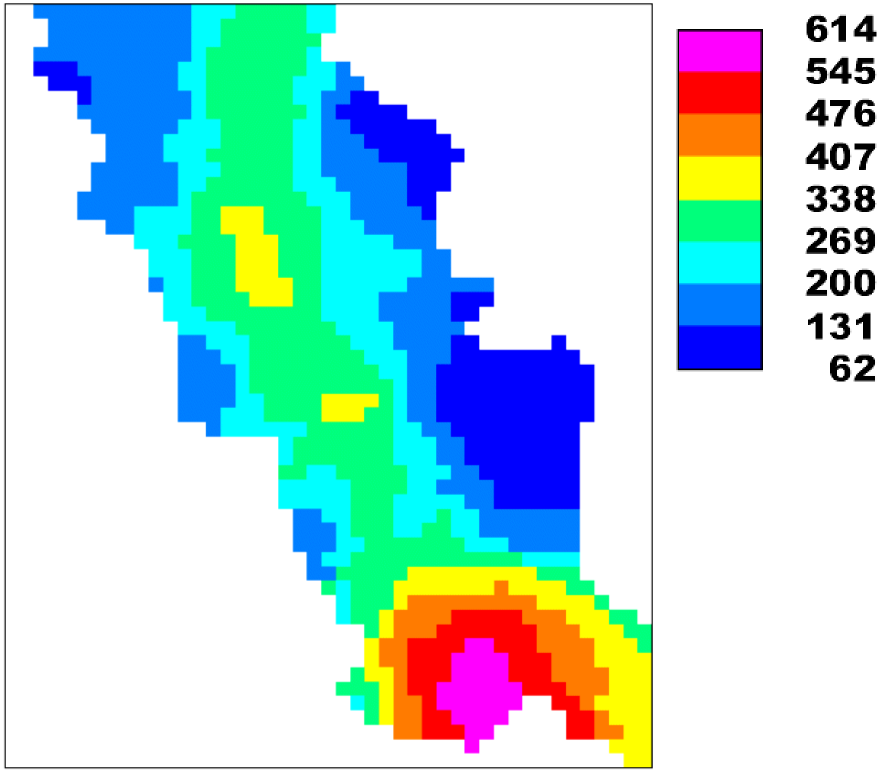
- Exploring error and uncertainty, what is it?
- Statistical modelling of uncertainty with probability distributions
- Uncertainty propagation in spatial analysis and environmental modelling
- Derivation of soil carbon stock from soil properties, with uncertainty propagation

Computer practical:

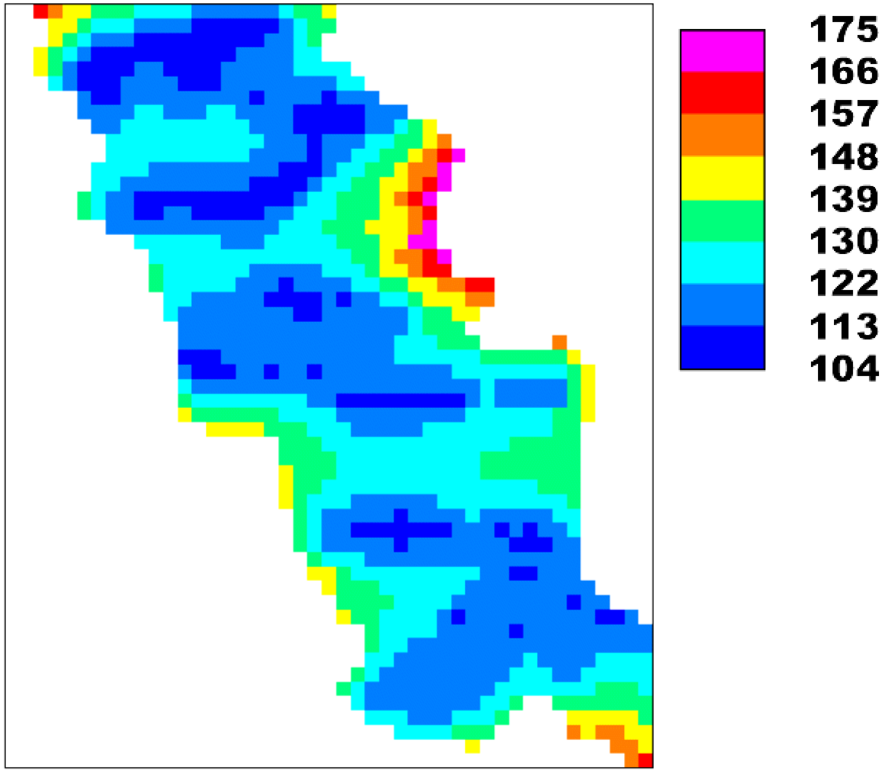
- Analyse how uncertainties in soil and other factors propagate through a very simple 'model' and may affect environmental decision making

Is this map error-free?

Pb prediction (mg/kg)



Pb st.dev. (mg/kg)

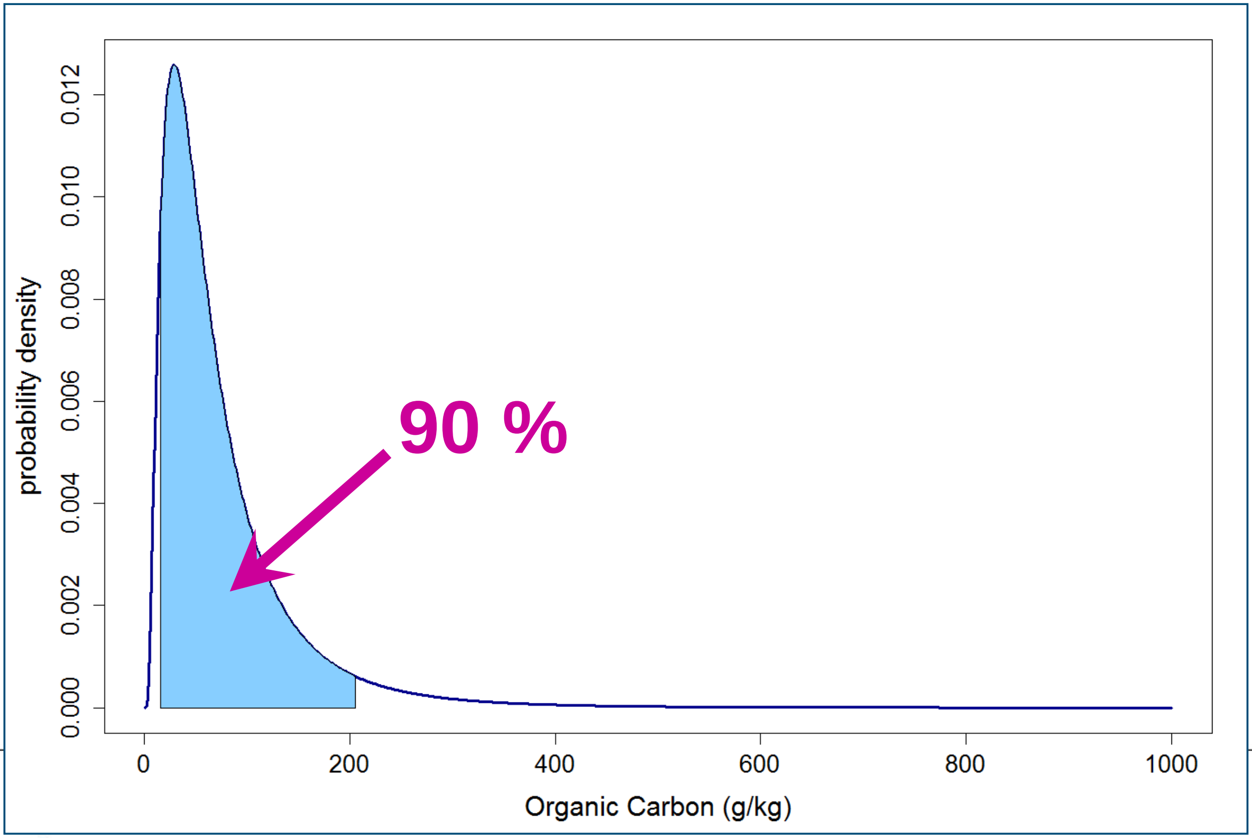


Error and uncertainty

- We are uncertain about a soil property if we do not know its true value
- We may have an estimate of it, but this estimate may well be in error (i.e. differ from the true value)
- For example, the pH of the soil at some location and depth may be 6.3, while according to a map it is 5.9. In this case the error is simply 6.3 – 5.9 = 0.4
- The problem is that usually we do not know the error, because if we knew it, we would eliminate it
- But in many cases we do know something:
  - We may know that the error has equal chances 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, often we are uncertain about the true state of the environment because we lack information, but we are not completely ignorant

How can we represent uncertainty statistically?

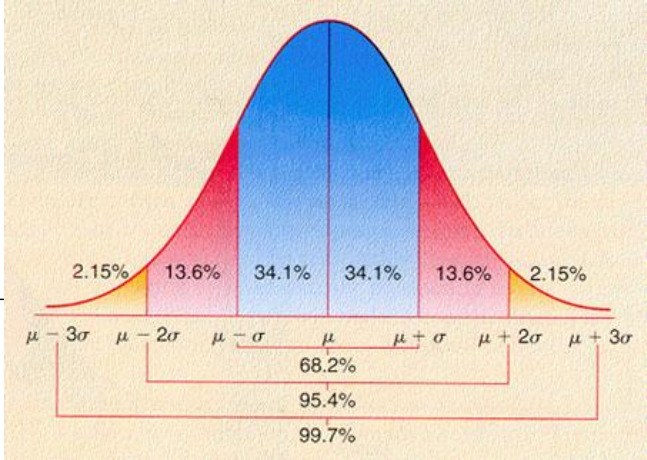
- In the presence of uncertainty, we cannot identify a single, true reality.. But perhaps we can identify all possible realities and a probability for each one
- In other words: we may be able to characterise the uncertain variable with a probability distribution



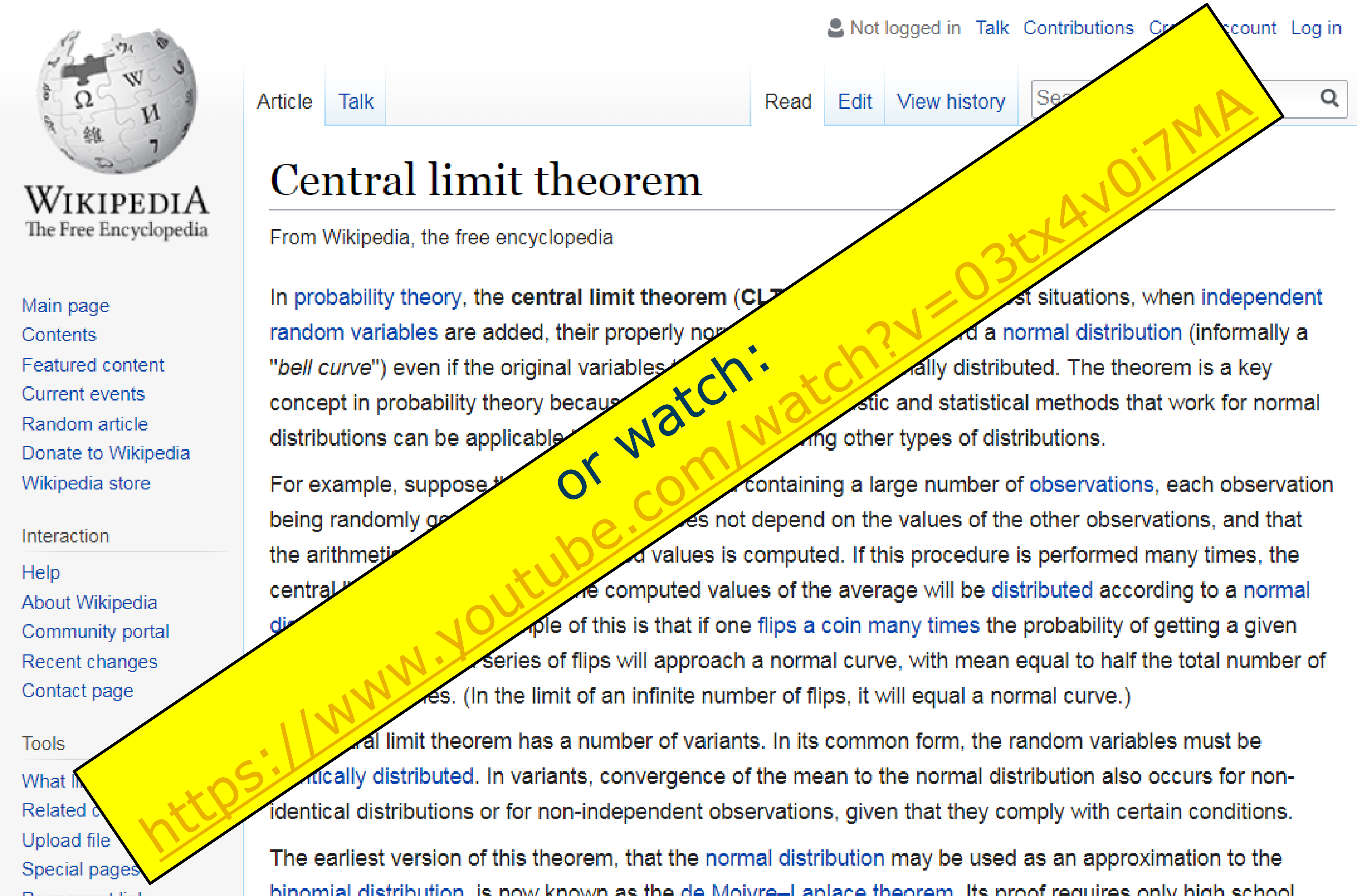
The probability distribution characterises uncertainty completely, it is all we need

- It is usually parametrised and thus reduced to a few parameters, such as the mean and standard deviation
- Common parametrisations: 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

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



Why (so often) the normal distribution?



