Mastering Machine Learning for spatial prediction II

GEOSTAT 2017
Friday 9-10:30

Bern University
of Applied Sciences

Madlene Nussbaum

Today: Overview

Selection with covariate importance

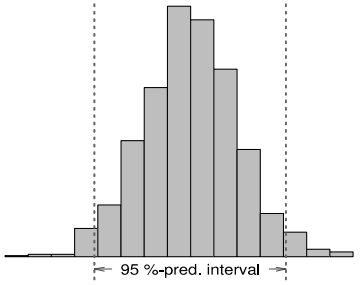
Model interpretation

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

Uncertainty

- bootstrapping
- evaluation

Exercises



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:

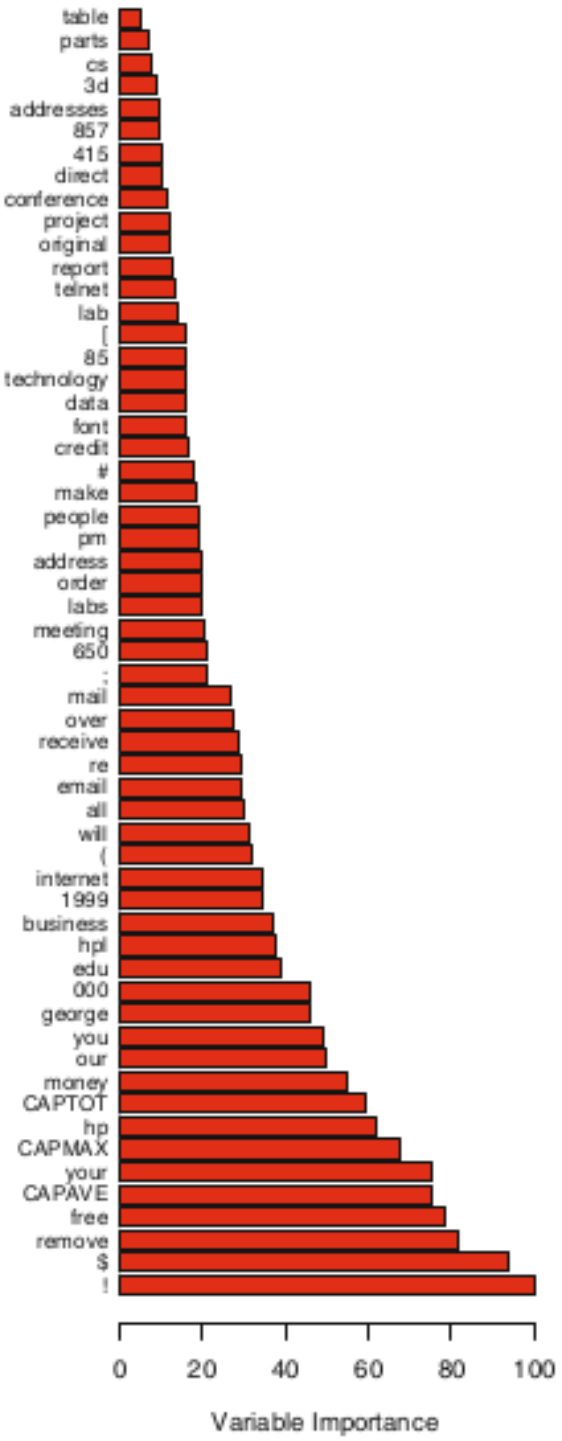
- Mean decrease in **accuracy** (`type=1` in `importance()`): oriented on fitting the data. How much do we reduce error by using this covariate at this split?

  → this importance is used for boosted trees.

