Aalto University
School of Engineering

Uncertainty and error propagation

Jaakko Madetoja
31.1.2019
Advanced Spatial Analytics
Slides on quality and uncertainty by Kirsi Virrantaus

Contents

- **Concepts in uncertainty; quality**
- **Error**
- **Error propagation**
  - Analytic methods
  - Stochastic methods
    - *Example*

Learning goals

After this lecture, you are able to

- **explain the difference between internal and external quality**
- **describe what analytic and stochastic error propagation mean**
- **run error propagation with your own data (no tools available, coding required!)**

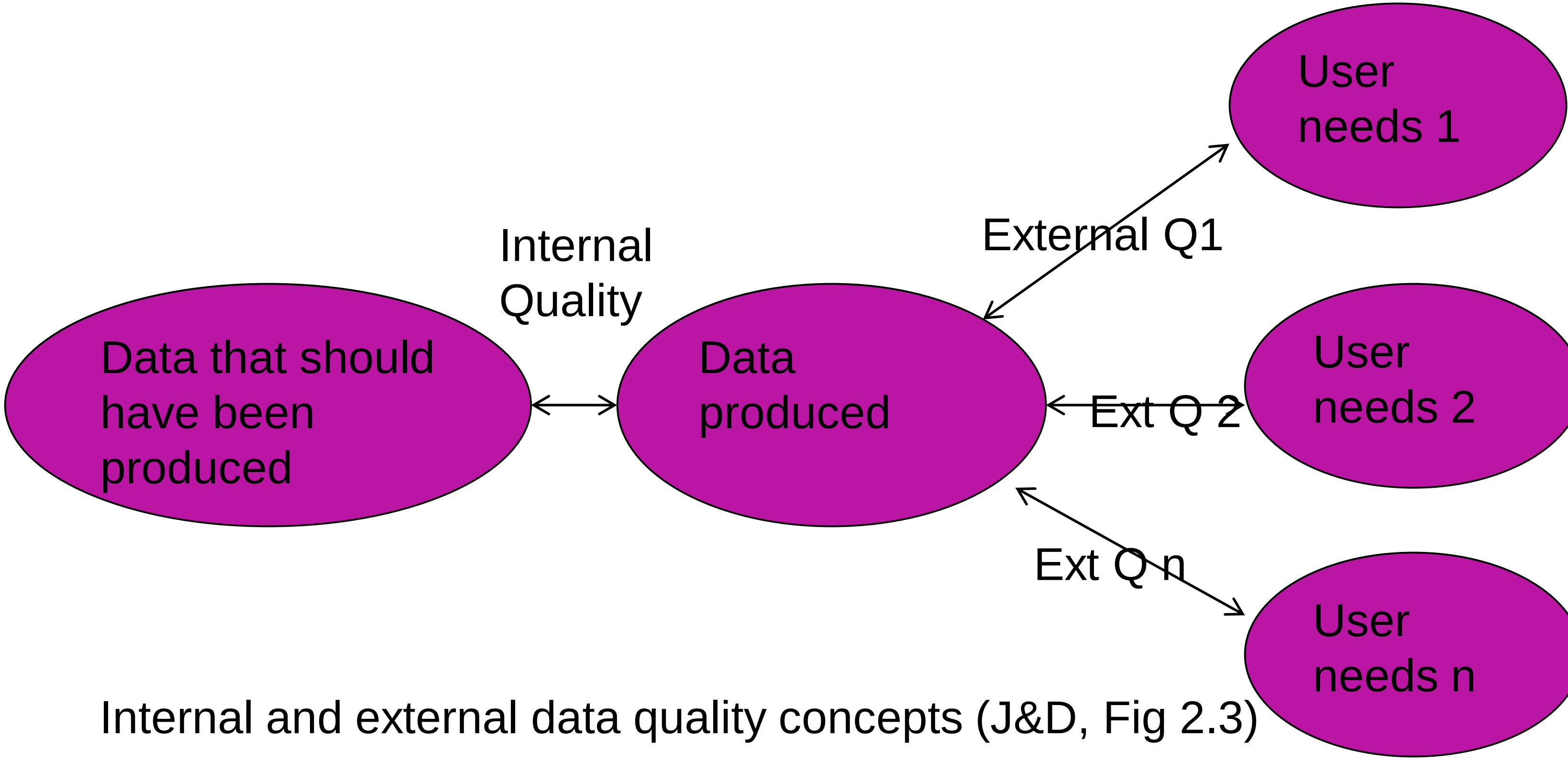
Reading materials

- **Madetoja, J. 2018. Error propagation in geographically weighted regression. Chapter 3.2-3.4. https://aaltodoc.aalto.fi/handle/123456789/29575**
- **Devillers,R., Jeansoulin,R. 2006. Fundamentals of Spatial Data Quality. Chapter 2 & 3.**

