Uncertainty Analysis and Statistical Validation of Spatial Environmental Models

PE&RC Course 9-13 December 2024

Gerard Heuvelink and Sytze de Bruin

WAGENINGEN UNIVERSITY
WAGENINGEN UR

This week's programme

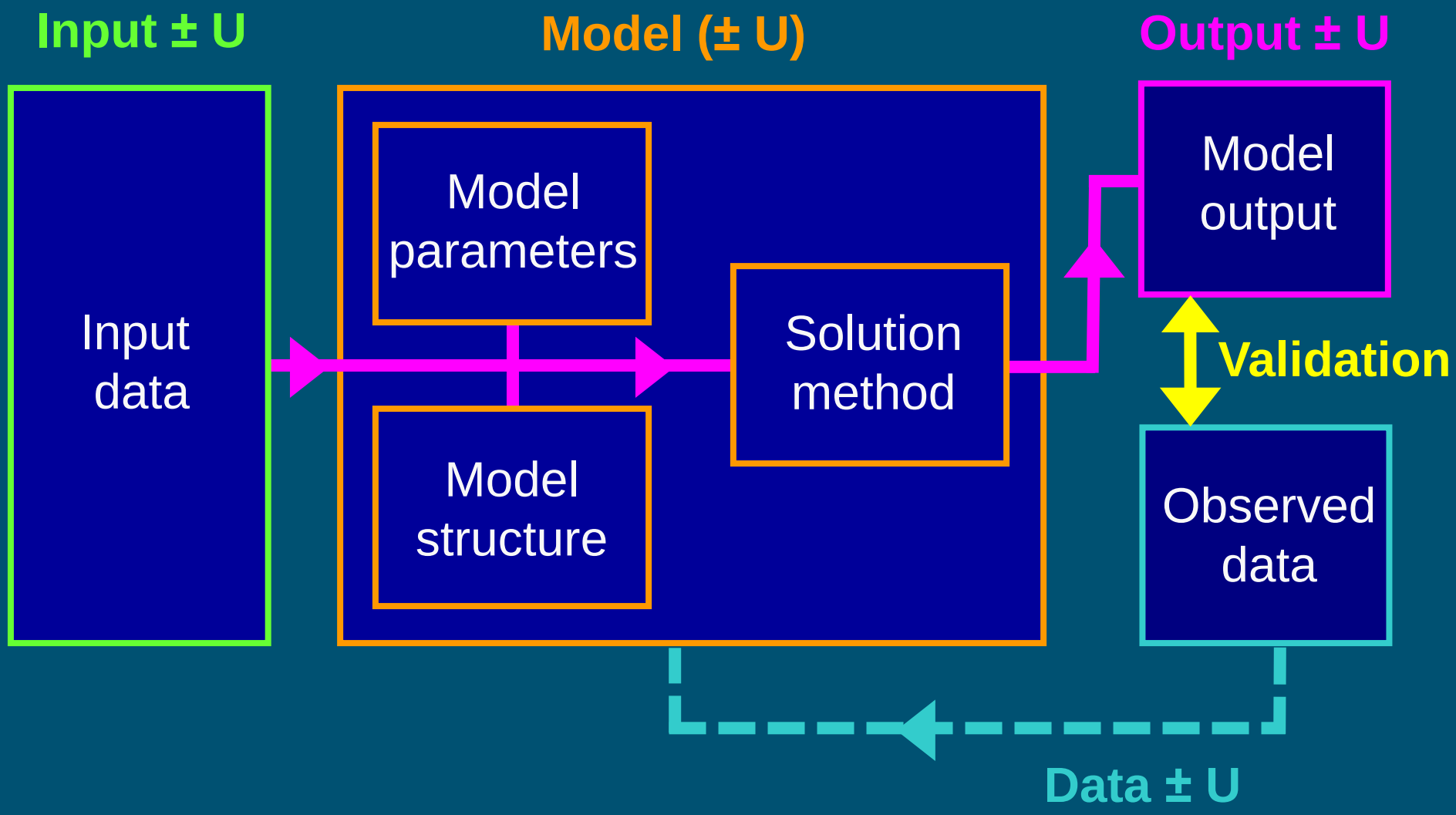
Post-graduate course

Uncertainty Analysis and Statistical Validation of Sp… …2024)

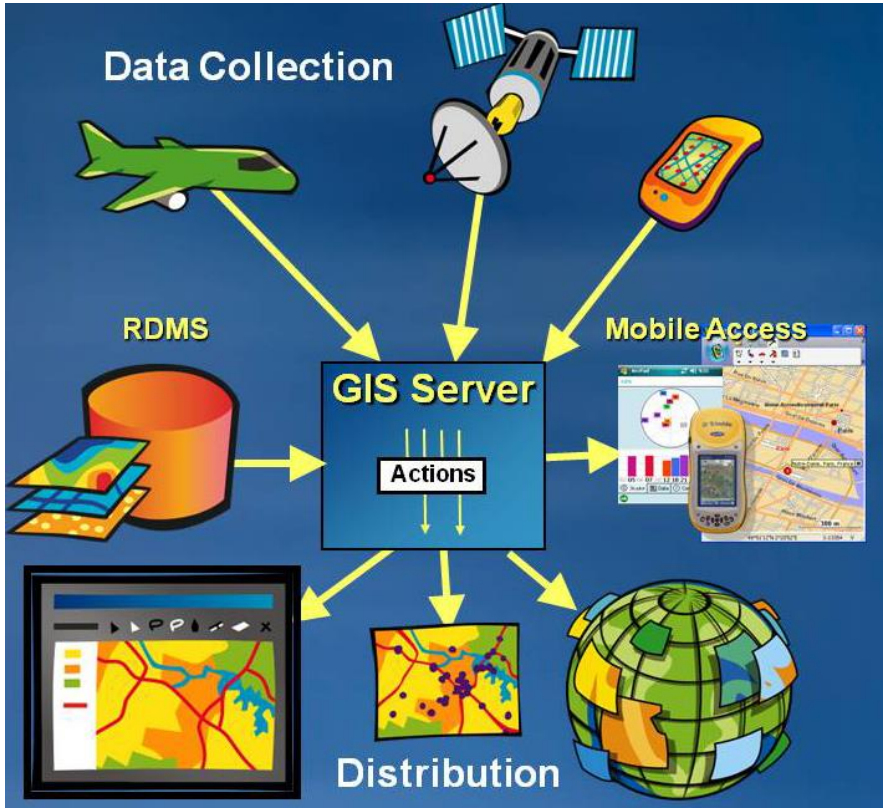
Day	9:00 – 12:00					16:15 – 17:00
Mon. 9 Dec.	Course overview, lecture … probabilistic modelli… Taylor…				Feedback computer practical	Continue practical, own data or advice
Tue. 10 Dec.			…cal Monte Carlo method: …vegetation indices		Feedback computer practical	Continue practical, own data or advice
Wed. 11 Dec.	Le… mod…	…nch break	Computer practical Bayesian calibration and propagation of model parameter uncertainty		Feedback computer practical	Continue practical, own data or advice
Thu. 12 Dec.	Lect… …rcises statistical validation… …cross-validation of spatial model outputs (maps), including sampling for validation, spatial cross-validation and reliability plots	Lunch break	Computer practical statistical validation and cross-validation of spatial model outputs		Feedback computer practical	Continue practical, own data or advice
Fri. 13 Dec.	Lecture and computer practical uncertainty of spatial averages and totals	Lunch break	Finish computer practical and feedback	Course evaluation	Uncertainty game	

Not to forget: Monday at 18:30 dinner in H41!

