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 Spatial Environmental Models (9-13 December 2024)

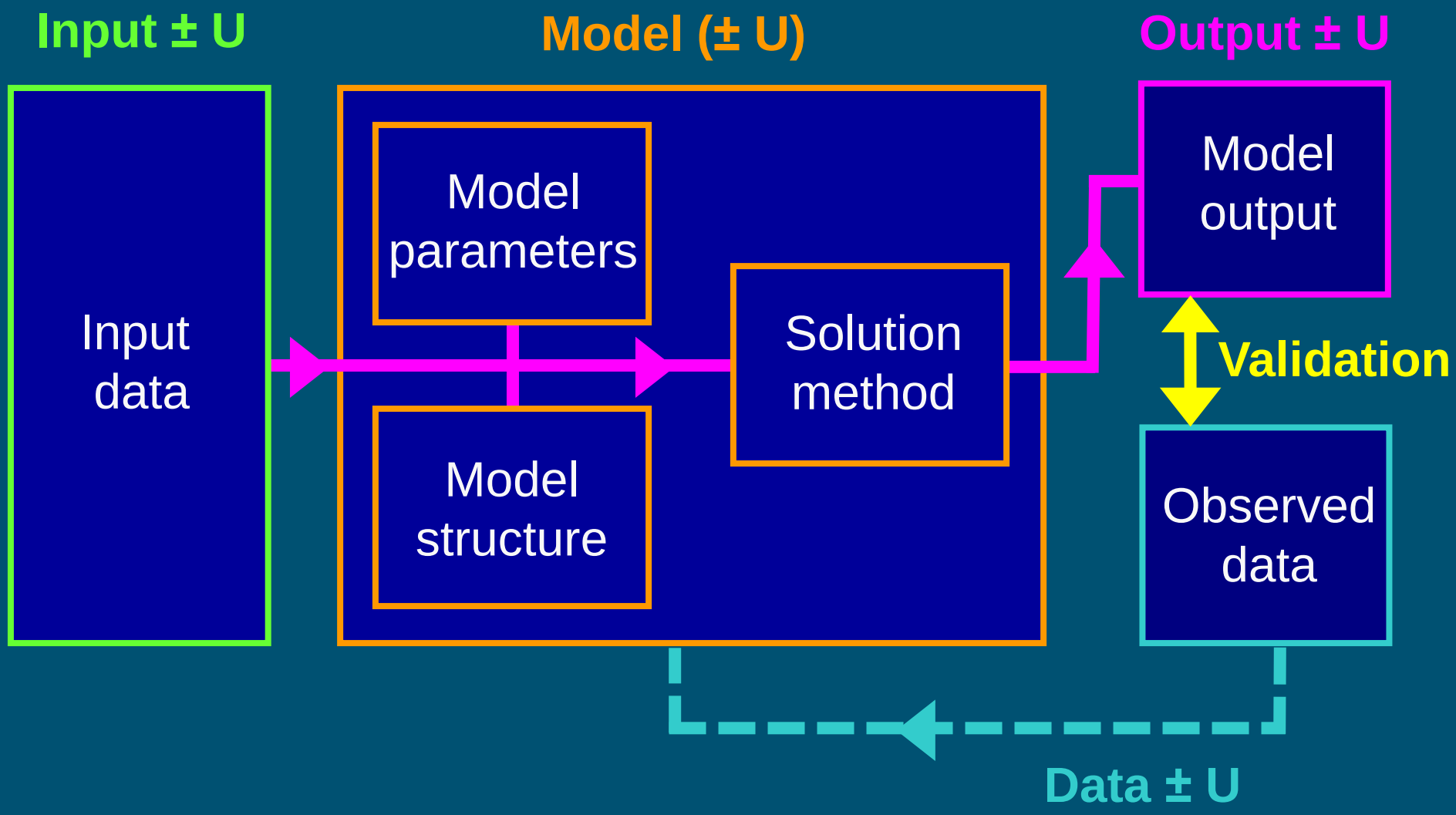
Day	9:00 – 12:00	12:00 – 13:30	13:30 – 15:45	15:45 – 16:15	16:15 – 17:00
Mon. 9 Dec.	Course overview, lecture and exercises: probabilistic modelling of uncertainty, Taylor series method	Lunch break	Computer practical Taylor series method: vegetation indices	Feedback computer practical	Continue practical, own data or advice
Tue. 10 Dec.	Lecture and exercises Monte Carlo method, including uncertainty source contributions and scenarios	Lunch break	Computer practical Monte Carlo method: vegetation indices	Feedback computer practical	Continue practical, own data or advice
Wed. 11 Dec.	Lecture and exercises uncertainty in model parameters and model structure, including Bayesian calibration	Lunch break	Computer practical Bayesian calibration and propagation of model parameter uncertainty	Feedback computer practical	Continue practical, own data or advice
Thu. 12 Dec.	Lecture and exercises statistical validation and 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		

We achieved a lot already

- We analysed how uncertainty in model inputs, model parameters and model structure **propagate** to a model's output
- In this way we could quantify the uncertainty of the model output at each and every location in our area of interest
- We could even quantify the **uncertainty contribution** of individual uncertainty sources
- But all this is based on the **assumption** that we characterised the uncertainty well with probability distributions of the model inputs, model parameters and model structure

Today we look at a very different way to quantify the accuracy of the model output: **statistical validation**. This has the important advantage that it is completely **model-free**

Uncertainty propagation and model validation overview



Statistical validation of spatial model output (maps)

- Suppose we have predictions $\hat{z}(s)$ for all locations s in our area of interest A (be it the output of a spatial model or a map derived in some other way, it does not matter how)
- How accurate are these predictions, how good is the map?
- Common validation metrics:

$$\textbf{Mean Error} = ME = \frac{1}{|A|}\int_{s\in A}(z(s)-\hat{z}(s))ds$$

$$\textbf{Root Mean Squared Error} = RMSE = \sqrt{\frac{1}{|A|}\int_{s\in A}(z(s)-\hat{z}(s))^2 ds}$$

$$\textbf{Model Efficiency Coefficient} = MEC = 1 - \frac{\int_{s\in A}(z(s)-\hat{z}(s))^2 ds}{\int_{s\in A}(z(s)-\bar{z})^2 ds}$$

where $\bar{z} = \frac{1}{|A|}\int_{s\in A} z(s)ds$

Many call this **"R-square"** but confusing!