Core concepts in uncertainty

Quality

- Internal quality – sisäinen laatu (producer)
  - *The relationship between the data that **has been produced** and the data that **should have been produced**; the error between them*
- External quality – ulkoinen laatu (user)
  - *The relationship between the data **that has been produced** and various **user needs**; there can be various users and varying user needs*



Internal and external data quality concepts (J&D, Fig 2.3)

Criteria for internal quality (from ISO standard)

Completeness (täydellisyys)

- presence and absence of features, attributes and relationships

Logical consistency (looginen eheys)

- degree of adherence to logical rules of data structure, attributes or relationships

Positional accuracy (sijaintitarkkuus)

- Accuracy of the position of the features

Temporal accuracy (ajallinen tarkkuus)

- Accuracy of the temporal attributes, temporal relationships or features

Thematic accuracy (ominaisuustietotarkkuus)

- Accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships

External quality concepts and criteria

- **the same data product can be of different quality to different users**
- **”fitness for use”** (in Juran and American standard NCDCDS), **”fitness for purpose”**
- **The qualitative quality elements**
  - *history (processing history, lineage), usage (for what purpose the data was created), use experiences(for what purposes the data has been used)*
- **Various sets of criteria to analyze external quality**
  - *See in Devillers & Jeansoulin, page 40*

Modeling the crisp and imprecise reality - more concepts

- ***models*** of reality are ***approximations***
  - *no-model is completely 1:1 accurate*
  - *”all models are wrong but some are useful” (Box 76)*
- there is always ***error*** at present
- the umbrella term is ***uncertainty***
  - *covers many different aspects*

- *when the truth is obtainable*
  - *the term used is ”accuracy” and ”error”*
- *when the truth can not be recovered*
  - *the term used is ”imprecision” (epätäsmällisyys)*

Examples

- true values can be found, for
  - *man-made objects like buildings, which have ”crisp” boundaries that can be defined by using points (coordinates)*

- true values can not be found, for
  - *boundaries of different vegetation areas or soil types; no crisp boundaries but ”transition zones” between*
  - *boundaries of some geographical concepts, like ”mountain”; where does it begin (slope?)*

- also there can be a confusion with definitions; one concept can have several definitions
  - *example: ”swamp”, at least three different definitions exist (agriculture, soil science, mapping)*

Error

- **Error = the difference between the real value and the modelled (or measured) value**

- **If the real value was known, we would simply add it to the data**

- **Real value unknown, but we have an idea about its distribution; error as random variable**

Taking error into account

- **Error affects every (?) analysis**

- **Problem: If we do not take error into account, we assume that it is zero.**

(Gilman et al., 2015)

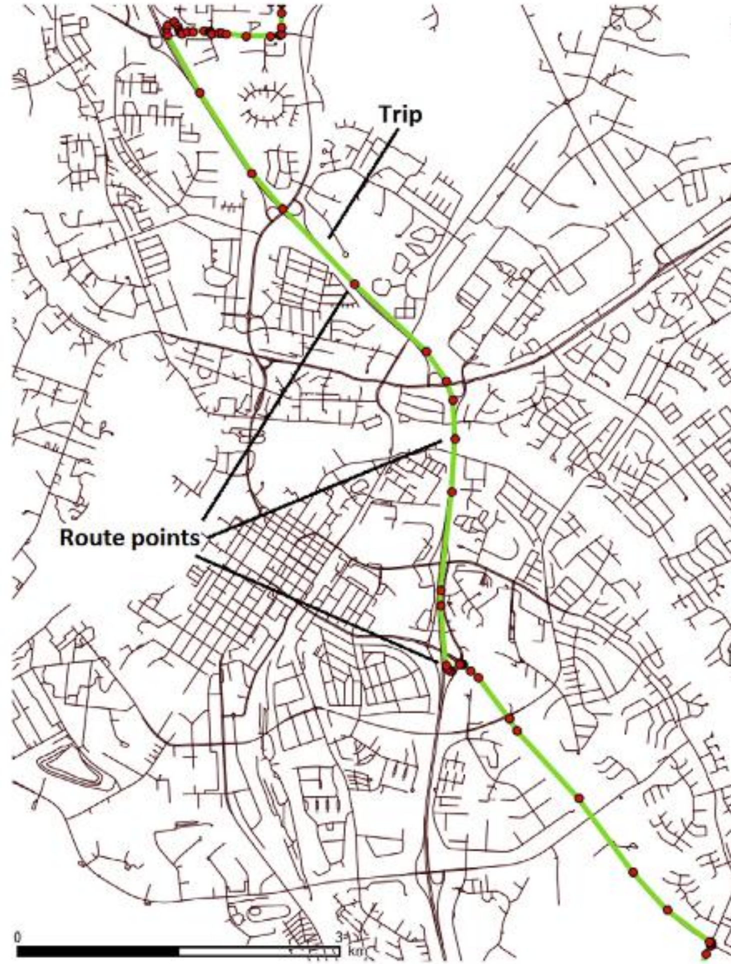
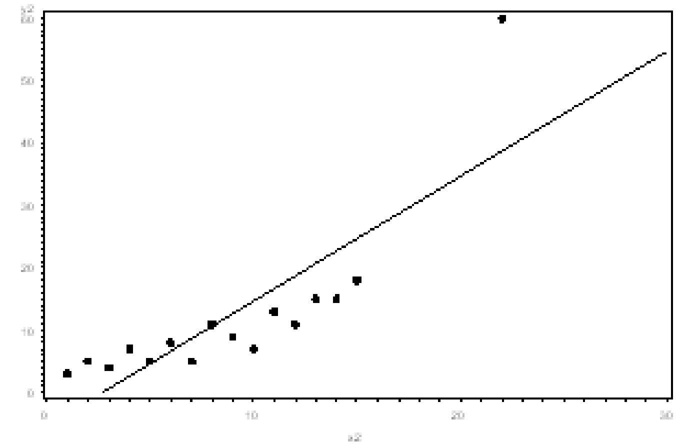


