Mastering Machine Learning for spatial prediction I

GEOSTAT 2017
Thursday 11-12:30

Bern University of Applied Sciences

Madlene Nussbaum

Objectives ...

- Get an **overview**, understand ML techniques
- Get to know quite different approaches in detail
- Move away from **ML = black box**
- Get to know how to compute and evaluate **uncertainty**
- **Be critical!**

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

Overview

Spatial modelling

- define requirements
- get overview

Get to know ..

- Lasso
- Gradient boosting
- Model averaging

Exercises

Literature

Books:

Very good and detailed book on ML, although quite complex:
Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning; Data Mining, Inference and Prediction, Springer, New York, 2 edn., 2009. with examples and data in R package ElemStatLearn, https://cran.r-project.org/web/packages/ElemStatLearn/index.html

Extended book on boostrapping:
Davison, A. C. and Hinkley, D. V.: Bootstrap Methods and Their Applications, Cambridge University Press, Cambridge, doi:10.1017/cbo9780511802843, 1997.

Very good book on categorical responses, mostly parametric methods, some ML described, comes with R package:
Tutz, G.: Regression for Categorical Data, Cambridge University Press, doi:10.1017/cbo9780511842061, 2012.

Useful book for validation measures including for uncertainty, see chapter 8 and R package “validation”:
Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, 3 edn., 2011.

Some articles the slides are referring to:

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.-X., and Scholten, T.: Hyper-scale digital soil mapping and soil formation analysis, Geoderma, 213, 578–588, doi:10.1016/j.geoderma.2013.07.031, 2014.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., and Edwards Jr., T. C.: Machine learning for predicting soil classes in three semi-arid landscapes, Geoderma, 239–240, 68–83, doi:10.1016/j.geoderma.2014.09.019, 2015.

Hothorn, T., Müller, J., Schröder, B., Kneib, T., and Brandl, R.: Decomposing environmental, spatial, and spatiotemporal components of species distributions, Ecological Monographs, 81, 329–347, 2011.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M., and Papritz: Evaluation of digital soil mapping approaches with large sets of environmental covariates, SOIL Discussions, 2017, 1–32, doi:10.5194/soil-2017-14, URL http://www.soil-discuss.net/soil-2017-14/, in review, 2017a.

Nussbaum, M., Walthert, L., Fraefel, M., Greiner, L., and Papritz, A.: Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models, SOIL Discussions, 2017, 1–32, doi:10.5194/soil-2017-13, URL http://www.soil-discuss.net/soil-2017-13/, in review, 2017b.

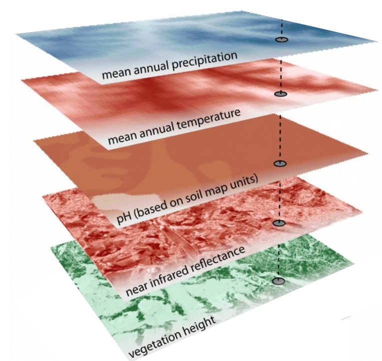
Spatial predictions ...

For example:
Digital soil mapping

texture
density
gravel
soil depth
drainage
pH, ECEC
SOC

300-1400
locations with soil properties in

2–4 soil depth
3 study areas



300-500
environmental covariates

48
statistical models

Requirements

A spatial prediction method should …

- model **nonlinear** relations
- consider **spatial** autocorrelation
- model continuous and categorical responses
- handle **numerous** correlated **covariates** without overfitting calibration data
- **automatically** build models with **good predictive power**
- preferably result in **sparse model**
- accurately quantify **accuracy** of **predictions**
- give prediction **uncertainty**

