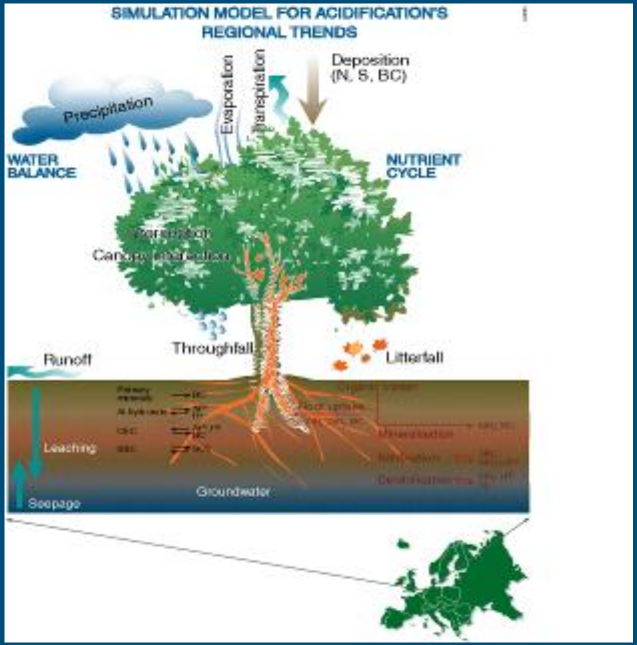
Accounting for spatial sampling effects in regional uncertainty propagation analyses

Gerard Heuvelink, Dick Brus and Gertjan Reinds
Wageningen University and Research Centre

ALTERRA
WAGENINGEN UR

Many environmental models are ‘point’ models

- Output at some location only depends on inputs at that same location
- Examples: evapotranspiration, crop growth, soil acidification, pesticide leaching to groundwater, greenhouse gas emission



ALTERRA
WAGENINGEN UR

Output y is some function of input u, consider case where interest is in the spatial average

$$y(x) = g(u(x)) \qquad x \in D$$

$$\overline{y} = \frac{1}{\|D\|} \int_D y(x)\, dx \quad \text{or} \quad \overline{y} = \frac{1}{M} \sum_{i=1}^{M} y(x_i)$$

When input is uncertain, uncertainty will propagate to output:

$$Y(x) = g(U(x))$$

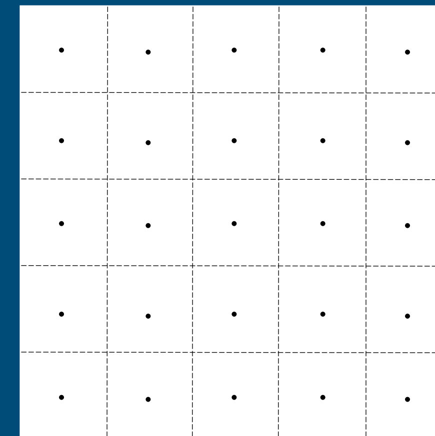
$$\overline{Y} = \frac{1}{M} \sum_{i=1}^{M} Y(x_i)$$

How large is uncertainty about the spatial average?

Can be solved using Monte Carlo simulation:

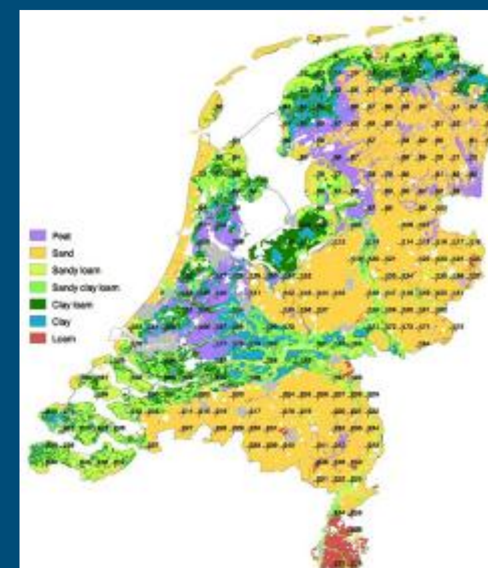
- Repeat n times:
  - Use pseudo-random number generator to draw a realisation from the probability distribution of (spatially correlated) input U(x) for all $x \in D$
  - Run model g for the simulated input, calculate spatial average of model output and store result
- Collection of n spatially averaged model outputs is a random sample from its probability distribution, uncertainty can be characterised using a measure of spread such as the variance
- Analysis requires $n \times M$ model runs (M very large, it may even be infinite)

In practice, geographic domain D is represented by a (small) sample



$$\hat{Y} = \frac{1}{m}\sum_{i=1}^{m} Y(x_i) \qquad m << M$$

- Kros *et al*. (Journal of Environmental Quality 1999) used m=25 where D was a 5×5 km^2 grid cell; Heuvelink *et al*. (Geoderma 2009) used m=258 where D was the entire Netherlands
- **Nice**: number of Monte Carlo runs n can be made much larger because computing costs are proportional to number of model runs n×m instead of n×M
- **Not so nice**: sampling error



Can sampling error be quantified, can sampling bias be corrected for, can optimum ratio of m and n be calculated?

