Bern University
of Applied Sciences

Mastering Machine Learning for Spatial Prediction II

Model selection and interpretation, uncertainty

OpenGeoHub
4/5 September 2019

Madlene Nussbaum

Objectives ...

- Know **2** ways of ...
  - of **model selection**
  - of **model interpretation**
  - computation of **uncertainty**

- Learn why we do model selection (or not)
- Learn that **ML != black box**
- Learn why we need uncertainty and how to validate your prediction intervals

Overview

Model Selection

- linear regression
- with lasso
- with covariate importance

Model interpretation

- partial residual plots
- partial dependence plot
- partial dependence maps

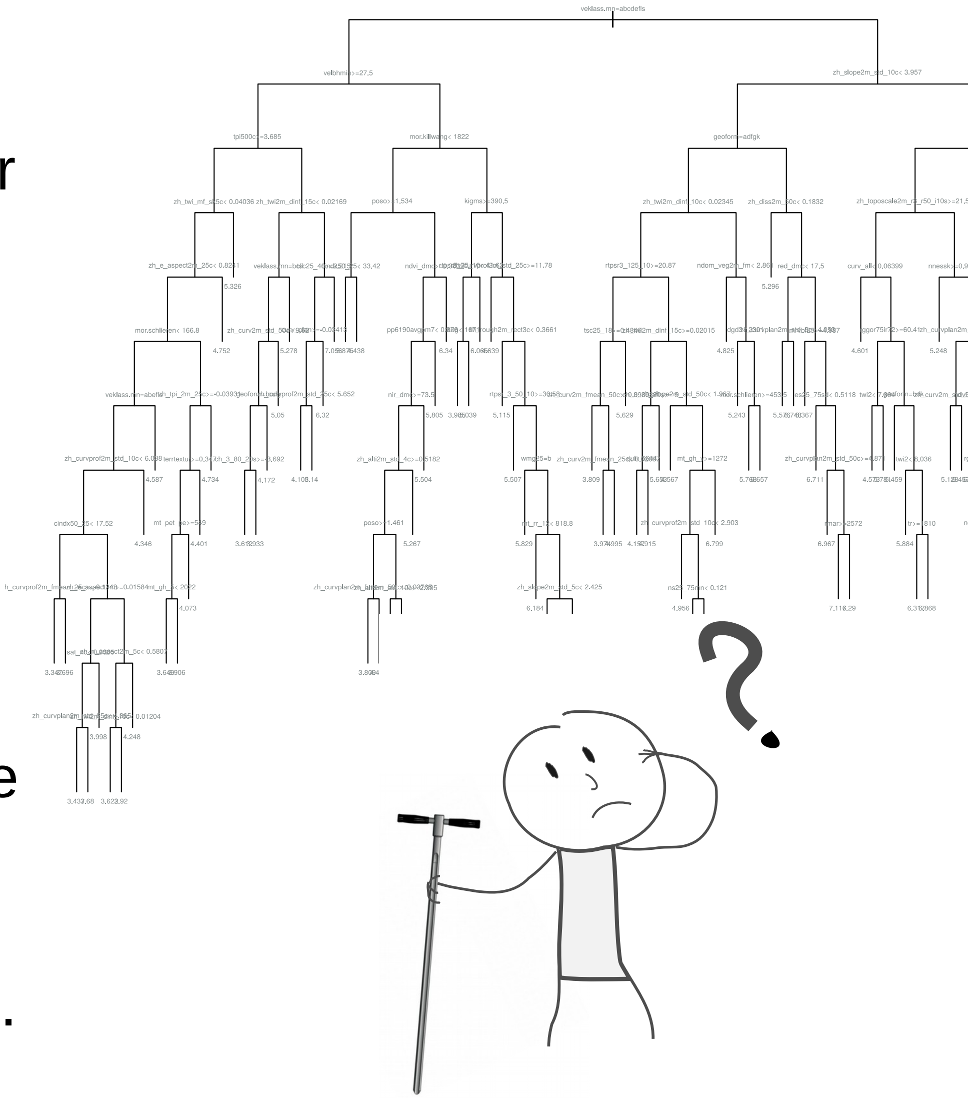
Uncertainty

- non-parametric bootstrap
- model-based bootstrap
- evaluation

Is there a reason for model selection?
Or is it enough to do model building?

Model selection = reduce the inital covariate set
Model building = find relationships between covariates and response

- ✓ Model interpretation
- ✓ Better just use relevant covariates for prediction
- ✓ Computational effort for predictions (just prepare 12 instead of 300 rasters)
- ✓ Maybe reduce effort for future data collection and modelling on same topic
- ✖ However, theoretical statisticians do not recommenced selection, because it is often biased, difficult to find the true model..
- ✖ We might loose prediction accuracy...



Model flexibility vs. model complexity

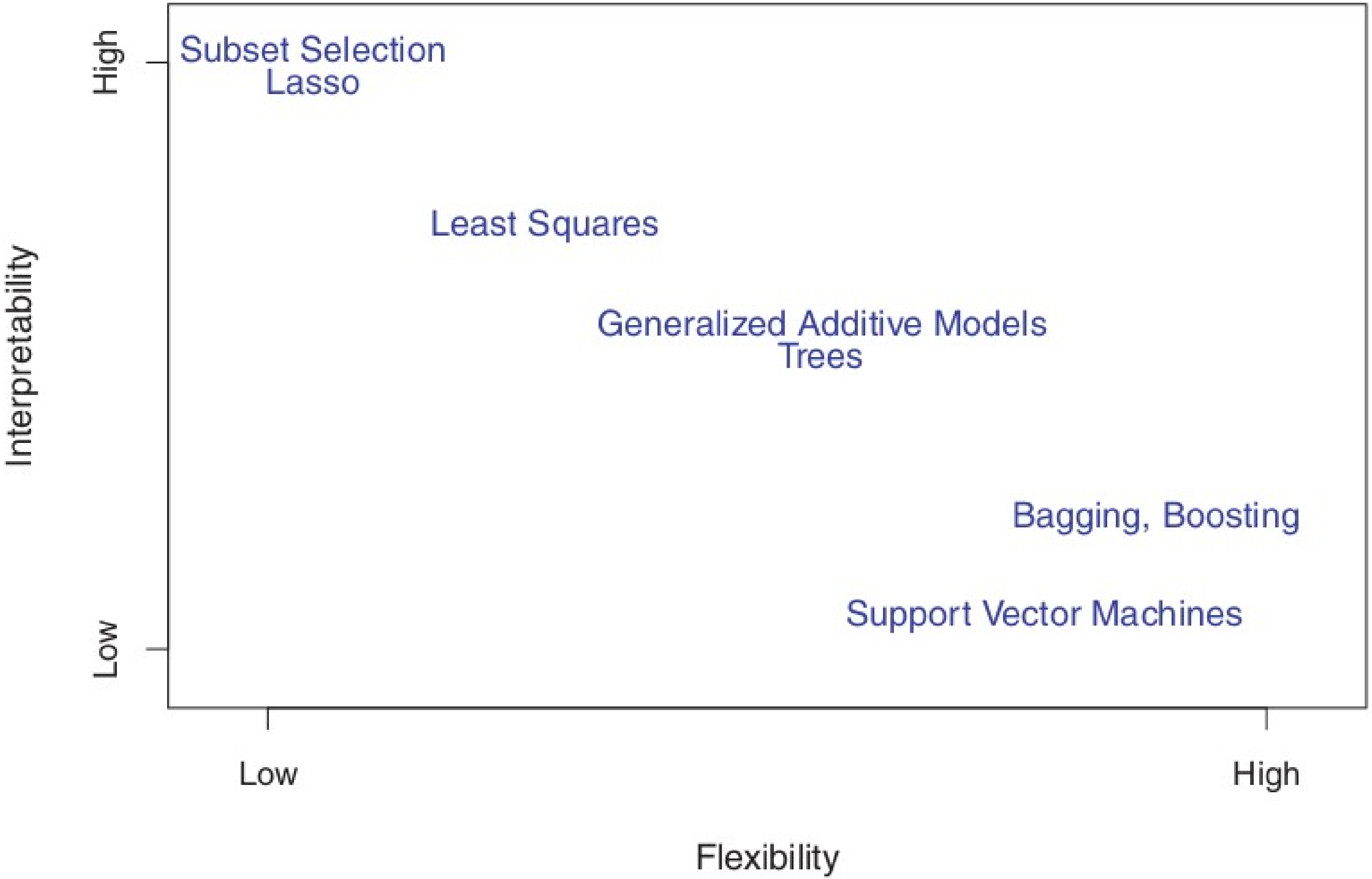
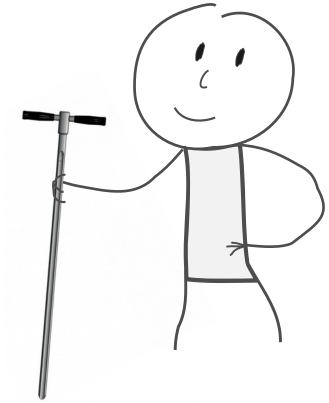


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Gareth et al. 2013, p. 25

Model selection – strategies



- Ask an soil surveyor that is familiar with the area

- Remove *n* worst covariates

- Stepwise addition or removal (see later)

- Test all possible models

- Shrinkage (e.g. by boosting or lasso, see later).

Model selection for linear regression

Usually used for linear models (e. g. OLS):

- Forward selection
  - Start with a model with just an intercept
  - Try all possible covariates and add the one that results in the best fit (e.g. R^2)
  - Add covariates until increase in R^2 is only very small (threshold) or evaluate the gained models by cross validation.