Bias-Variance tradeoff

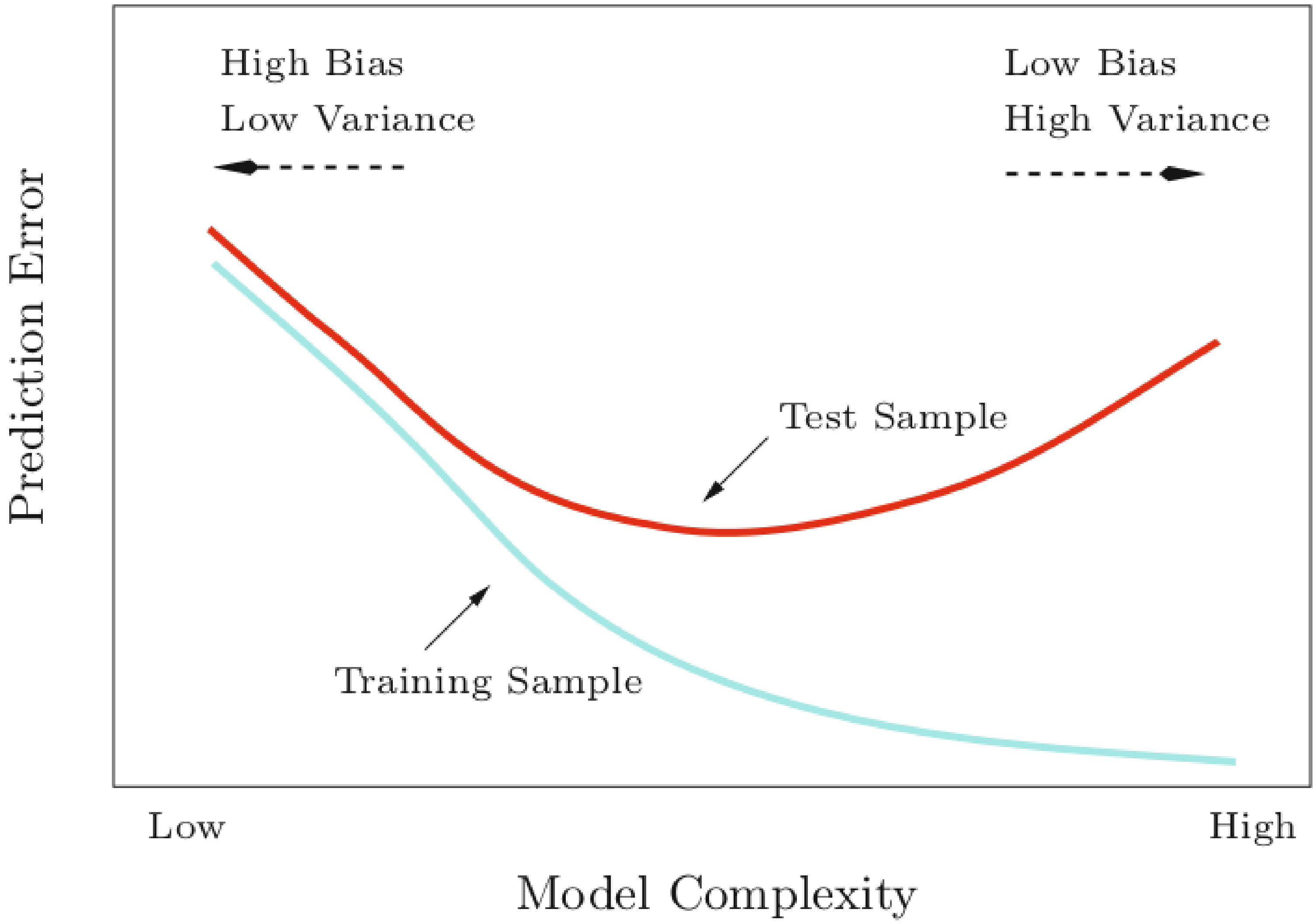
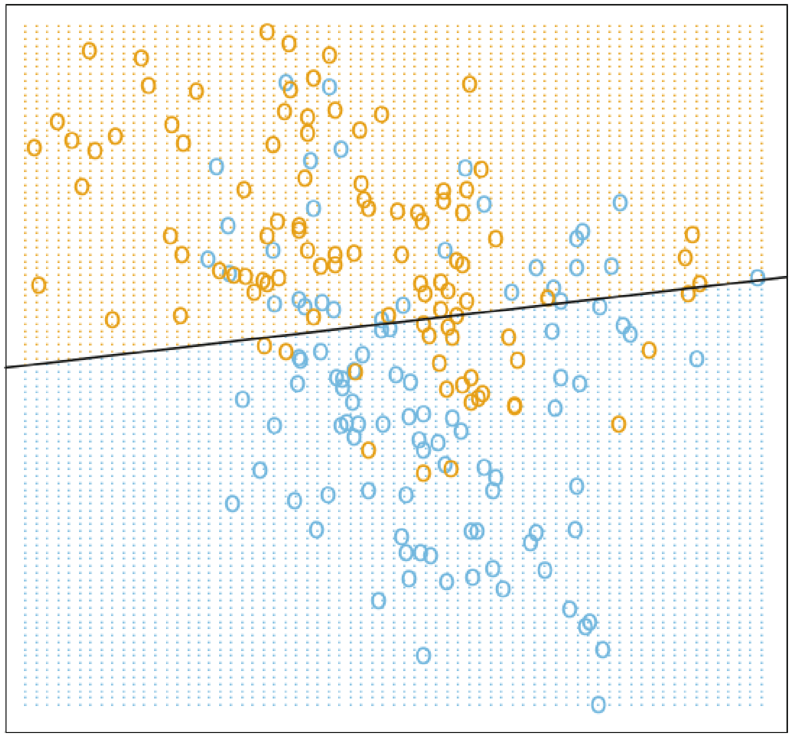
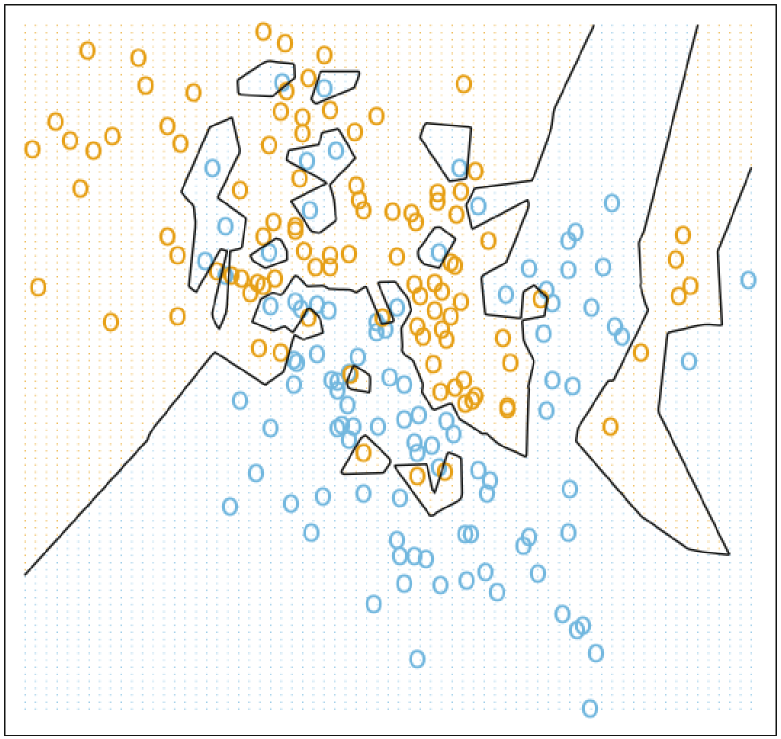


FIGURE 2.11. *Test and training error as a function of model complexity.*

Hastie et al. 2009, p. 38.