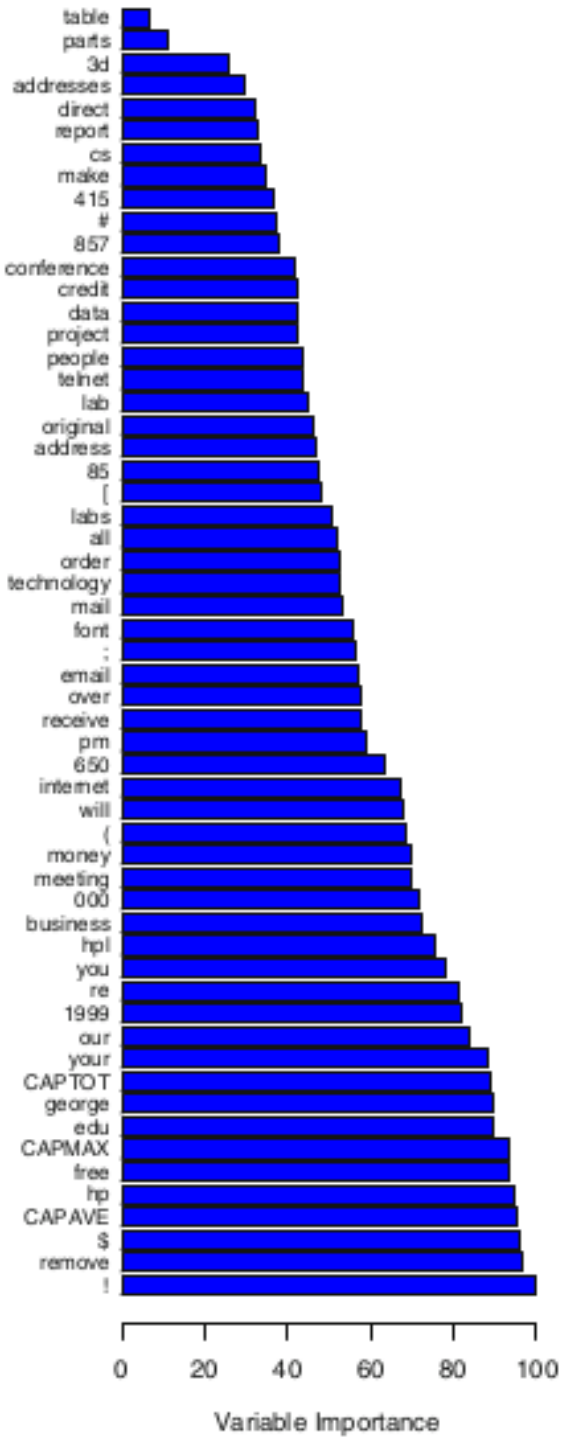
- Mean decrease in **node impurity** (`type=2` in `importance()`), permutation based, oriented on predictions: How much worse do OOB predictions get if we randomly permutate a covariate?

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

Loss in accuracy



OOB, Permutation



Selection, very simple:

Recursive backward elimination

1) Remove covariate(s) with lowest importance

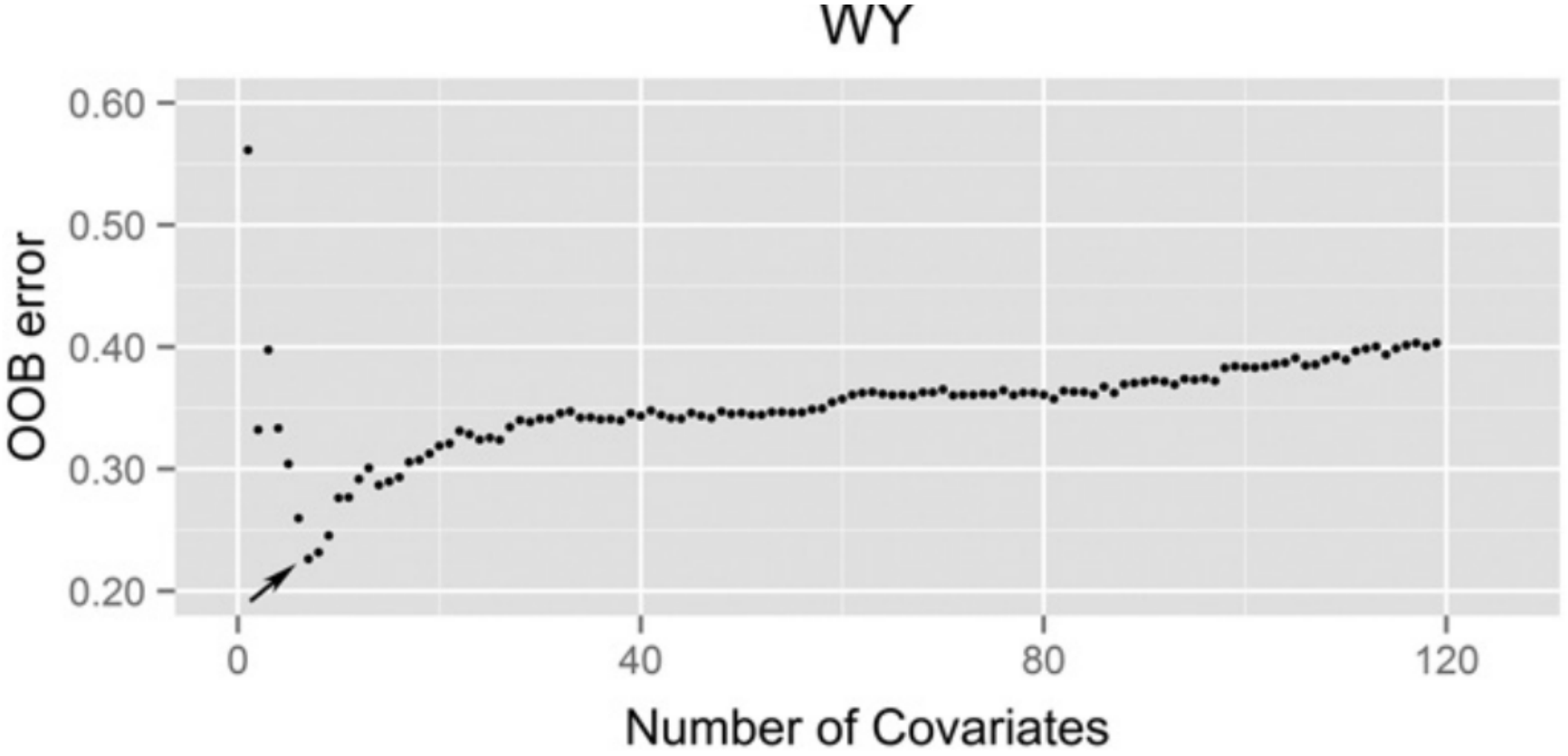
2) Refit random forest with remaining

Find optimum number of covariates by minimizing OOB error.

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

Hastie et al. 2009

Covariate selection for random forest
(or boosted trees)



Brungard et al. 2015

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

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

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:

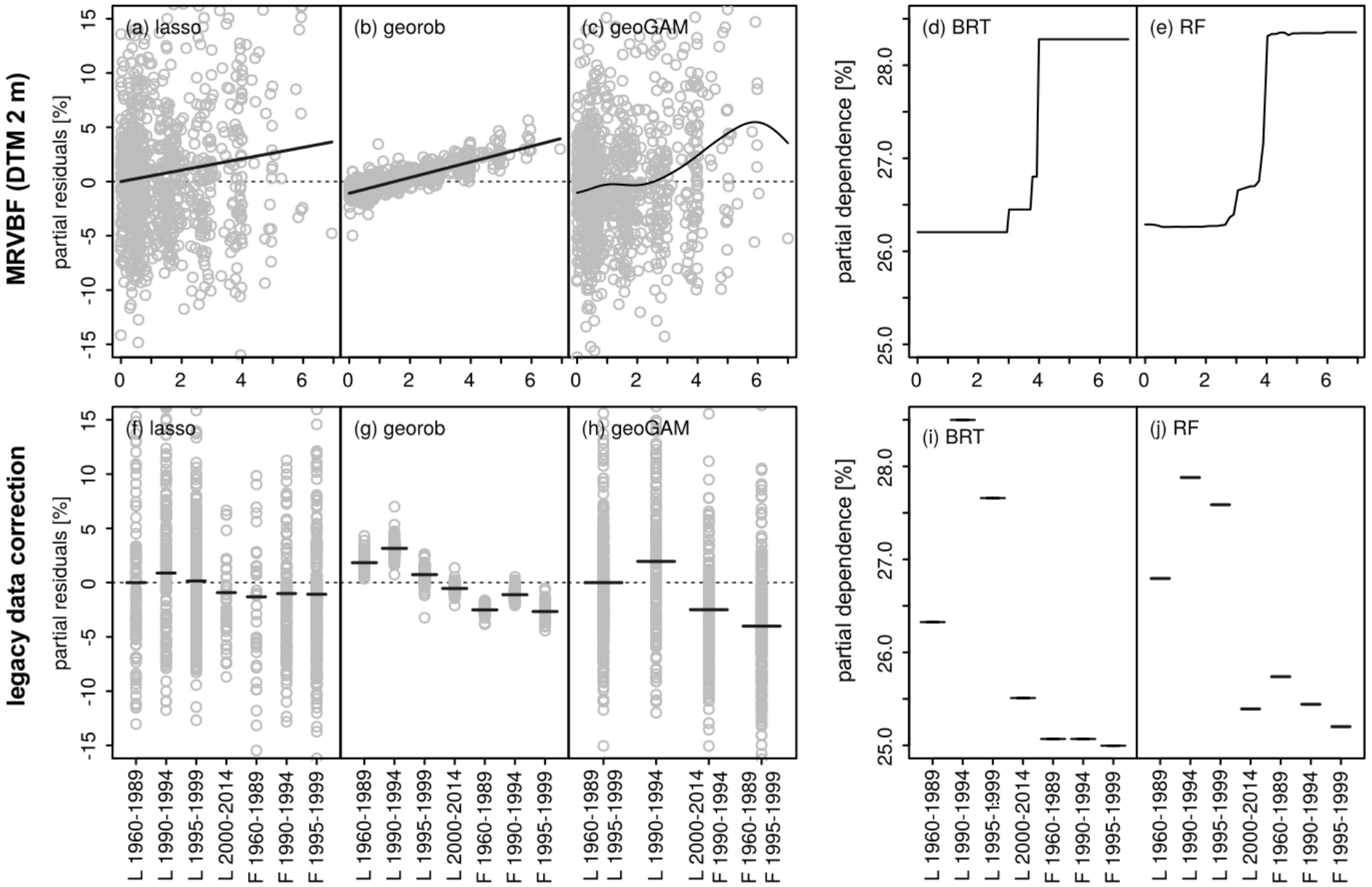
$$\bar{f}_S(X_S) = \frac{1}{N}\sum_{i=1}^{N} f(X_S, x_{iC}),$$

