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

Bern University
of Applied Sciences

Overview Soil Property Maps for Forests of Switzerland

predicted by machine learning based model averaging

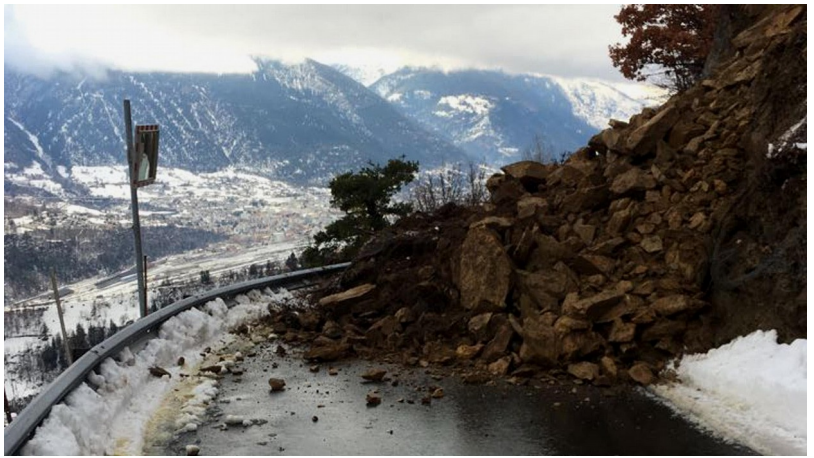
Madlene Nussbaum, Andri Baltensweiler, Lorenz Walthert

Forests under pressure

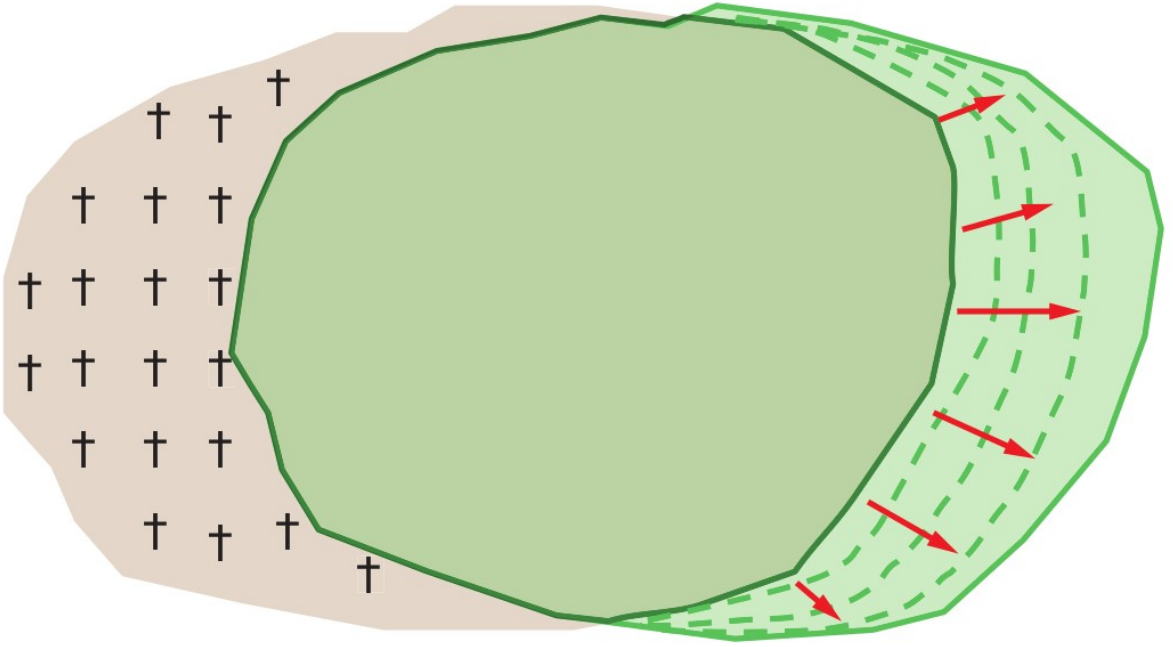
Oaks, 1. August 2018, nearby Zurich, Switzerland