Bias-Variance tradeoff





Hastie et al. 2009, Chap. 2.3.

Linear model
high bias, but stable

1-nearest neighbours
low bias, high variance

$$\mathrm{E}\left[\left(y-\hat{f}(x)\right)^2\right]=\mathrm{Bias}\left[\hat{f}(x)\right]^2+\mathrm{Var}\left[\hat{f}(x)\right]+\sigma^2$$

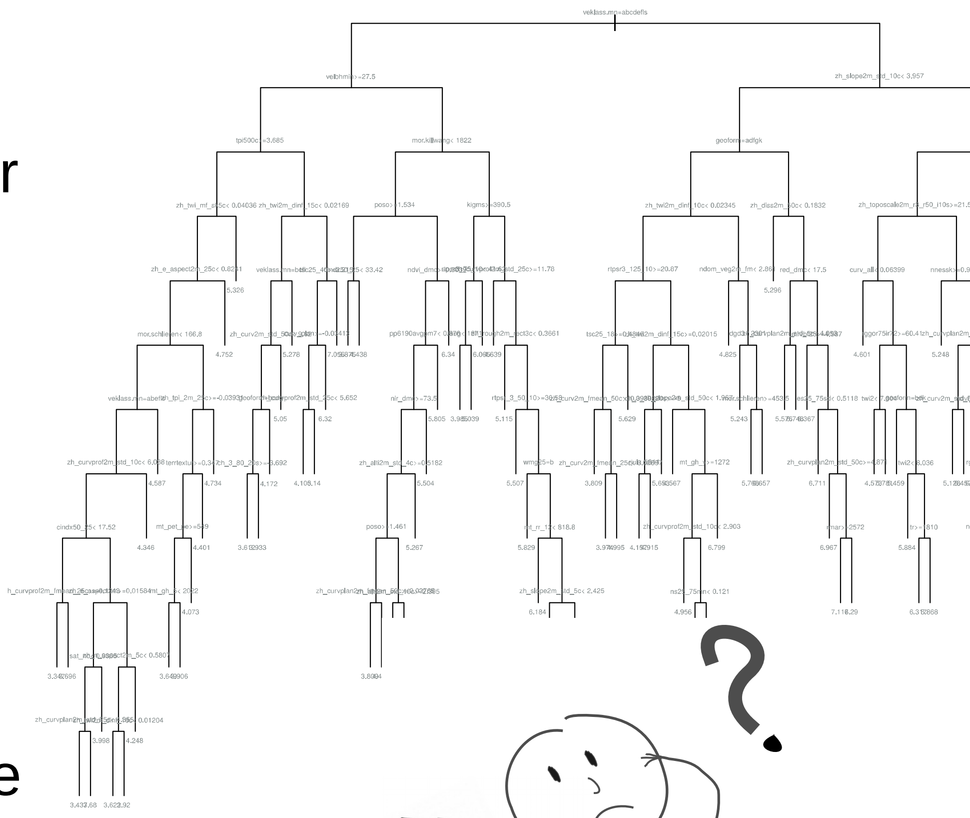
Bias: erroneous assumptions in the model, miss relevant relationship (underfitting).

Variance: sensitivity to small fluctuations in the calibration data, algorithm models random noise in calibration data, instead of just relevant relationship (overfitting).

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...



I tried to tidy up ...

- linear regression
- geostatistical methods
  external-drift kriging, regression kriging
- additive models (GAM)
- machine learning
  classification and regression trees (CART), support vector machines, neural nets
- ensemble machine learners
  random forest, boosted regression trees
- model averaging

parametric (rely on distribution assumptions), solve some likelihood function.

Drawback: transformations, extrapolation, lack of stability with collinear covariates, with many covariates → **how to select trend?** No fit for *n* > *p.*

based on algorithms, stepwise procedure to build up model.

For (spatial) prediction: **supervised learning**

response ← model trained
on covariates

I tried to tidy up ...

- linear regression
- geostatistical methods
  external-drift kriging, regression kriging
- additive models (GAM)
- machine learning
  classification and regression trees (CART),
  support vector machines, neural nets
- ensemble machine learners
  random forest, boosted regression trees
- model averaging

lasso

Boosting with splines (geoGAM)