Requires probability sampling of the locations: locations become stochastic as well

$$\hat{\bar{Y}} = \frac{1}{m}\sum_{i=1}^{m} g(U(X_i))$$

In case of simple random sampling in attribute and geographic space, variance of spatial mean satisfies

$$V(\bar{Y}) = V_{\xi p}(\hat{\bar{Y}}) - E_{\xi}[V_{p|\xi}(\hat{\bar{Y}})]$$

(ξ refers to stochasticity in U, p to stochasticity in X)

ALTERRA
WAGENINGEN UR

Estimating the variance of the spatial mean with n×m model runs

$$V(\overline{Y}) = V_{\xi p}(\hat{\overline{Y}}) - E_{\xi}[V_{p|\xi}(\hat{\overline{Y}})]$$

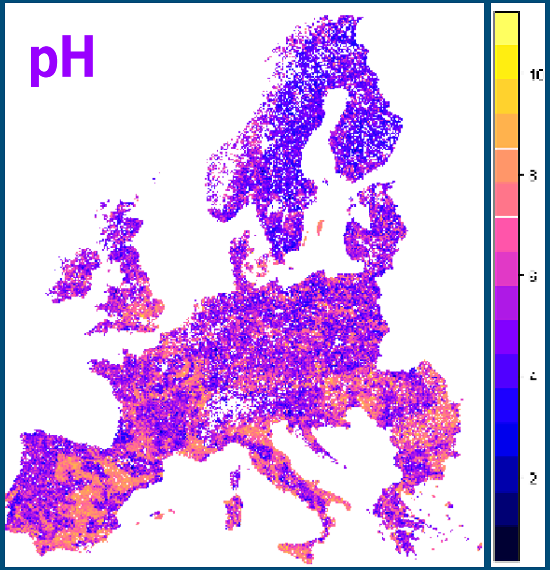
- Repeat n times:
  - draw m locations
  - simulate inputs and run model at these locations
  - calculate mean of m model outputs
- Calculate variance of n model means

- Repeat n times:
  - draw m locations
  - simulate inputs and run model at these locations
  - calculate variance of sampling error
- Calculate mean of n sampling error variances

Real-world application: N_2O emission from soil in non-agricultural areas for EU25

$$\log(N_2O(x)) = a_0 + a_1 \cdot N_{dep}(x) + a_2 \cdot \mathrm{Clay}(x) + a_3 \cdot C_{soil}(x) + \\ + a_4 \cdot \mathrm{Temp}(x) + a_5 \cdot C_{soil}(x) \cdot \mathrm{Temp}(x) + a_6 \cdot \mathrm{Prec}(x) + \\ + a_7 \cdot C_{soil}(x) \cdot \mathrm{Prec}(x) + a_8 \cdot \mathrm{pH}(x) + a_9 \cdot \mathrm{TreeSpecies}(x)$$