Fig. 6. Example of a trip (green) and its route points (red).[14]

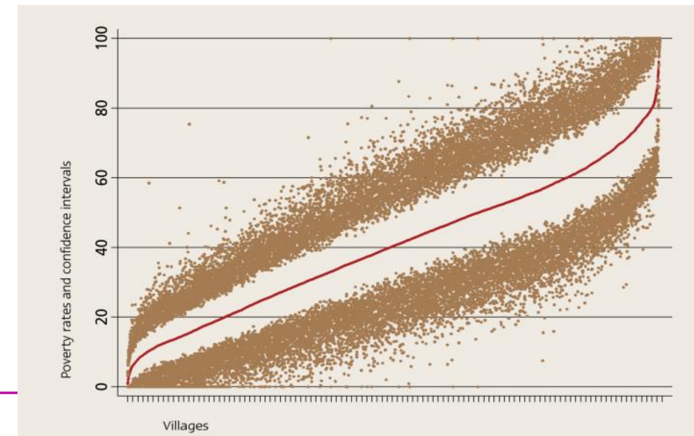
Taking error into account

Common ways to consider error in (spatial) analysis:

- **Preprocessing: Delete clearly erroneous observations**



- **Estimate and report the amount of error in the input data**



”must be used very cautiously”
(Epprecht et al. 2008)

Error propagation (virheen kasautuminen)

- **The effect of errors of input data on output values of an operation or function (compare with sensitivity analysis)**

- **Methods can be divided in two:**
  - Analytic methods
    - *Exact*
    - *Approximate*
  - Stochastic methods

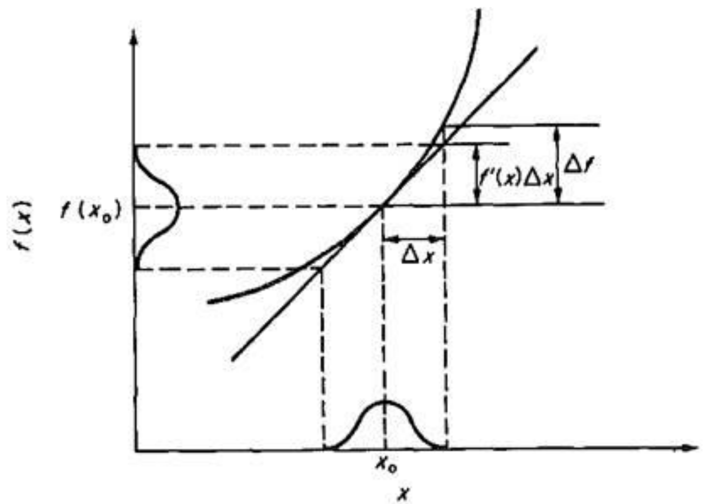
Analytic error propagation

- **Calculate the effect of variation in input data to the result of an analysis (or function) using algebra**
  - For example, variables A and B, and their standard deviations σ_A and σ_B
  - $f = A + B, {\sigma_f}^2 = {\sigma_A}^2 + {\sigma_B}^2$
  - $f = A^B, {\sigma_f}^2 \approx \left[\left(\frac{B}{A}\sigma_A\right)^2 + (\ln(A)\,\sigma_B)^2\right]$

- **Algebra gets really difficult with complex analysis or functions**

Analytic error propagation

- **Solution to complex functions: approximate function around a value with Taylor series and calculate error propagation using a simpler function**
- **Pros:**
  - Simpler formulas
- **Cons:**
  - Original function needs to differentiable
  - Only an approximation, result gets worse away from the value