Random Forest

Lasso: ML for linear models

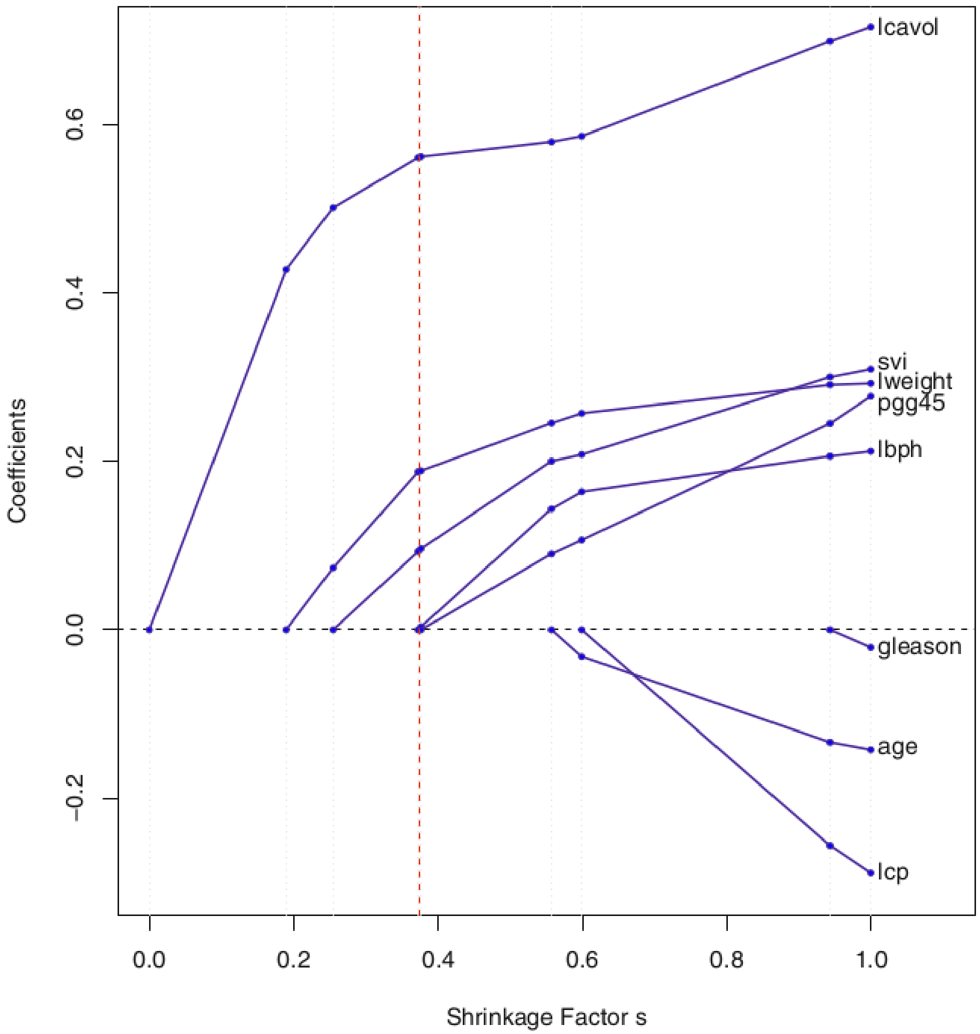
- Select linear regression with stepwise forward/backward, best subset: Most often does not find true model, does overfit, selection is binary – either in or out
- **Shrinkage:** include a covariate, but with smaller / downweighted coefficients
- Different approaches (ridge regression etc.), most promising: Lasso: least absolute shrinkage and selection operator

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \boxed{\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2} + \boxed{\lambda \sum_{j=1}^{p} |\beta_j|} \right\}.$$

OLS **Lasso penalty**

- Thus the lasso does a kind of continuous subset selection.
- Tuning Parameter λ, find by cross validation