We can only estimate the validation metrics

- We cannot measure the 'true' z everywhere but must do with a limited number of observations in the study area
- Let us assume we measured z at n randomly sampled locations s_i $(i = 1 \ldots n)$ in A, so that we can compute:

$$\widehat{ME} = \frac{1}{n}\sum_{i=1}^{n}(z(s_i) - \hat{z}(s_i))$$

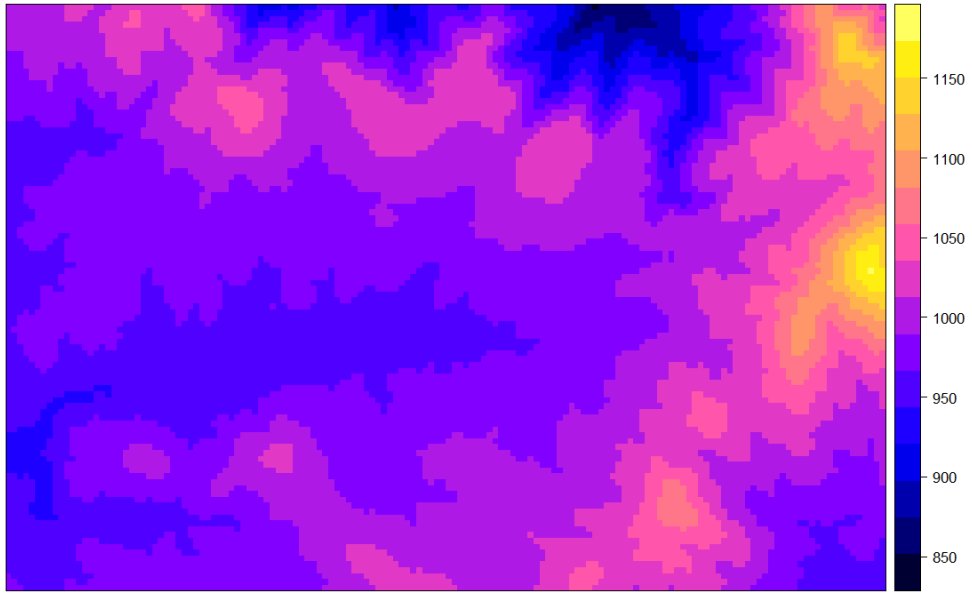
$$\widehat{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z(s_i) - \hat{z}(s_i))^2}$$

$$\widehat{MEC} = 1 - \frac{\sum_{i=1}^{n}(z(s_i) - \hat{z}(s_i))^2}{\sum_{i=1}^{n}(z(s_i) - \bar{z})^2}$$

where $\bar{z}$ is now the mean of the $z(s_i)$, that is $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z(s_i)$