Not logged in Talk Contributions Cr... ccount Log in

Article Talk Read Edit View history Sea...

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What li...
Related c...
Upload file
Special pages

Central limit theorem

From Wikipedia, the free encyclopedia

In probability theory, the **central limit theorem** (**CL**... ...st situations, when independent random variables are added, their properly no... ...d a normal distribution (informally a "*bell curve*") even if the original variables ... ...ally distributed. The theorem is a key concept in probability theory becau... ...stic and statistical methods that work for normal distributions can be applicable ... ...ng other types of distributions.

For example, suppose t... ...containing a large number of observations, each observation being randomly ge... ...es not depend on the values of the other observations, and that the arithmeti... ...d values is computed. If this procedure is performed many times, the central ... ...e computed values of the average will be distributed according to a normal di... ...ple of this is that if one flips a coin many times the probability of getting a given ... series of flips will approach a normal curve, with mean equal to half the total number of ...es. (In the limit of an infinite number of flips, it will equal a normal curve.)

...al limit theorem has a number of variants. In its common form, the random variables must be ...tically distributed. In variants, convergence of the mean to the normal distribution also occurs for non-identical distributions or for non-independent observations, given that they comply with certain conditions.

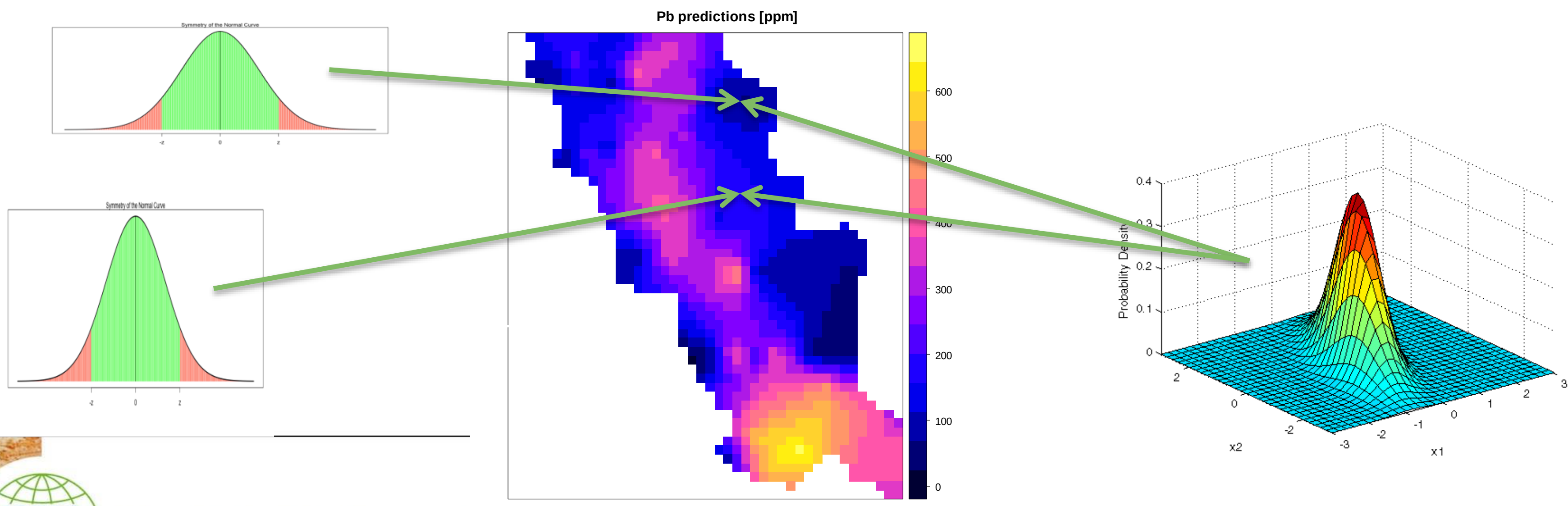
The earliest version of this theorem, that the normal distribution may be used as an approximation to the binomial distribution, is now known as the de Moivre–Laplace theorem. Its proof requires only high school

or watch:

https://www.youtube.com/watch?v=03tx4v0i7MA

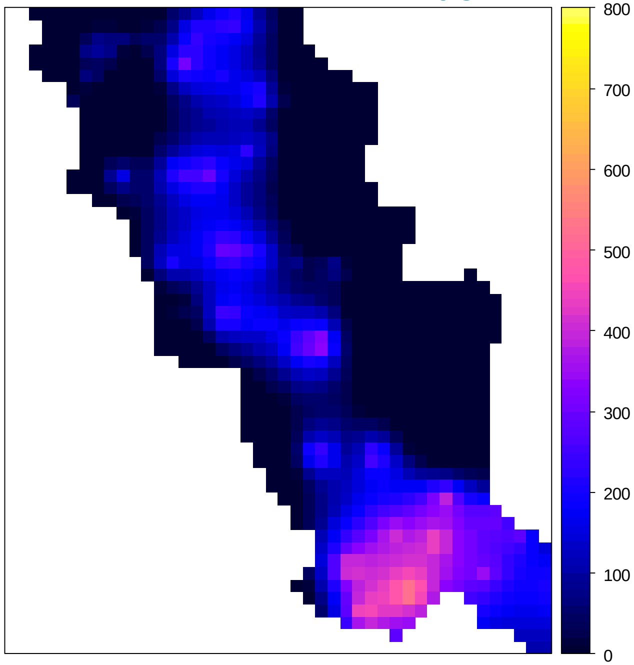
Extending the univariate probability distribution function (pdf) to a spatial pdf

- We need a univariate pdf at each and every location in the study area and we need to know how the errors are correlated spatially
- We can derive all this with kriging: the true value has a pdf whose mean equals the kriging prediction and whose standard deviation equals the kriging standard deviation; the spatial correlation can be derived from the semivariogram

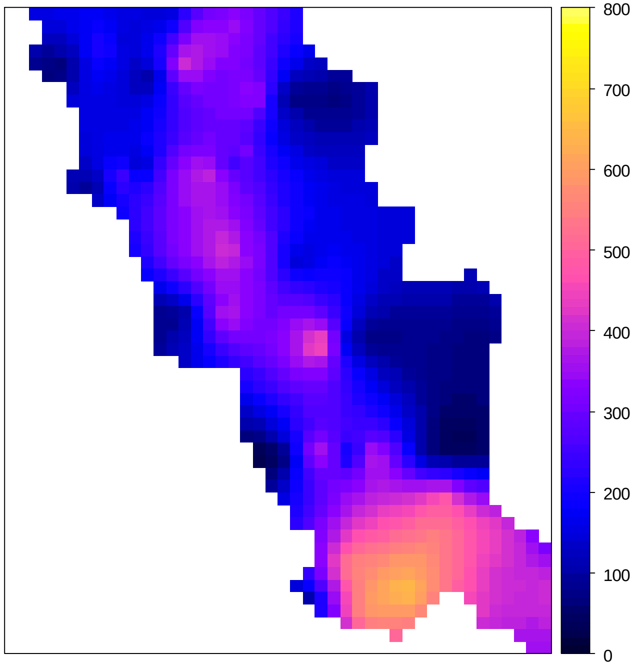


We can derive lower and upper limits of the 90% prediction interval from kriging prediction and standard deviation maps