Lasso: ML for linear models

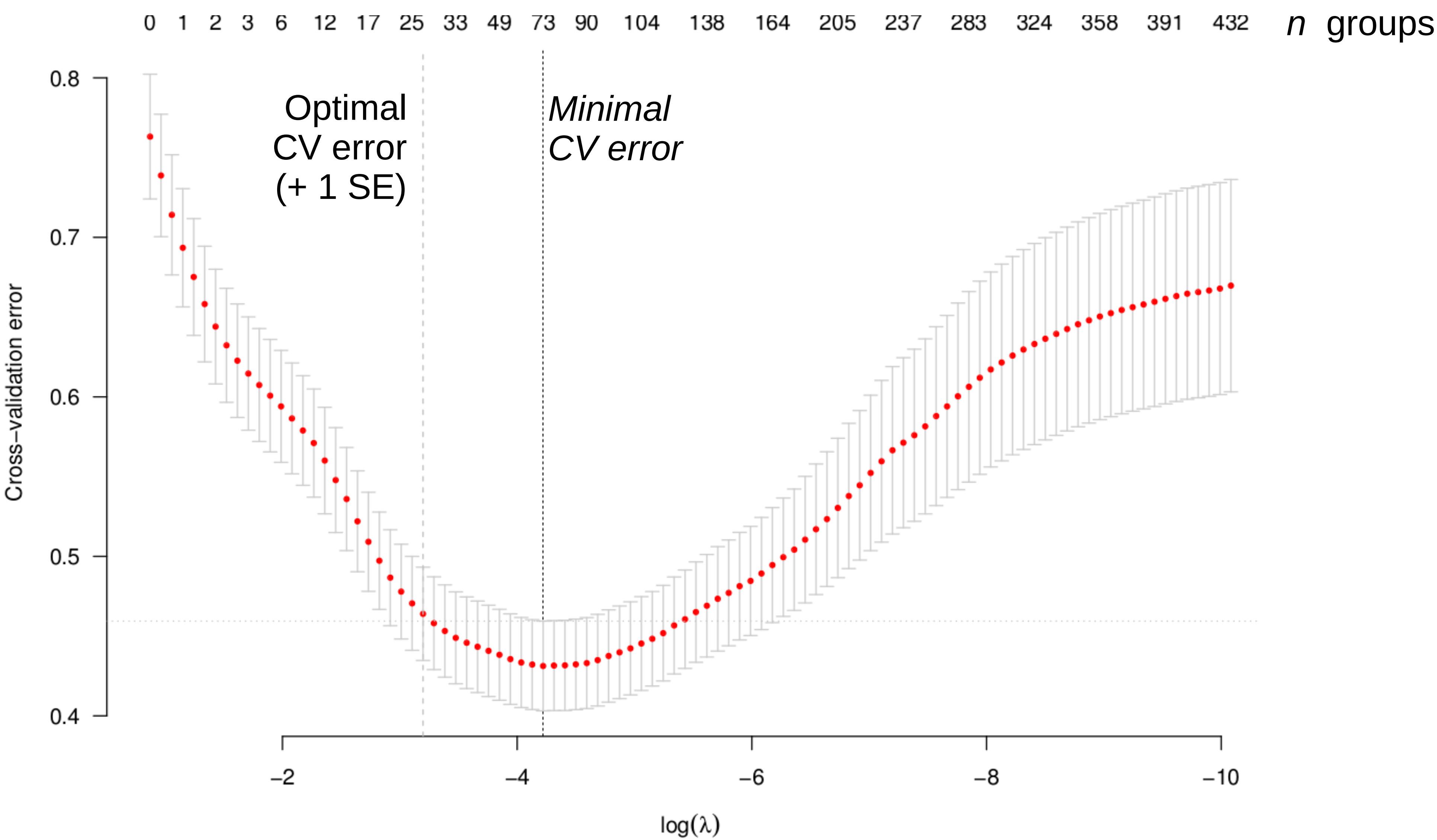


Path of coefficents for increasing tuning parameter

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

Lasso: ML for linear models



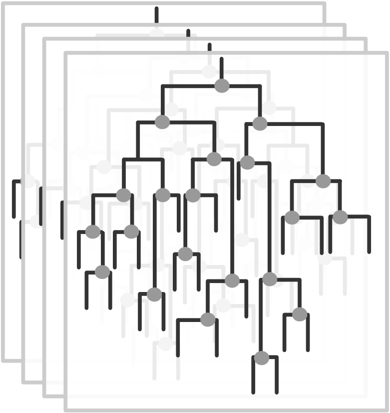
Berne data set, subspoil pH, >400 partly higly correlated and noisy covariates

Lasso: ML for linear models

- ✔ Very fast
- ✔ Selects covariates
- ✔ No problems with colinearity
- ✔ Easy interpretation (linear relationships)
- ✔ Linear regression with a lot of covariates, even *n>>p*

- ✖ Linear only, no interactions if not added explicitly
  (if n>>>p becomes nonlinear again)
- ✖ Take care, not always stable
- ✖ Rather underfitting
  (possible solution: relaxed Lasso with a second fit on non-zero covariates only)
- ✖ Standard errors not defined, prediction uncertainty only with bootstrap
- ✖ No direct spatial modelling, only via workaround

Ensemble Machine Learners



- Combine predictions of several learners
  (any method)
- Meaningful for low-bias, high-variance procedures

  Strategies:

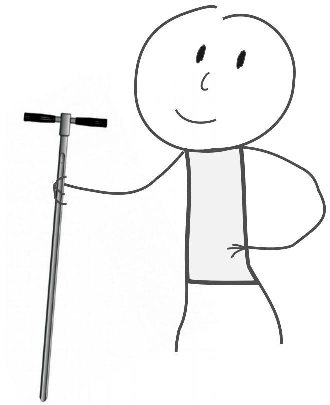
  – *Bagging = bootstrap aggregation.*
    Uniform *r*esampling the data with replacement (no change of response distribution), fit the data to each resampled set, prediction = average of all single predictions

    Random forest = bagged trees?

  – Gradient boosting
    Adaptive updating strategy, shrunken stepwise forward selection, fits on residuals → change of distribution