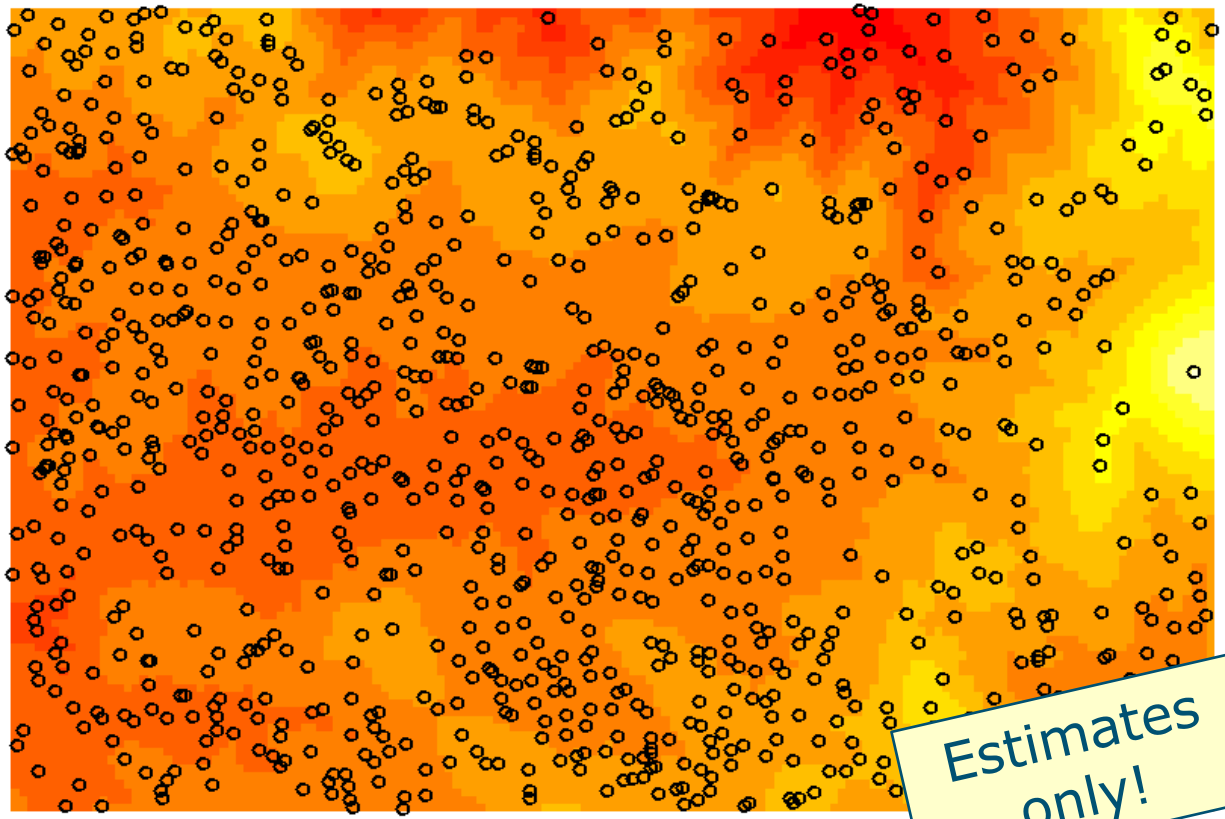
Example: DEM of Zlatibor area, Serbia

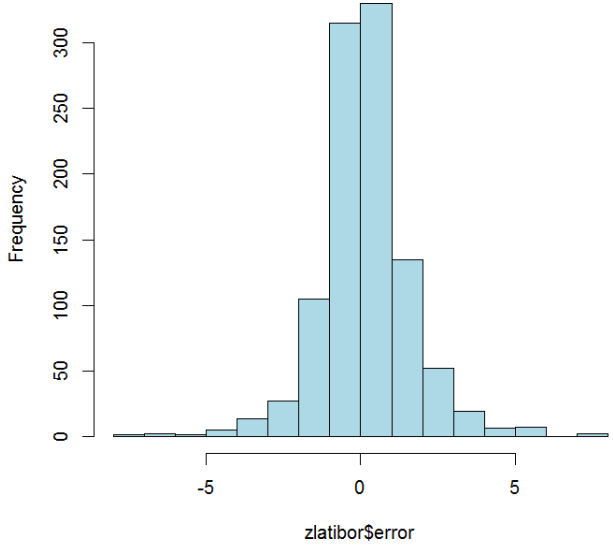
- 30 m resolution
- 4.5 km x 3.0 km



WAGENINGEN UNIVERSITY
WAGENINGEN UR

‘True’ elevation measured at 1020 randomly selected control points





Estimates only!