- Backward elimination
  - Fit the full model with all covariates
  - Remove the covariate that will cause the smallest decrease in R^2
  - Continue removing until the drop in R2 gets too big (threshold) or evaluate the fitted models by cross validation

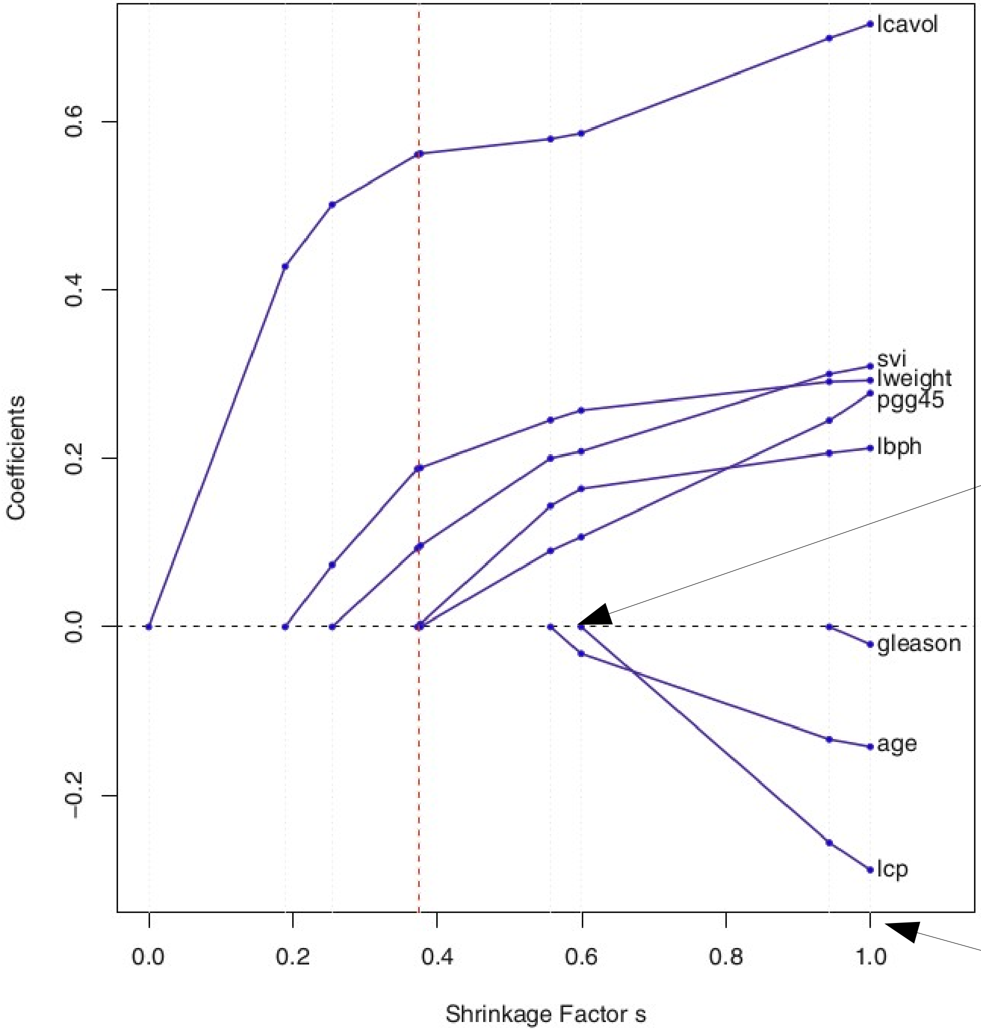
Model selection for linear regression

Evaluation of forward and backward selection:

- ✓ Straightforward, easy to understand
- ✓ Follows a selection path, hence much more efficient and less arbitrary than best subset selection
- ✖ Binary selection: A covariate is either in or out.
  Being in by e.g. 30 % is not possible.
- ✖ Multi-collinearity, unstable fits!
  We need to remove correlated covariates beforehand.
- ✖ Likely overfits the data, biased model selection.
  Most often does not find true model.
- ✖ We can not fit p > n

→ Possible solution: regularization / shrinkage with e.g. **lasso**.

Model selection with lasso



Path of coefficients for increasing tuning parameter

Coefficient becomes 0, meaning the covariate is removed from the model

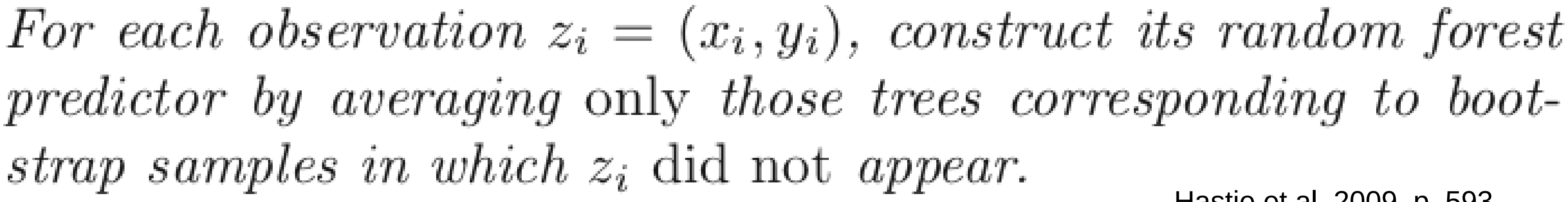
With lambda = 1, there is no shrinkage, and we have the normal ordinary least squares linear model fit

FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

Hastie et al. 2009, p. 70

Covariate selection for random forest
(or boosted trees)

What again was this **out-of-bag (OOB) error**?



For each observation $z_i = (x_i, y_i)$, *construct its random forest predictor by averaging* only *those trees corresponding to bootstrap samples in which* z_i did not *appear.*

Hastie et al. 2009, p. 593

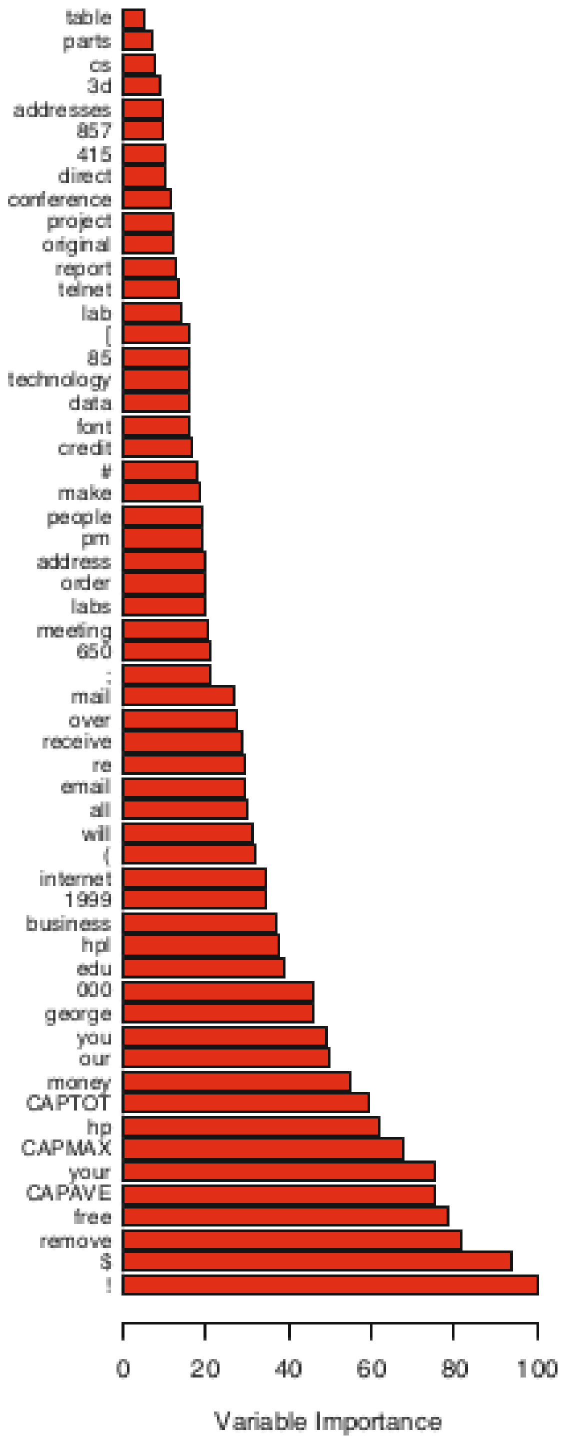
2 types of covariate importance:

- Mean decrease in **accuracy** (`type=1` in `importance()`): oriented on fitting the data. How much do we reduce error by using this covariate at this split?