  – Model averaging
    Fits on the same response by different methods

Gradient boosting: Algorithm

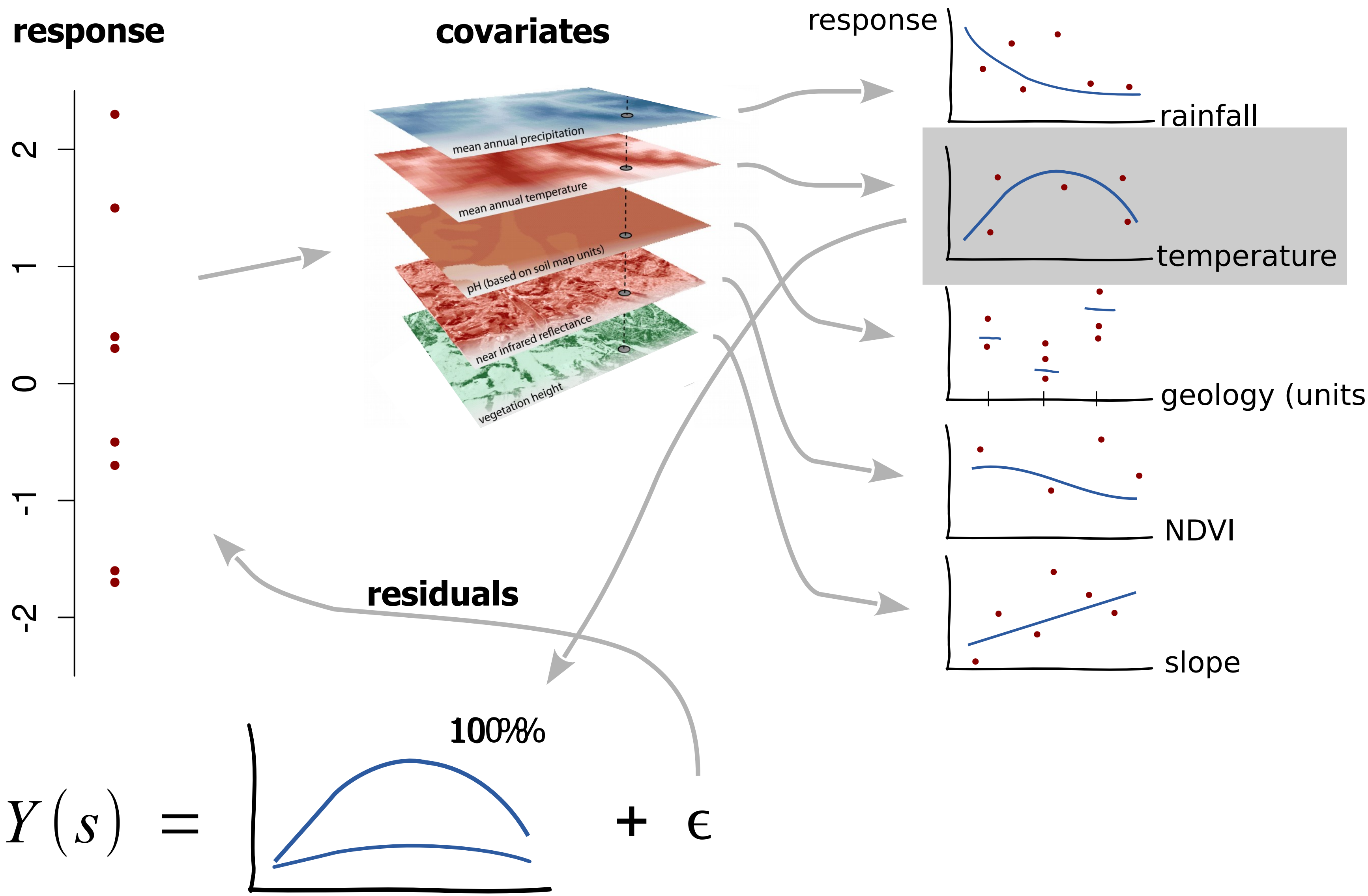


model selection for
high-dimensional regression

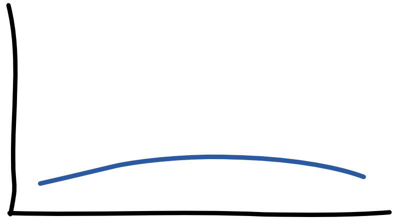
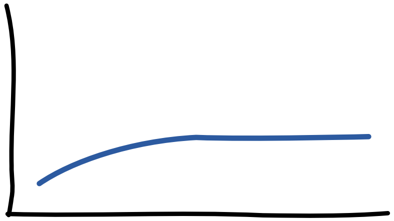
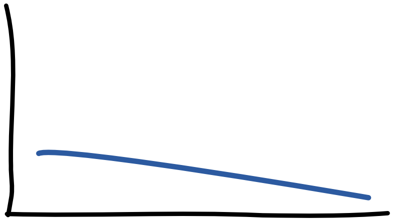
1. compute residuals
(gradient)

small step size v
= “weak” learner
(again shrinkage!)

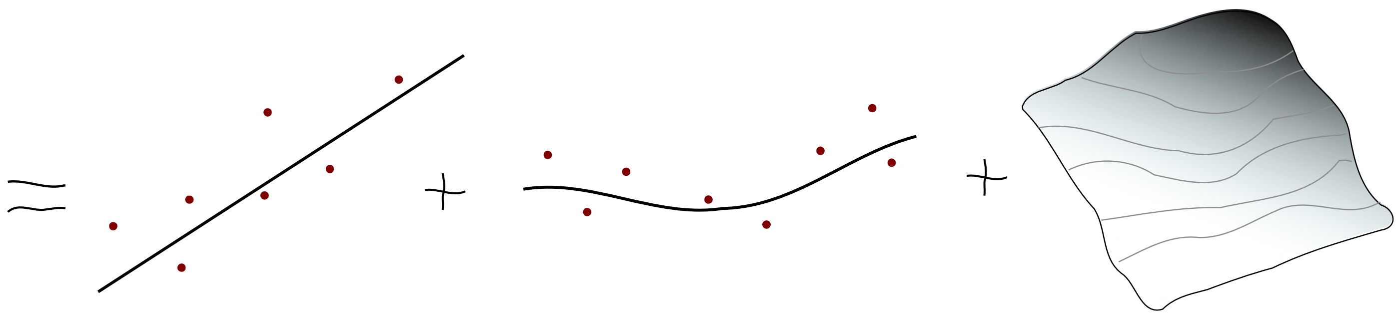
Gradient boosting: mini example



Gradient boosting: mini example

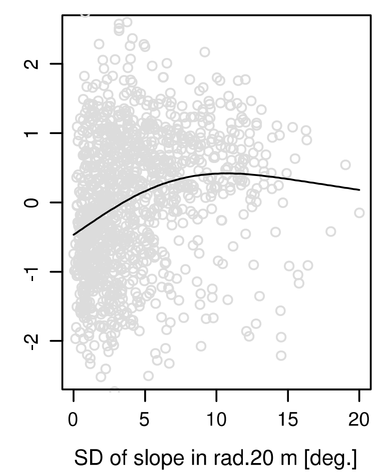
$Y(s) =$  +  +  +

Gradient boosting:
linear, splines and spatial baselearners



linear non-linear spatial

$$Y(s) = f_{env}(X) + f_s(s) + f_{ns}(X, s) \ldots + \epsilon$$



partial residuals

see e.g. Hothorn et al. 2011

Gradient boosting: Spatial modelling with splines

Spatial autocorrelation can be modelled by including a „smooth spatial surface" as baselearner, non-stationary effects by creating interactions with the spatial surface