ME = 0.18 m

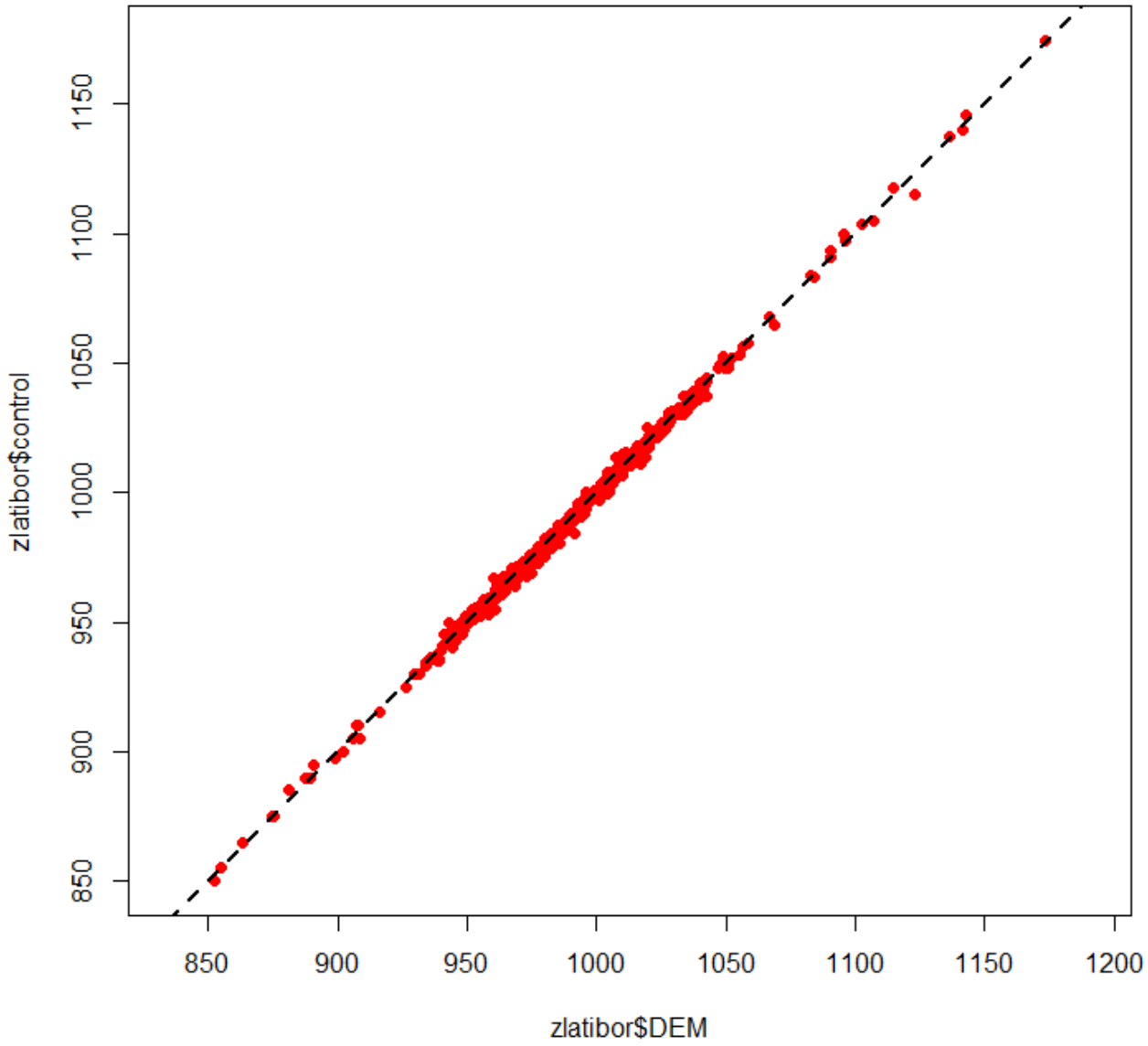
RMSE = 1.49 m

MEC = 0.9977

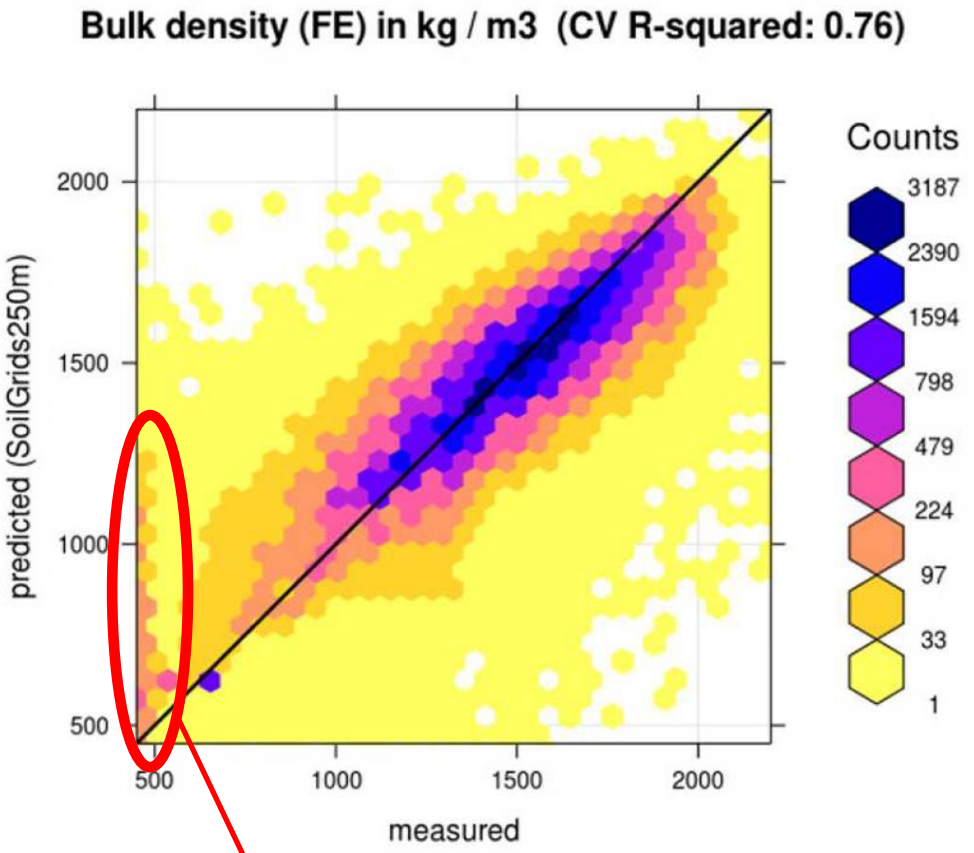
- That’s a very good map!
- Note these are only estimates

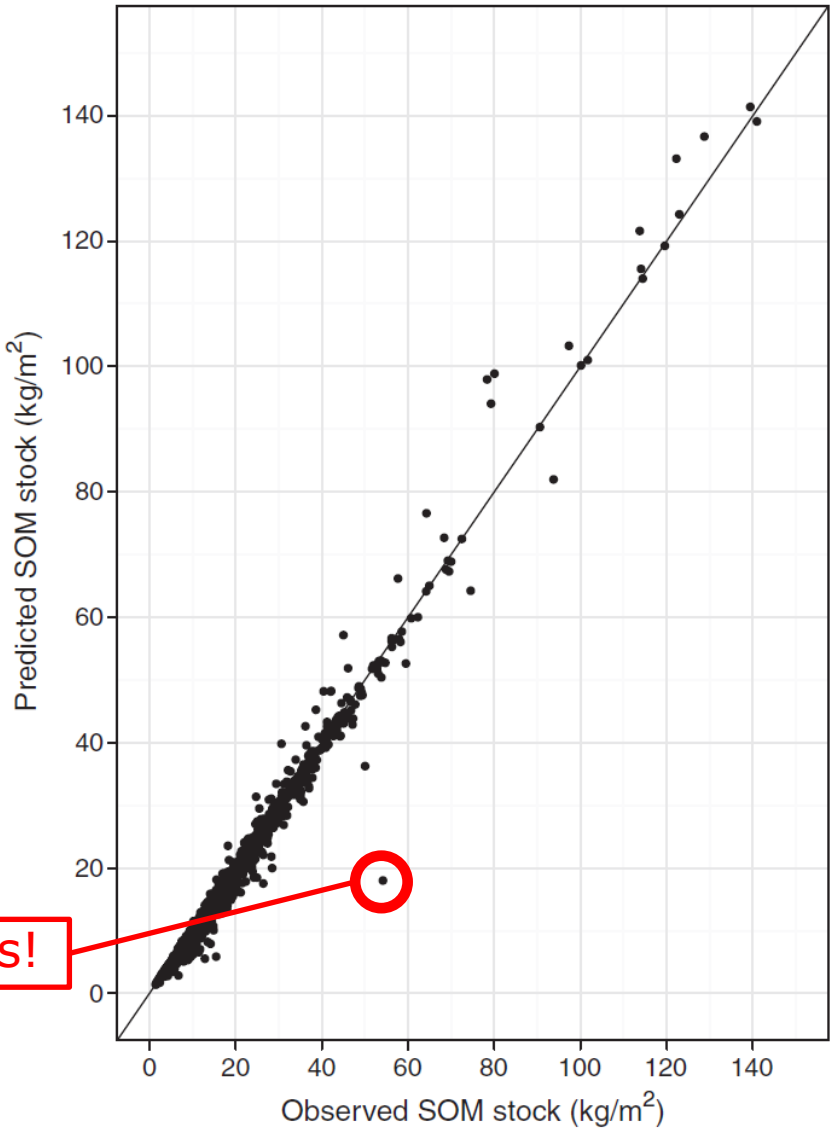
WAGENINGEN UNIVERSITY
WAGENINGEN UR

Indeed almost all variance explained. If only all models were that good!