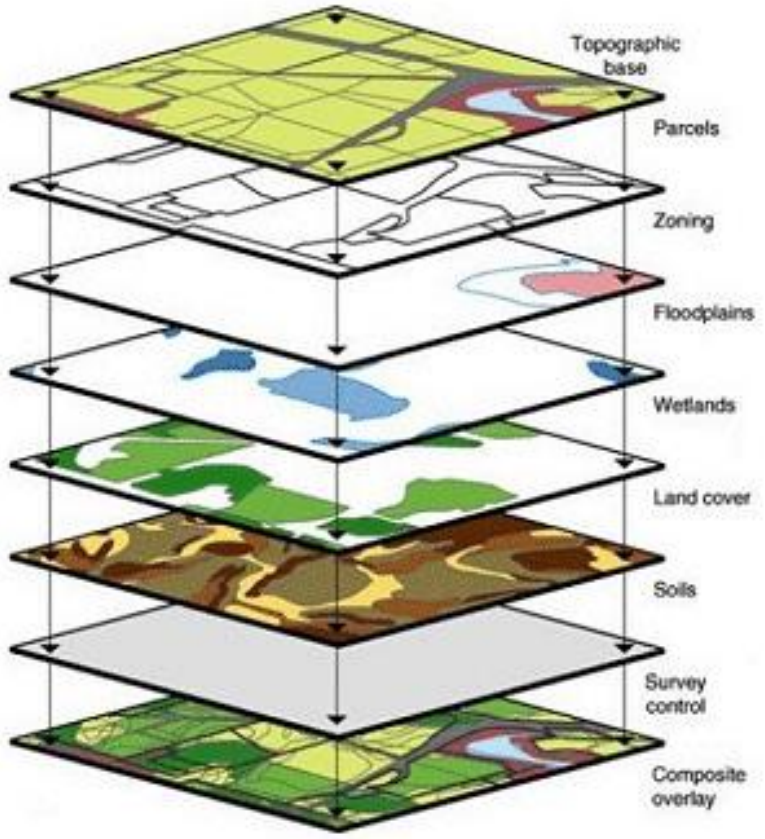
Uncertainty propagation and model validation overview



GIS = software tool for storage, analysis and presentation of geographical date



Analysis includes deriving new maps from existing maps



input map A_1

input map A_2

⋮

input map A_m

↓

output map U

output map = g(input maps)

$$U = g(A_1, A_2, \ldots, A_m)$$

For example:

- slope angle = g(elevation)
- erosion risk = g(landuse, slope, soil type)
- soil acidification = g(deposition, soil physical and chemical characteristics)
- crop yield = g(soil properties, water availability, fertilizer rate)
- evapotranspiration = g(temperature, tree type, tree density, soil moisture)

From Burrough, 1986

Quote from more than 30 years ago but still very relevant: **"The quality of GIS products is often judged by the visual appearance of the end-product ..... uncertainties and errors are intrinsic to spatial data and need to be addressed properly, not swept away under the carpet of fancy graphics displays"**

Fact of life: maps stored in the GIS database are rarely if ever error-free

Causes: generalisation, digitisation, measurement, classification and interpolation errors

Consequence: errors will propagate through GIS operations and spatial models

Key research question: given the errors in the inputs to the GIS operation, how large are the errors in the output?

Error propagation analysis involves three steps:

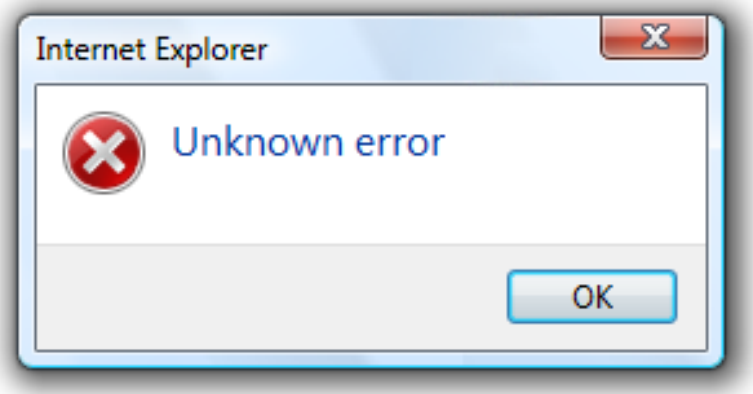
1. DEFINITION of a (statistical) error model for spatial objects and attributes

2. IDENTIFICATION of the error model (estimate its parameters)

3. Perform the actual ERROR PROPAGATION ANALYSIS

Before addressing these three steps, let us first take a step back and discuss:

- What is error?
- What is uncertainty?
- Is there a difference between error and uncertainty? If yes, what is it?
- How can we represent uncertainty statistically?



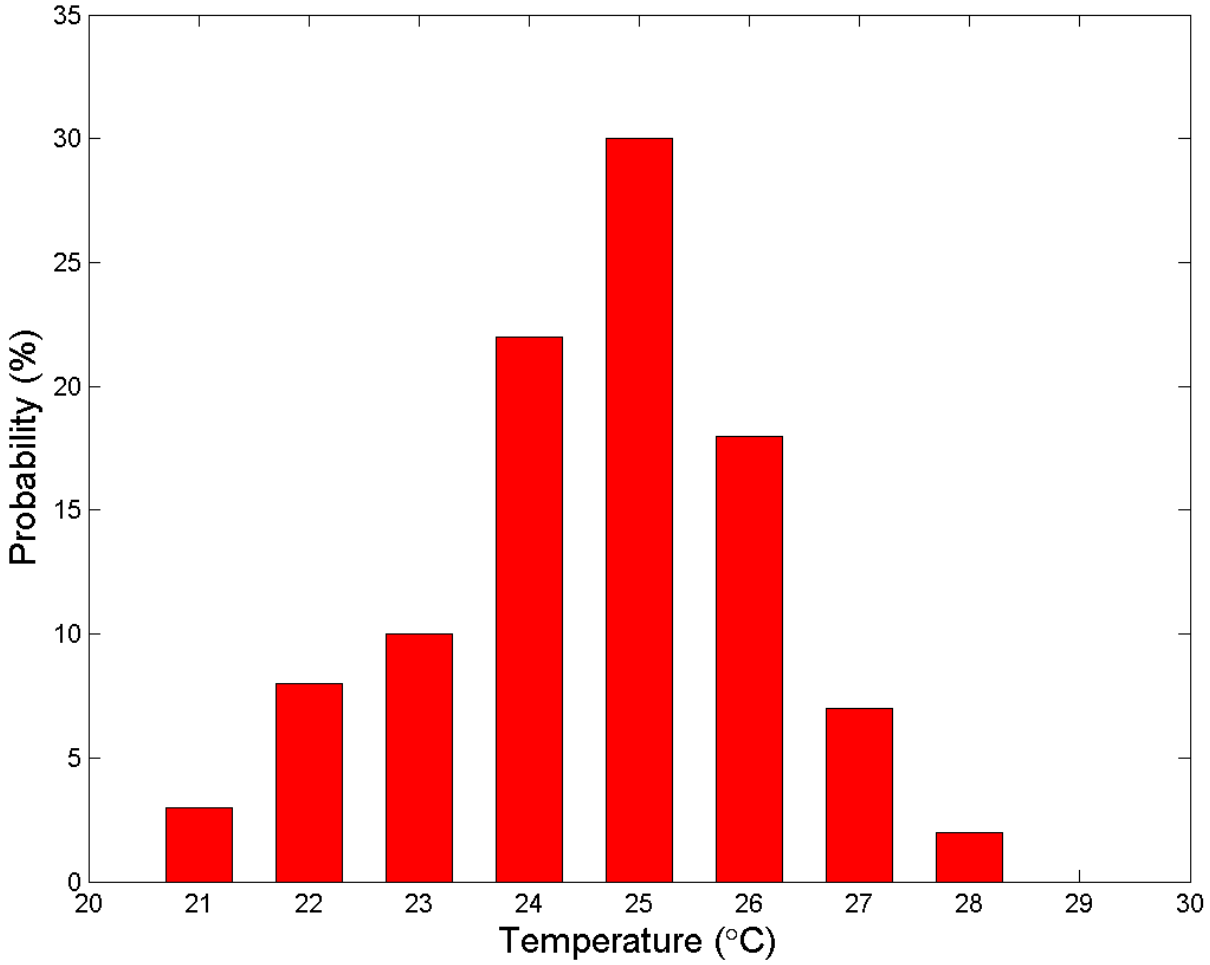
What is error?