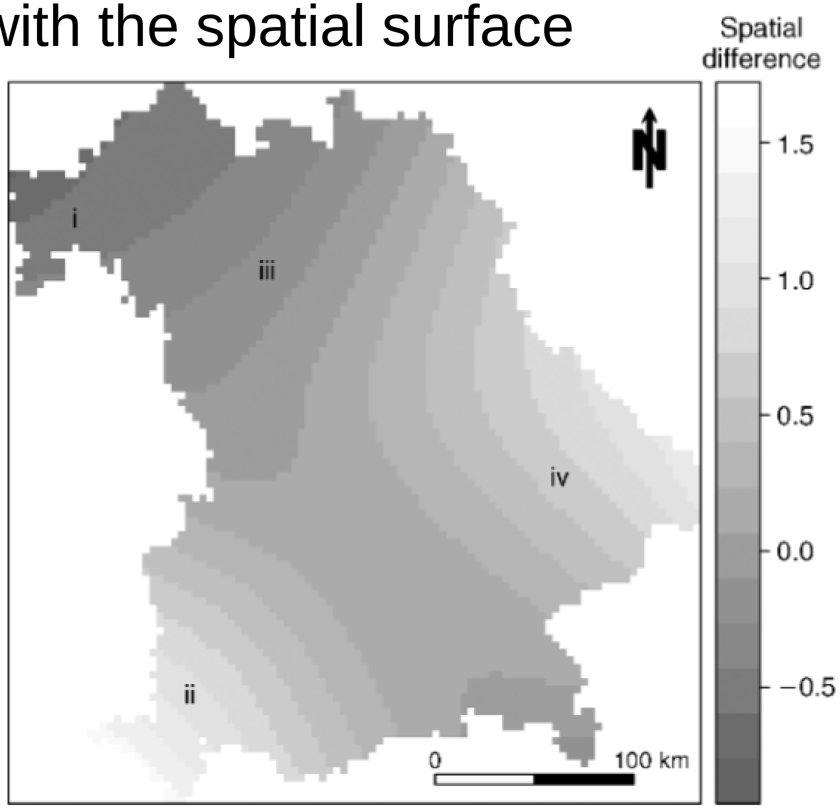


FIG. 6. Spatial difference in Red Kite breeding between 1979–1983 and 1996–1999 for model (add/vary). The breeding probabilities in the northwestern part decreased, while the southwestern part goes with increased breeding probabilities. For the four selected areas [(i) Unterfranken, (ii) Schwaben, (iii) Mittelfranken, and (iv) Niederbayern], the variability of the estimated spatial difference is shown in Fig. 7. Spatial differences can be interpreted as difference in log-odds ratios.

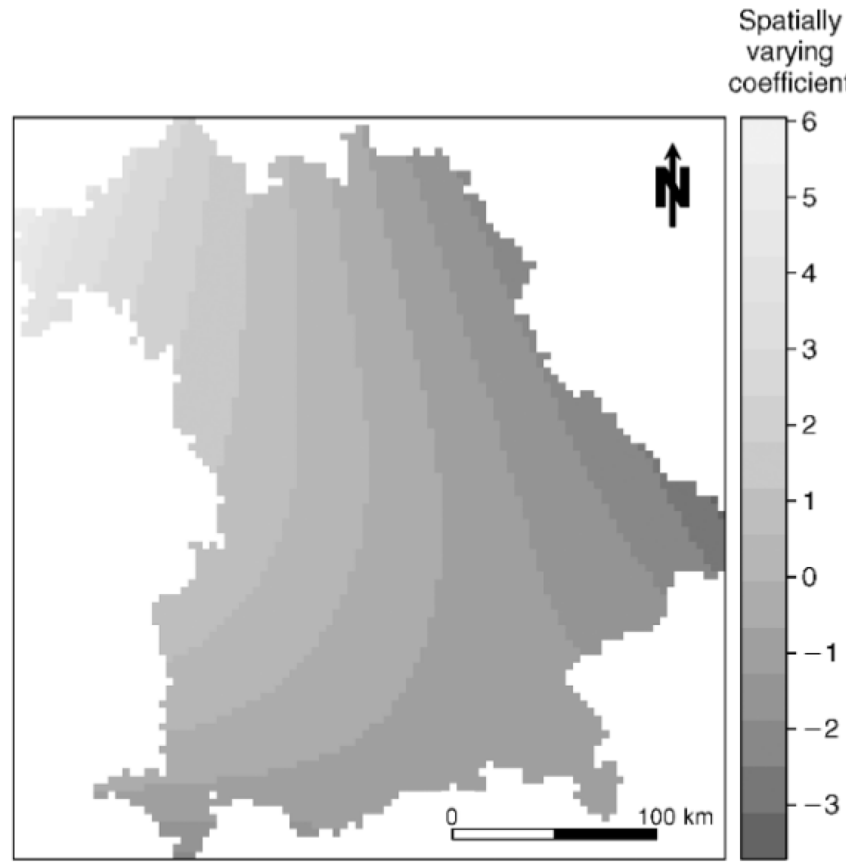


FIG. 8. Spatially varying coefficients for altitude in Red Kite breeding model (add/vary); here altitude was standardized to the unit interval. Altitude has a positive effect in the western and northwestern part, while its effect is zero or even negative in the rest of Bavaria.

Hothorn et al. 2011

Gradient boosting: with splines baselearner

- ✔ Finally a ML method that explicitly models spatial surfaces and non-stationarity!
- ✔ Selects covariates (but not very rigorous)
- ✔ Simple Interpretation of non-linear relationships
- ✖ Not so fast, needs a lot of setup for fitting ★
- ✖ Unfair/biased selection of categorical covariates ★
- ✖ Interpretation of covariate importance difficult, if no strong selection ★
- ✖ Parametric method: transformations, extrapolation errors
- ✖ Prediction uncertainty only with bootstrap

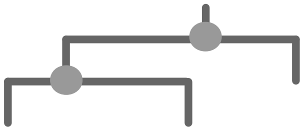
★ **geoGAM**

- ✔ Strong covariate selection (after boosting), improves interpretation
- ✔ Simple application for prediction problems (binary, ordinal, continuous) with roughly fair covariate selection
- ✖ Reduced model performance
- ✖ Spatial surface too coarse to capture small scale variability
- ✖ Selection stability?

R package geoGAM,
Nussbaum et al. 2017a

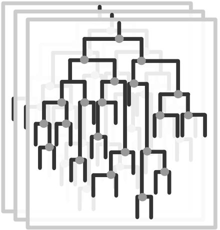
Should I use boosted trees or random forests?