Graphical analysis very useful: scatter (density) plot of predicted against observed





not well-noticed by validation metrics!

Exercise 1

- Work alone or in groups, whatever you prefer
- Open file 'PERC uncertainty Thursday exercise 1.pdf' from MS-Teams and copy the datafile needed for this exercise
- Compute confusion matrix and map purity in R, for example using functions "table()", "sum()" and "sum(diag())"

In case of random sampling we not only estimate the validation metrics, but we can also quantify the associated estimation accuracy:

- **Sampling theory** from statistics tells us:
  - The estimates are **unbiased** (no systematic over- or underestimation)
  - The accuracy of the estimates is also known, in other words we can derive **confidence intervals**
- And all this without making any assumptions because we took a '**design-based**' approach instead of a '**model-based**' approach
- Well, to be honest, three assumptions were made:
  1. The validation data are a **random sample** from the population
  2. Measurement errors of validation data are **negligible**
  3. Validation data are **completely independent** and not in any way used to make the map or calibrate and run the model (no data leakage)

Confidence interval for the Mean Error

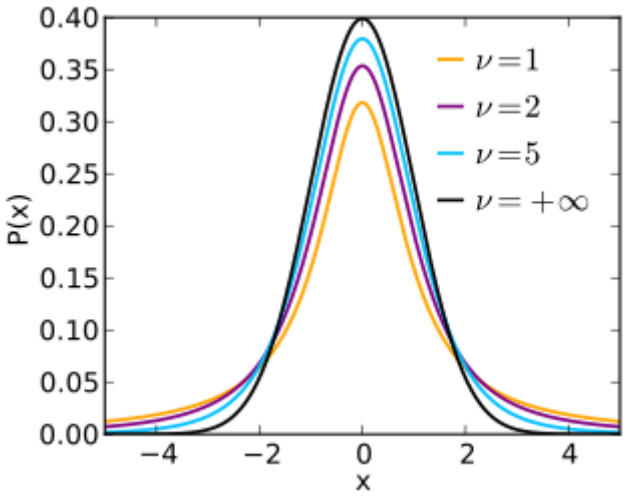
If the validation data are a simple random sample from the area of interest, then the **lower and upper limits** of a (100-α)% confidence interval for the Mean Error are given by:

$$\widehat{ME} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2,n-1}$ is the value of the **t-distribution** at the $\alpha/2$ quantile and $n-1$ degrees of freedom and s is the standard deviation of the n errors

For Zlatibor DEM:

- ME_{lower} = 0.11 m
- ME_{upper} = 0.26 m



Confidence interval for the Root Mean Squared Error

We can do this in two ways:

1. Same approach as for Mean Error, but now we do not estimate the spatial mean of the error, but the spatial mean of the **squared error**. In this way we get lower and upper limits for MSE by computing:

$$\widehat{MSE} \pm t_{\alpha/2,n-1} \cdot \frac{s_{squared\ error}}{\sqrt{n}}$$

   The RMSE limits are easily computed from the MSE limits by taking their square root

2. **Bootstrap sampling** (next slide)

For Zlatibor DEM:

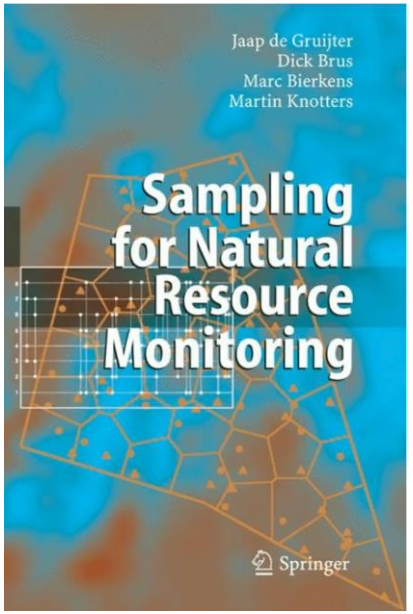
- $RMSE_{lower}$ = 1.40 m
- $RMSE_{upper}$ = 1.59 m

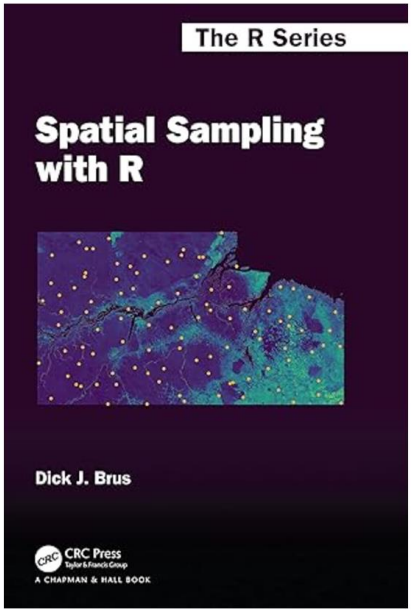
Confidence intervals using bootstrap sampling