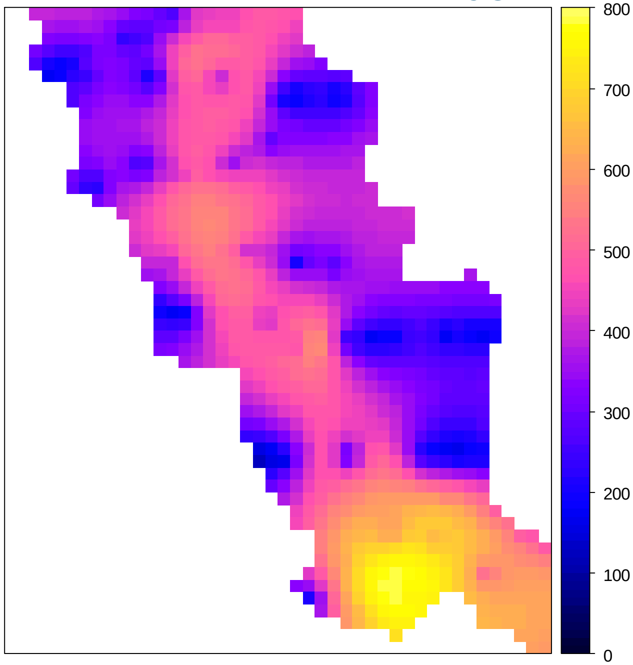
lower limit p_{05}

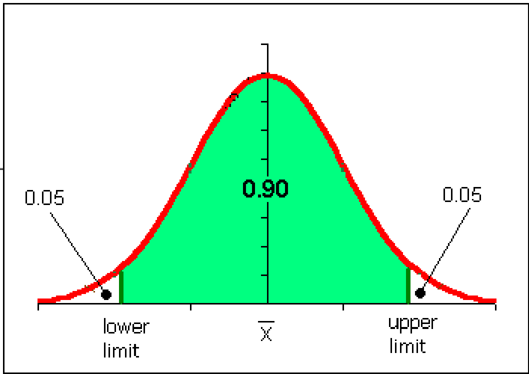


prediction

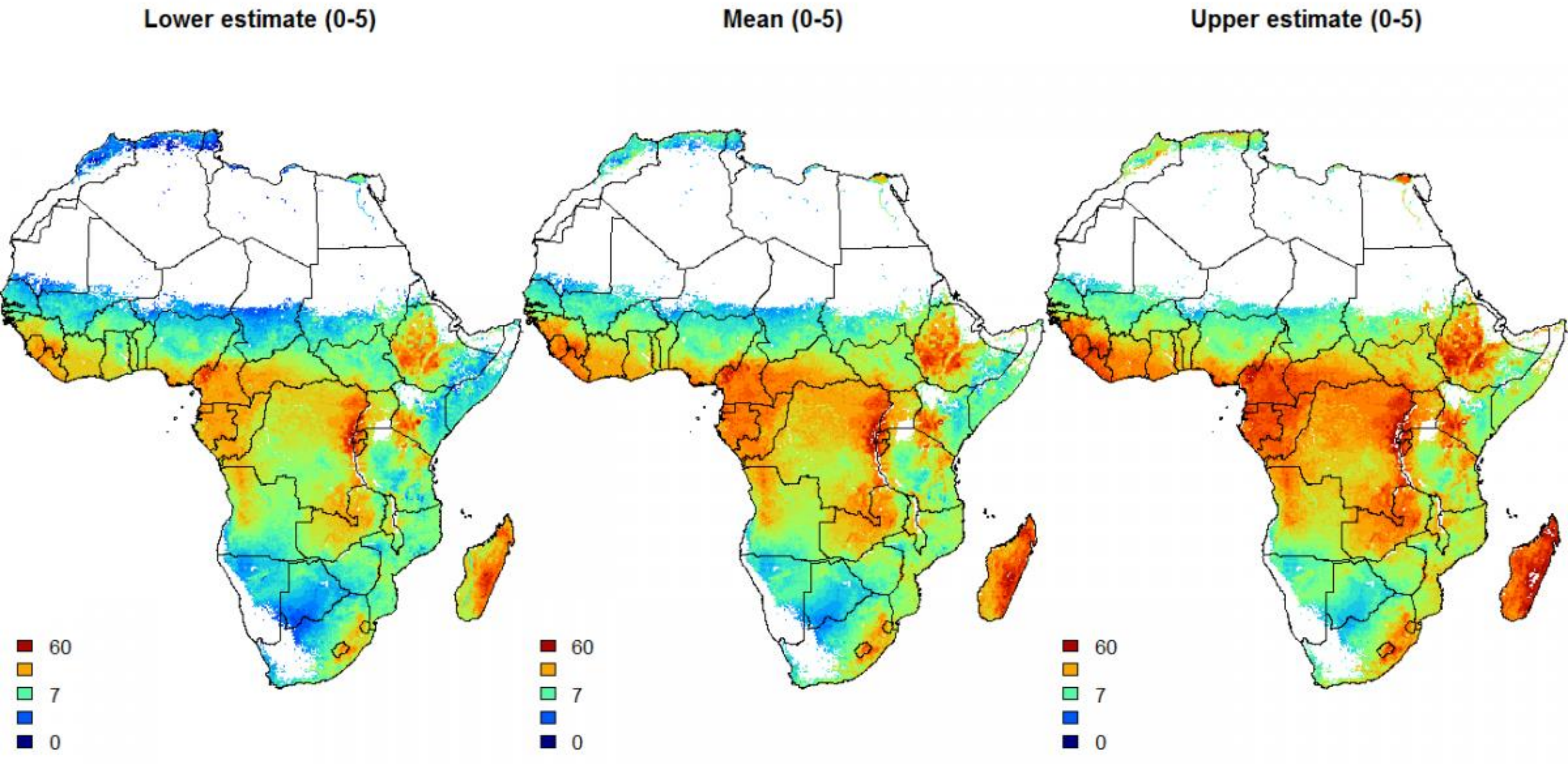


upper limit p_{95}

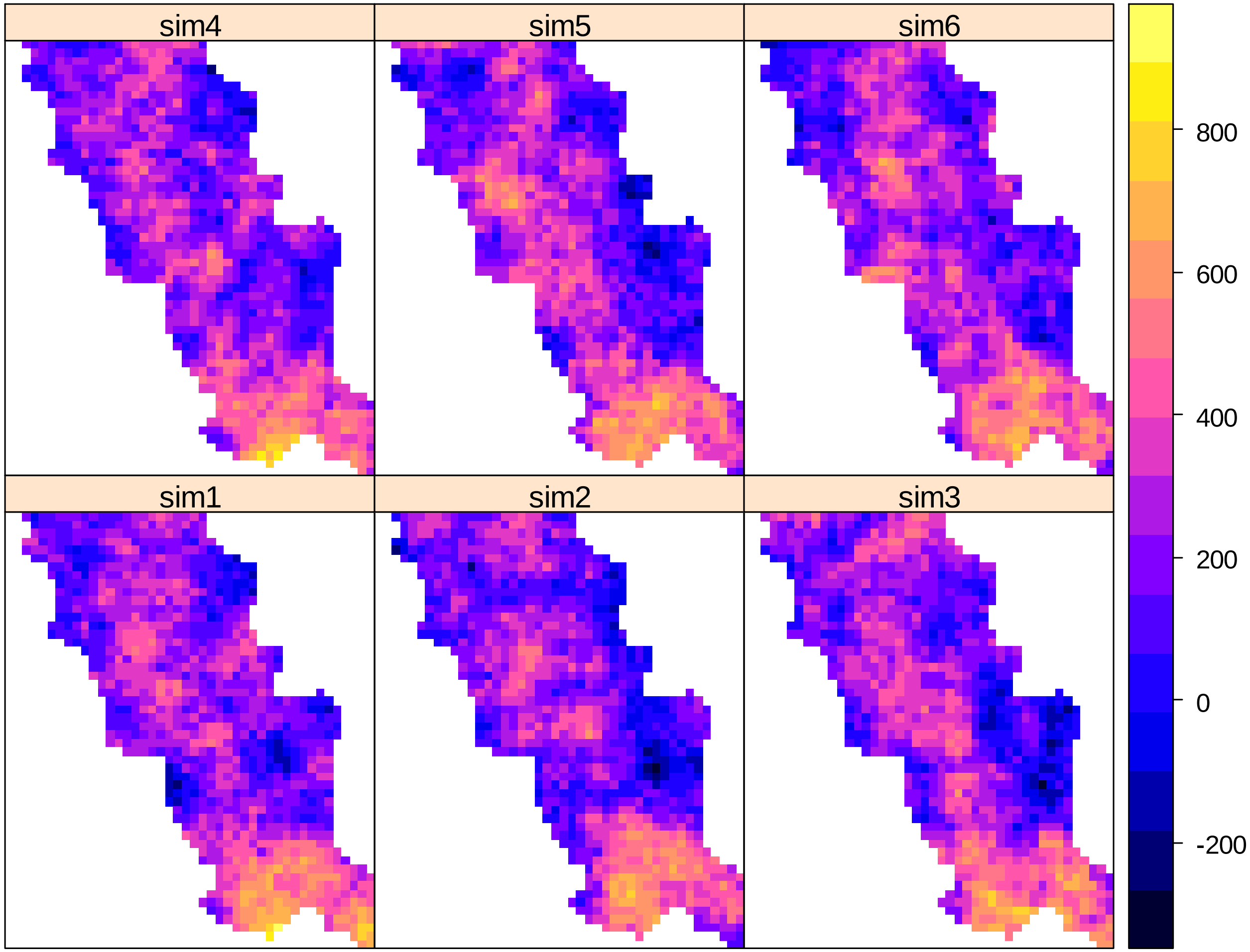




Uncertainties in SoilGrids1km also ‘automatically’ quantified because we used a geostatistical approach



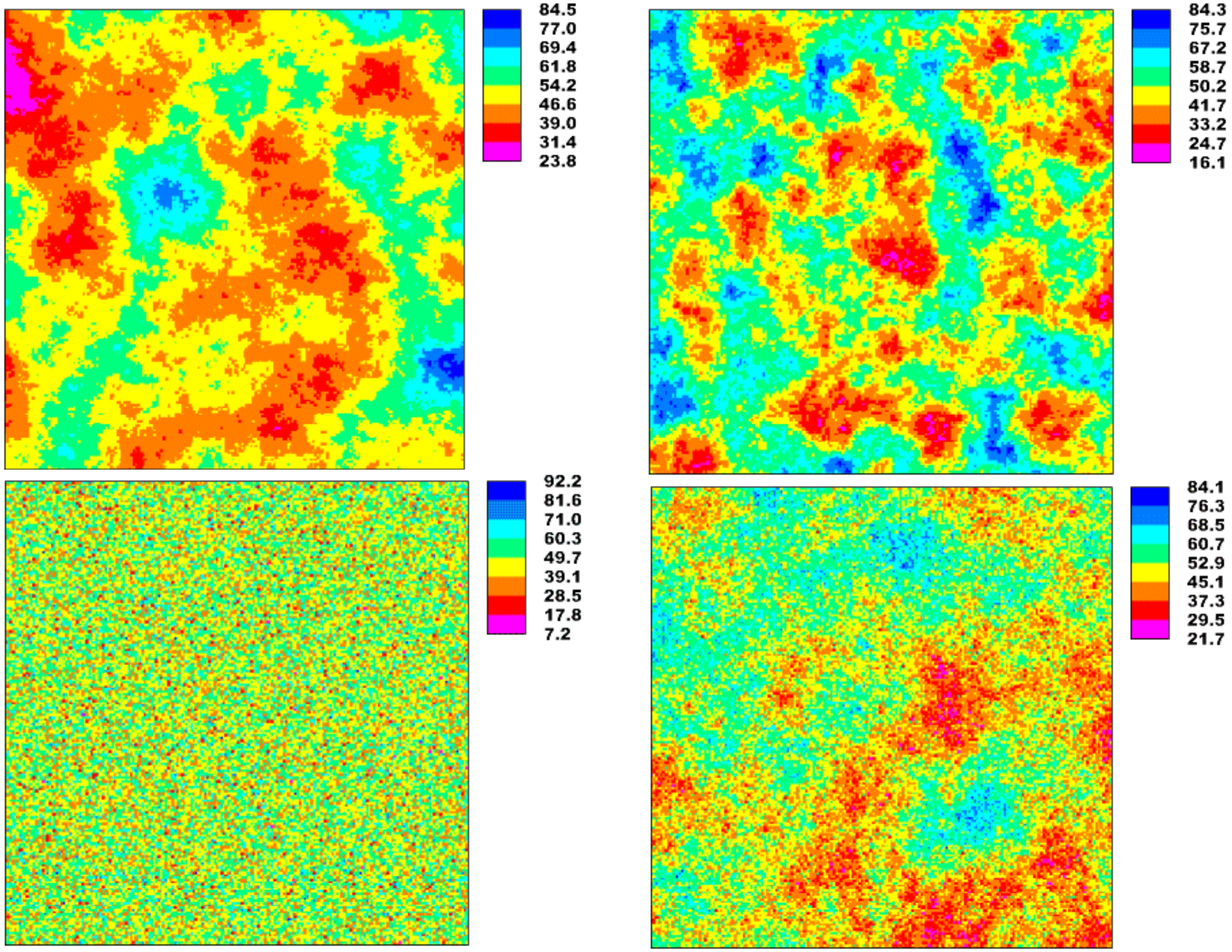
We can also sample from the spatial probability distribution using a random number generator



Spatial stochastic simulation

- Kriging makes optimal predictions: it yields the most likely value at any location
- But it is only a prediction. The real value is uncertain, we treat it as stochastic, it has a probability distribution
- In spatial stochastic simulation we do not compute a prediction but instead we generate a possible reality, by simulating from the probability distribution (using a random number generator)
- Simulation must take into account that errors at (nearby) locations are correlated
- Examples shown on Monday were created in this way

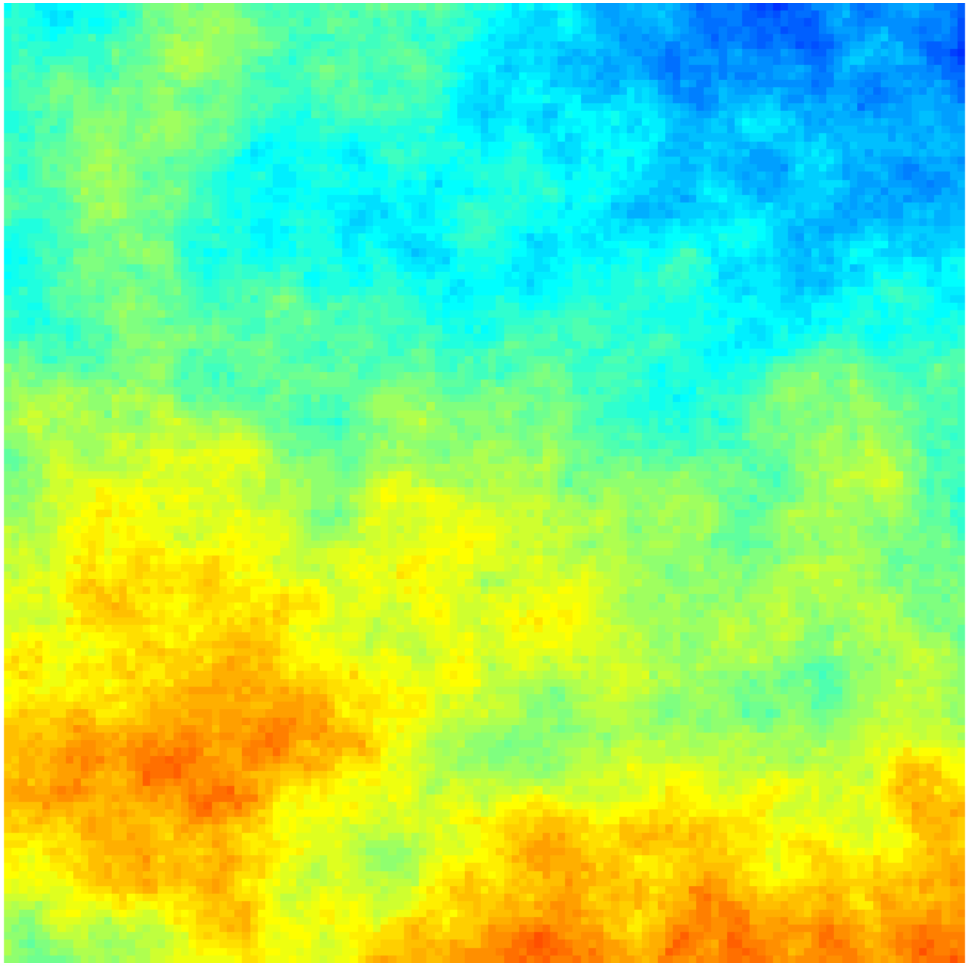
Spatial stochastic simulation, examples



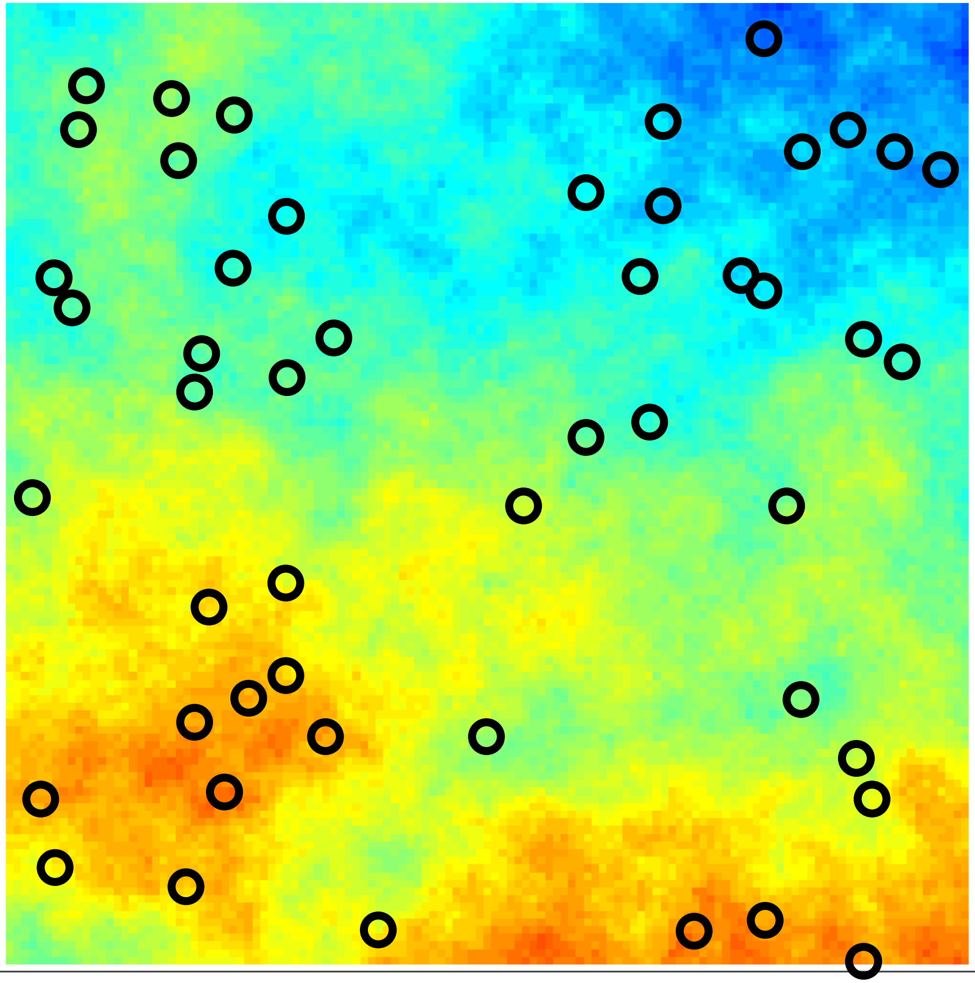
Sequential Gaussian simulation, how does it work

1. Visit a location that was not measured
2. Krige to the location using the available data, this yields a probability distribution of the target variable
3. Draw a value from the probability distribution using a random number generator and assign this value to the location
4. Add the simulated value to the data set, and move to another location
5. Repeat the procedure above until there are no locations left