Stochastic error propagation

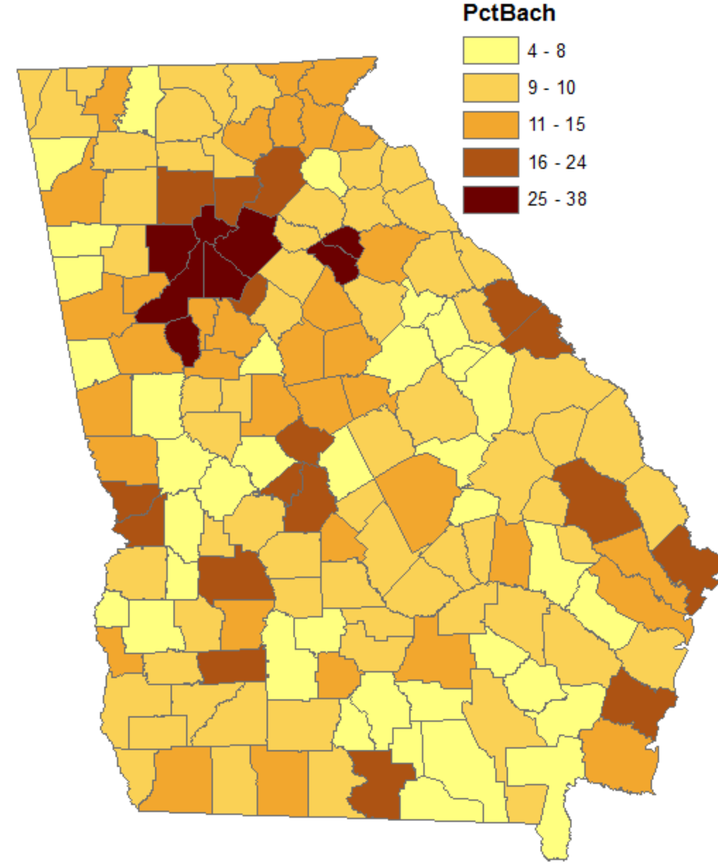
- **Solution to the problems with analytic methods**

- **Monte Carlo simulation:**
  1) Take a realization from the distribution of input values
  2) Calculate the analysis using the realization and store the result
  3) Repeat n times

- **Used often in geoinformatics; can be done with complex processes**

Example

- The GWR tutorial from the exercises: explaining the proportion of educated population using other variables
- How do the results change when error in the input data is taken into account?
- Madetoja, J. 2018. Error propagation in geographically weighted regression



1) Error distribution

- **First step: defining error distributions**

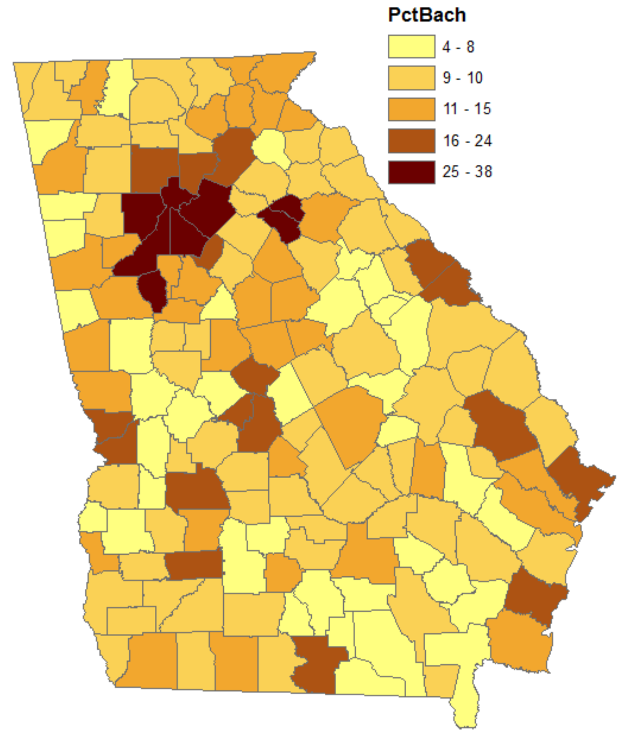
- **Can be done**
  - using algebra
  - using simulations
  - using expert knowledge
  - comparing with more accurate data

- **Metadata relating to accuracy is often insufficient**

Example

- **Positional accuracy:** unknown error for borders of polygons; likely small

- **Attribute accuracy:**
  - Data from 1990 census; information on sampling and response rates available
  - Assuming random sampling and response, standard deviation $\sigma = \sqrt{\frac{1}{n}p(1-p)}$