- A generic and powerful approach, albeit approximate only:
  1. Repeat N times (say N=1000):
     a. Create a bootstrap sample from the original sample of the n map errors (sample n times with replacement)
     b. Calculate validation metric for this bootstrap sample and store it
  2. Compute 0.05 and 0.95 quantiles of the N validation metrics, these represent the lower and upper limits of a 90% confidence interval for the metric
- For Zlatibor DEM we then get:
  - $RMSE_{lower}$ = 1.40 m, $RMSE_{upper}$ = 1.58 m (methods agree!)
  - MEC_{lower} = 0.9973, MEC_{upper} = 0.9981

Other probability sampling designs

- Stratified simple random sampling
- Cluster random sampling
- Systematic random sampling
- … (many more)





Each has its own statistical inference, both for estimating the validation metric andcalculating the associated confidence interval

You will apply **stratified simple random sampling** in the practical this afternoon

See also (in your literature folder):

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

Summary

The increase in digital soil mapping around the world means that appropriate and efficient sampling strategies are needed for validation. Data used for calibrating a digital soil mapping model typically are non-random samples. In such a case we recommend collection of additional independent data and validation of the soil map by a design-based sampling strategy involving probability sampling and design-based estimation of quality measures. An important advantage over validation by data-splitting or cross-validation is that model-free estimates of the quality measures and their standard errors can be obtained, and thus no assumptions on the spatial auto-correlation of prediction errors need to be made. The quality of quantitative soil maps can be quantified by the spatial cumulative distribution function (SCDF) of the prediction errors, whereas for categorical soil maps the overall purity and the map unit purities (user's accuracies) and soil class representation (producer's accuracies) are suitable quality measures. The suitability of five basic types of random sampling design for soil map validation was evaluated: simple, stratified simple, systematic, cluster and two-stage random sampling. Stratified simple random sampling is generally a good choice: it is simple to implement, estimation of the quality measures and their precision is straightforward, it gives relatively precise estimates

Uncertainty propagation and statistical validation quantify the accuracy of model outputs in very different ways

- Uncertainty propagation quantifies uncertainty at each and every location in the area of interest, statistical validation only gives **summary measures**
- Uncertainty propagation makes lots of assumptions, statistical validation is **model-free**
- Statistical validation requires an **independent probability sample of the target variable,** uncertainty propagation not (and can thus be used in cases where the target variable cannot be measured, such as when we model the future)
- Uncertainty propagation is a **lot of work** (because input and model uncertainty must be quantified with probability distributions), statistical validation requires **field- and labwork**
- Uncertainty propagation can be **computationally demanding**, statistical validation is not
- Uncertainty propagation can quantify the **contribution of uncertainty sources**, statistical validation cannot

Both have merit, ideally you do them both!

What if we have no probability sample?

In such case you can still compute validation metrics, **BUT**:

1. You cannot prove **unbiasedness**
2. You cannot compute **confidence intervals**

In practice we rarely have a probability sample for validation, especially when we evaluate 'mapping models' (e.g. kriging or machine learning): we have just one dataset and would like to use this one data set for **three purposes**: model calibration, prediction and validation

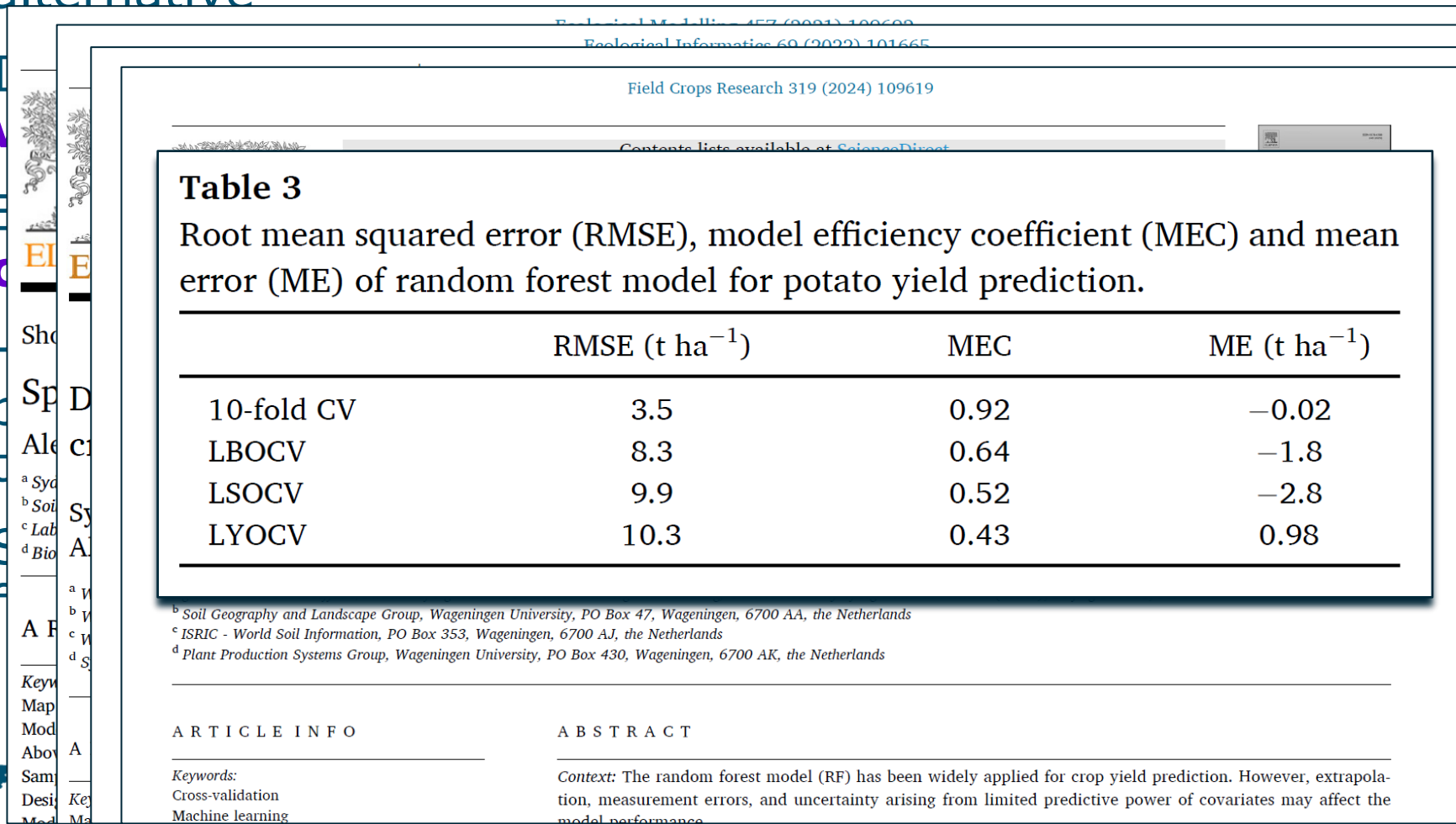
This can be done using **cross-validation**