- Error is the difference between reality and our representation of reality (assuming reality exists and is clearly defined)
- Example: population size of the European Union. We may estimate it as 442 million. Perhaps in reality it is 449,632,483; hence the error is 7,632,483
- Error is usually not known because reality is not known. If we knew the error, we would simply eliminate it!
- Definition of error needs slight modification for categorical variables. For example: what is the dominant car brand in Europe? My guess would be Toyota, but perhaps it is Volkswagen. We cannot subtract Toyota from Volkswagen!

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 itself and this makes us uncertain (about the reality)
- Uncertainty is often subjective; one person can be more uncertain than another (but can you name an example where everybody will be equally uncertain?)
- 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: maximum temperature Wageningen 1 June 2030

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



Exercise 1

- Work in groups of two or three persons
- Discuss in your group: how old is Gerard?
- Make an estimate but also quantify the uncertainty of your estimate
- Let us take 5-10 minutes for this exercise

(never too old to learn something new) The more information the better: adding information can only reduce uncertainty?

- Not true!
- Example: lost keys

Mathematical explanation (you may forget):

It is indeed true that:

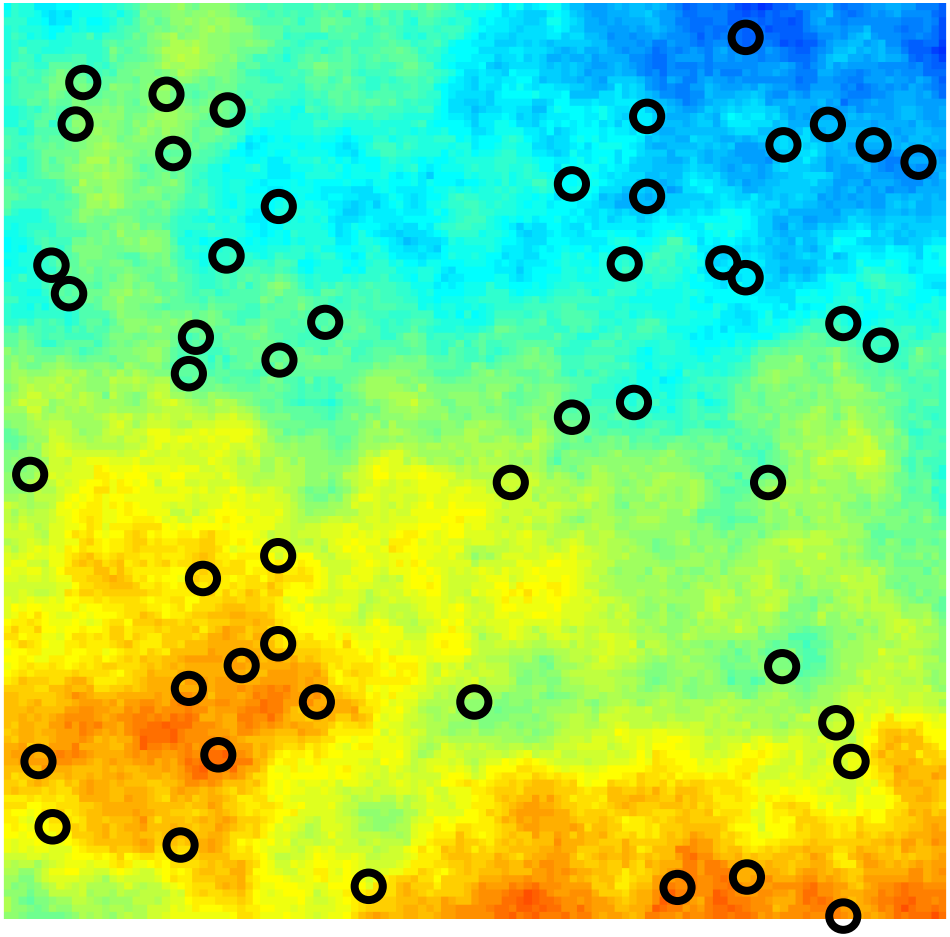
$$Var(Y) = E_X[Var(Y|X)] + Var_X(E[Y|X])$$
$$\Rightarrow Var(Y) \geq E_X[Var(Y|X)]$$

But there might be outcomes x of X for which:

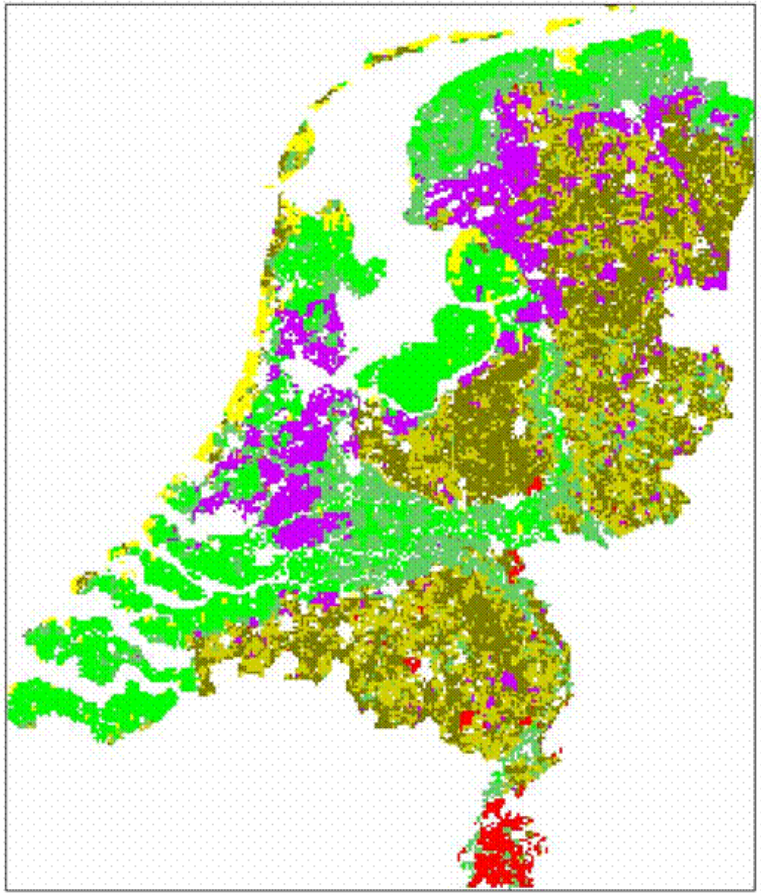
$$Var(Y) < Var(Y|X = x)$$

Once we have a pdf, we can sample from it.

Possible realities of an uncertain spatial attribute:

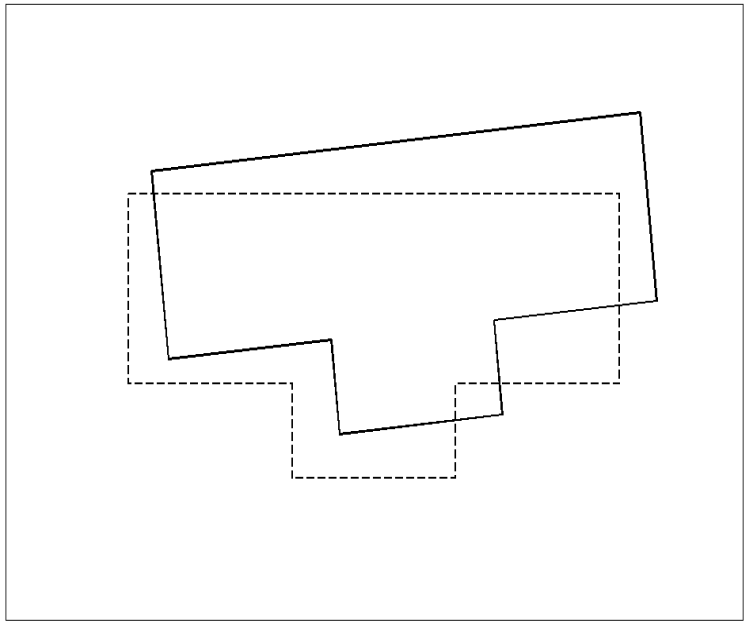


Example of an uncertain categorical spatial variable: possible realities of soil type in the Netherlands