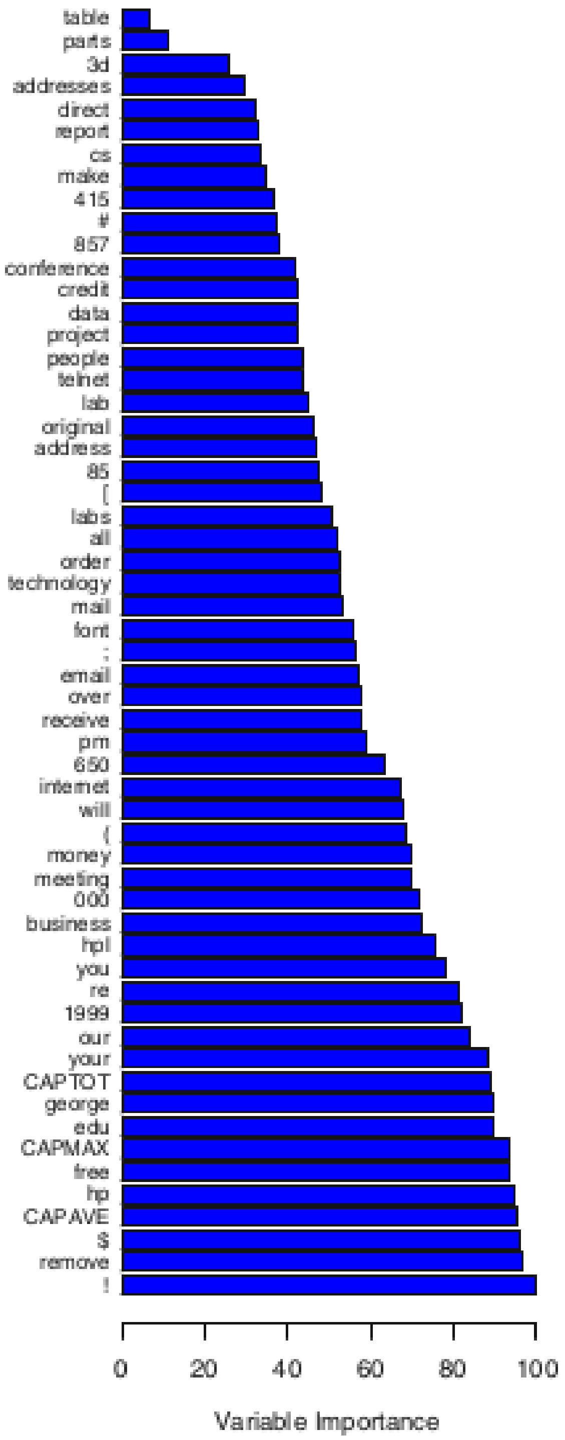
  → this importance is used for boosted trees.

- Mean decrease in **node impurity** (`type=2` in `importance()`), permutation based, oriented on predictions: How much worse do OOB predictions get if we randomly permutate a covariate?

  → removing a covariate is not the same, other correlated covariate could replace its “predictive capacity”



Loss in accuracy



OOB, Permutation

Hastie et al. 2009

Selection, very simple:

Recursive backward elimination

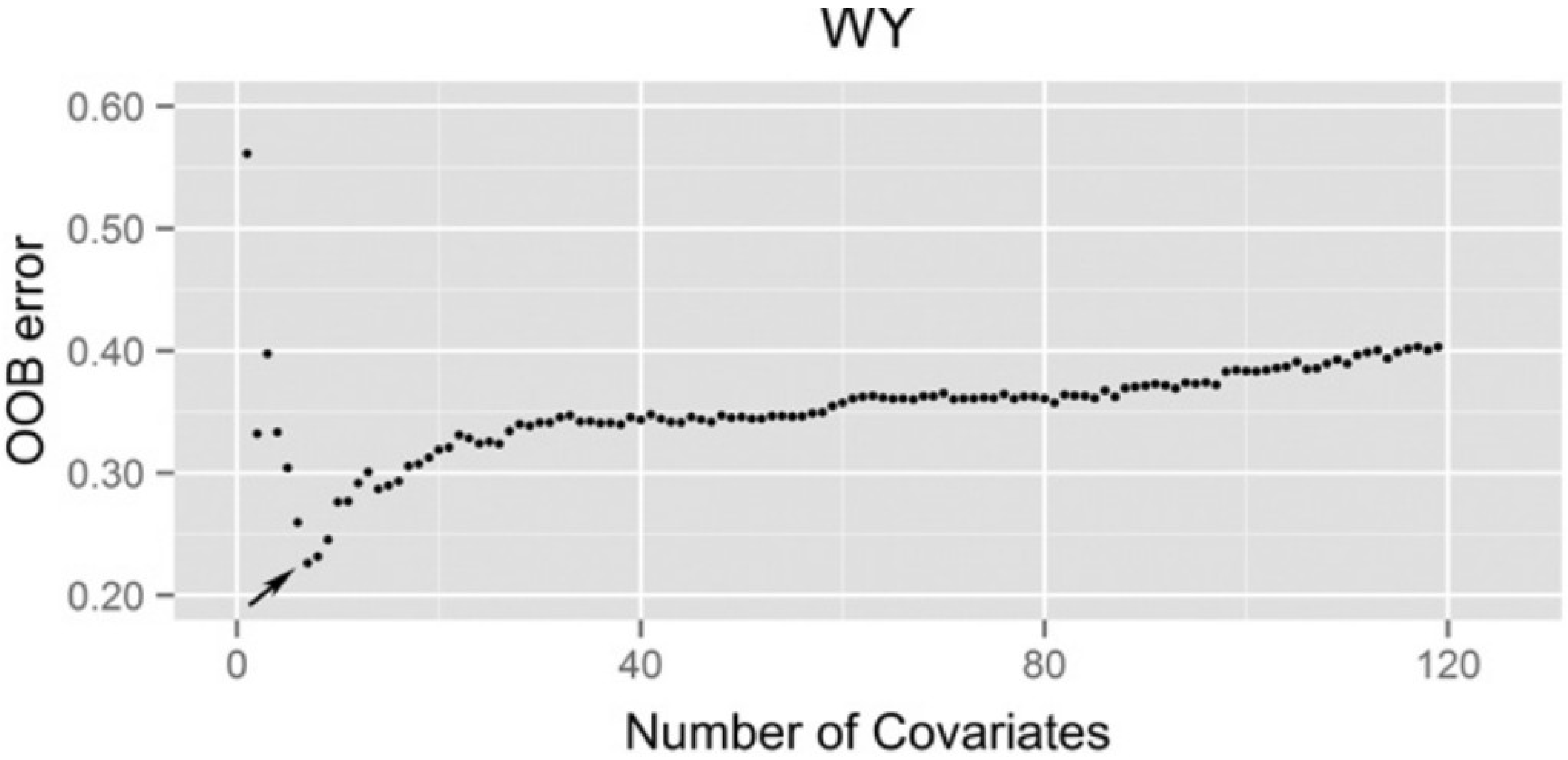
1) Remove covariate(s) with lowest importance

2) Refit random forest with remaining

Find optimum number of covariates by minimizing OOB error.

Problem:
Correlated covariates remain, because of randomisation at each split (m_{try}) . Interpretation needs to account for that.

Covariate selection for random forest
(or boosted trees)



Brungard et al. 2015

Model selection with random forest

- ✓ Straightforward approach
- ✓ Easy to implement
- ✓ From my experience: efficient
  (meaning a lot of covariates are removed)

- ✗ Correlated covariates remain in the model
- ✗ Artefacts in predictions possible
- ✗ Stability unclear?
  (Small changes in covariate or response values might change result drastically)
- ✗ Biased?
  (maybe biased as other backward elimination methods)
- ✗ Time consuming
  (iterative method, no paralell computing possible)
- ✗ Possibly: A lot of effort for a small result

Overview

Model Selection

- linear regression
- with lasso
- with covariate importance

Model interpretation

- partial residual plots
- partial dependence plot
- partial dependence maps

Uncertainty

- non-parametric bootstrap
- model-based bootstrap
- evaluation

Covariate interpretation for any model

Partial residual plots (see e.g. Wikipedia)
Regression based methods, plot *Residuals* of full model plus the covariate effect $\hat{\beta}_i X_i$ against the values of covariate X_i

$$Residuals \;+\; \hat{\beta}_i X_i \quad versus \quad X_i$$

Partial dependence plots Hastie et al. 2009, chapt. 10.13.2
Any “black box” learning model, dependence of covariate on response after *accounting* (not *ignoring*) for the effects of all other covariates.
Approximation of function by:

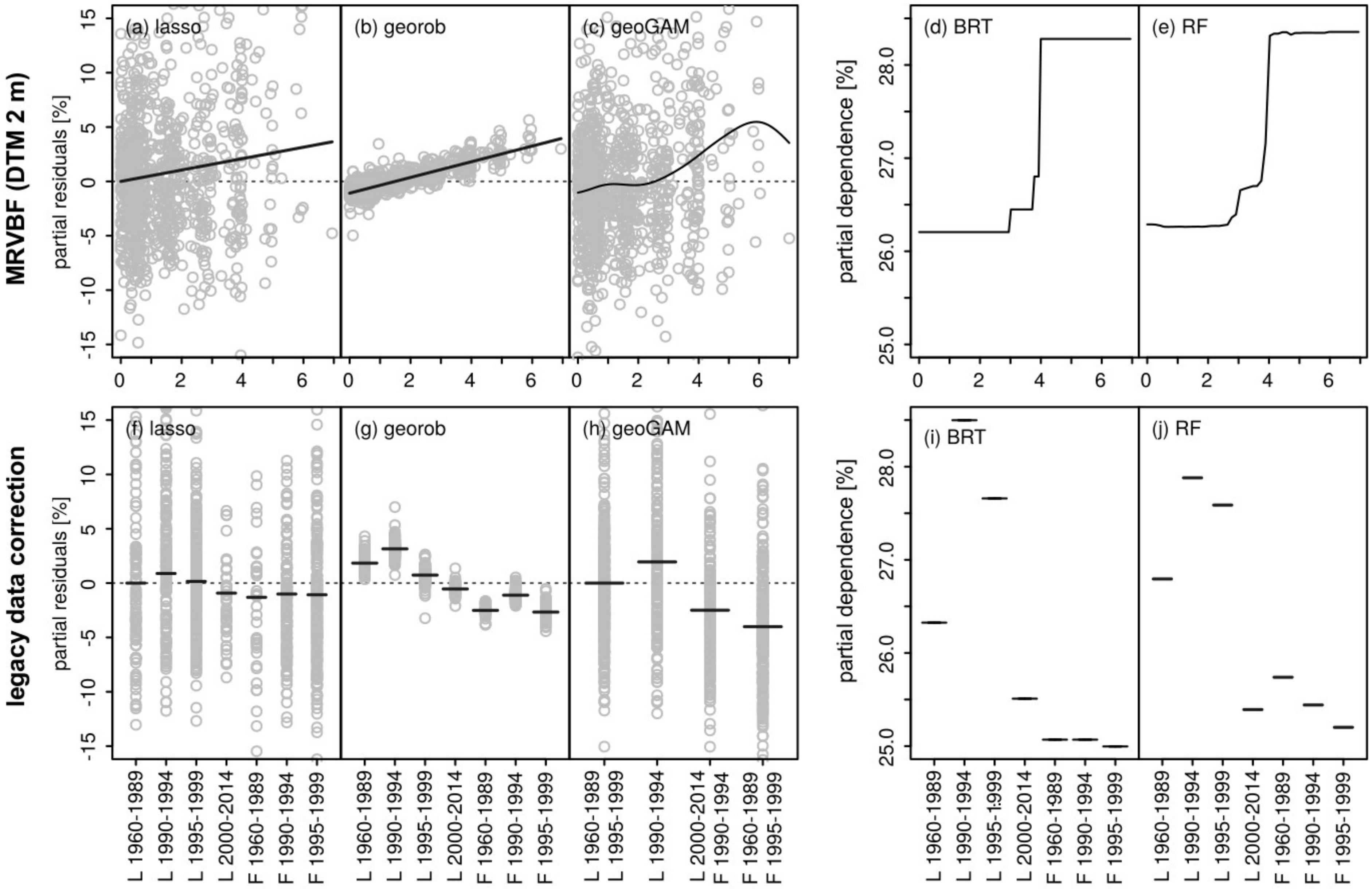
$$\bar{f}_S(X_S) = \frac{1}{N}\sum_{i=1}^{N} f(X_S, x_{iC}),$$

Take care with interpretation: If many covariates it is difficult to choose which to interpret. If collinearity in data set, covariates might replace each other …

Covariate interpretation for any model

Partial residual plots

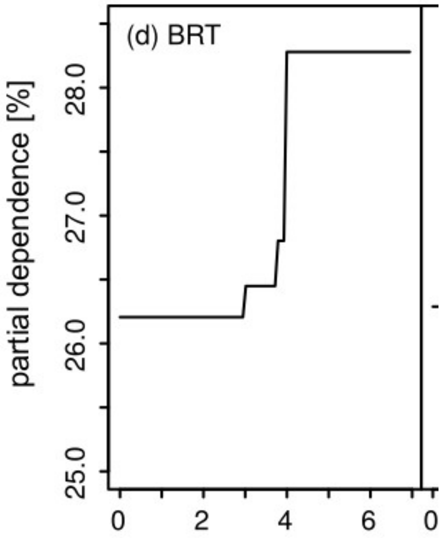
Partial dependence plots



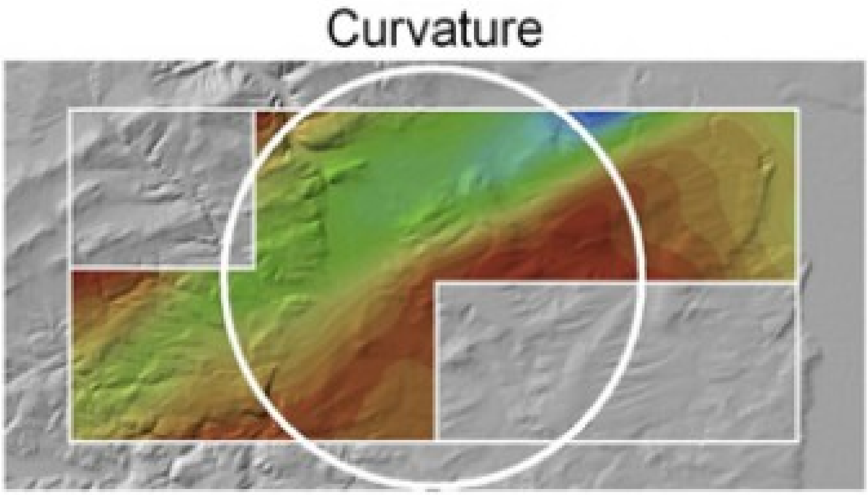
Nussbaum et al. 2017b

Spatial covariate interpretation

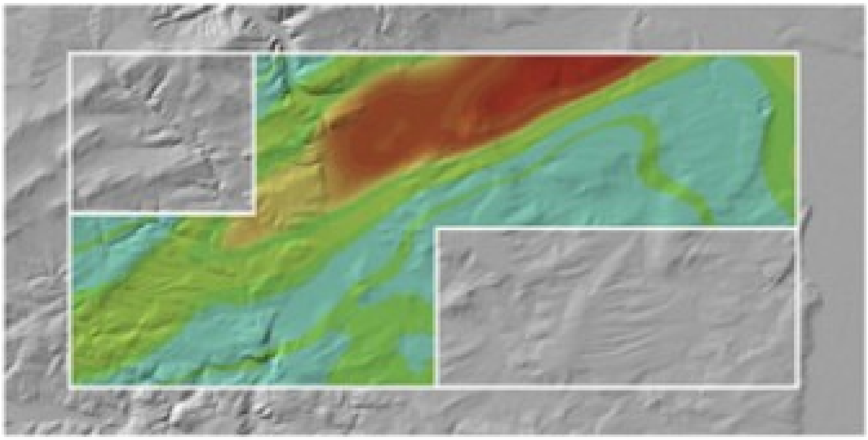
Create maps from relationship:



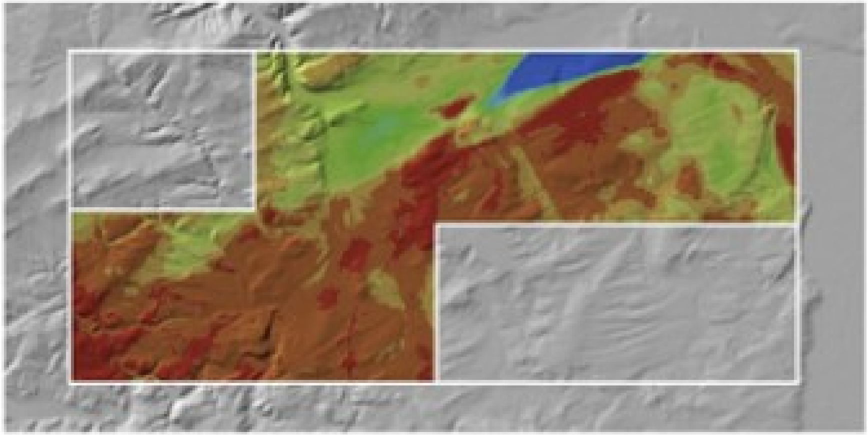
Original covariate



Partial dependence



Local importance



Behrens et al. 2014

Overview

Model Selection

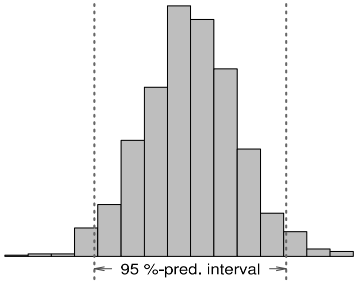
- linear regression
- with lasso
- with covariate importance

Model interpretation

- partial residual plots
- partial dependence plot
- partial dependence maps

Uncertainty

- non-parametric bootstrap
- model-based bootstrap
- evaluation



Confidence intervals vs. prediction intervals

Confidence intervals

Intervals of confidence for the estimate of a population mean. Considers uncertainty in our estimation of β (based on standard error of coefficient).

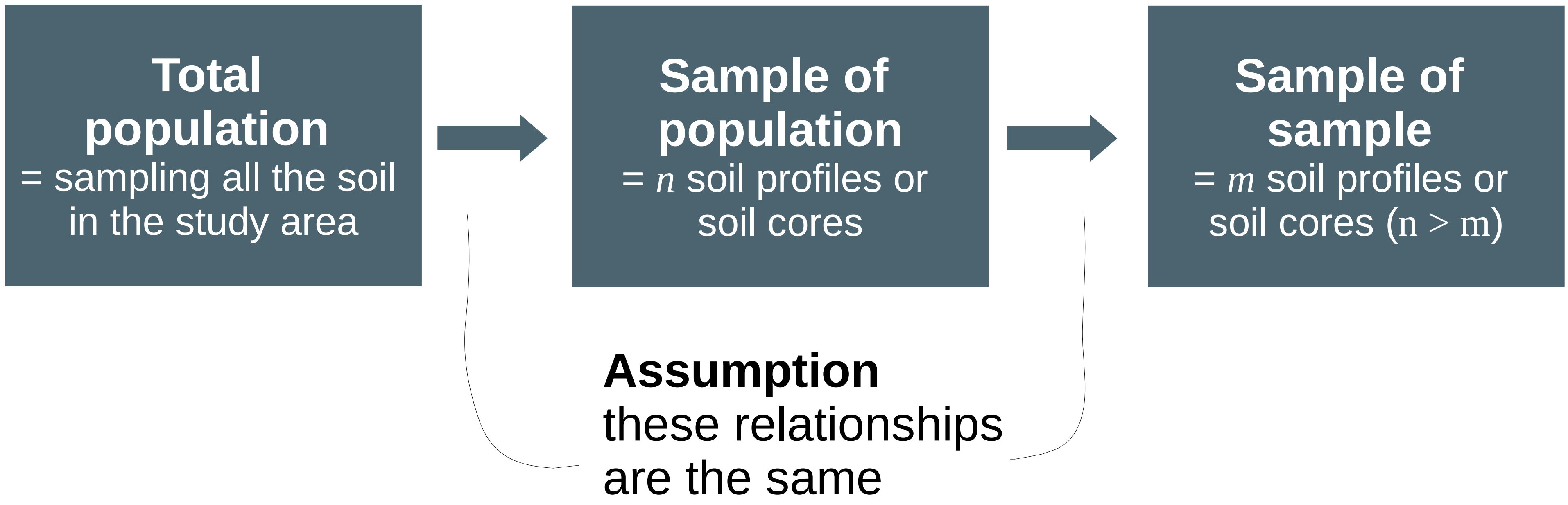
$$y = X\hat{\beta} + \epsilon$$

Prediction intervals

Intervals of confidence for the estimate of a new observation. Considers uncertainty estimated β and the variation the model does not account for (based on standard error of coefficient and residual error).