Cross-validation, how does it work?

1. Visit all n observation locations **one by one**
2. Put the observation **aside** and use the remaining $n-1$ observations to **calibrate** the mapping model
3. Use the calibrated model to **predict** at the location of the put-aside observation
4. Once done for all locations we have n **independent observations and predictions**: we can compute ME, RMSE and MEC in the usual way
5. This procedure is known as **leave-one-out cross-validation**; if model calibration and prediction are computatially demanding we might opt for **K-fold cross-validation**: randomly split dataset in K equally-sized folds, each time putting one fold aside and using the others for model calibration and prediction
6. Cross-validation is better than one-time **data-splitting** (but **Monte Carlo cross-validation**, which repeats data-splitting many times, is fine)

Spatial cross-validation

- Random cross-validation has been criticized because it ignores **spatial autocorrelation** of map errors
- **Spatial cross-validation** was proposed as an alternative



Field Crops Research 319 (2024) 109619

Table 3

Root mean squared error (RMSE), model efficiency coefficient (MEC) and mean error (ME) of random forest model for potato yield prediction.

	RMSE (t ha^{-1})	MEC	ME (t ha^{-1})
10-fold CV	3.5	0.92	−0.02
LBOCV	8.3	0.64	−1.8
LSOCV	9.9	0.52	−2.8
LYOCV	10.3	0.43	0.98

[b] Soil Geography and Landscape Group, Wageningen University, PO Box 47, Wageningen, 6700 AA, the Netherlands
[c] ISRIC - World Soil Information, PO Box 353, Wageningen, 6700 AJ, the Netherlands
[d] Plant Production Systems Group, Wageningen University, PO Box 430, Wageningen, 6700 AK, the Netherlands