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

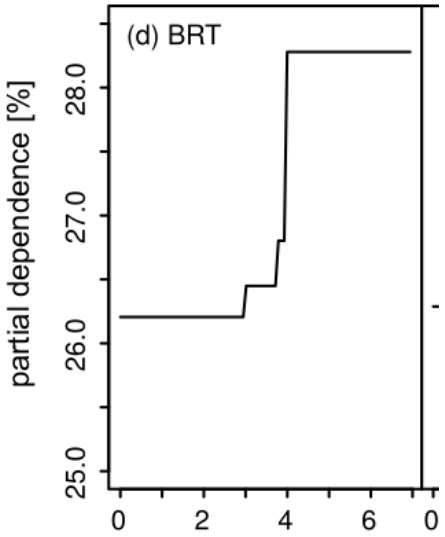
Partial dependence plots



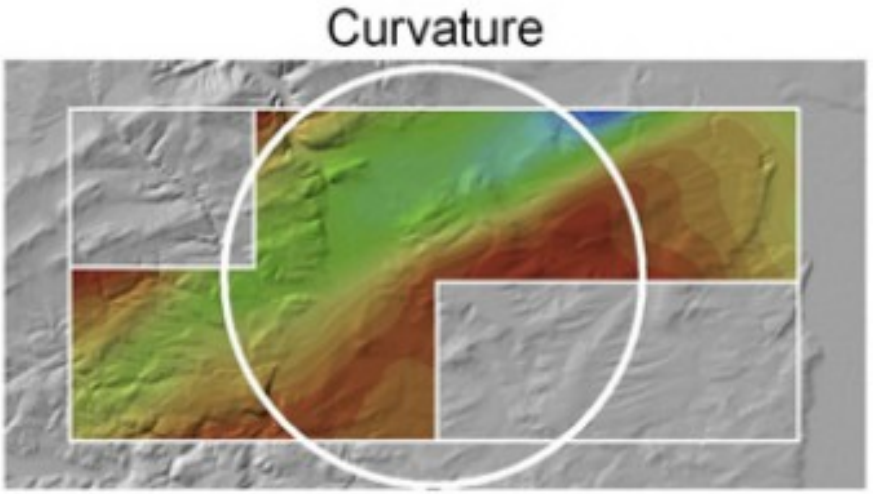
Nussbaum et al. 2017b

Spatial covariate interpretation

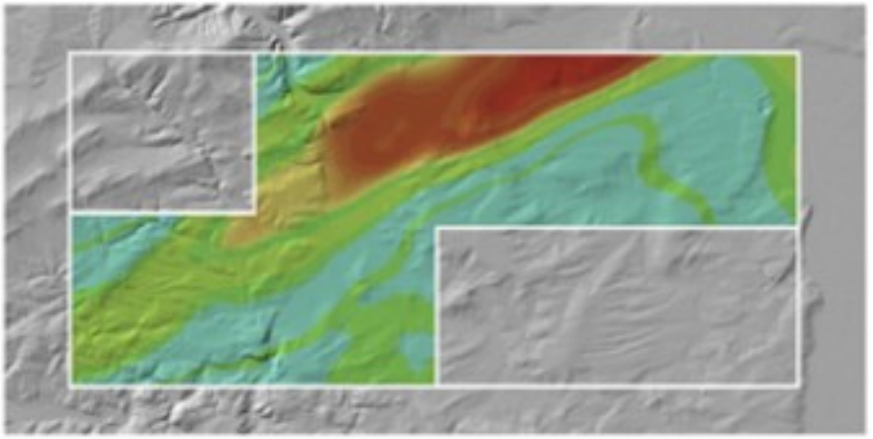
Create maps from relationship:



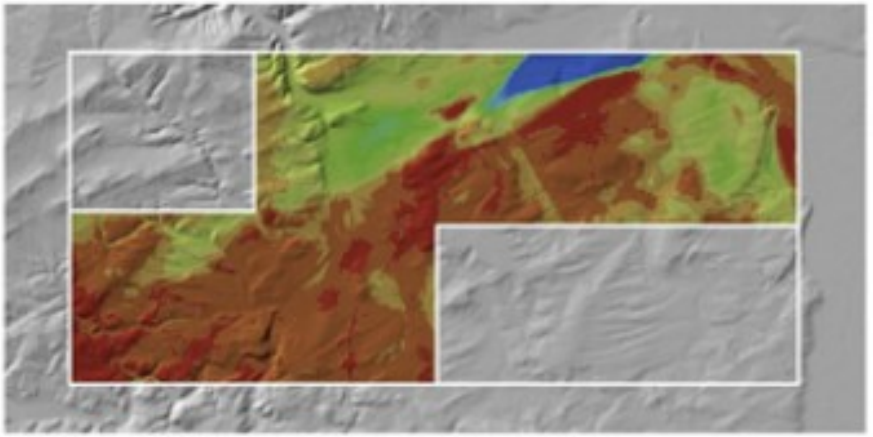
Original covariate



Partial dependence



Local importance



Behrens et al. 2014

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).