$$y = X\hat{\beta} + \epsilon$$

Non-parametric bootstrap – idea



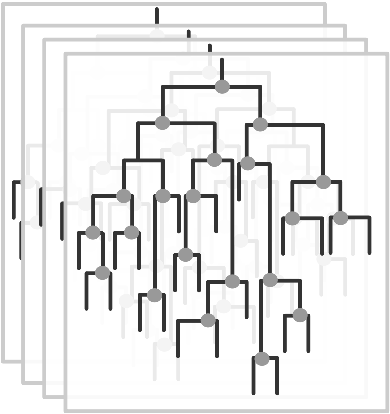
Total population
= sampling all the soil in the study area

Sample of population
= n soil profiles or soil cores

Sample of sample
= m soil profiles or soil cores ($n > m$)

Assumption
these relationships are the same

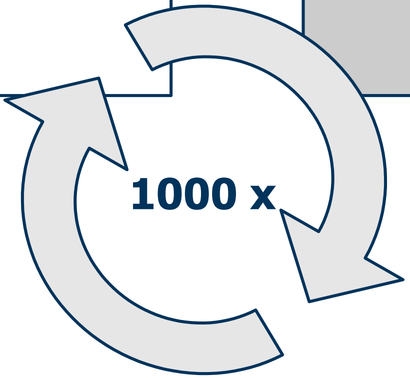
- Idea: by **resampling** our sample many times (e.g. 1000x) we can approximate properties of the **distribution** of the total population.
- Useful to
  - create model ensembles
    bagging = boostrap aggregation
  - estimate uncertainty for any model

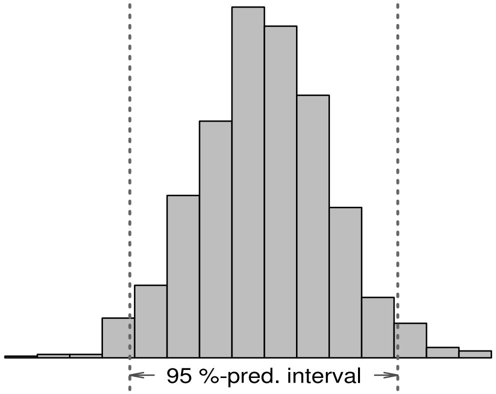


Simulate predictive distribution non-parametric bootstrapping ("model-free")

1. resample data with replacement to get a dataset with the same n

2. fit model to resampled data, compute predictions for new observations





95 % prediction interval

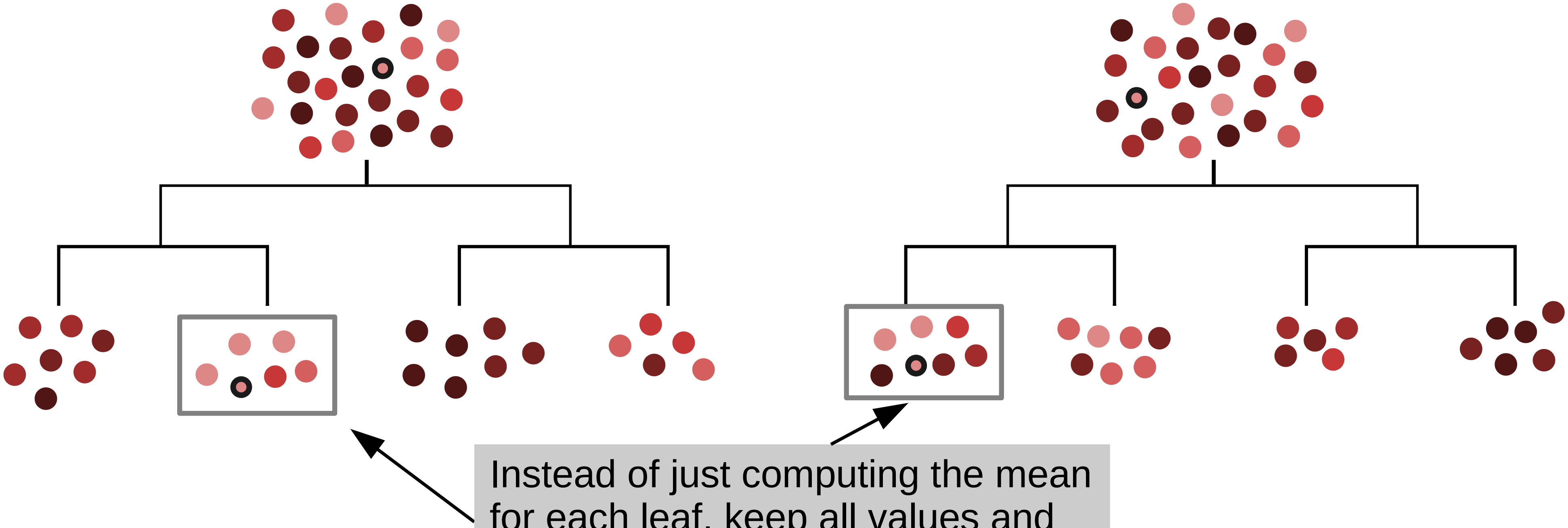
“Quantile regression forest” as non-parametric bootstrap

Citation: Meinshausen N, 2006. Quantile regression forests. Journal of Machine Learning Research, 7, 983–999.

- Bootstrap aggregation in random forest: each tree is fitted to a resampled dataset
- Keep all observations in the final tree leaves
- Get distribution from observations that were in leaves together

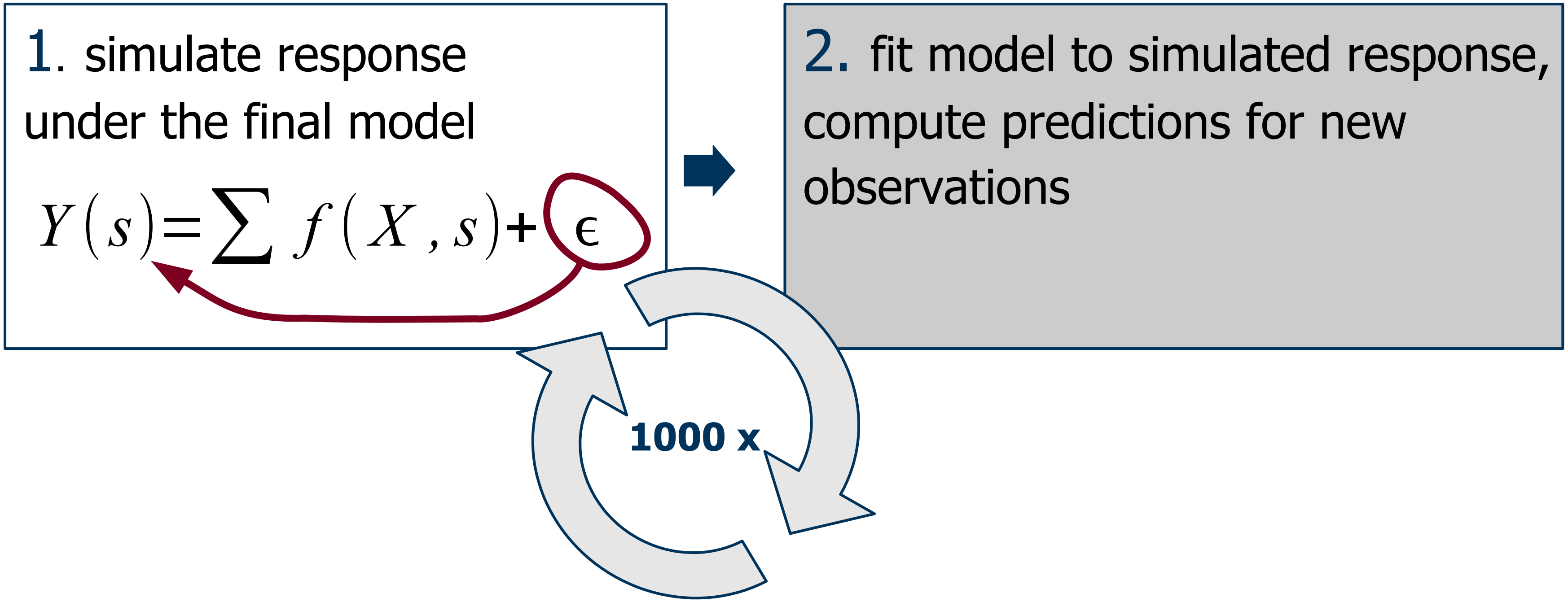
1. tree on resampled data

2. tree on resampled data .....