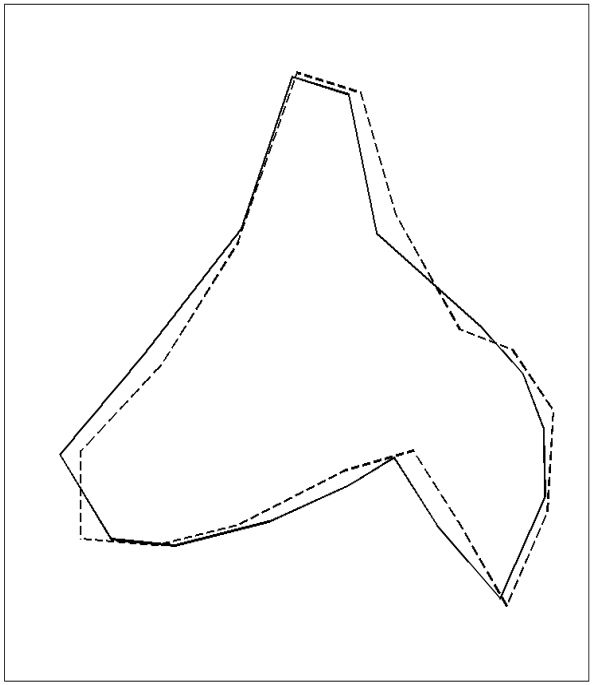


Possible realities of spatial objects that have positional uncertainty

Rigid object



Deformable object



Error and uncertainty both propagate through spatial models

- But because we do not know the error we cannot calculate how it propagates
- We do know the uncertainty (that is to say: we have characterised it by a probability distribution), so we can calculate how this propagates
- From here on we therefore use 'uncertainty propagation' instead of 'error propagation'

Uncertainty propagation analysis involves three steps:

1. DEFINITION of a (statistical) uncertainty model for spatial objects and attributes

2. IDENTIFICATION of the uncertainty model (estimate its parameters)

3. Perform the actual UNCERTAINTY PROPAGATION ANALYSIS

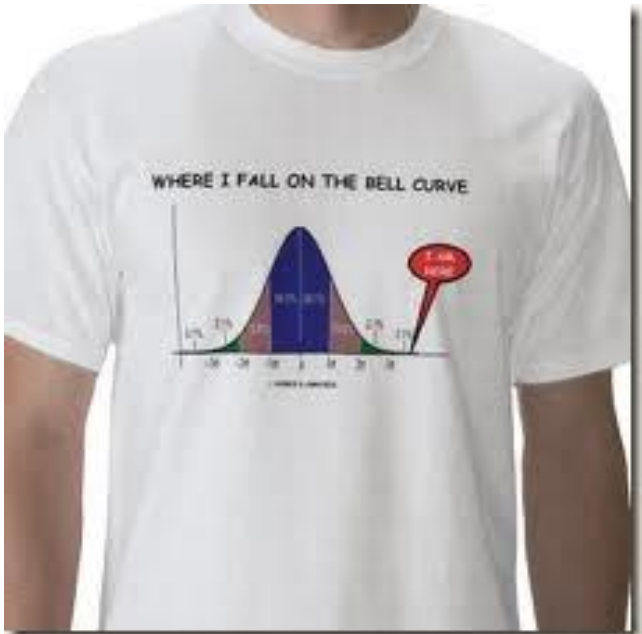
1. DEFINITION of the uncertainty model (concentrate on quantitative attributes)

$$A_i(x) = b_i(x) + V_i(x)$$

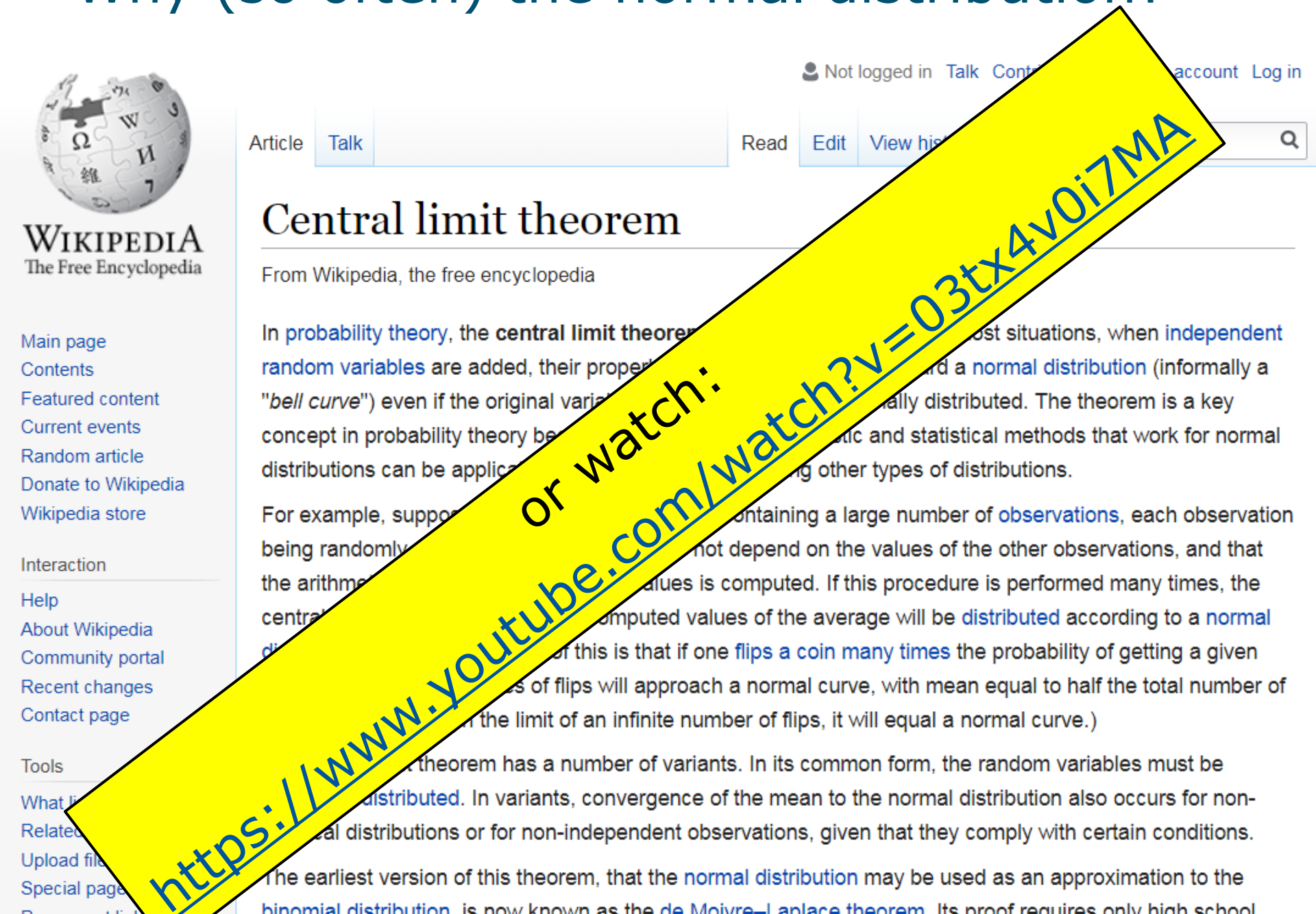
$b_i(x)$ is our (deterministic) representation of the attribute (i.e. the map as stored in the GIS), $V_i(x)$ is the uncertainty about it (typically zero mean, but non-zero variance and spatially correlated)

Error is difference between reality and our representation of reality:

$$V_i(x) = A_i(x) - b_i(x)$$



Why (so often) the normal distribution?



Central Limit Theorem in action...