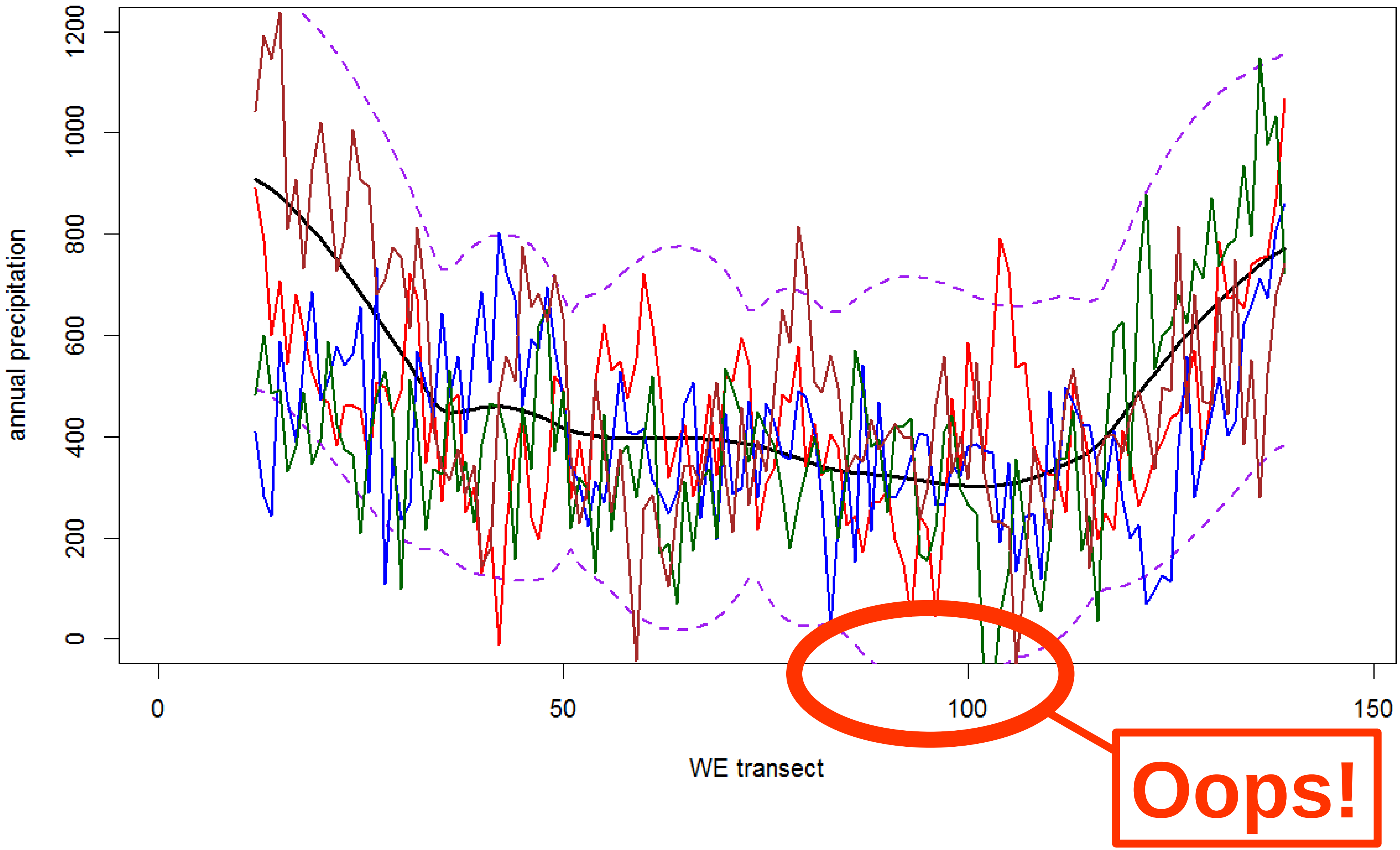
Simulations, obtained by sequential Gaussian simulation



Conditional simulation 'honour' the data at observation points



For illustration: Kriging predictions and simulations of annual rainfall along a West-East transect in Turkey



Programme of this module

Lecture:

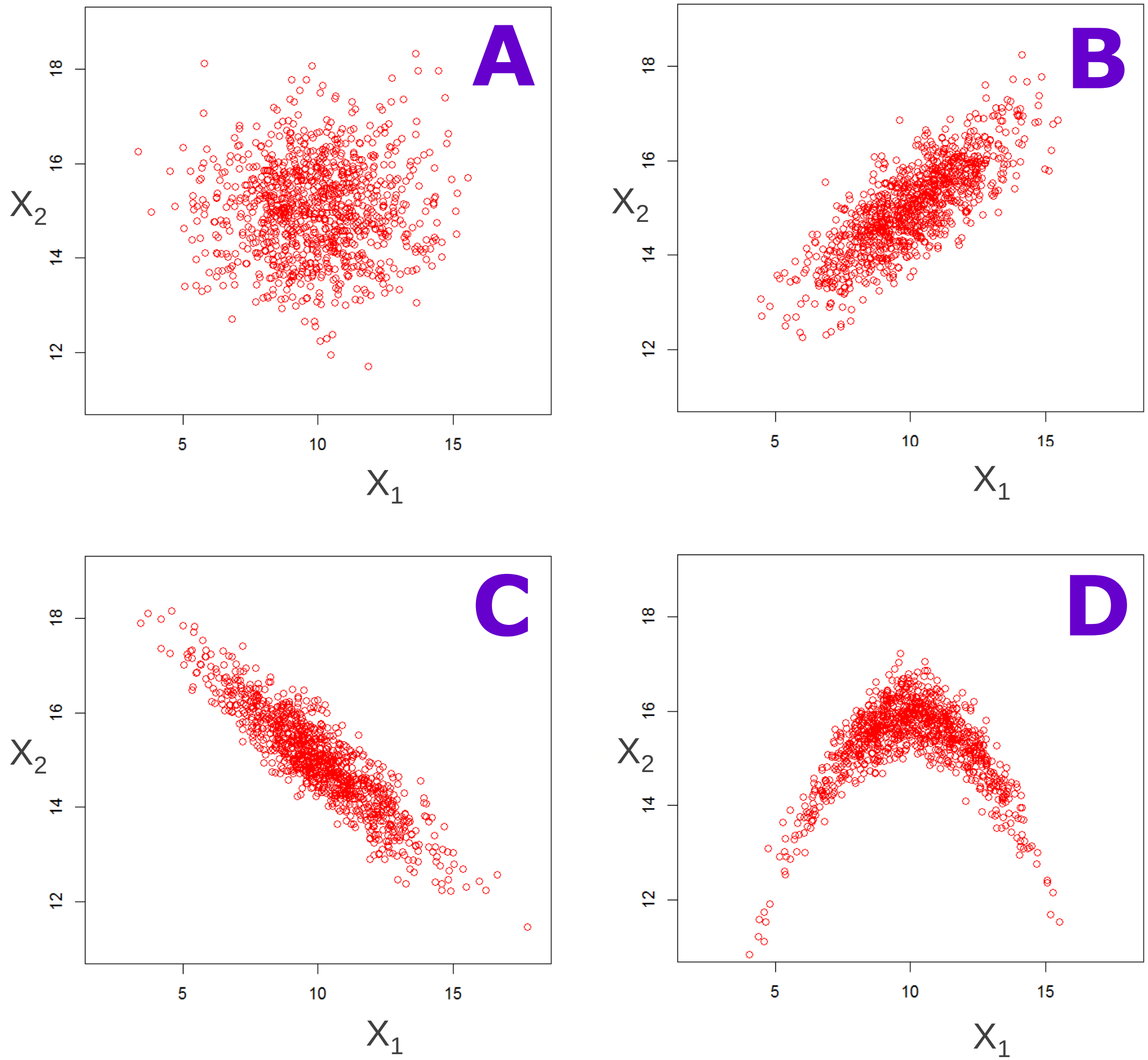
- Exploring error and uncertainty, what is it?
- Statistical modelling of uncertainty with probability distributions
- Uncertainty propagation in spatial analysis and environmental modelling
- Derivation of soil carbon stock from soil properties, with uncertainty propagation

Computer practical:

- Analyse how uncertainties in soil and other factors propagate through a very simple ‘model’ and may affect environmental decision making

Which statement is incorrect?

1) We get the largest output uncertainty for $Y=X_1+X_2$ in situation B.
2) We get the largest output uncertainty for $Y=X_1-X_2$ in situation C.
3) We get the largest output uncertainty for $Y=X_1*X_2$ in situation A.
4) We get the smallest output uncertainty for $Y=X_1/X_2$ in situation B.



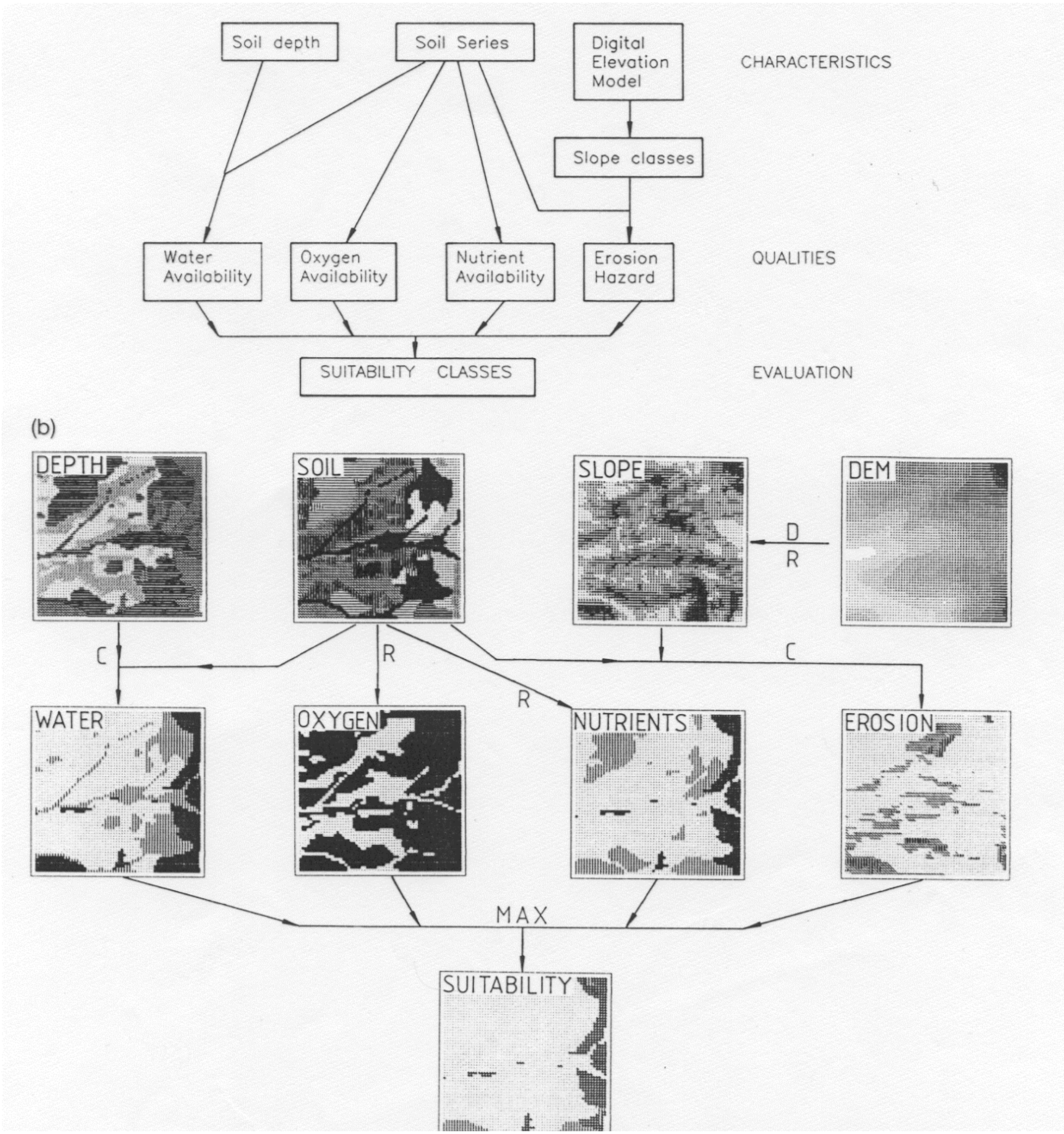
submit your answer to www.menti.com

output map = f(input maps)

input → environmental model → output

For instance:

- slope angle = f(elevation)
- erosion risk = f(landuse, slope, soil type)
- soil acidification = f(deposition, soil physical and chemical characteristics)
- crop yield = f(soil properties, water availability, fertilization)