Instead of just computing the mean for each leaf, keep all values and compute e.g. 95 % intervals

Simulate predictive distribution by <u>model-based</u> bootstrapping

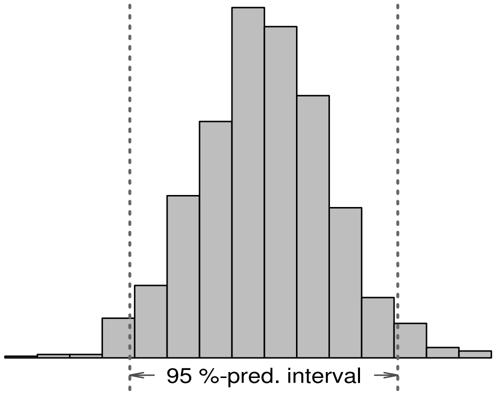


1. simulate response under the final model

$$Y(s)=\sum f(X,s)+\epsilon$$

2. fit model to simulated response, compute predictions for new observations

1000 x



95 % prediction interval

Model-based bootstrapping
(also “parametric bootstrap”)

Davidson and Hinkley, 1997

For 1000 repetitions, do:

1. Simulate new response *Y*(s)* with the fitted value *f(x(s))* plus a randomly chosen residual sampled from ε (or from normal distribution with same σ, μ as residual distribution):

$$Y(\mathbf{s})^* = \hat{f}(\mathbf{x}(\mathbf{s})) + \epsilon$$

2. Fit model to new response *Y*(s)*

3. Compute prediction error for new location s+ with again randomly sampled ε

$$\delta^*_+ = \hat{f}(\mathbf{x}(\mathbf{s}_+))^* - (\hat{f}(\mathbf{x}(\mathbf{s}_+)) + \epsilon)$$

Two-sided prediction intervals: $[\hat{f}(\mathbf{x}(\mathbf{s}_+)) - \delta^*_{+\,(1-\alpha)}\,;\ \hat{f}(\mathbf{x}(\mathbf{s}_+)) - \delta^*_{+\,(\alpha)}].$

Which bootstrap should I use?

non-parametric

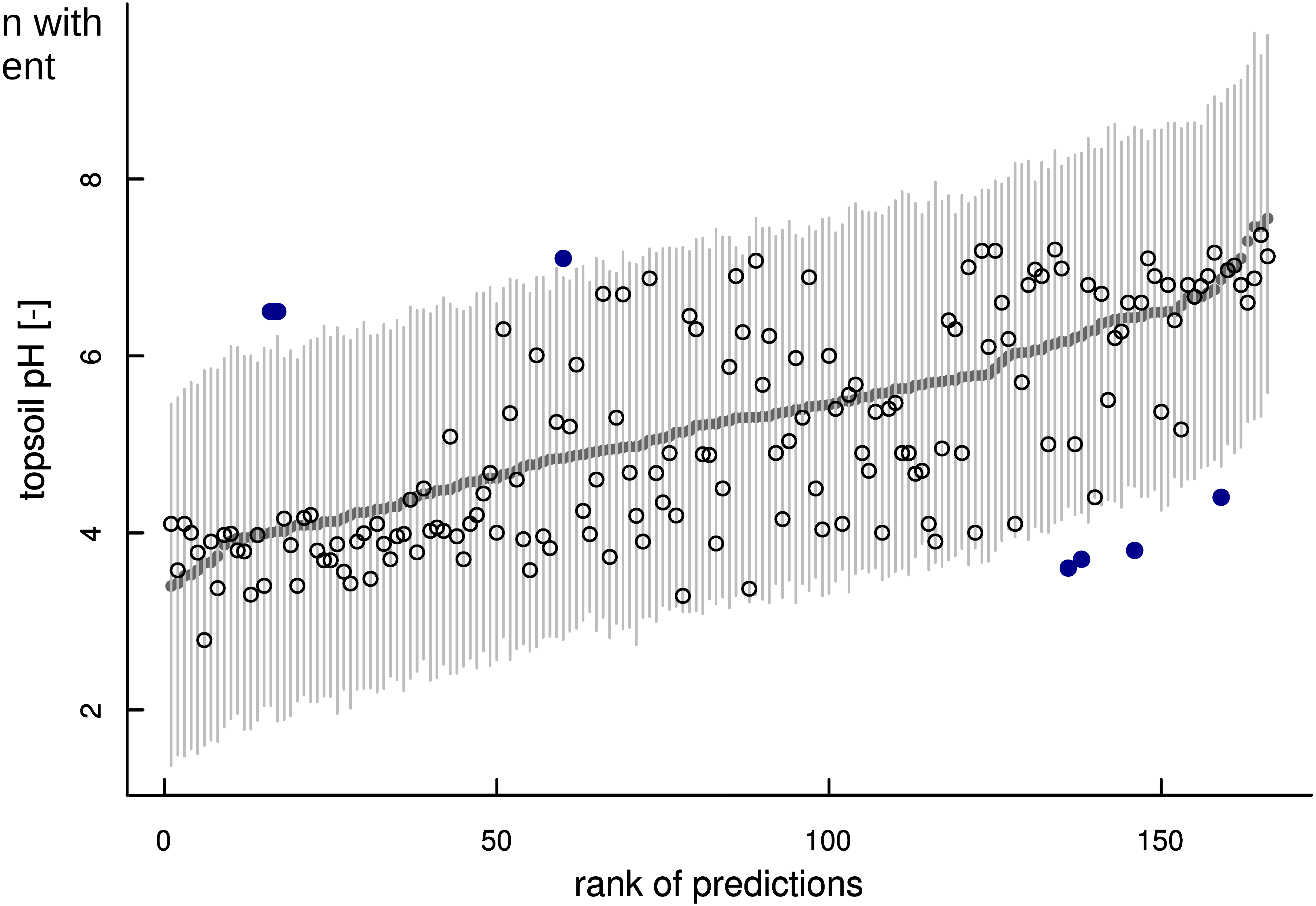
- ✓ No assumptions about a model or error distributions
- ✗ Likely to fail with small datasets:
  - underestimates variance
  - lacks to depict distribution of full population
- ✗ mostly no software
- ✓ easy to implement
- ✗ computationally intensive (mainly CPU)

model-based / parametric

- ✗ Assumes e.g. Gaussian errors
- ✗ Prediction intervals are given this model even if model is wrong
- ✓ More suitable for small datasets
- ✗ mostly no software
- ✗ a bit tricky to implement
- ✗ computationally intensive (mainly CPU)

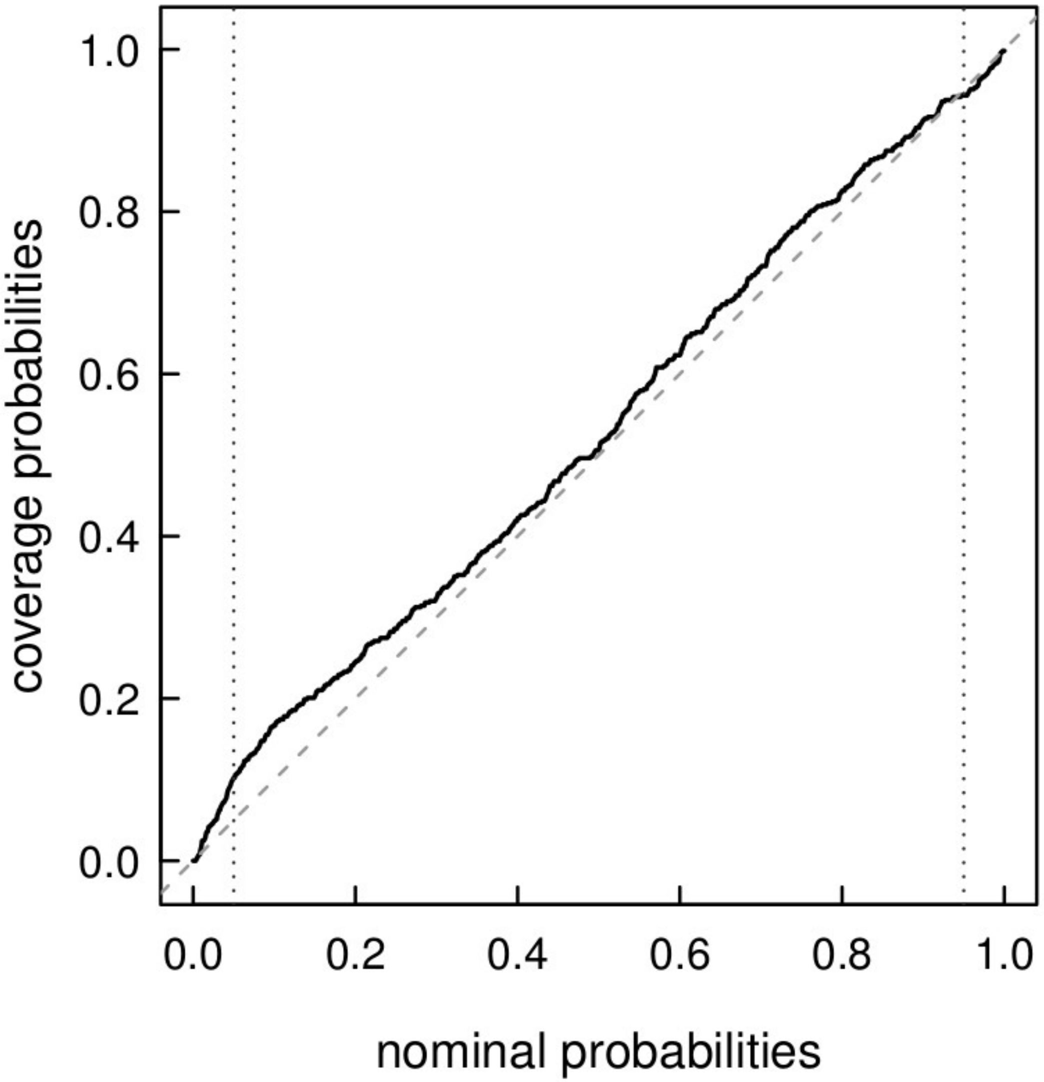
Performance plots for e.g. 95% prediction intervals