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

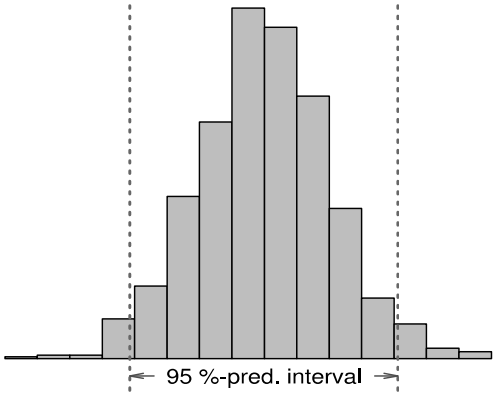
Simulate predictive distribution by model-based bootstrapping

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

Simulate predictive distribution by model-based bootstrapping

Davidson and Hinkley, 1997

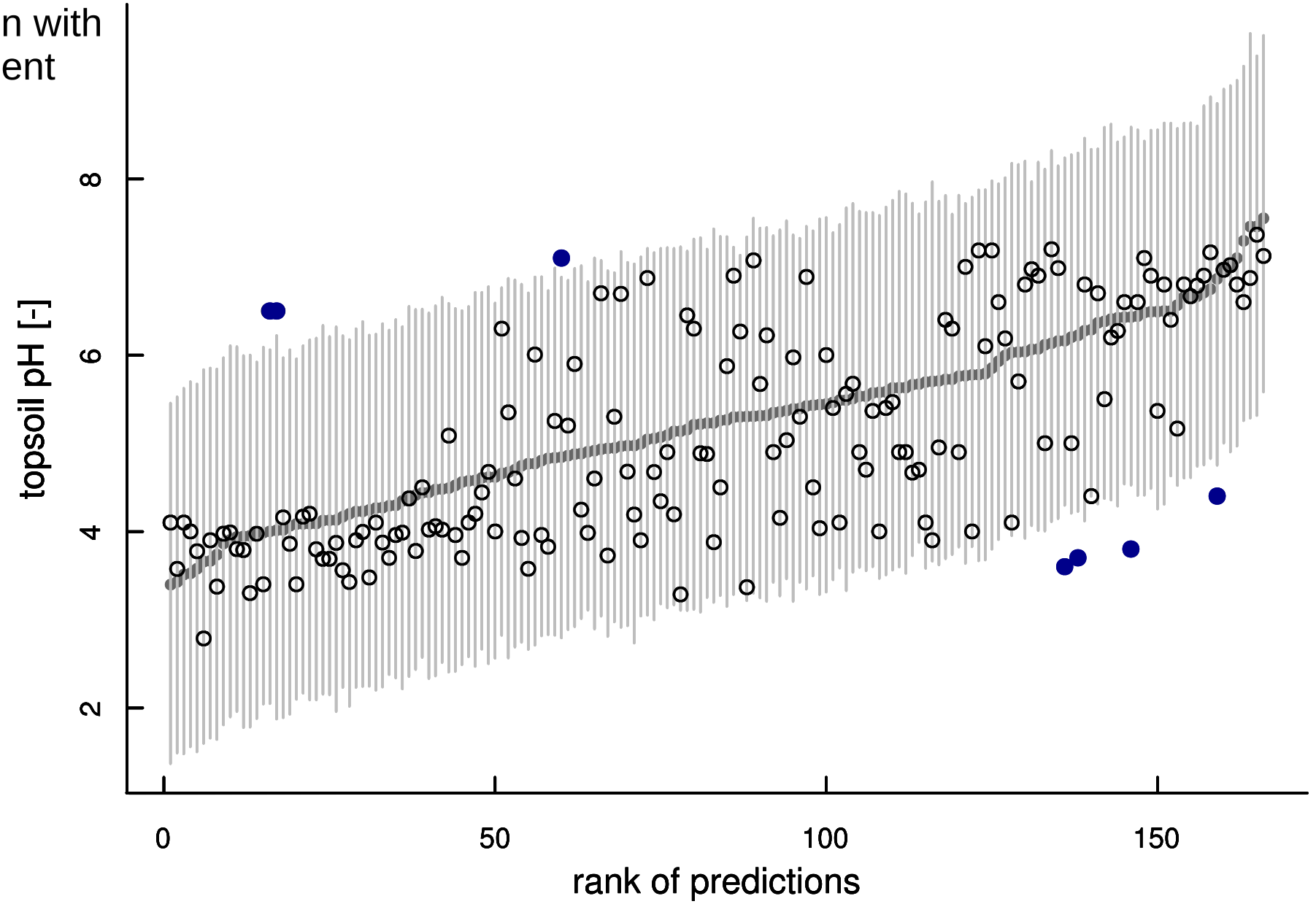
For 1000 repetitions, do:

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

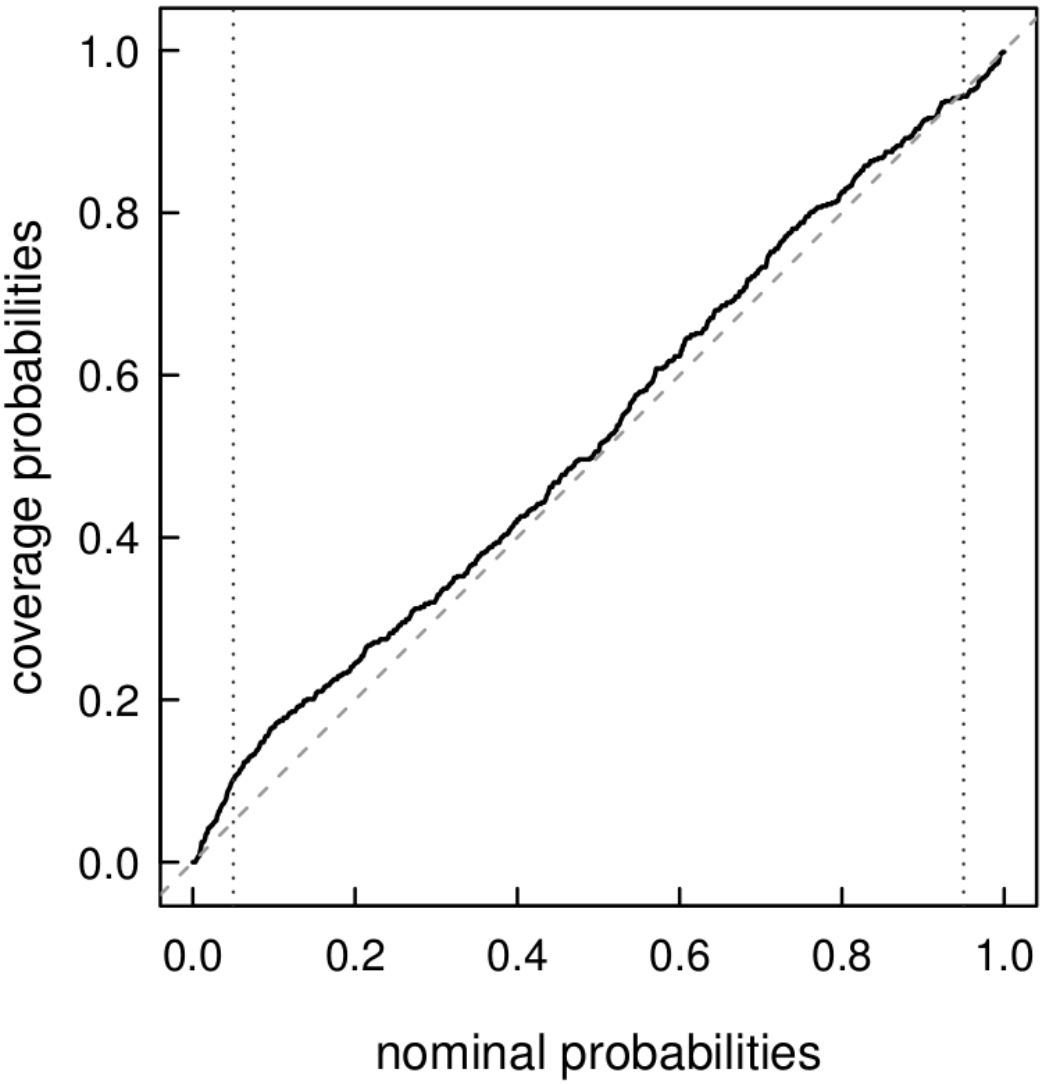
Performance plots for e.g. 95% prediction intervals