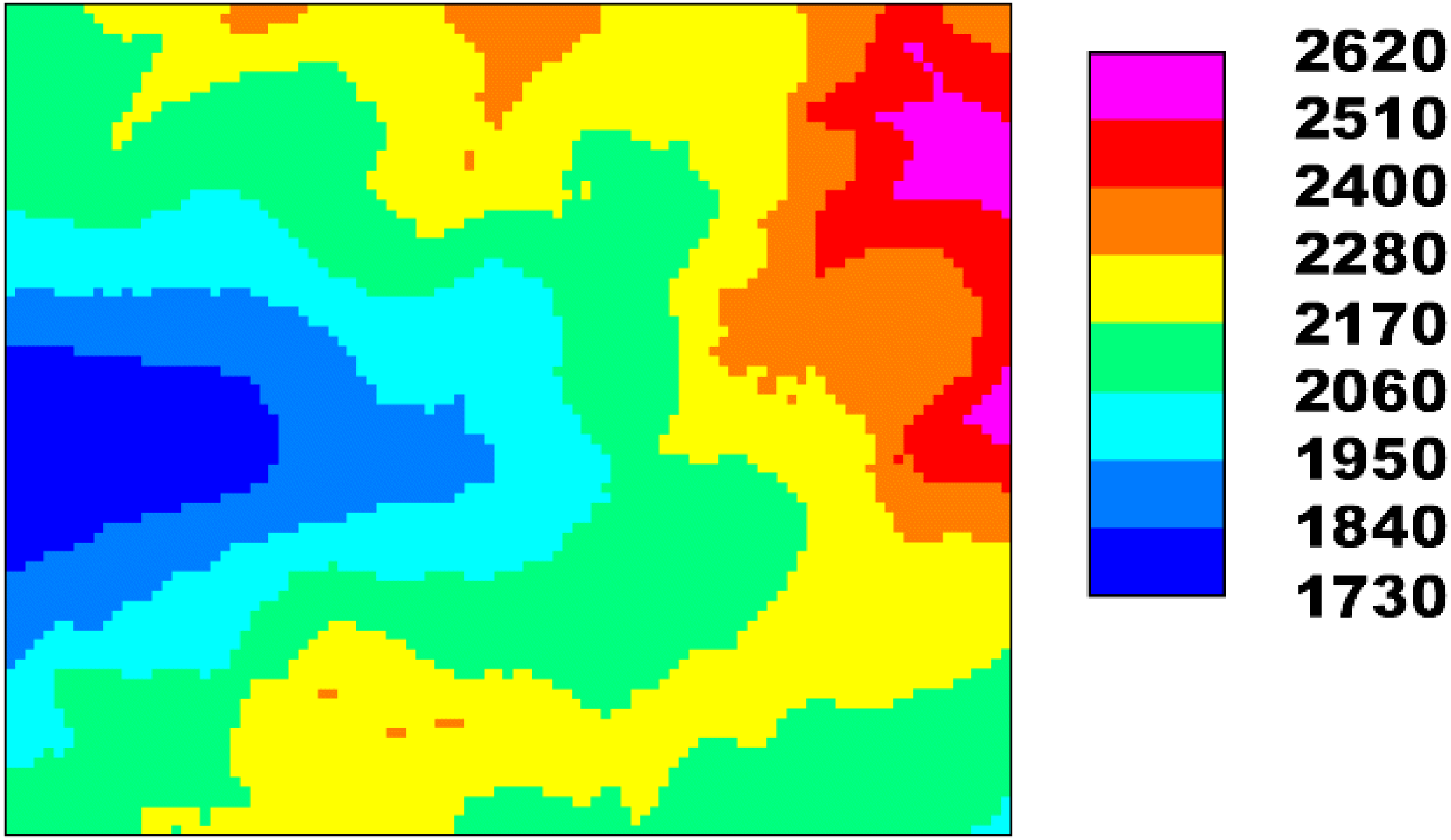


from Burrough, 1986

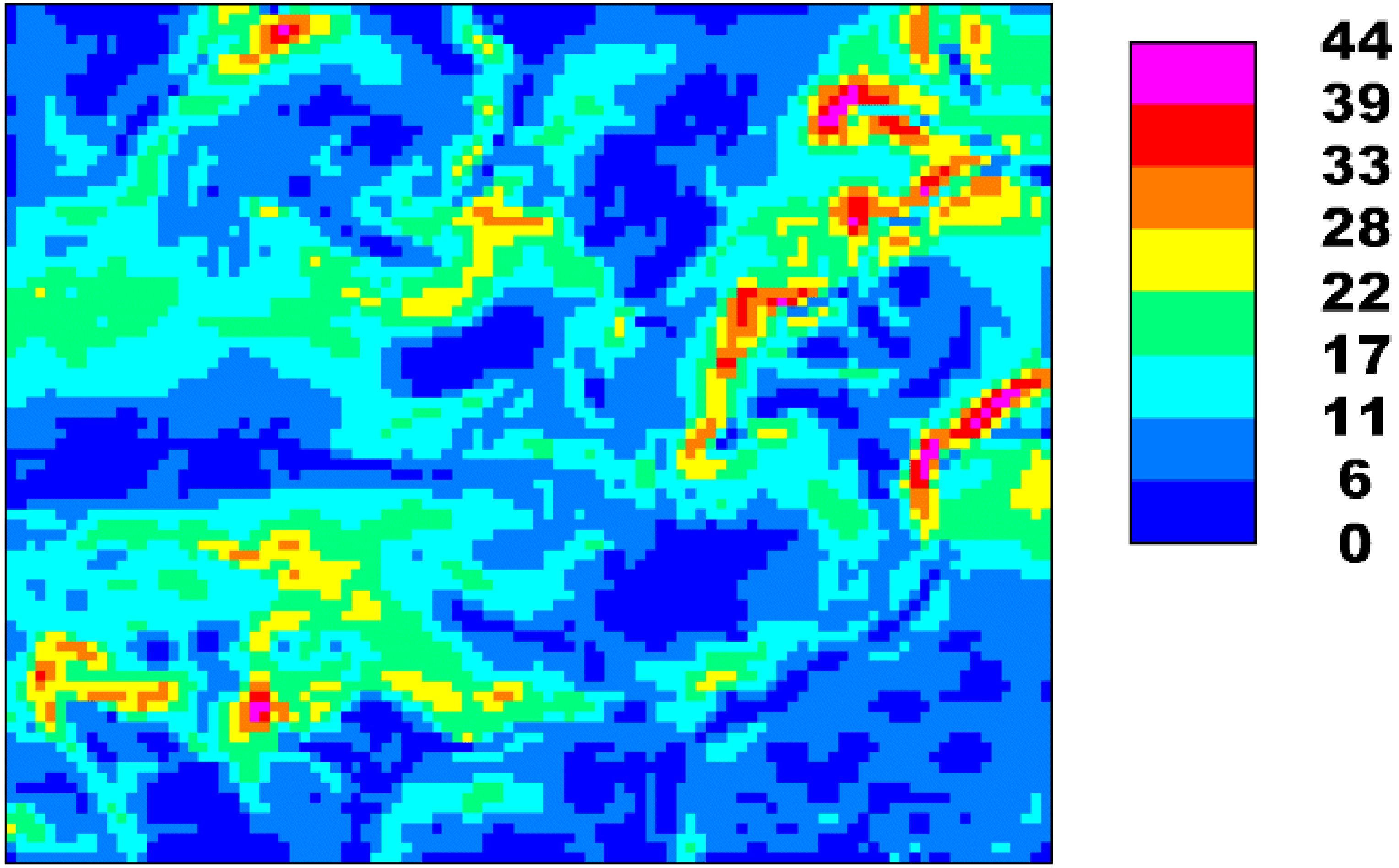
Monte Carlo method for uncertainty propagation

Explain by means of an example

Example: computing slope from DEM for a 2 by 2.5 km area in the Austrian Alps

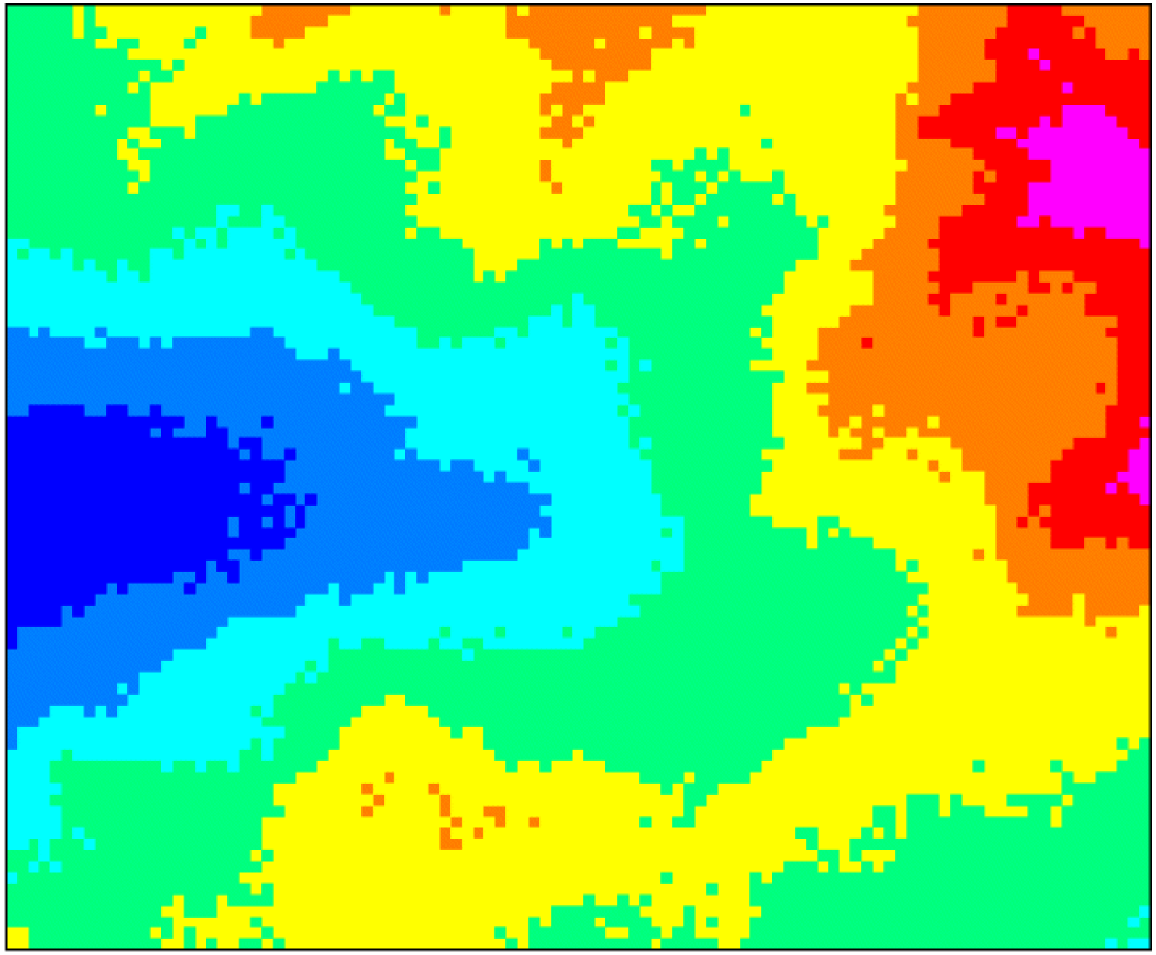


Slope map computed from the DEM (percent):