Statistical model of $V_i(x)$ must include:

- Marginal pdf at each location (both its shape and parameters)
- Spatial correlation (correlogram or semivariogram, see geostatistics course)
- Temporal correlation (for dynamic variables)
- Cross-correlation with other uncertain inputs
- It boils down to the full joint pdf, see Heuvelink et al. (2007) in literature folder:

An uncertain continuous numerical variable, V, that varies in space and/or time is characterized by its (cumulative) jpdf:

$$F_V(v_1, s_1 \ldots, v_n, s_n) = P(V(s_1) \leq v_1, \ \ldots, \ V(s_n) \leq v_n) \qquad (6)$$

where the s_i are coordinates (i.e. s_i may comprise x_i, y_i, z_i, and t_i), and n may assume any integer value. F_V must be known for each and every combination of the v_i and s_i. The corresponding jpdf for a discrete numerical or categorical variable is:

$$F_V(v_1, s_1, \ \ldots, \ v_n, s_n) = P(V(s_1) = v_1, \ \ldots, \ V(s_n) = v_n) \qquad (7)$$

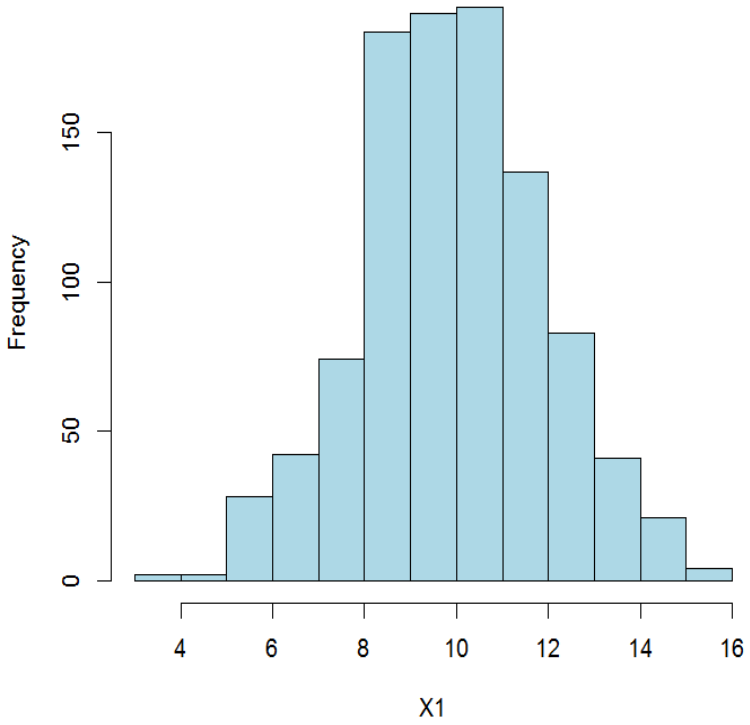
where v_i denotes integers or categories, respectively.

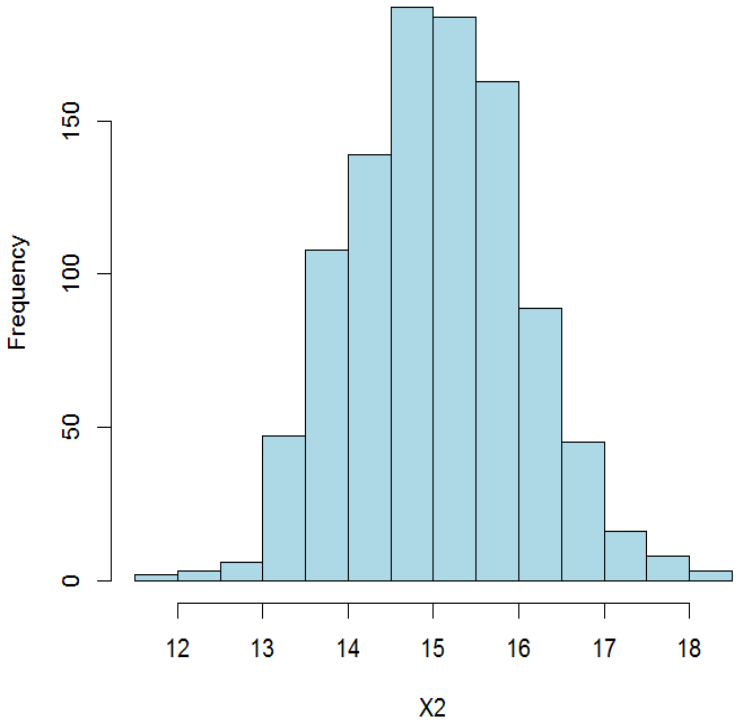
Example of the effect of cross-correlation

Consider two variables X_1 and X_2, both uncertain:

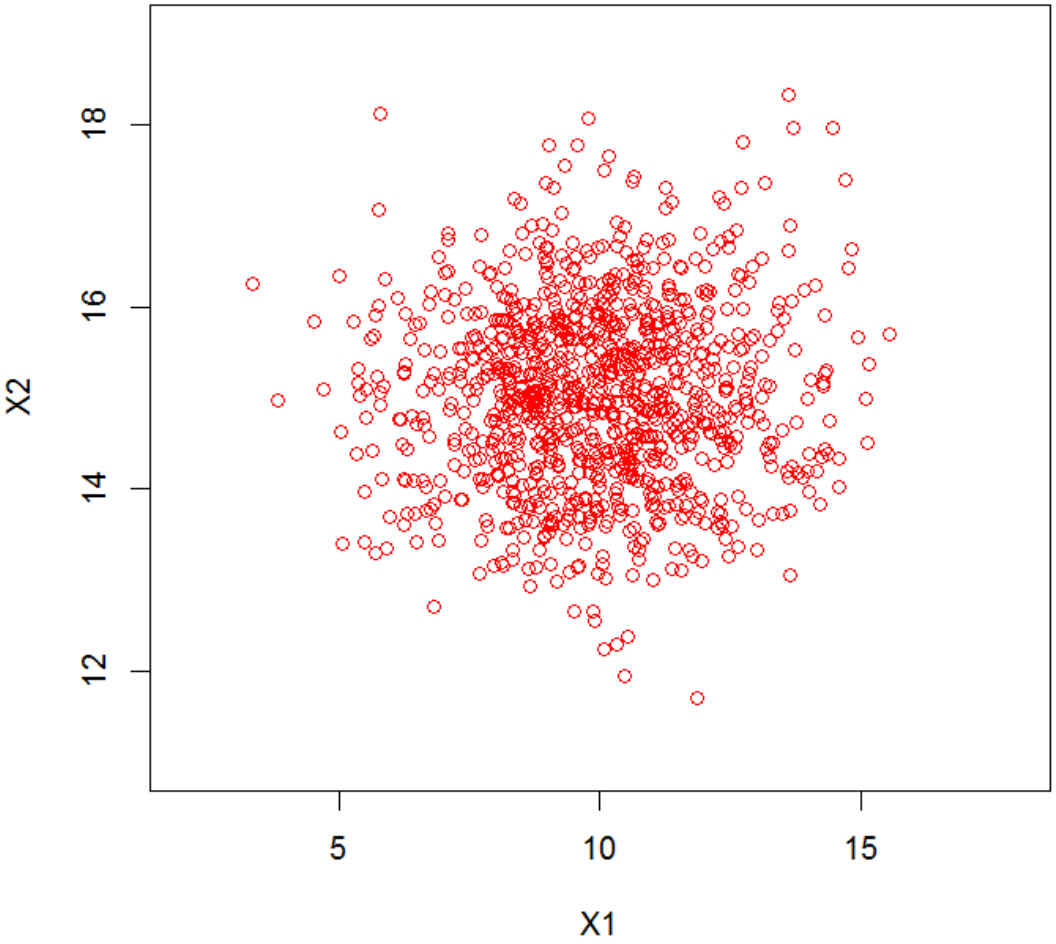
$X_1 = 10 \pm 2$

$X_2 = 15 \pm 1$

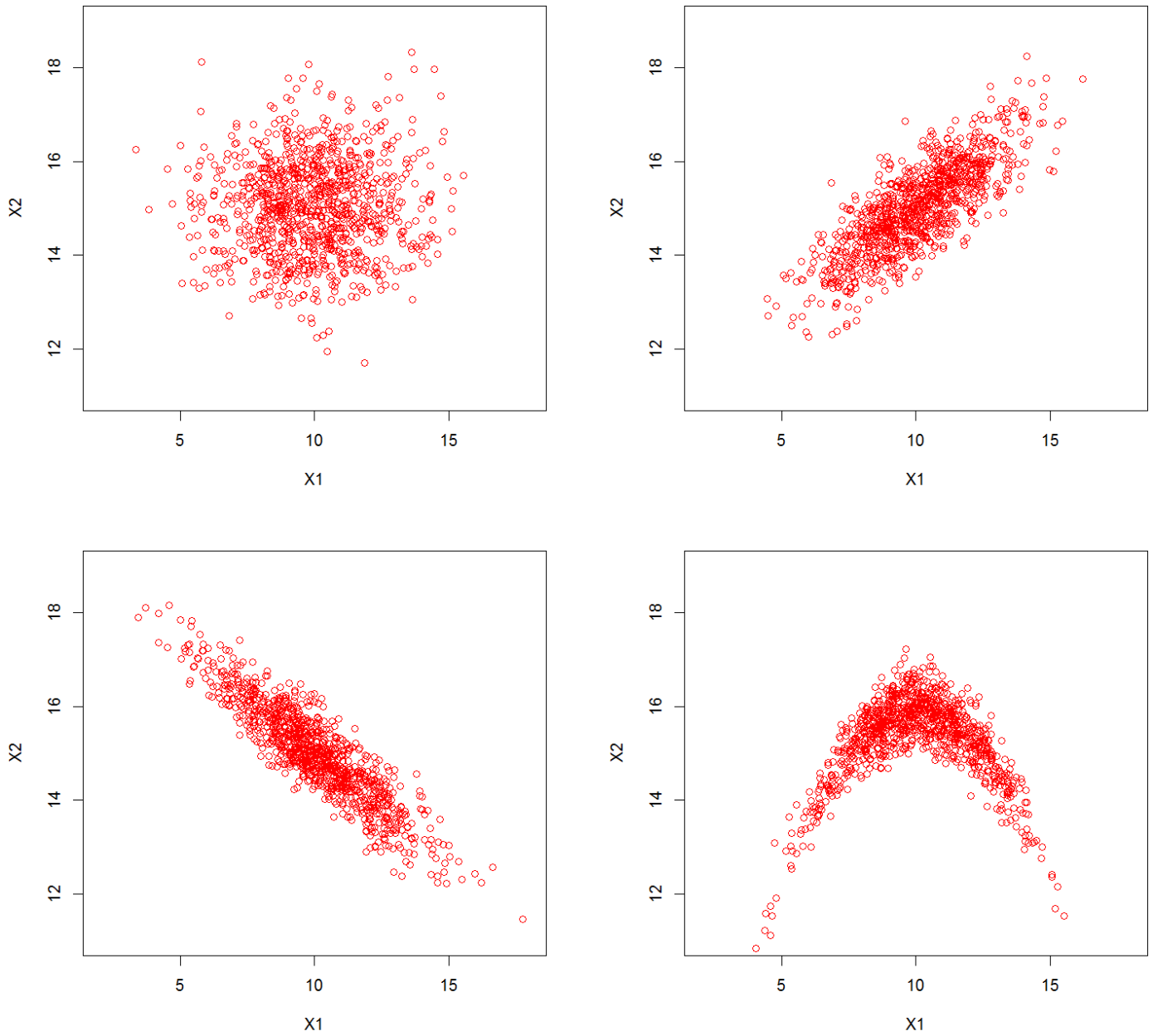




We can plot the sampled values from X_1 and X_2 against eachother

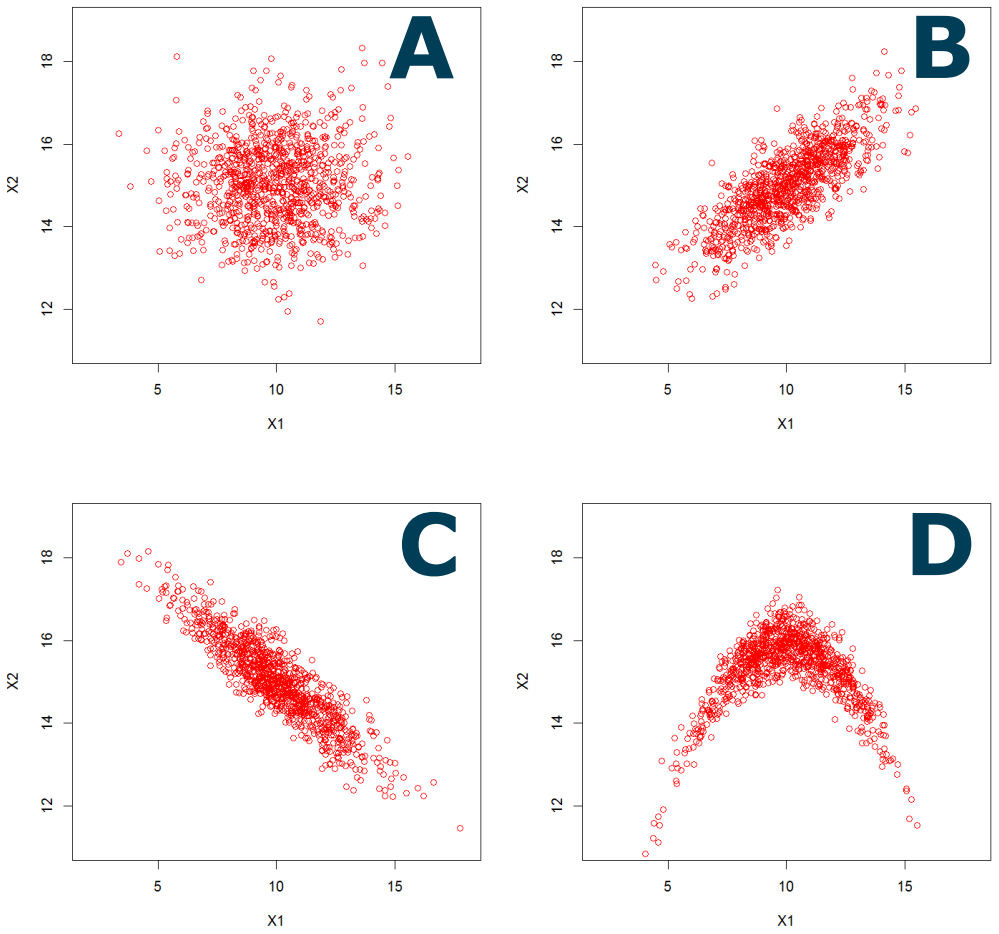


But uncertainty may be correlated



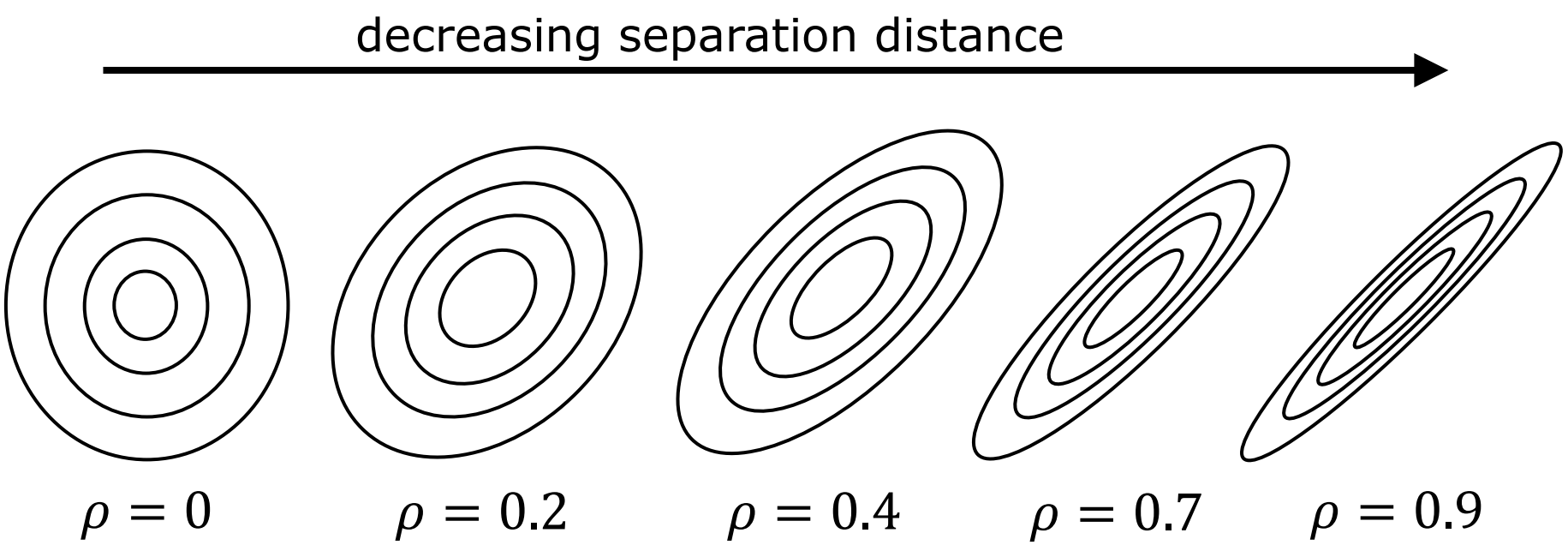
Exercise 2: Which statement is wrong?

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.



As before, discuss in your group, let's take 5 minutes

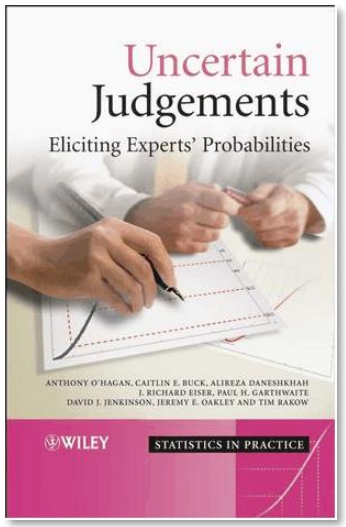
Spatial correlation is common: as locations get closer values become more similar, that is their correlation increases and probability density contour lines become narrower. The first law of Geography also applies to errors!



2. IDENTIFICATION of the uncertainty model

- Measurement error from instrument and lab specifications, or by taking replicates (always let laboratories analyse duplicates, without telling them!)
- Sampling error using sampling theory from statistics (e.g. standard error of the mean, confidence intervals)
- Use of ground truth verification data (e.g. control points)
- Interpolation error using geostatistics (kriging)
- Errors in transfer functions such as regression: R-square
- Classification error using multivariate statistics (e.g. maximum likelihood classification remote sensing imagery)
- Expert judgement or expert elicitation (last resort?)

In practice, this step is often the most difficult step of the entire uncertainty propagation analysis, so let us quickly move on to step 3 ☺