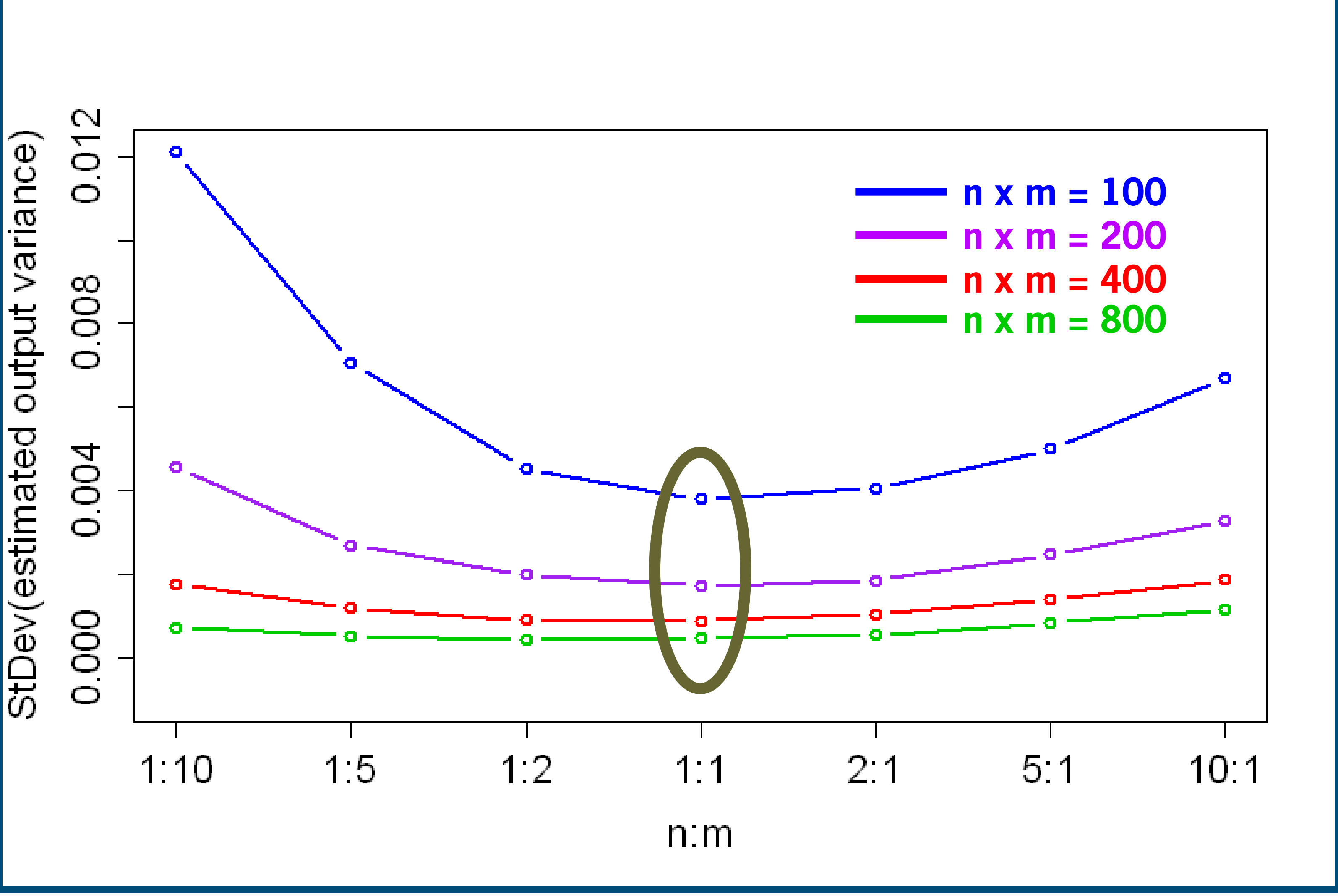
- Consider only uncertainty in C_{soil} and pH (carbon content and pH of topsoil)
- Both soil properties modelled geostatistically using European soil map and data from WISE/SPADE database



Numerical experiments

- Use four values for the total number of model runs $n \times m$ (100, 200, 400, 800)
- Use seven values for the ratio n:m (10:1, 5:1, 2:1, 1:1, 1:2, 1:5, 1:10)
- Estimate variance of spatial mean for all 28 cases with $n \times m$ model runs
- Do this many times (e.g. 1000 times) and compute the standard deviation of the many estimates for each of the 28 cases: measure of how accurately the variance of the spatial mean is estimated

Standard deviation of estimated variance



ALTERRA
WAGENINGEN UR

Conclusions (1/3)

- Propagation of input uncertainty to spatially averaged model output is often based on results for a (small) spatial sample
- Sampling error and sampling bias are usually ignored but may be substantial
- Spatial probability sampling must be employed to assess sampling error and eliminate sampling bias: can be done and does not inflate computation time
- Educated guess of spatial sample size is risky: too small sample yields non-negligible sampling error and bias, too large spatial sample is inefficient

Conclusions (2/3)

- Calculation of optimum ratio of Monte Carlo and spatial sample sizes is computationally demanding because it requires an additional loop
- The optimum ratio is likely case-specific (as yet unclear what triggers the optimum ratio)
- In the case study the optimum ratio was stable for different values of $n \times m$: if this holds more generally then for a given (new) case the ratio need be determined only once for moderate size of $n \times m$ and used in the final uncertainty propagation analysis with large $n \times m$

Conclusions (3/3)

- Spatial sampling cannot be used with models that involve spatial interactions (e.g. flow, diffusion). For such models, the spatial resolution may perhaps be decreased, but that is another issue

Thank you