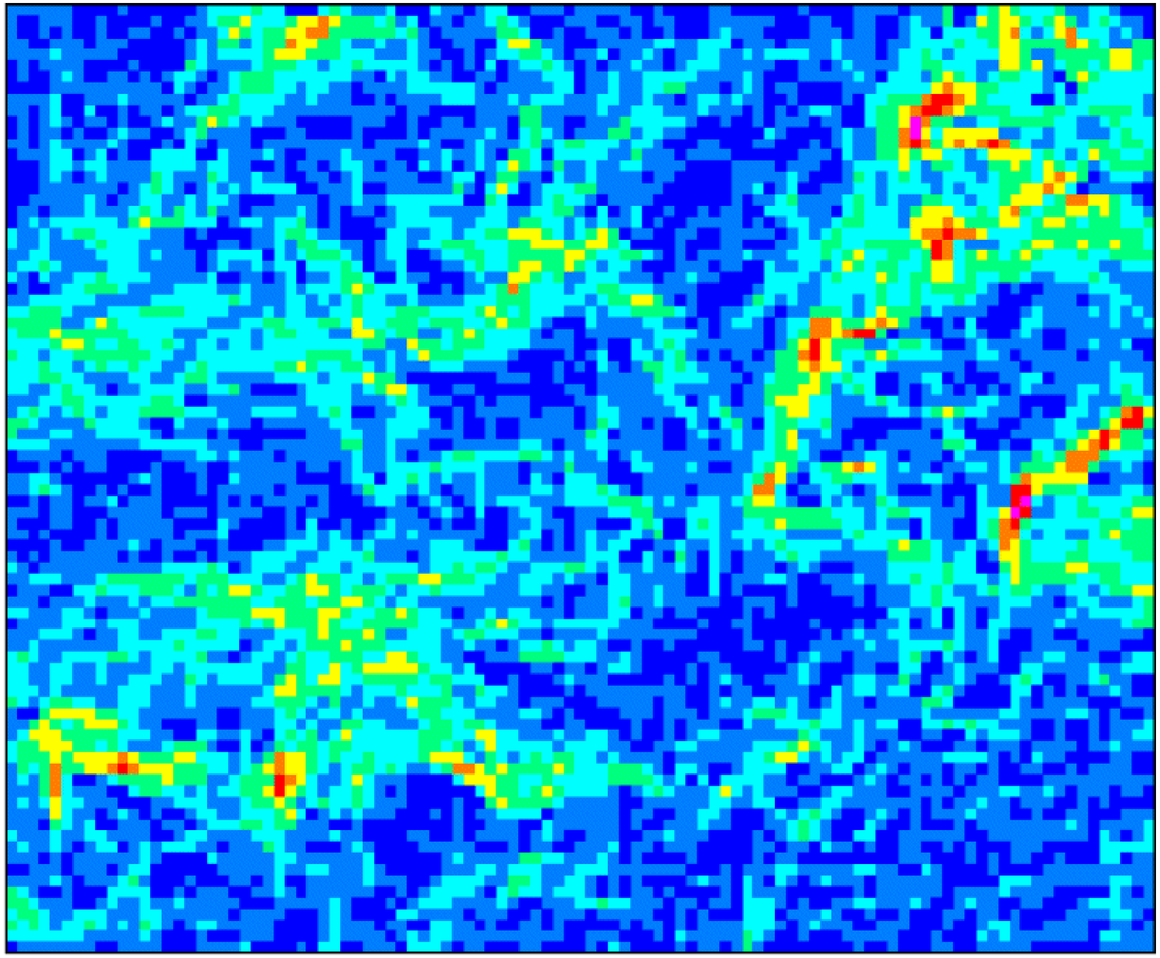


Now let the error in the DEM be ±10 meter

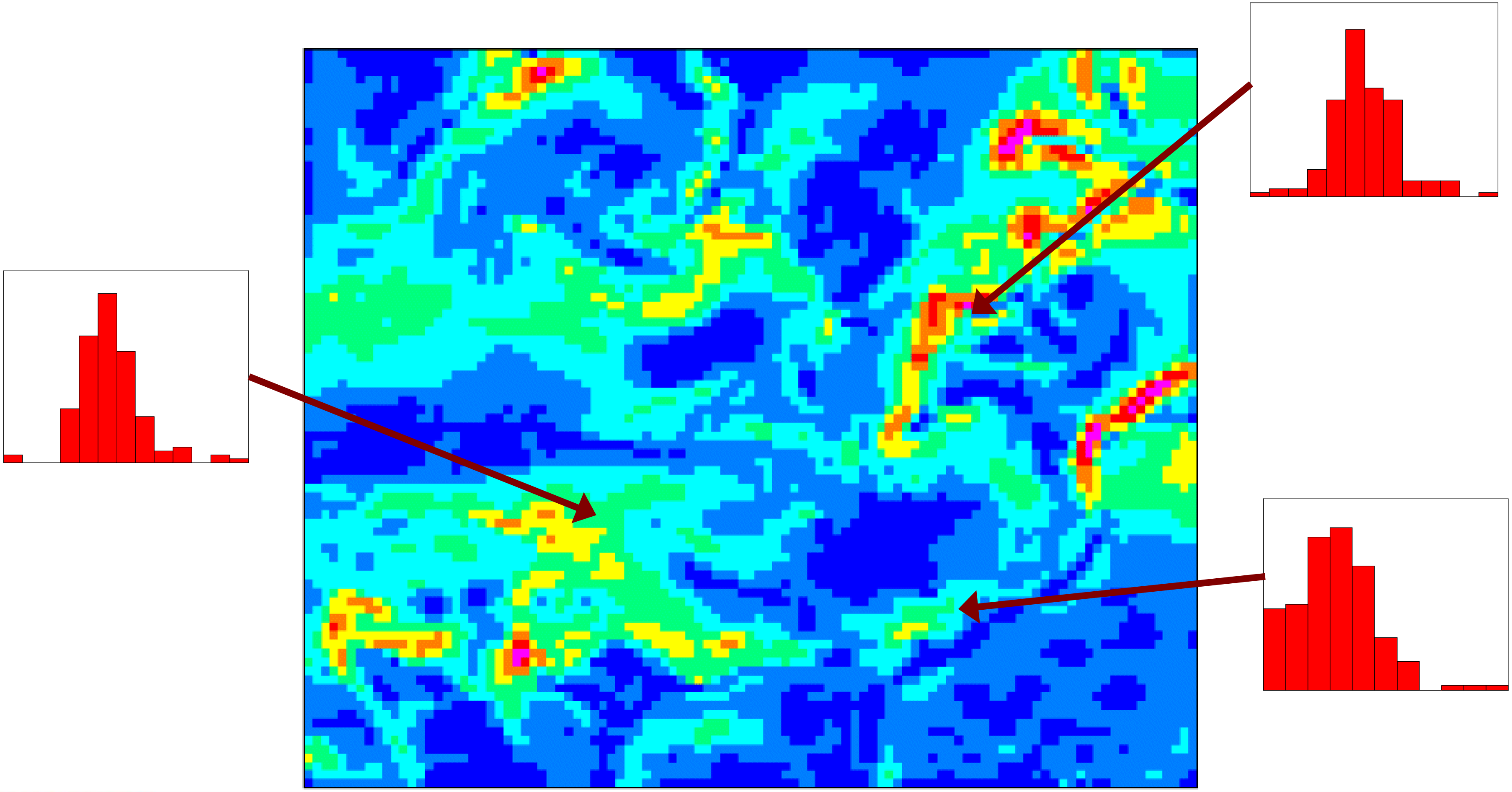
Realisations of uncertain DEM:



Corresponding uncertain slope maps:



Histograms capture uncertainty in slope:

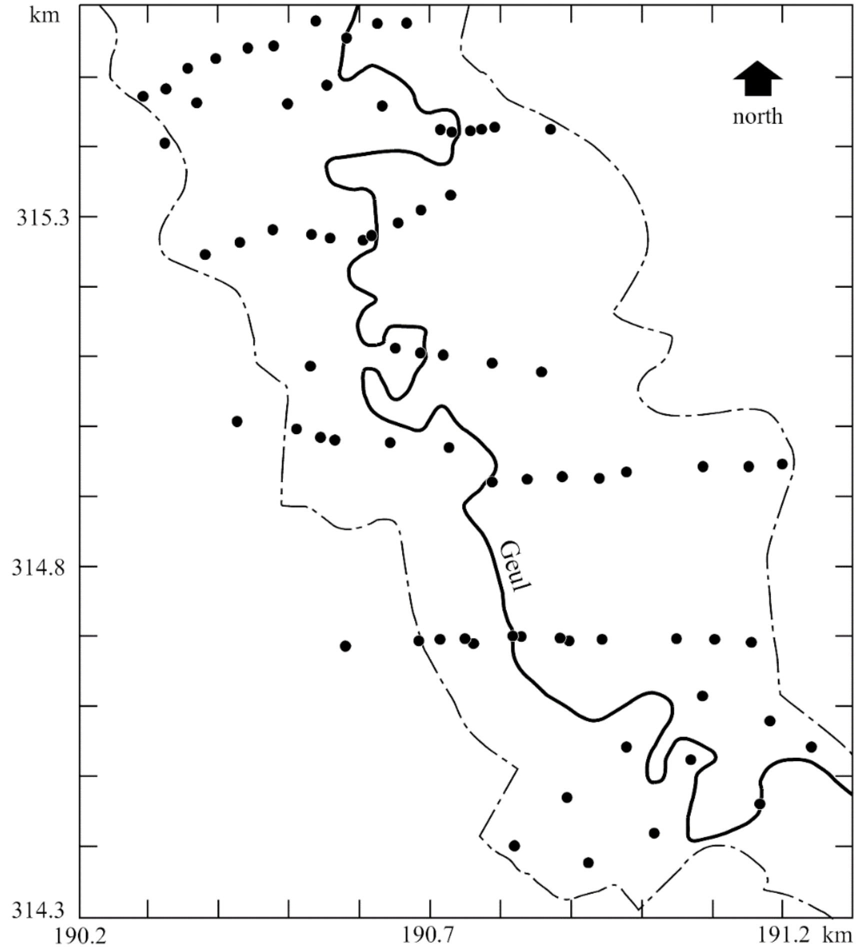


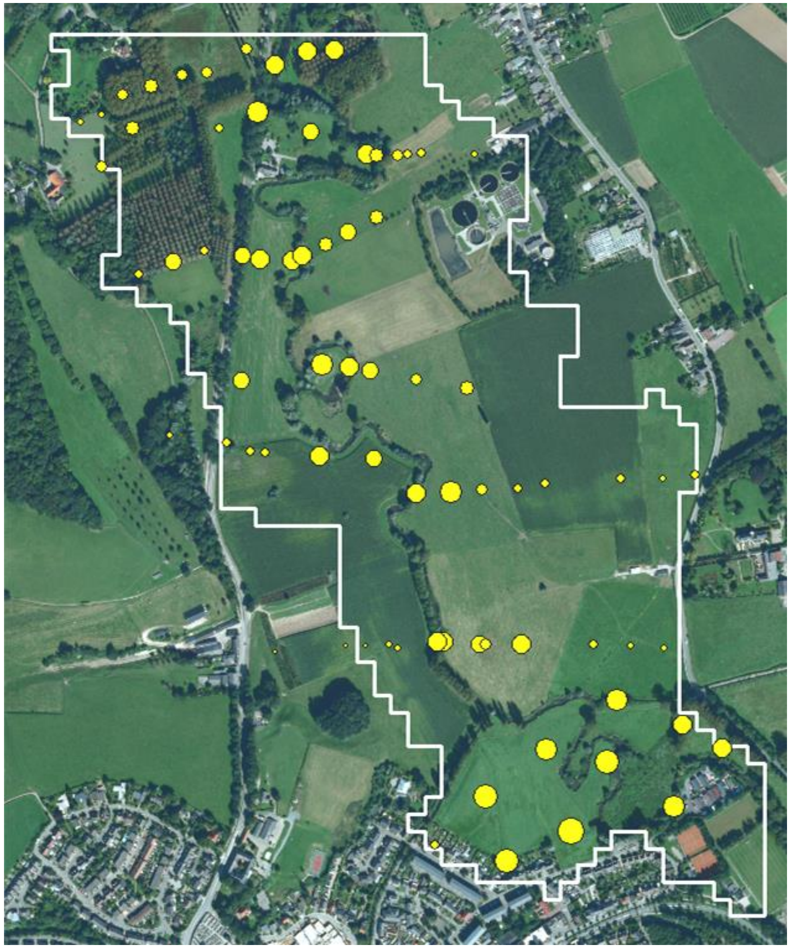
Monte Carlo algorithm

1. Repeat N times (N>100):
   a. Simulate a realisation from the probability distribution of the uncertain inputs
   b. Run the model with these inputs and store result

2. Analyse the N model outputs by computing summary statistics such as the mean and standard deviation (the latter is a measure of the output uncertainty)

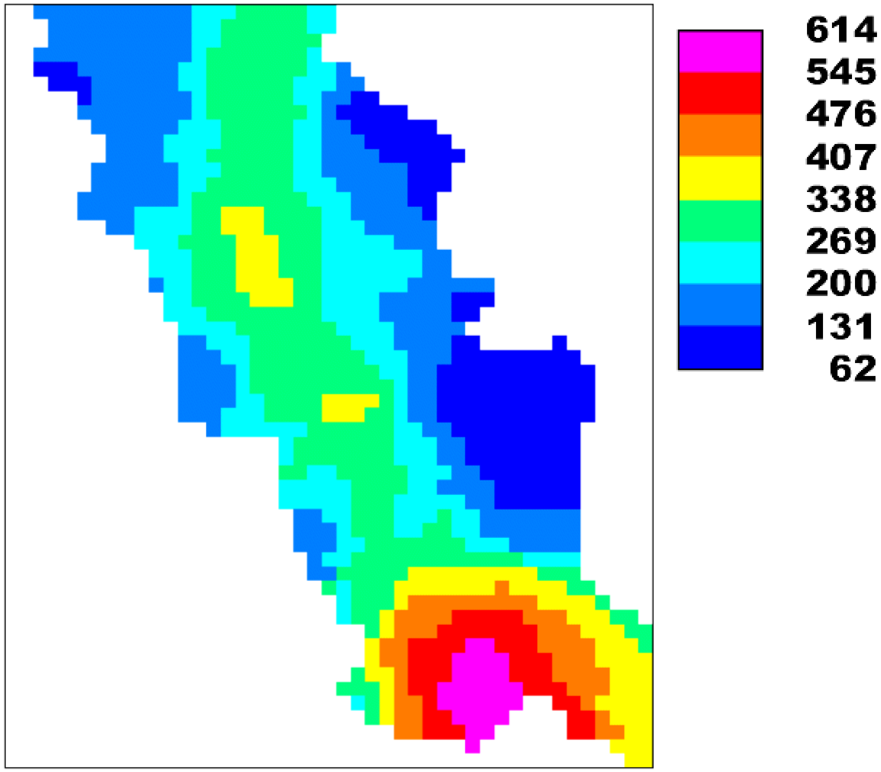
Return to Geul lead pollution example



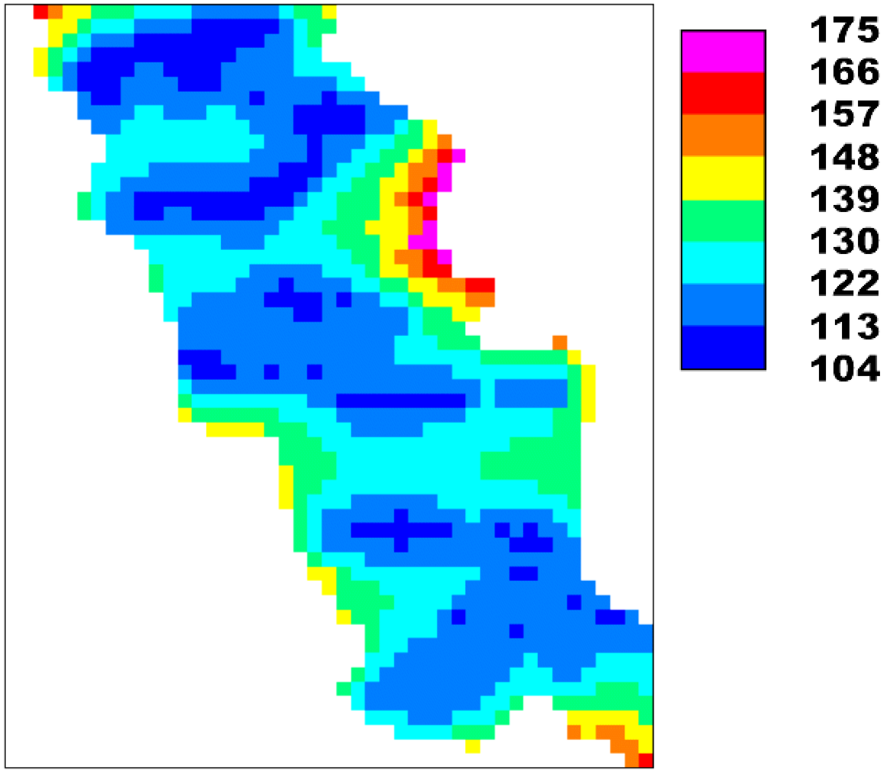


On Monday we interpolated the topsoil lead concentration with ordinary kriging

kriging prediction (mg/kg)



kriging st. dev. (mg/kg)



Playing children ingest lead

$$I = PB \cdot S$$

where:

I = lead ingestion

PB = lead concentration of soil

S = soil consumption

How do errors in mapped lead concentration of soil and soil consumption propagate to lead ingestion?