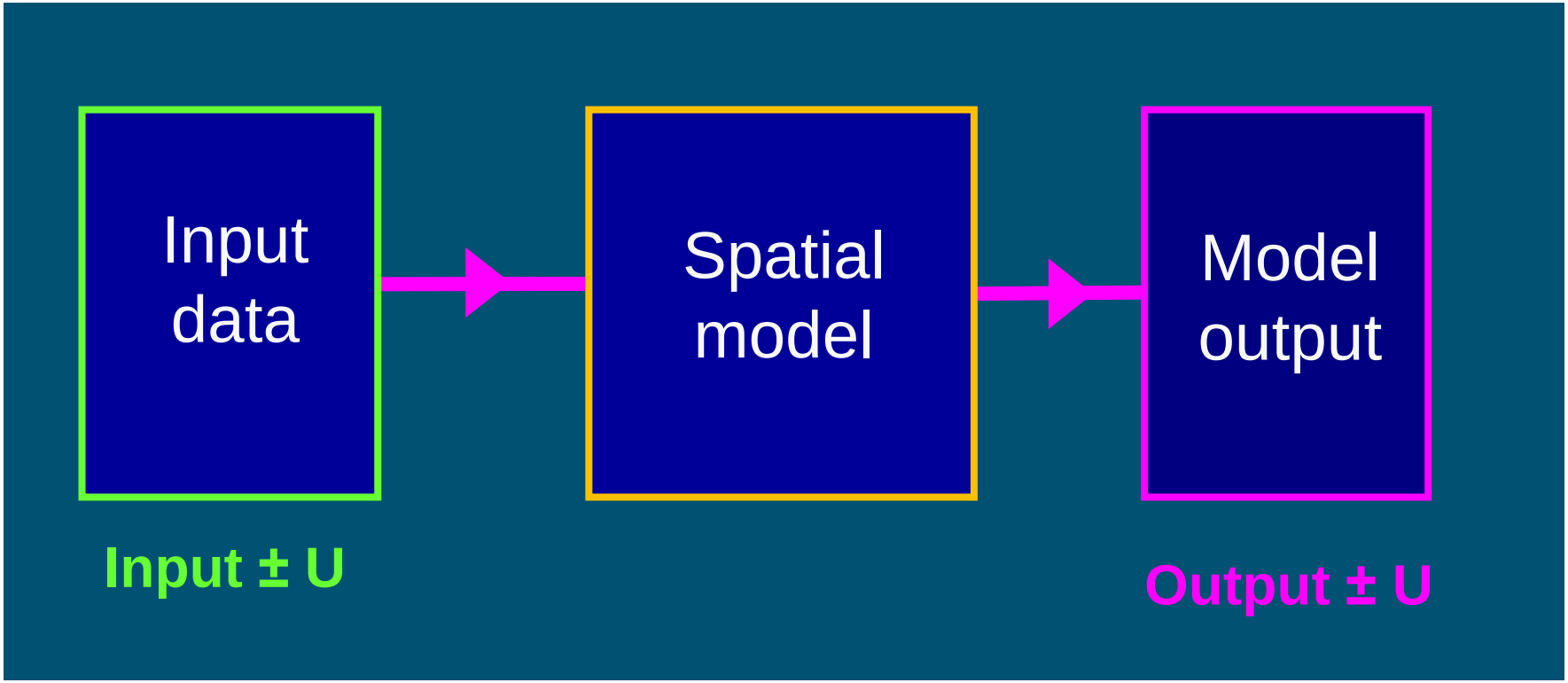


3. Actual UNCERTAINTY PROPAGATION ANALYSIS

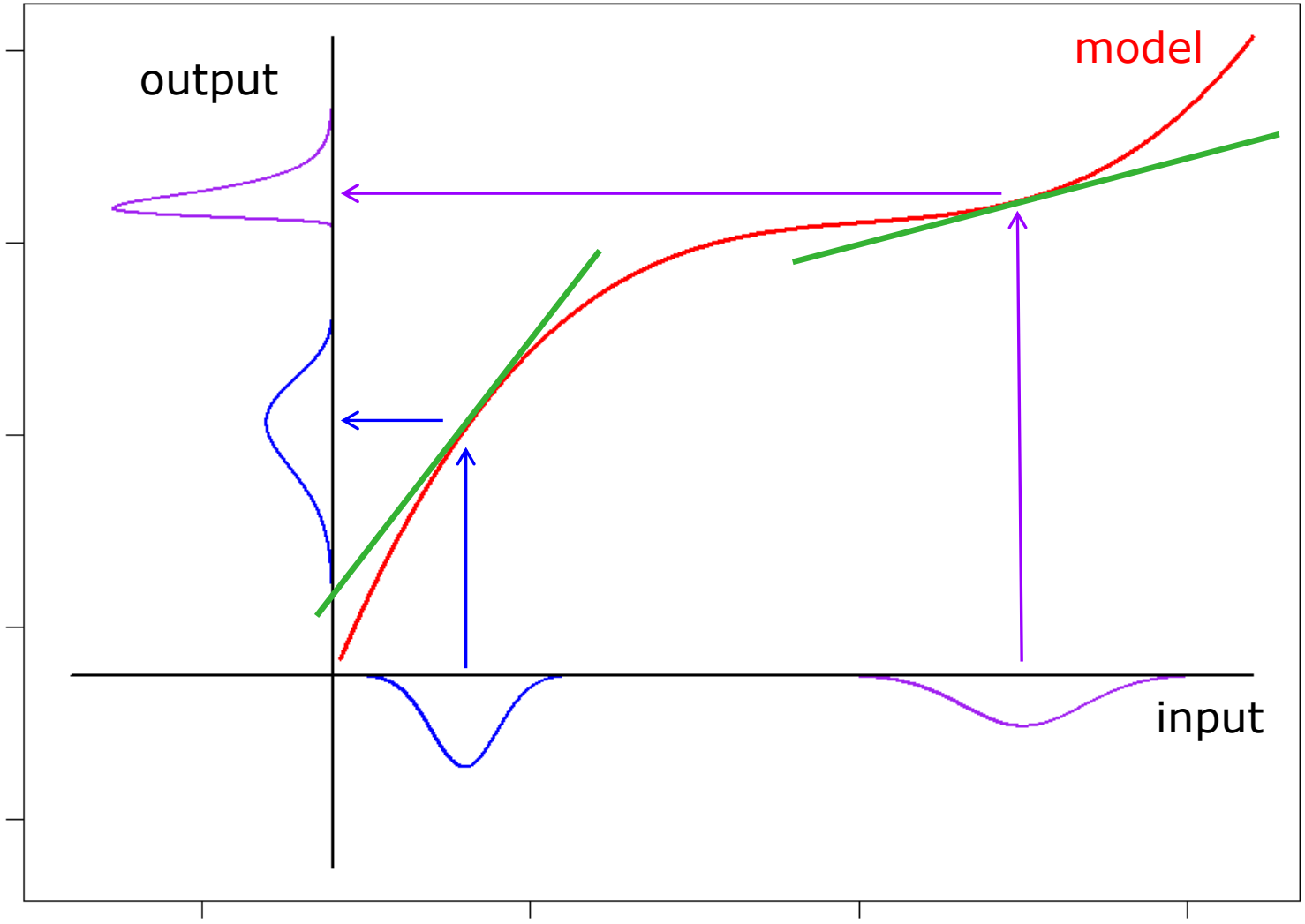
We discuss two methods:

1. Taylor series approximation (today)
2. Monte Carlo simulation (tomorrow)

Today and tomorrow we concentrate on propagation of input uncertainty only



Taylor series approximation, graphical illustration one-dimensional case (m=1)



Taylor series approximation

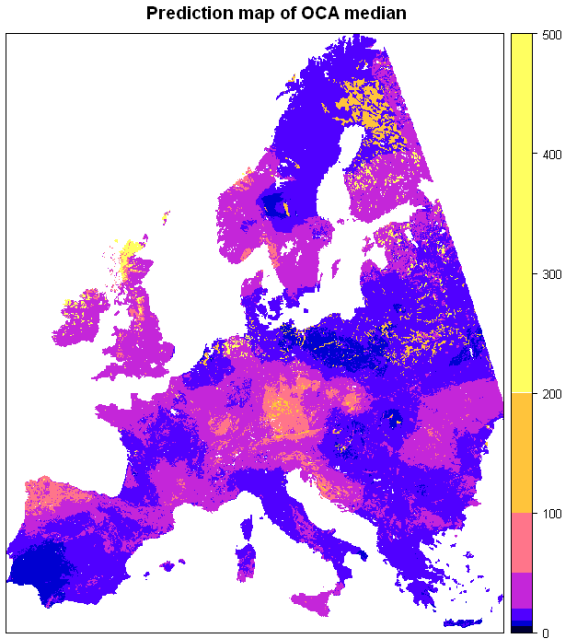
By linearising the GIS operation g we get:

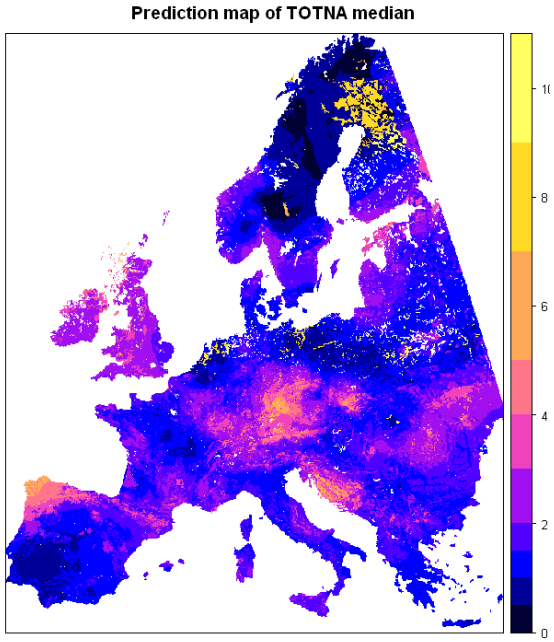
$$Var(U) = \sum_{i=1}^{m} Var(A_i) \cdot \left(\frac{\partial g}{\partial A_i}\right)^2$$

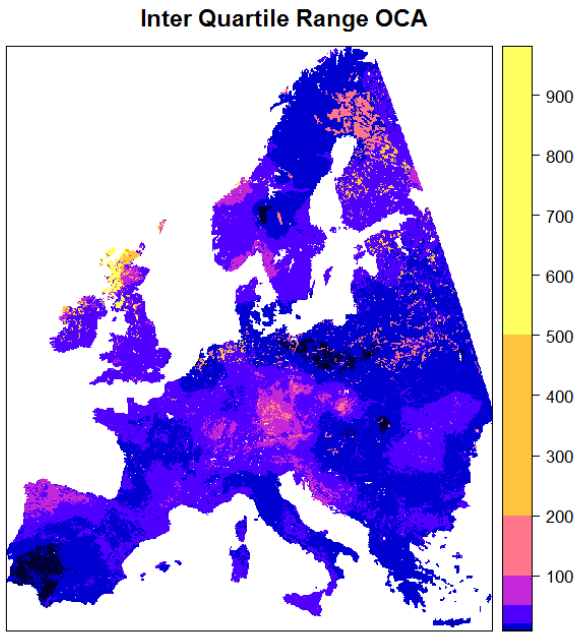
magnitude of input error matters

but also sensitivity of model to input

Example: calculating the topsoil CN ratio







- Maps of topsoil organic Carbon and Nitrogen for Europe created using geostatistical interpolation (kriging)
- This yields prediction maps but also associated uncertainties
- Let, for some location $b_C = \mu_C = 120\ g/kg$, $b_N = \mu_N = 10\ g/kg$, $sd(C) = \sigma_C = 40\ g/kg$ and $sd(N) = \sigma_N = 2\ g/kg$
- What is the predicted CN ratio, how large its uncertainty (standard deviation)?

Incorporating correlation between uncertain inputs

$$Var(U) = \sum_{i=1}^{m} Var(A_i) \cdot \left(\frac{\partial g}{\partial A_i}\right)^2 + 2 \sum_{i=1}^{m} \sum_{j=i+1}^{m} Cov(A_i, A_j) \cdot \left(\frac{\partial g}{\partial A_i}\right) \cdot \left(\frac{\partial g}{\partial A_j}\right)$$

- Covariance terms needed too, recall from statistics classes that covariance is correlation multiplied with the two standard deviations
- Will positive correlation between uncertainties in C and N lead to increase, decrease or same uncertainty in CN ratio?
- Think first and then check by calculating uncertainty in CN for the case $\rho_{CN} = 0.8$

Exercise 3

- Work in groups as before
- Open file ‘PERC uncertainty Monday exercise 3.pdf’ from MS-Teams
- Address al questions, perhaps partly work on your own but also discuss regularly within your group
- Let us take 30 minutes for this exercise (if we run out of time we will consider it ‘homework’, with feedback tomorrow)