Boosted trees



- ✔ Selects covariates weakly
- ✔ Covariate importance
  for interpretation and maybe selection
- ✖ Predictive accuracy slightly lower than random forest
- ✖ Prediction uncertainty only by bootstrapping
- ✔ Reduces bias by fitting on residuals

Random forest



- ✖ Does not select covariates
- ✔ Covariate importance
  for interpretation and maybe selection
- ✔ From my datasets on average best performance
  (up to 50 different responses tested)
- ✔ Prediction uncertainty with quantile regression forest
- ✖ Always fits on data with same distribution

Speed?

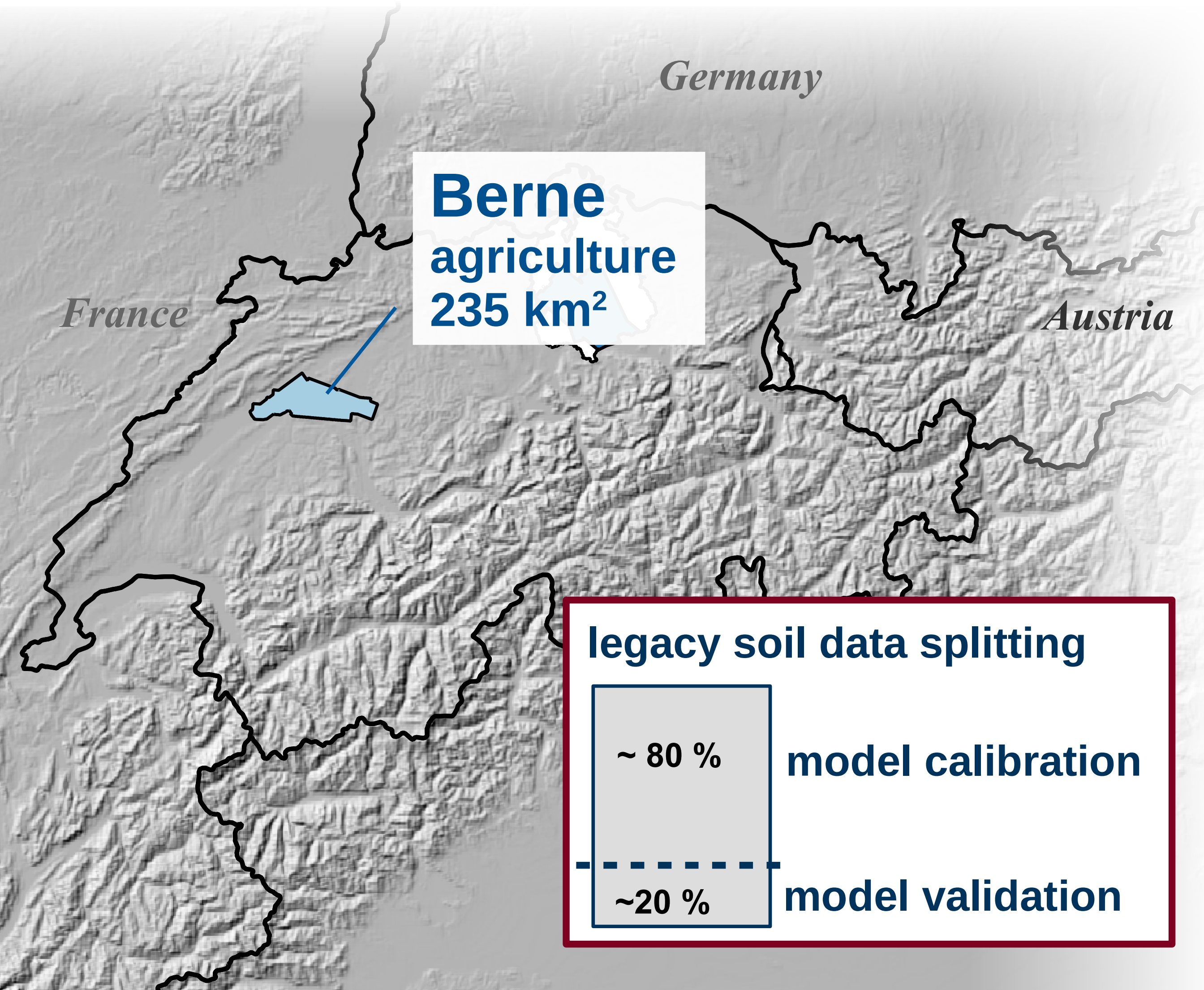
Do some benchmarking if interested ;-)

Model averaging

- Create predictions from different (ML) methods and **combine them**.
- Idea: each (ML) method as a mean of **reducing dimensions** in the dataset capturing different properties of the dataset → used methods should not be similar.
- Mathematical proofs show that combinations of different linear models result always in better performance. For other methods thats not a priori given, but very likely.
- Strategies
  - just take mean for every prediction
  - weighted mean, weights from model performance e.g. $\frac{1}{MSE}$
  - local weights with uncertainties of each method and prediction
  - linear fit with predictions as covariates and original data as response → but take car, never fit on validation set!!
  - or stacked generalisation, Bayesian approach

Exercise:
Berne soil mapping study

~ 1000 sites with legacy soil data from 1970-1980
Nussbaum et al. 2017b



Numerous covariates

Climate
different data sets
(monthly resolution)

Soil
soil overview map
historic wetlands
anthropogenic soil interventions
drainage networks

Parent material
(hydro)geological maps
and derivates

Vegetation
Landsat, SPOT5, DMC mosaic
forest vegetation map and
species composition

Terrain
90 derived attributes
(multiple scales)