Evaluation with
independent
test data



Performance plots for complete predictive distribution

one-sided prediction intervals of bootstrapped distribution against the nominal probabilities

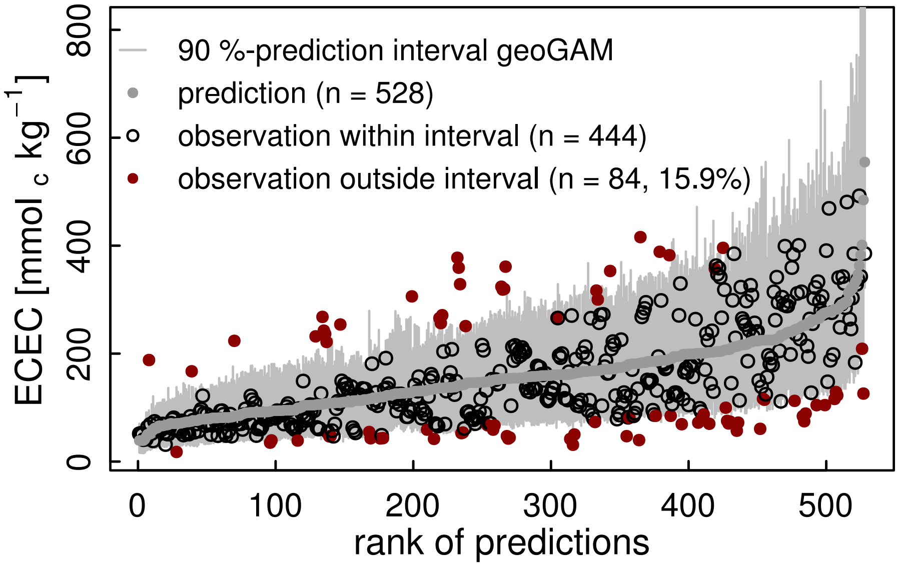


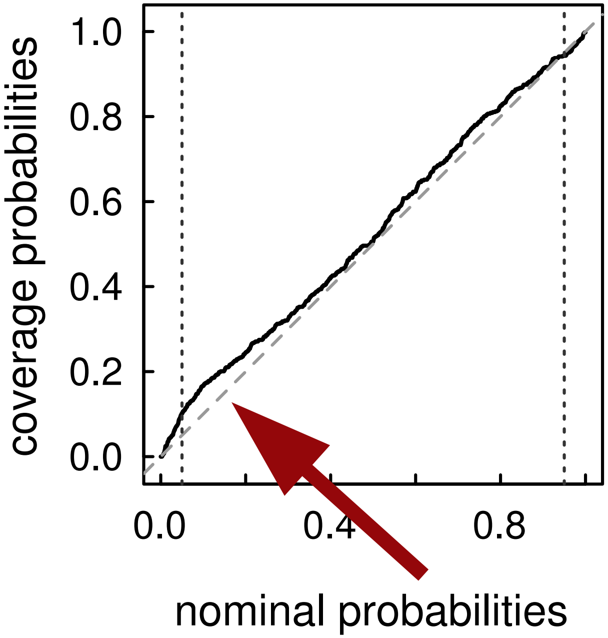
Evaluation of prediction intervals

coverage 90 %-intervals

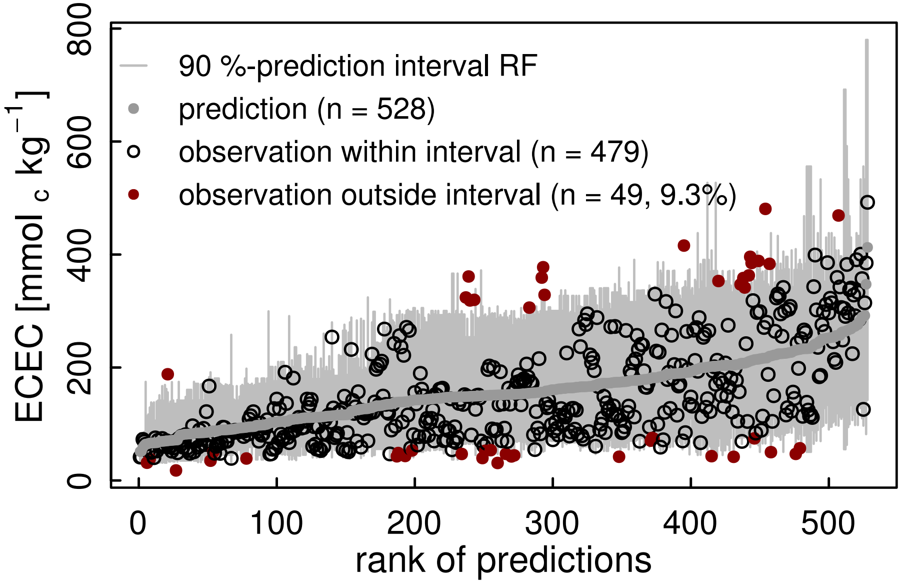
coverage one-sided intervals

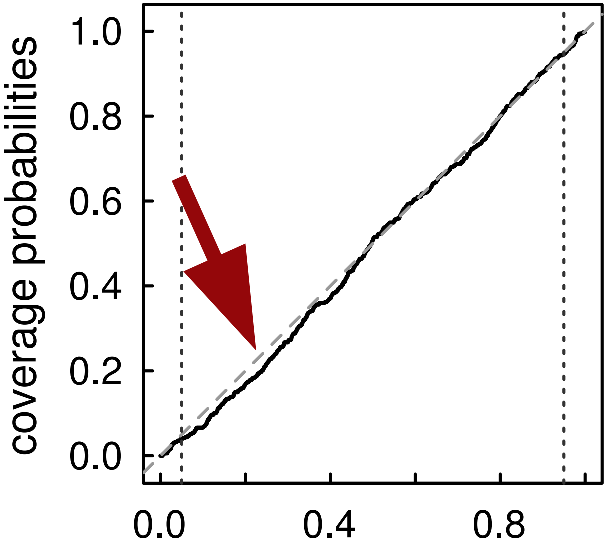
geoGAM bootstrap

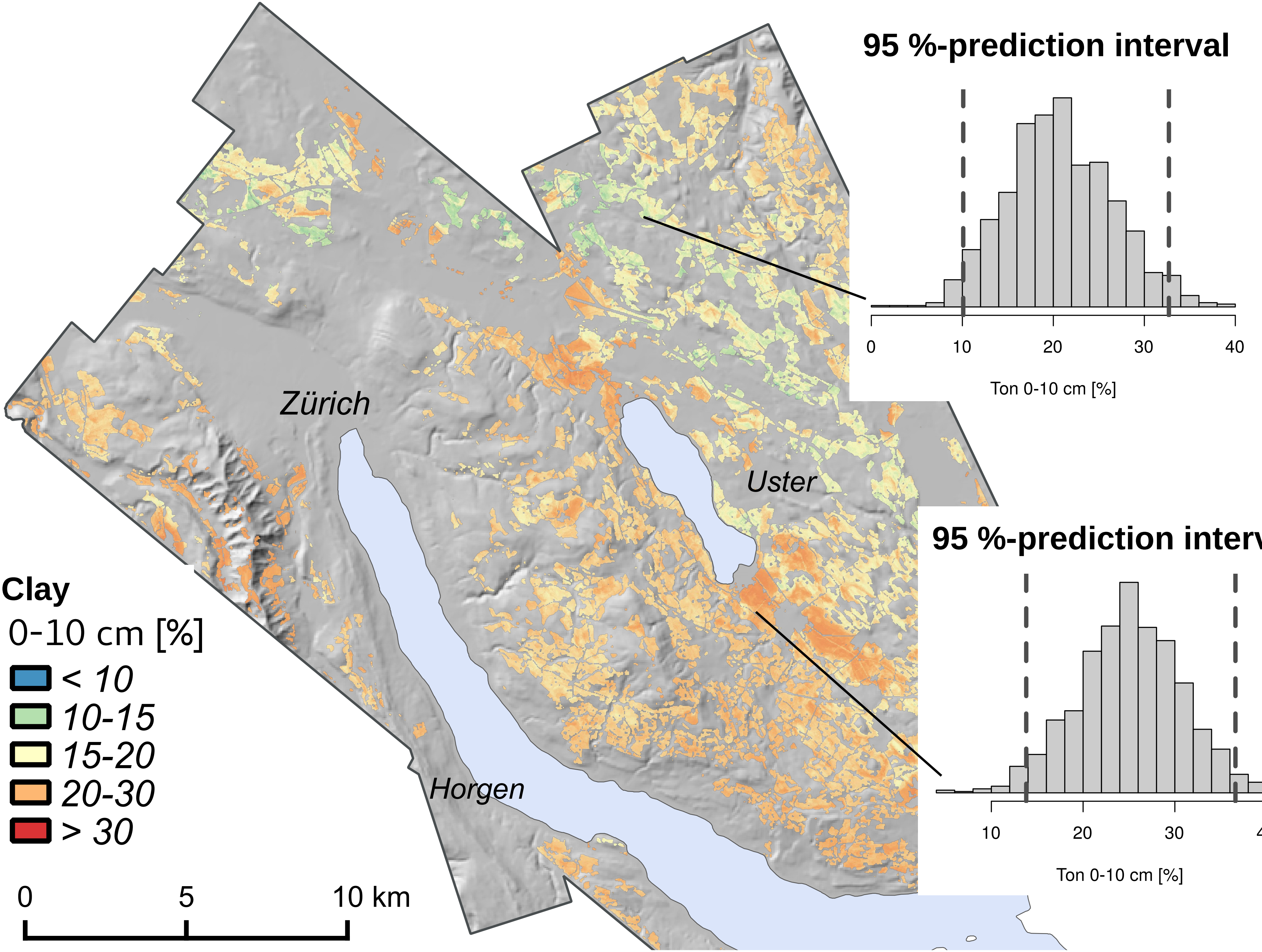




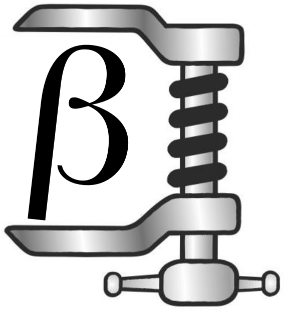
quantile regression forest

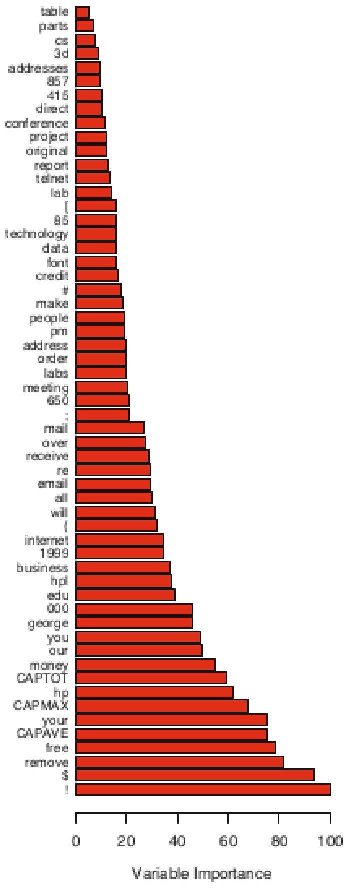


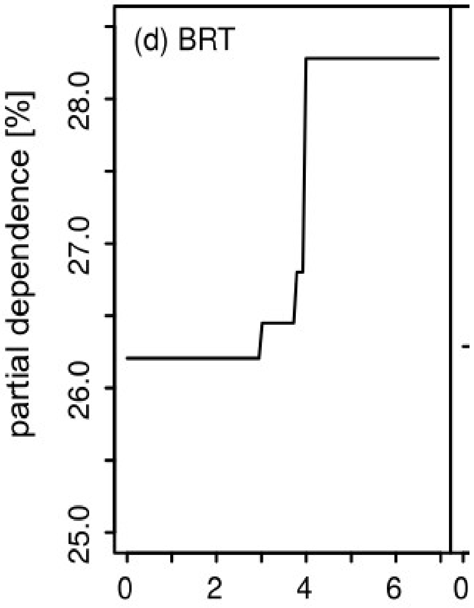


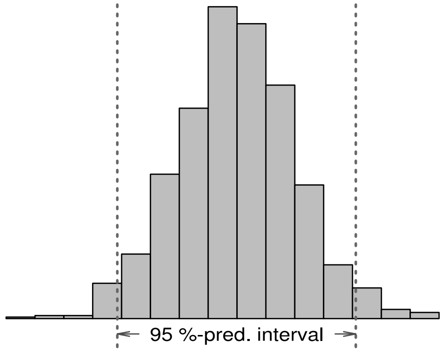

95 %-prediction interval
0
10
20
30
40
Ton 0-10 cm [%]
Zürich
Uster
95 %-prediction interv
10
20
30
4
Ton 0-10 cm [%]
Horgen
Clay
0-10 cm [%]
< 10
10-15
15-20
20-30
> 30
0
5
10 km

Summary of lecture









My personal tips
for your applied machine learners career:

- **Do not believe there is the ONE solution** (e.g. one model that never fails.). So make sure you understand advantages and disadvantages for your application.
- **Do not neglect classical statistics.** Without understanding linear models you will never master machine learning!
- If a method is hard to understand to you, **please use a simpler one**, that you feel secure with! Your credibility and credibility of your data product is at stake.