- Uncertainty in lead concentration and soil consumption propagate both to lead ingestion
- Uncertainty in lead concentration caused by interpolation error and quantified with ordinary kriging
- Research by Medical Faculty University Maastricht indicates that soil consumption may be assumed lognormally distributed with mean 0.120 g/day and st.dev. 0.250 g/day
- Uncertainty propagation can be analysed with the Monte Carlo method. In the computer practical you will investigate:
  - In which parts of the study area is the Acceptable Daily Intake of 50 μg/day not exceeded?
  - Which parts are we 95% certain that the ADI is not exceeded?
  - Which is the main source of uncertainty?

Validation for model-free accuracy assessment

Uncertainty quantification by the kriging standard deviation makes model assumptions (e.g. stationarity, normal distribution), can we assess the accuracy of soil property maps also objectively, without making assumptions?

YES WE CAN!

It will all be explained in this afternoon's module

Programme of this module

Lecture:

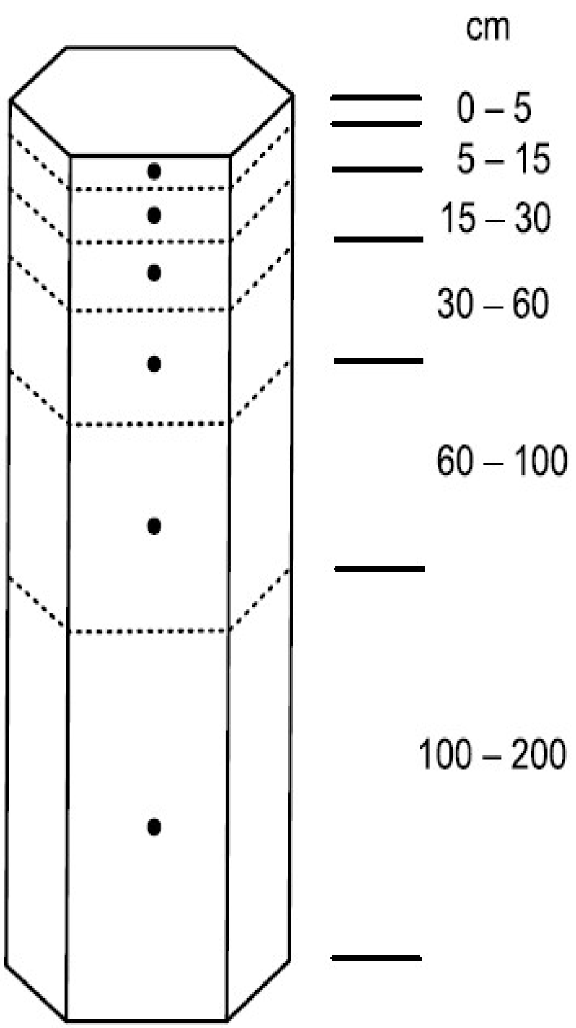
- Exploring error and uncertainty, what is it?
- Statistical modelling of uncertainty with probability distributions
- Uncertainty propagation in spatial analysis and environmental modelling
- Derivation of soil carbon stock from soil properties, with uncertainty propagation

Computer practical:

- Analyse how uncertainties in soil and other factors propagate through a very simple ‘model’ and may affect environmental decision making

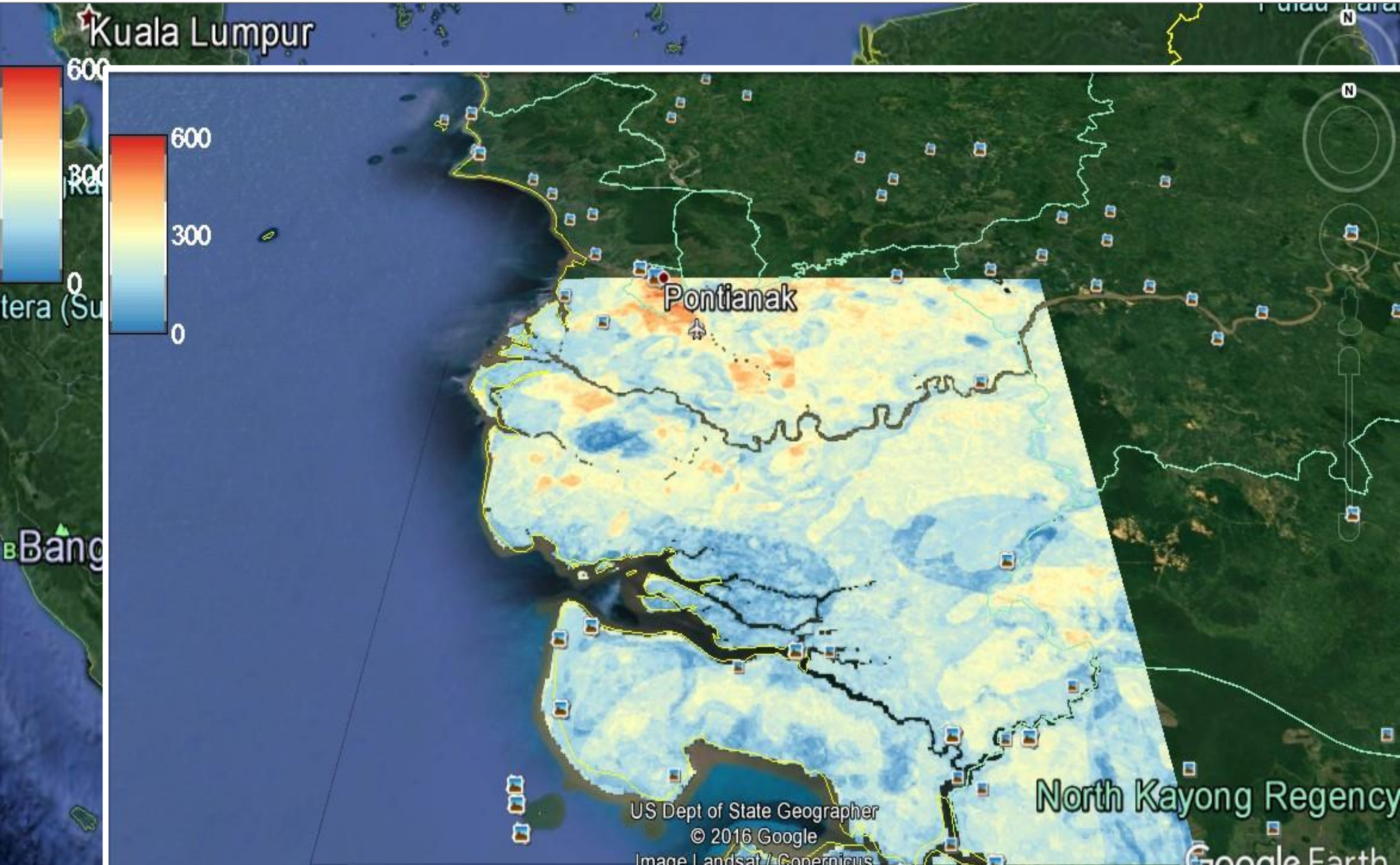
First compute SOC stock per layer:

$$OCS\ [\mathrm{kg\ m^{-2}}] = \frac{ORC}{1000}\ [\mathrm{kg\ kg^{-1}}] \cdot \frac{HOT}{100}\ [\mathrm{m}] \cdot BLD\ [\mathrm{kg\ m^{-3}}] \cdot \frac{100 - CRF\ [\%]}{100}$$

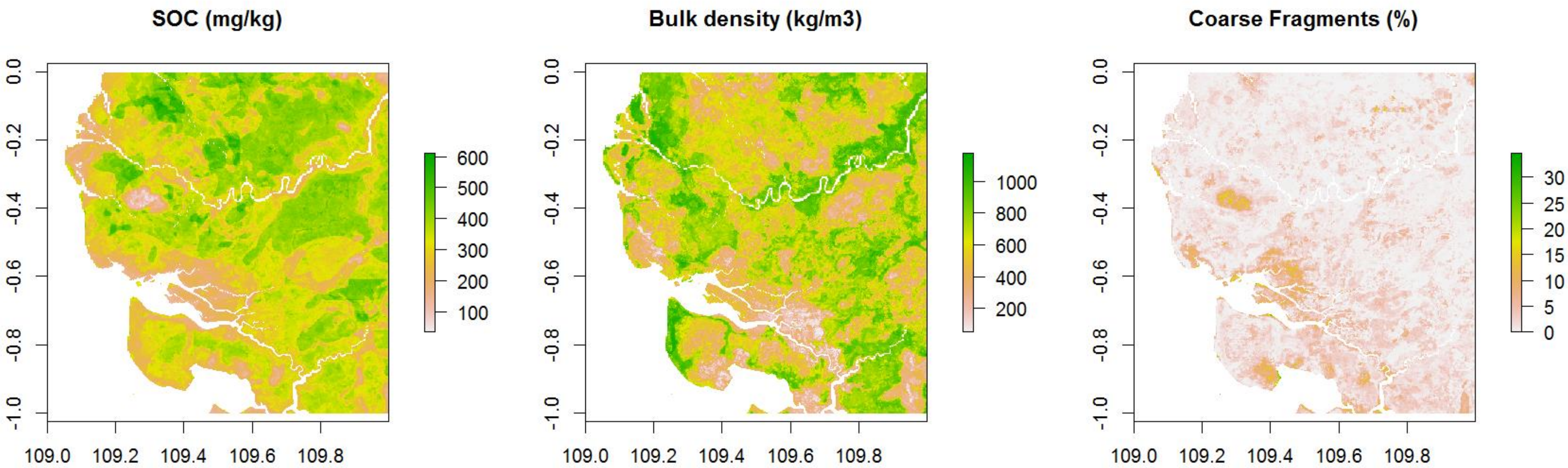


Next aggregate over layers:

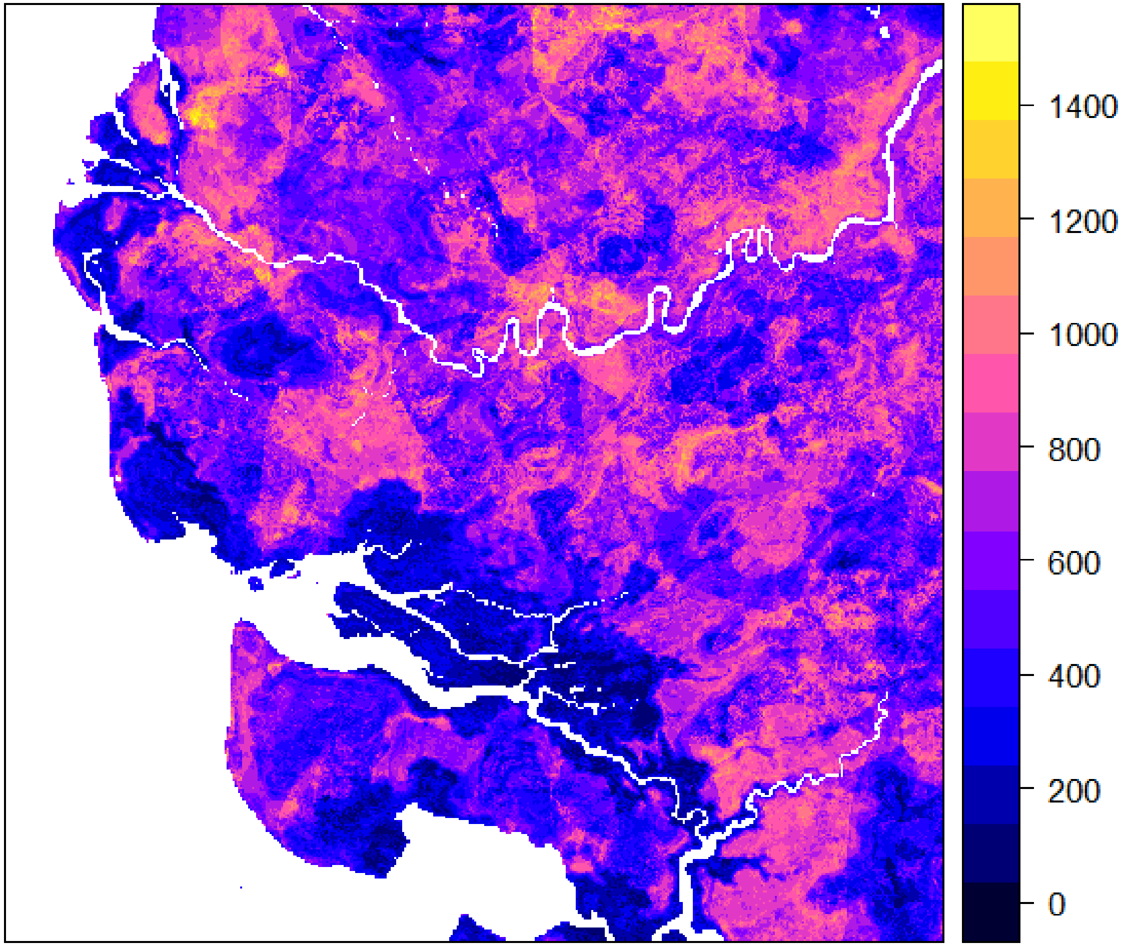
Example: 1x1 degree tile in Borneo, Indonesia, input data from SoilGrids



Input maps topsoil (0-30 cm)



Output SOC stock map (ton/ha) 0-30 cm



How uncertain is this map?