2) Analysis, example

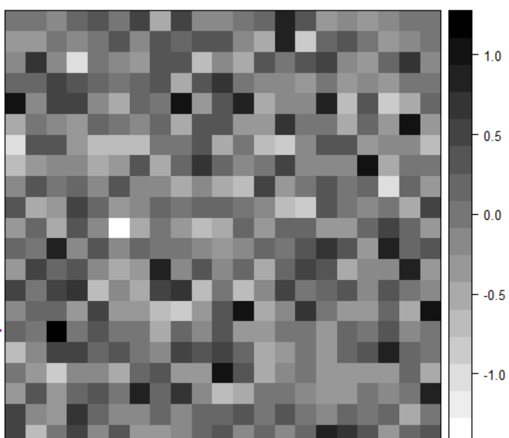
1) **Realizations: take a random value using standard deviation, and add it to the data**
   - Errors assumed to follow normal distribution, but with min. 0 % and max. 100 %

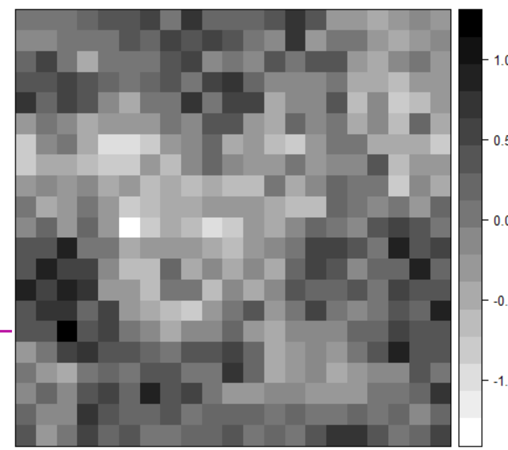
2) **Calculate the entire GWR process (incl. OLS) for the realization**

3) **Save the results; repeat from step 1**

Spatially autocorrelated errors

- **If errors are spatially autocorrelated, utilize SAR-process:**
  $\boldsymbol{X} = \rho \boldsymbol{W} \boldsymbol{X} + \boldsymbol{\varepsilon} \quad => \quad \boldsymbol{X} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{\varepsilon}$
  - $\boldsymbol{\varepsilon}$ is the realizations, $\boldsymbol{W}$ is weights matrix, ρ is autocorrelation parameter
- **The method has originally been developed for raster data, but can be applied to irregular points if $\boldsymbol{W}$ is normalized**