- **Do not neglect domain knowledge**. Most likely consulting an expert of your field might improve your model more than the 100redst tuning of the latest fancy method, e.g. by inlcuding a neglected covariate!

Additional literature ...

Davison & Hinkley 1997, on bootstrap methods, a bit technical though.

Davidson, A. C. and Hinkley, D. V.: Bootstrap Methods and Their Applications, Cambridge University Press, Cambridge, doi:10.1017/cbo9780511802843, 1997.

Tutz, 2012, very good book on categorical responses, mostly parametric methods, some ML described, comes with many examples and a R package:

Tutz, G.: Regression for Categorical Data, Cambridge University Press, doi:10.1017/cbo9780511842061, 2012.

Wilks 2011, Useful book for validation measures including for uncertainty, see chapter 8 and R package "verification":

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, 3 edn., 2011.

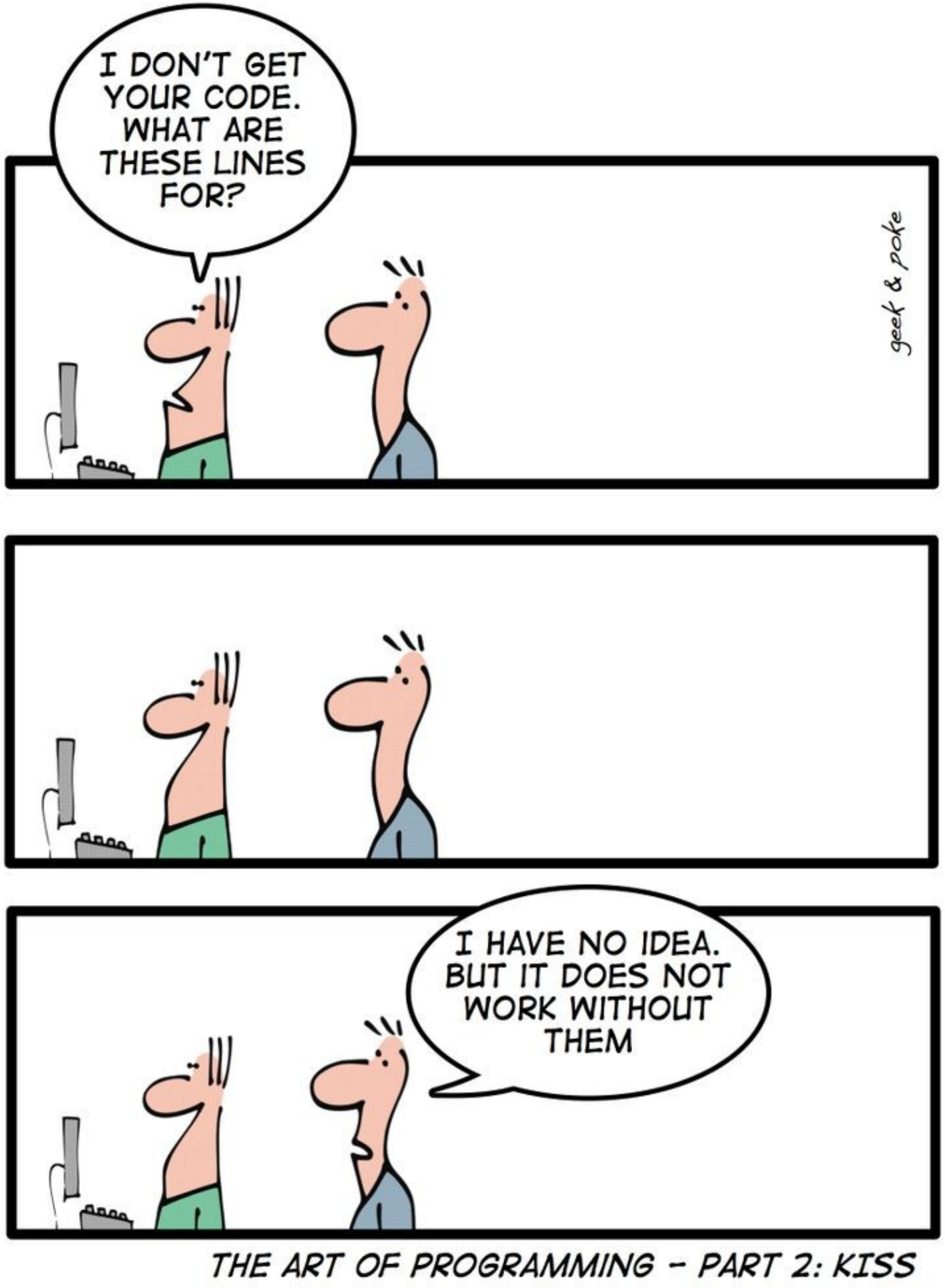
Practical training

You will learn:

model selection

model interpretation

and much more!


I DON'T GET YOUR CODE. WHAT ARE THESE LINES FOR?
I HAVE NO IDEA. BUT IT DOES NOT WORK WITHOUT THEM
geek & poke
THE ART OF PROGRAMMING - PART 2: KISS

http://geekandpoke.typepad.com/geekandpoke/2009/07/the-art-of-programming-part-2.html