- Uncertainties in SOC concentration, bulk density and coarse fragments will propagate to SOC stock
- We can analyse the uncertainty propagation using the Monte Carlo method, as before
- But this is a nice example to show how uncertainty propagation can also be analysed using the Taylor series uncertainty propagation method
- One problem still to be solved is quantification of the input uncertainties (i.e., uncertainty in SOC, bulk density and coarse fragments)

$$O = f(U_1, U_2, \ldots, U_m)$$

By linearising the model f we get:

$$Var(O) \cong \sum_{i=1}^{m} Var(U_i) \cdot \left(\frac{\partial f}{\partial U_i}\right)^2$$

magnitude of input uncertainty matters

but also sensitivity of model to input

Taylor series method, graphical illustration in case of a single uncertain input (m=1)

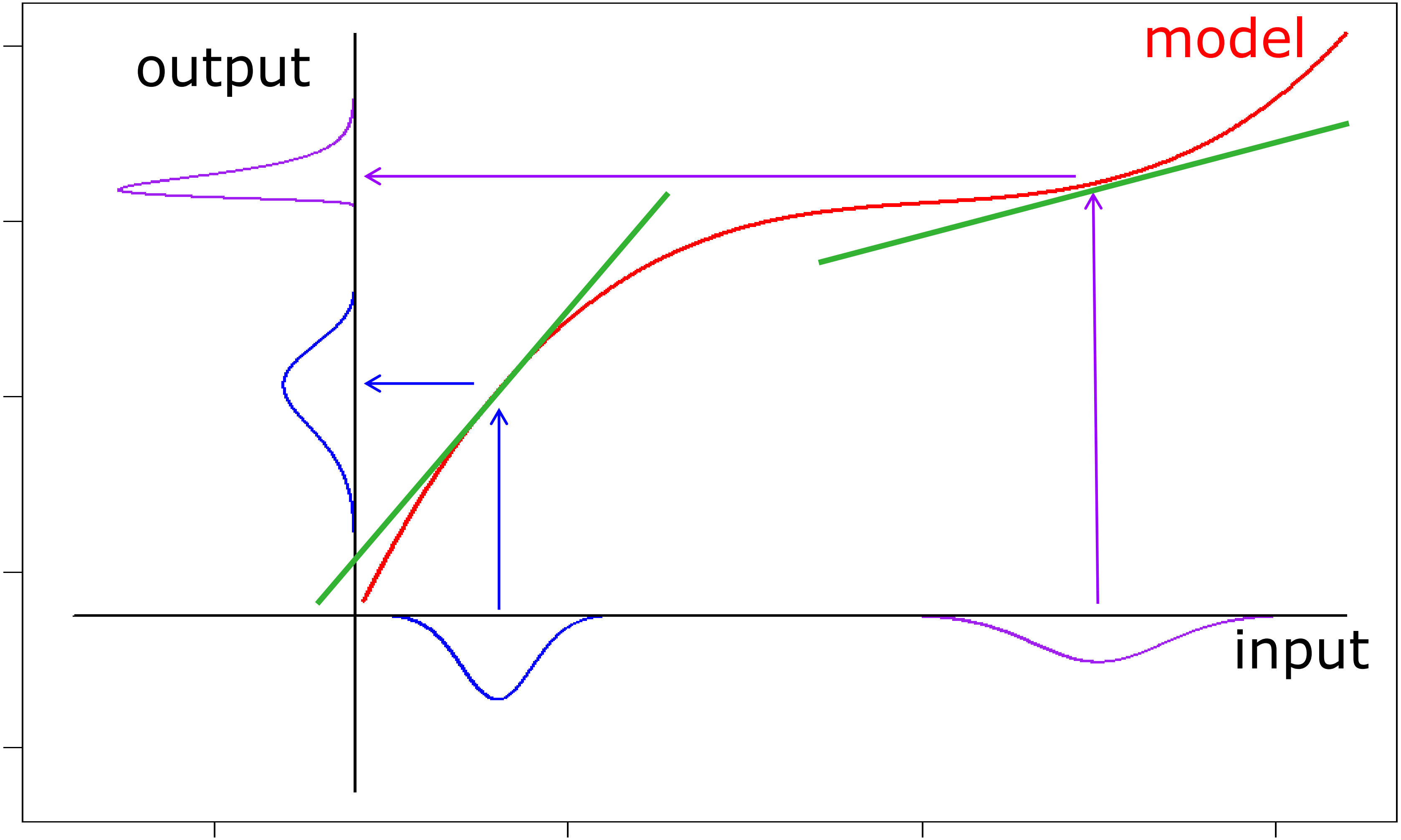
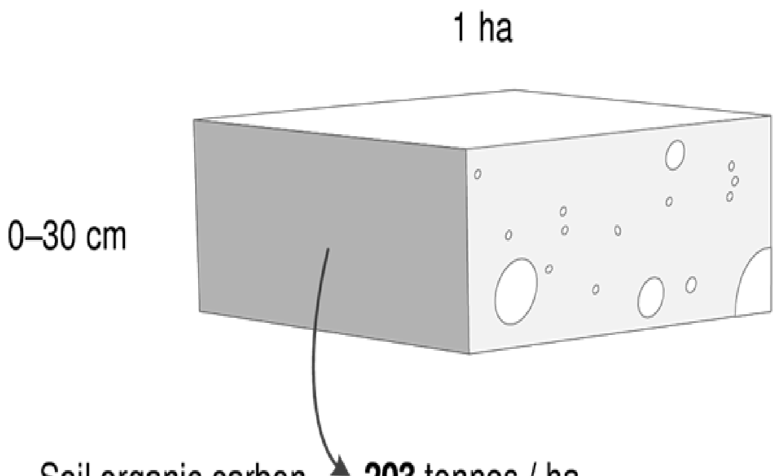


Illustration how this works out for SOC stock uncertainty propagation



Soil organic carbon stock **203** tonnes / ha (±44 tonnes / ha)

Total fine-earth soil **4050.0** tonnes / ha

Bulk density (BLD): 1500 kg / m^3 (s.d. = ±100)

Organic carbon (ORC): 50‰ (s.d. = ±10)

Coarse fragments (CRF): 10% (s.d. = ±5)

Total volume of the block (HOT): 30 cm (· 1 ha)

Soil organic carbon stock (OCS): 203 tonnes / ha (±44)

$$\mathbf{OCS} = ORC/1000 \cdot BLD \cdot (100\text{-}CRF)/100 \cdot HOT/100$$
$$= 1/10{,}000{,}000 \cdot ORC \cdot BLD \cdot (100\text{-}CRF) \cdot HOT$$
$$= 1/10{,}000{,}000 \cdot 50 \cdot 1500 \text{ kg / m}^3 \cdot (100\text{-}10) \cdot 30 \text{ cm}$$
$$= 20.25 \text{ kg / m}^2 = 203 \text{ tonnes / ha}$$

$$\mathbf{OCS.sd} = 1/10{,}000{,}000 \cdot HOT \cdot \mathrm{sqrt}(\, BLD^2 \cdot (100 - CRF)^2 \cdot ORC.sd^2 +$$
$$+ BLD.sd^2 \cdot (100 - CRF)^2 \cdot ORC^2 + BLD^2 \cdot CRF.sd^2 \cdot ORC^2\,)$$
$$= 4.4 \text{ kg / m}^2 = 44.1 \text{ tonnes / ha}$$

Input uncertainty crudely derived from global SoilGrids RMSE statistics

PLOS ONE

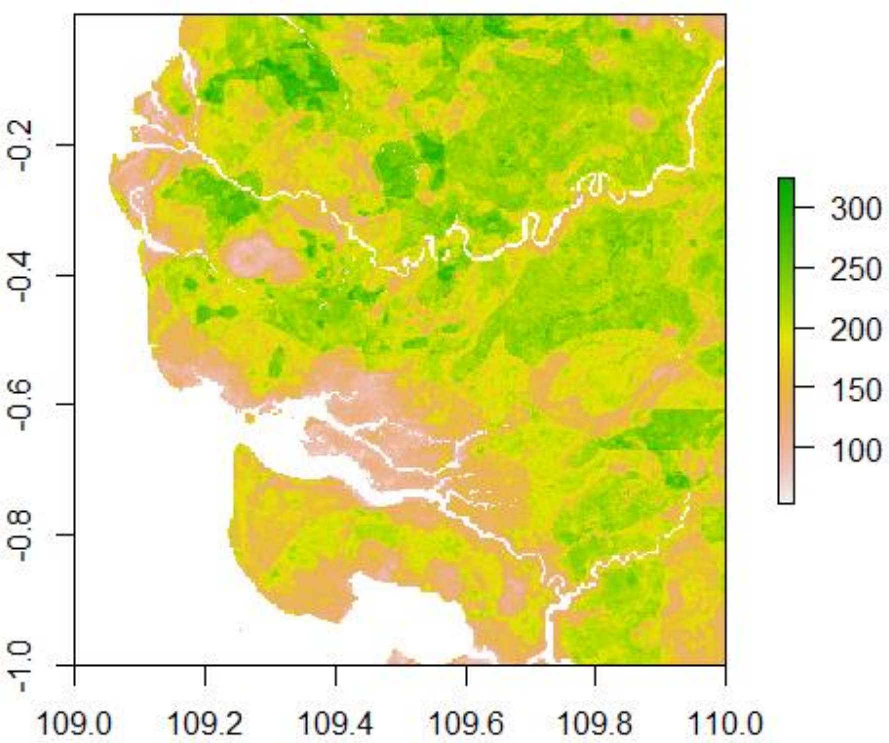
Table 1. SoilGrids average prediction error for key soil properties based on 10–fold cross-validation. N = "Number of samples used for training", ME = "Mean Error", MAE = "Mean Absolute Error", RMSE = "Root Mean Squared Error" and R-square = "Coefficient of determination" (amount of variation explained by the model). For variables with a skew distribution, such as organic carbon, coarse fragments and CEC, the accuracy statistics are also provided on log-scale⊗.

Variable name	N	Min	Max	ME	MAE	RMSE	R-square	RMSE⊗	R-square⊗
Soil organic carbon (gravimetric)	605,054	0	520	-0.292	10.2	32.8	63.5%	0.715	68.8%
pH index (H_2O solution)	604,019	2.1	11.0	-0.002	0.4	0.5	83.4%		
Sand content (gravimetric)	616,762	1%	94%	-0.037	9.0	13.1	78.6%		
Silt content (gravimetric)	613,750	2%	74%	0.023	6.7	9.8	79.4%		
Clay content (gravimetric)	625,159	2%	68%	-0.102	6.6	9.5	72.6%		
Coarse fragments (volumetric)	303,139	0%	89%	-0.104	5.5	10.9	55.9%	1.185	64.3%
Bulk density (fine earth fraction)	140,596	250	2870	-1.574	108.3	164.7	75.8%		
Cation-exchange capacity (fine earth fraction)	393,585	0	234	-0.071	5.5	10.3	64.5%	0.483	67.0%
Depth to bedrock (in cm)	1,580,798	0	125,000	-29	678	835	54.0%	1.12	42.8%

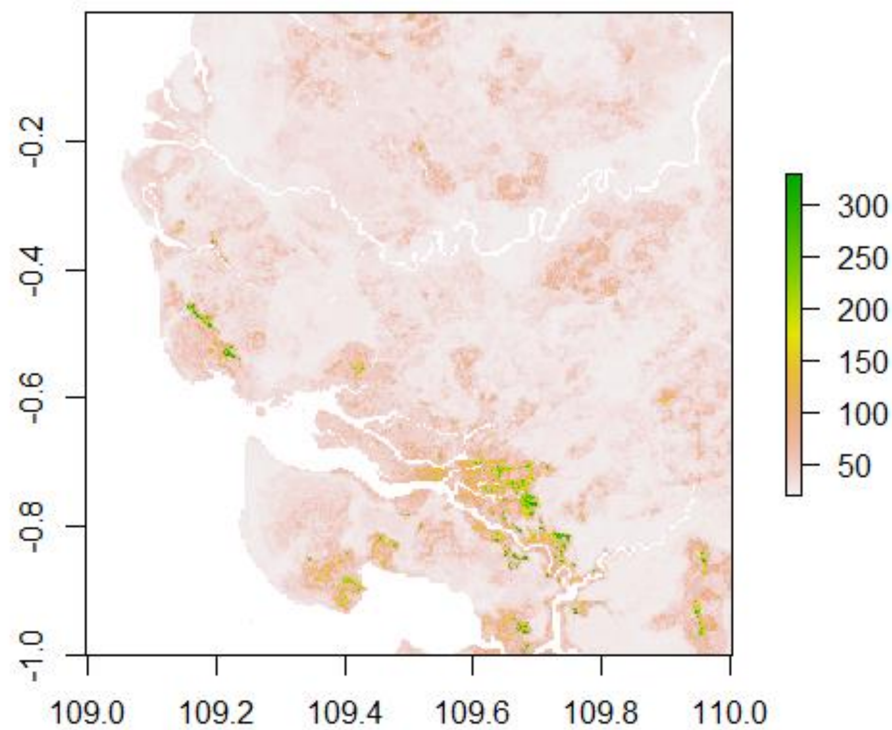
doi:10.1371/journal.pone.0169748.t001

Applicatuion of Taylor method shows that uncertainty SOC stock locally very high

standard deviation (ton/ha)



relative error (%)



World Soil Information

Programme of this module

Lecture:

- Exploring error and uncertainty, what is it?
- Statistical modelling of uncertainty with probability distributions
- Uncertainty propagation in spatial analysis and environmental modelling
- Derivation of soil carbon stock from soil properties, with uncertainty propagation

Computer practical:

- Analyse how uncertainties in soil and other factors propagate through a very simple 'model' and may affect environmental decision making

TIME FOR A BREAK