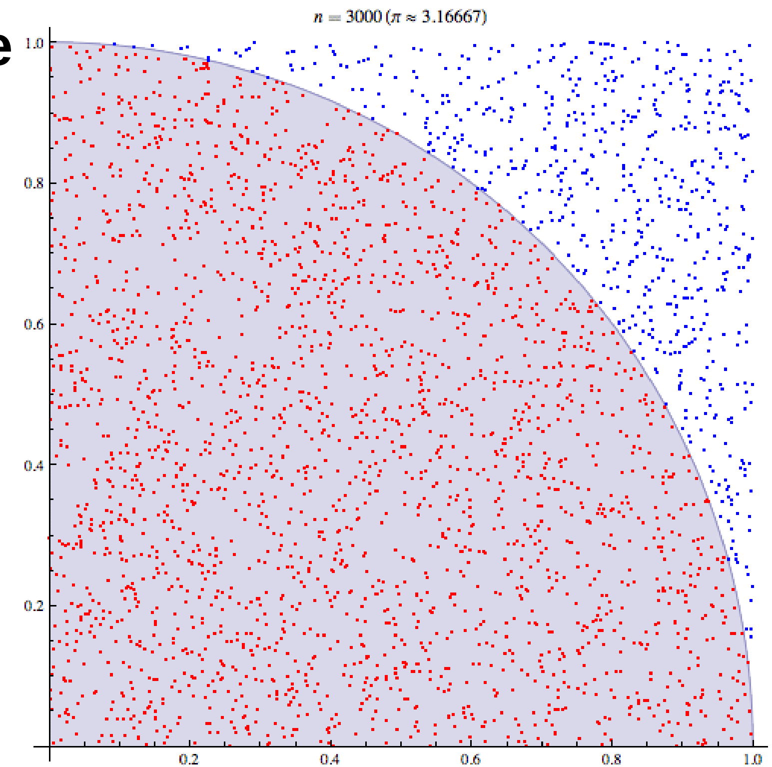




3) Choosing n, accuracy of Monte Carlo simulation

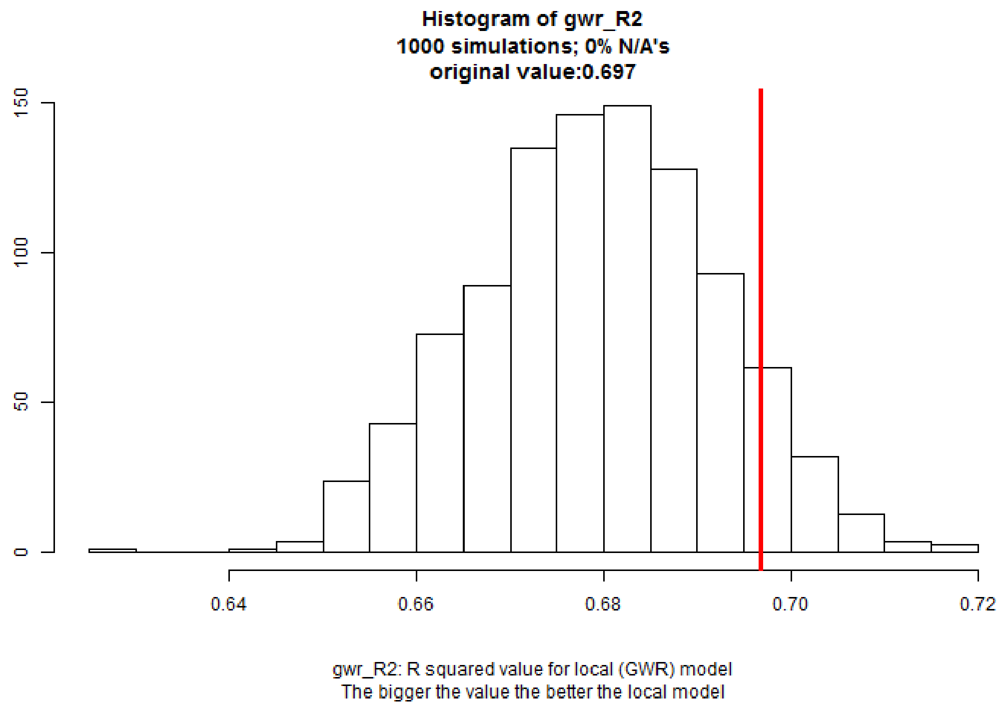
- The bigger the n, the more accurate the simulation, i.e. the closer the calculated uncertainty to the real uncertainty

- The bigger the n, the longer the simulation takes



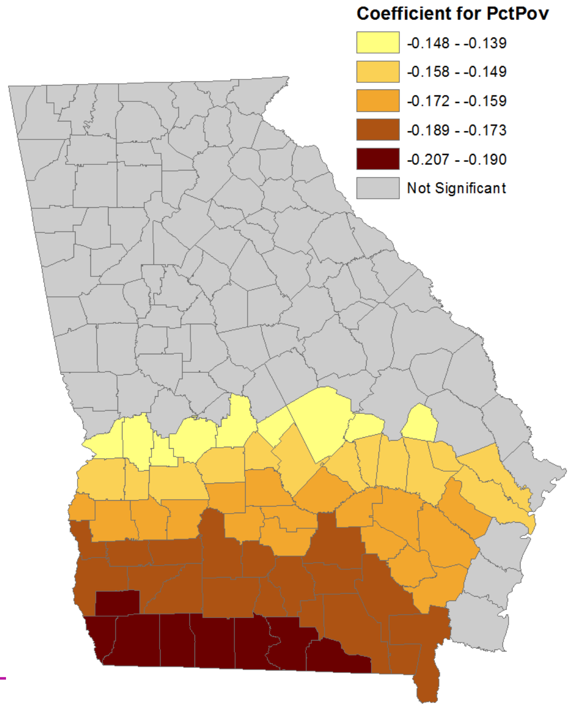
Visualization of the results

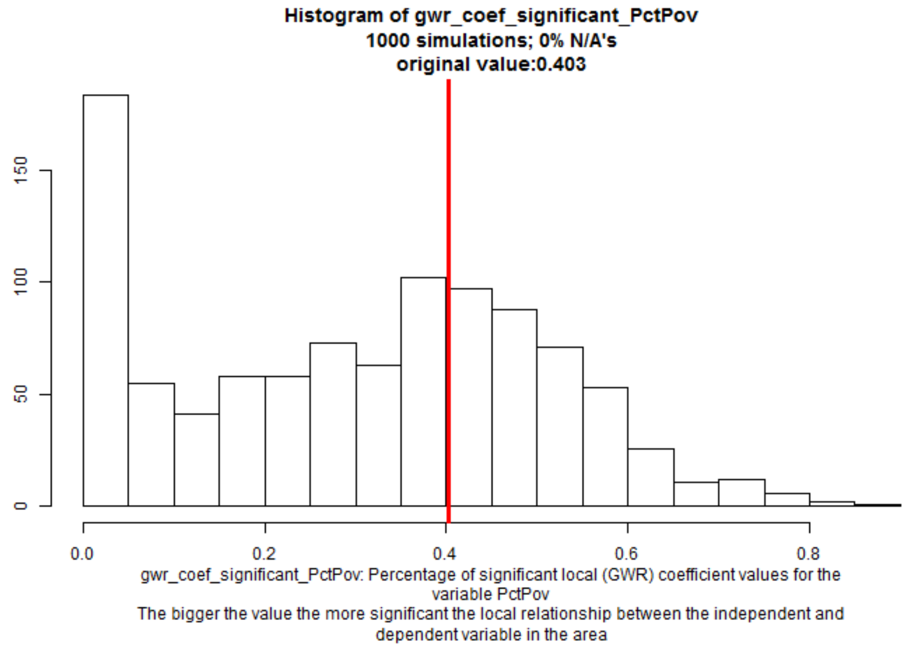
- **When the result of the analysis is only one value, the examination is easy**