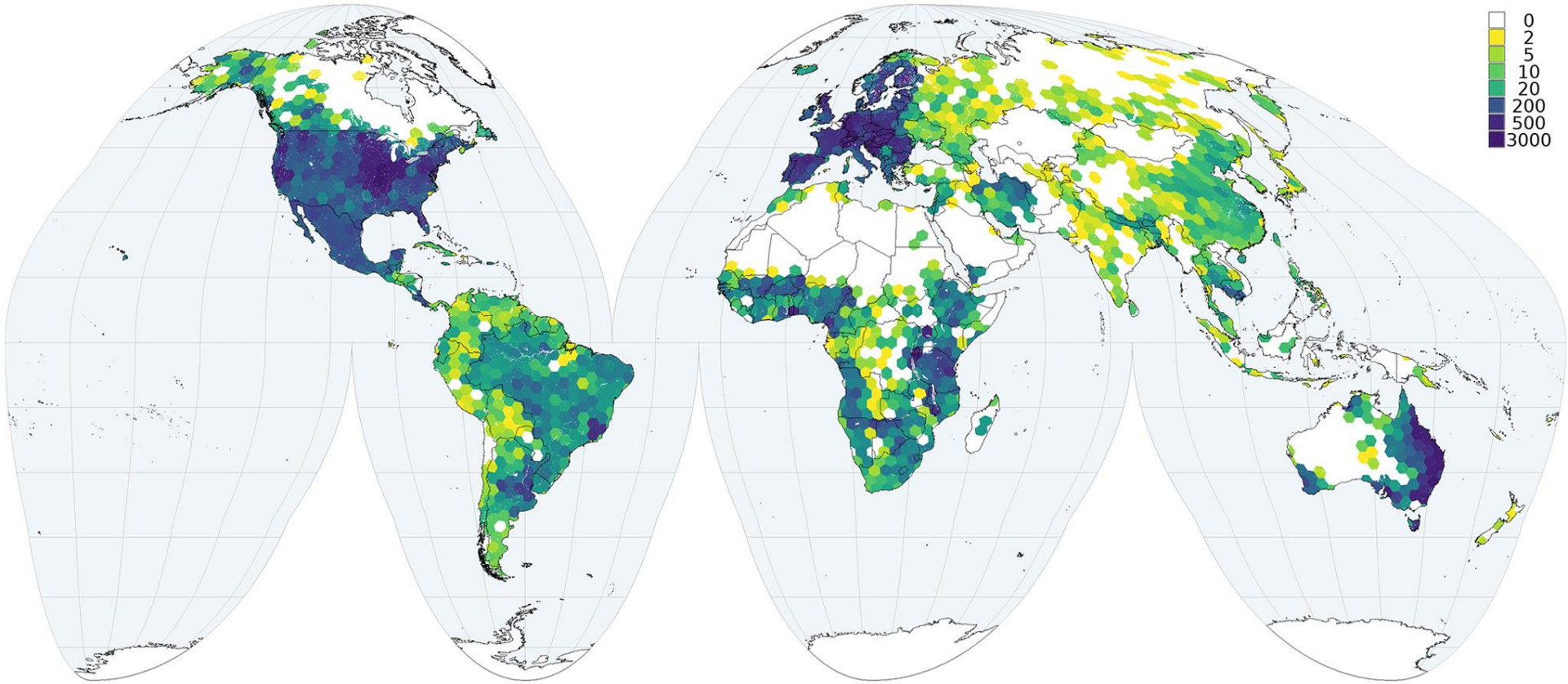
ARTICLE INFO

Keywords:
Cross-validation
Machine learning

ABSTRACT

Context: The random forest model (RF) has been widely applied for crop yield prediction. However, extrapolation, measurement errors, and uncertainty arising from limited predictive power of covariates may affect the

Spatial clustering is not uncommon: can we predict the Sahara? And can we quantify the accuracy of those predictions?



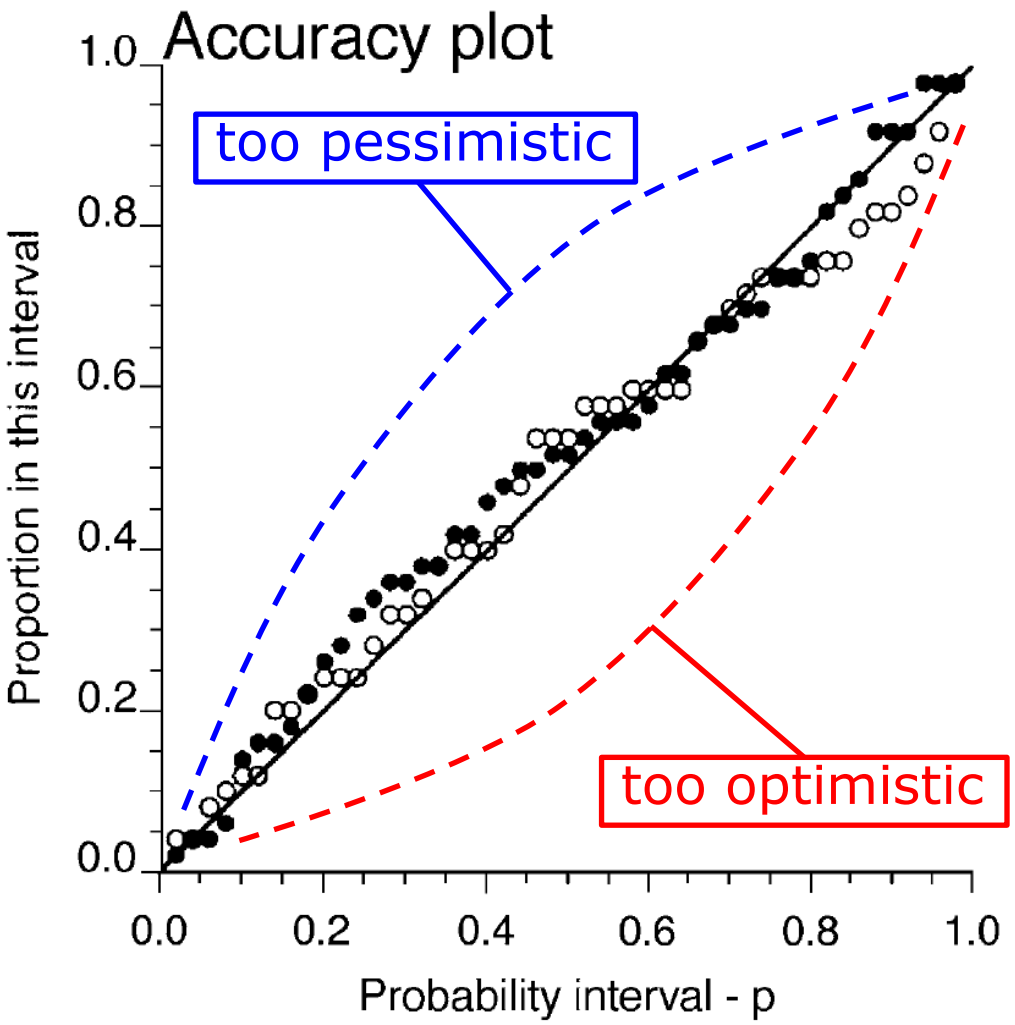
SoilGrids calibration data from WoSIS (Poggio et al. 2021)

Very interesting: we can also use (cross-) validation to check whether our uncertainty assessment was correct

Prediction Interval Coverage Probability (PICP): co[illegible]t what percent[illegible] the va[illegible]are ins[illegible]α) % pred[illegible]nterval

Link with Tuesday's **SRMSE**!

Accuracy/Reliability plot: Plot of PICP for all values of α between 0 and 100



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Summary of today's lecture

- Statistical validation of spatial model outputs (maps) is quite **complementary** to spatial uncertainty propagation analysis: both have their pros and cons
- Ideally the validation data are a **probability sample** from the area of interest, if not you might get biased (cross-)validation results, especially in case of spatially **clustered data**
- Statistical validation is usually used to evaluate model prediction performance (ME, RMSE, MEC) but we can also evaluate whether the predicted uncertainty was right (using the **PICP**)
- Validation data should not in any way have been used for model selection, model calibration and prediction: always be on your guard for **data leakage**
- One thing we did not talk about yet is the spatial '**support**' of the observations and predictions (defined as "the area or volume over which an observation or prediction is made") and how this influences uncertainty assessment: that's for tomorrow!