Evaluation with independent test data



Performance plots for complete predictive distribution

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

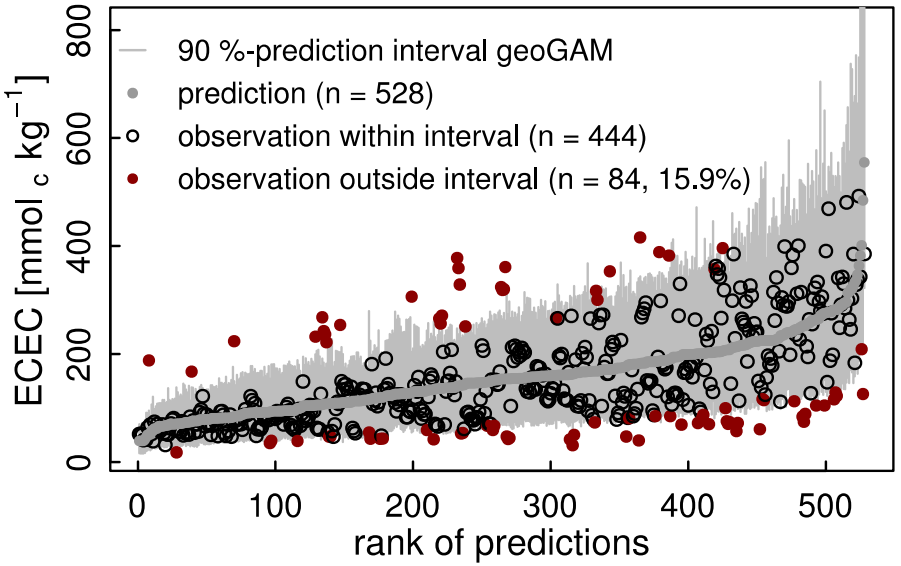


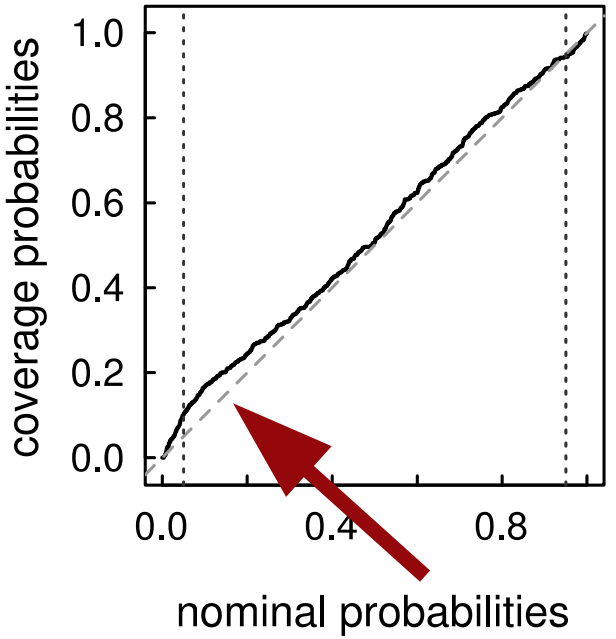
Evaluation of prediction intervals

coverage 90 %-intervals

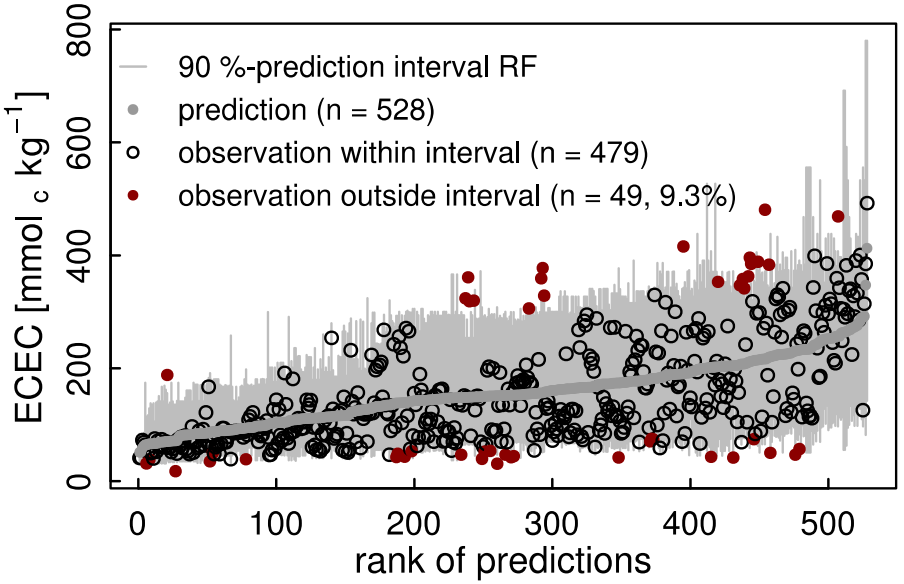
coverage one-sided intervals

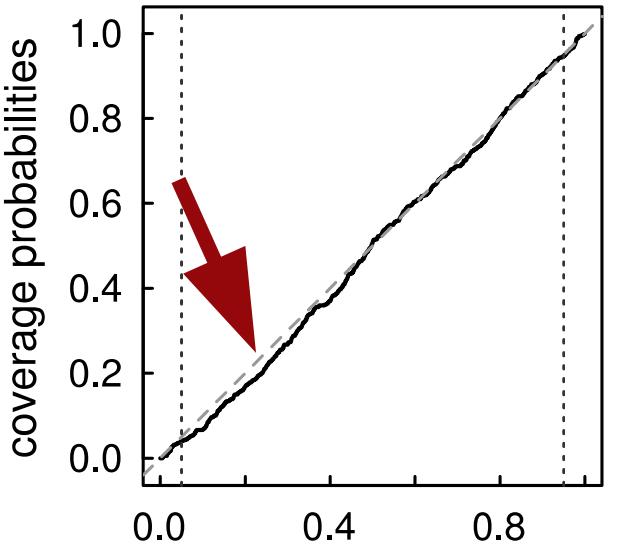
geoGAM bootstrap

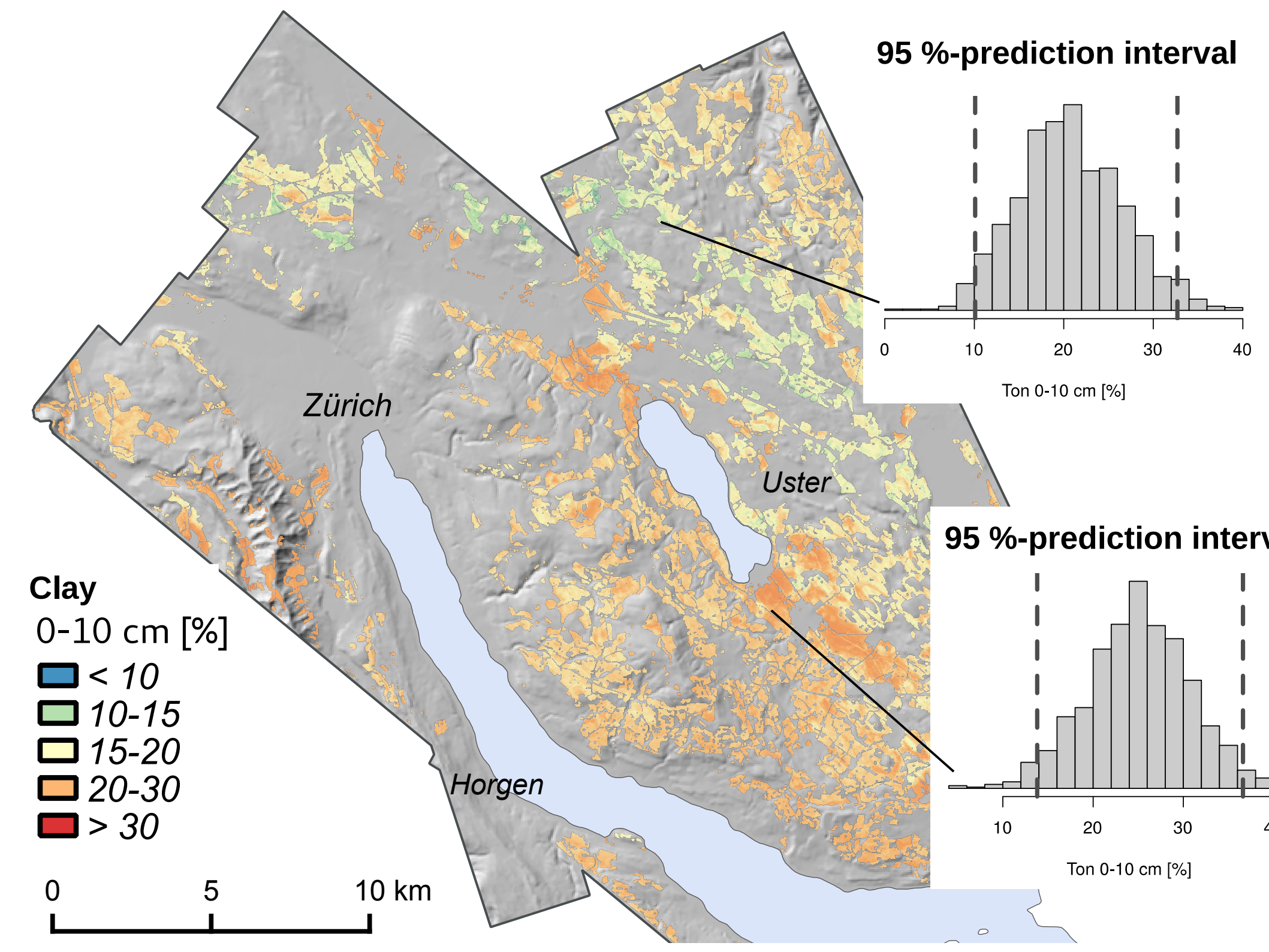




quantile regression forest






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

Summary

Spatial modelling

- define requirements
- get overview

Get to know ..

- Lasso
- Gradient boosting
- Model averaging

Selection with covariate importance

Model interpretation

- partial residual plots
- partial dependence plot & maps

Uncertainty

- bootstrapping
- evaluation

Be able to judge if computing model averaging on 78 methods found in Package caret is a sensible thing to do …