Visualization of the results

- **When the result is more complex, more thought is required**

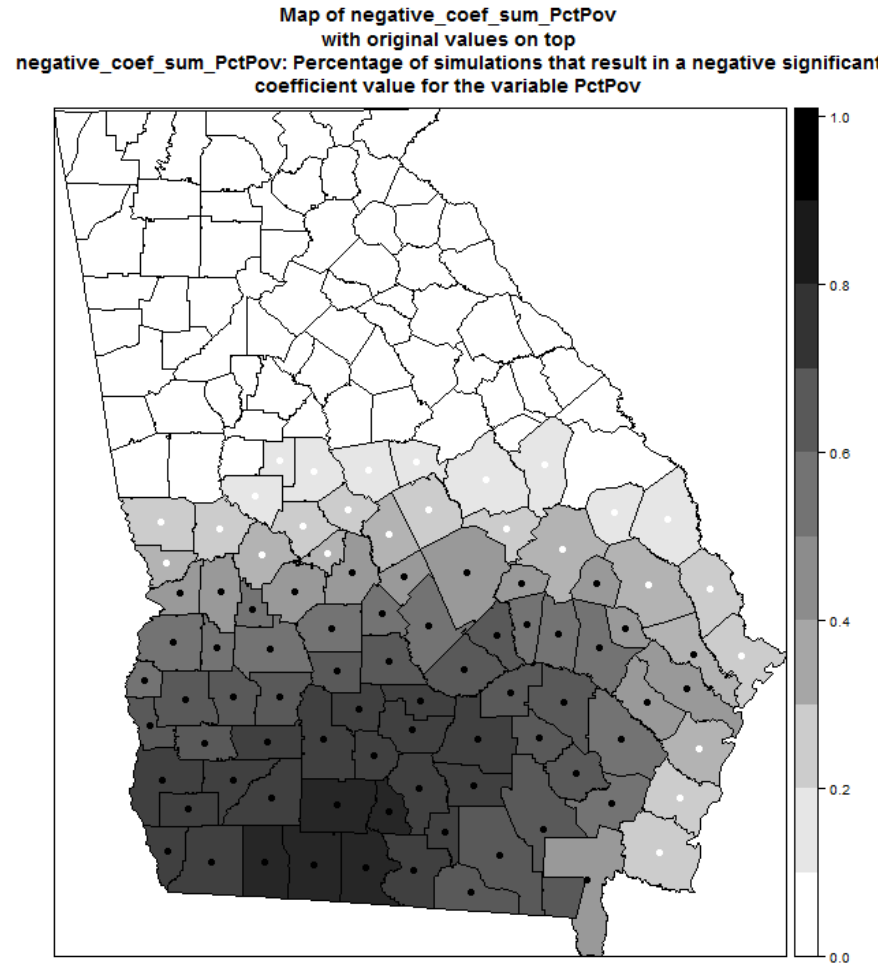




transform the map into a value

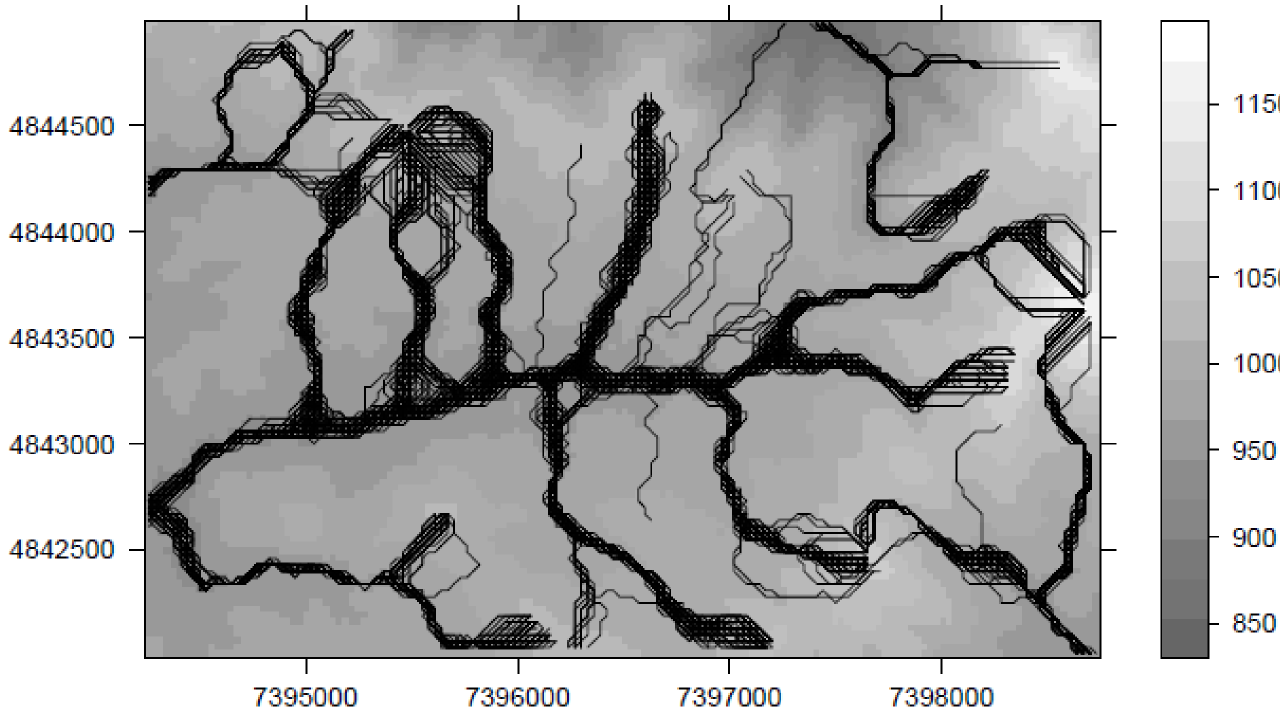
Visualization of the results

- **How about location?**



Visualization of the results

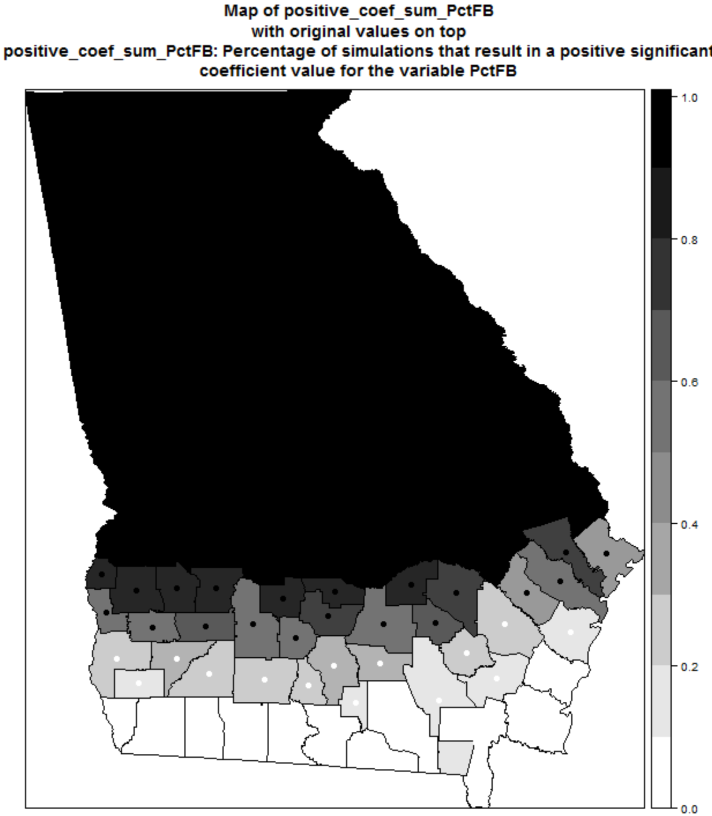
- **Calculation of river network from DEM, 100 simulations (Hengl et al., 2010)**



Final result

- The distribution of output needs to be compared to the accepted level of uncertainty

- The previous examples were the largest uncertainties in the study; generally the results were fairly accuracte



Computing in practise

- **Research was made using the statistical software R**
- **The Georgia example above took 6 minutes for 1000 simulations; Laos case (9000 points, 20 variables) took 7 days for 100 simulations**
- **R code published online https://github.com/jaakkomadetoja/epgwr**

References

- Heuvelink, G. B. (1998). Error propagation in environmental modelling with GIS. CRC Press.

- Epprecht, M., Minot, N., Dewina, R., Messerli, P., & Heinimann, A. (2008). The geography of poverty and inequality in the Lao PDR. Swiss National Centre of Competence in Research (NCCR) North-South, Geographica Bernensia, and In-ternational Food Policy Research Institute (IFPRI).
- Gilman, E., Keskinarkaus, A., Tamminen, S., Pirttikangas, S., Röning, J., & Riekki, J. (2015). Personalised assistance for fuel-efficient driving. Transportation Research Part C: Emerging Technologies, 58, 681-705.
- Hengl, T., Heuvelink, G. B. M., & Loon, E. (2010). On the uncertainty of stream networks derived from elevation data: the error propagation approach. Hydrology and Earth System Sciences, 14(7), 1153-1165.