Forests under pressure

Oaks, 1. August 2018, nearby Zurich, Switzerland

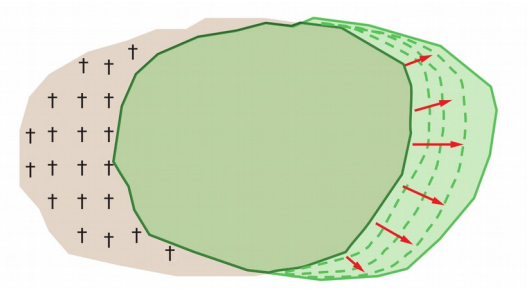


Climate change: shift in tree species

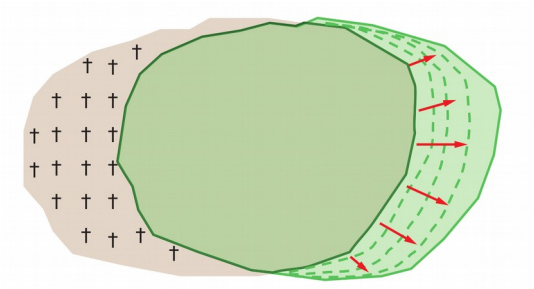


Zimmermann et al. 2016

Climate change: shift in tree species

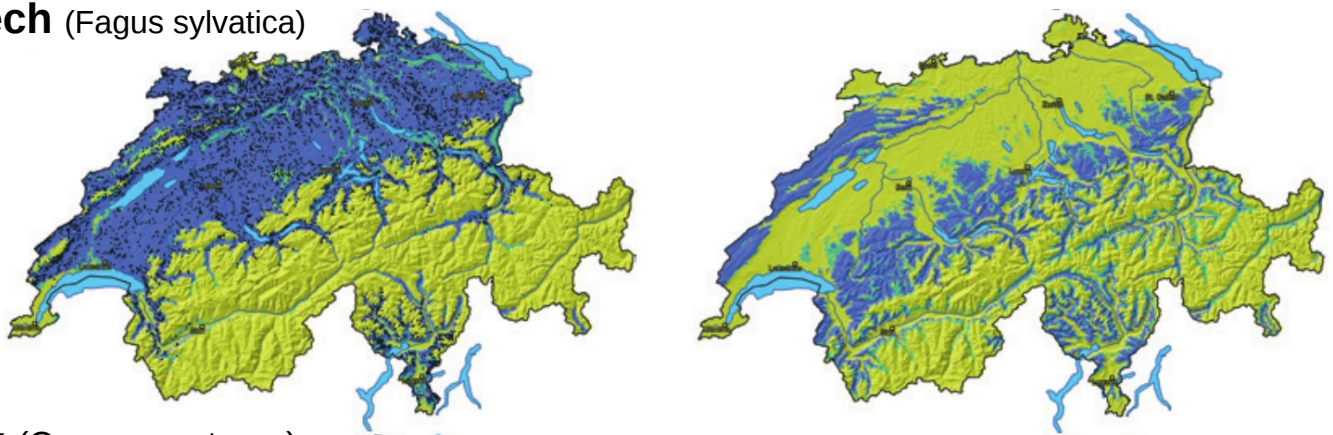


Climate change: shift in tree species

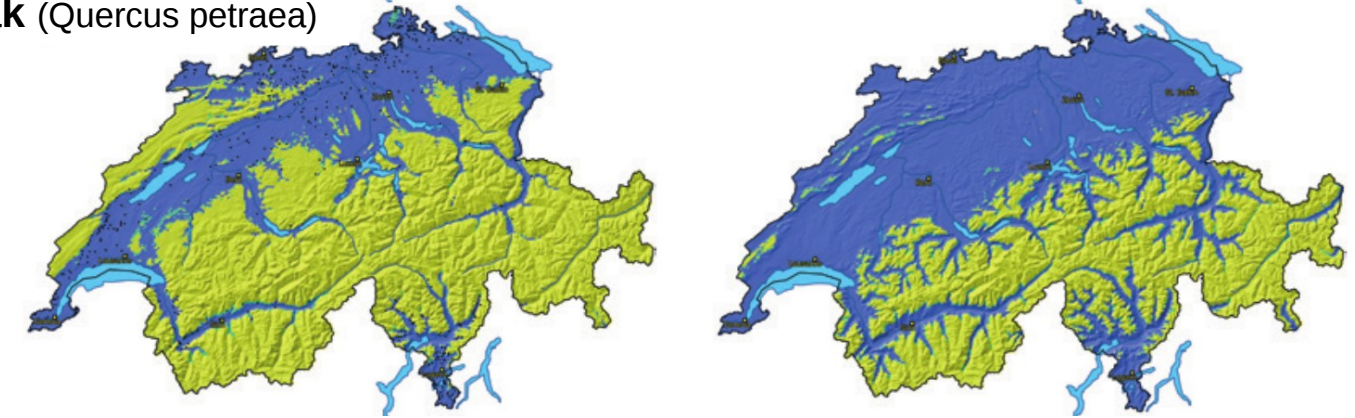


Today (1981–2010) **2051–2080**

Beech (Fagus sylvatica)



Oak (Quercus petraea)



Simulated presence

- <30 %
- 30–60 %
- >60 %

Zimmermann et al. 2014

Soil map in these models:

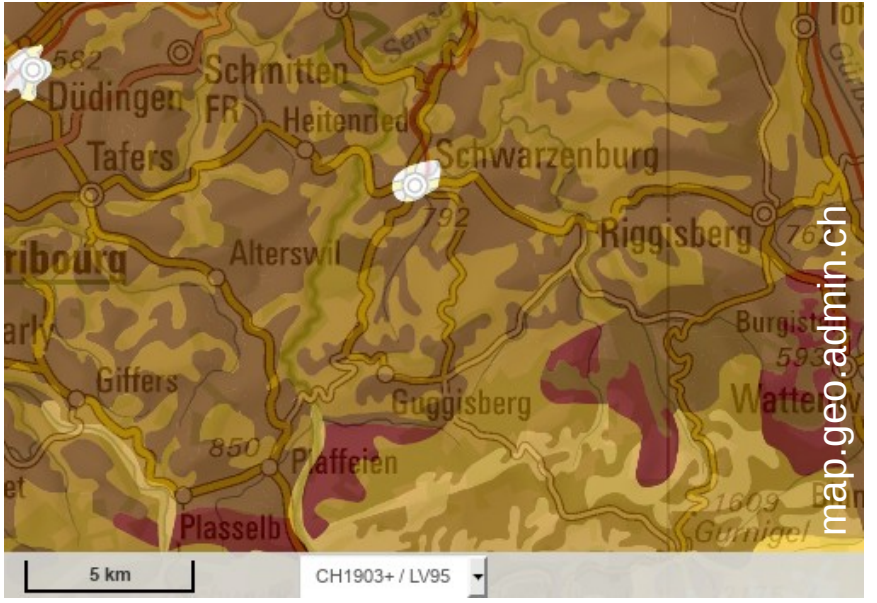
Soil map in these models:

Best soil map available:

Overview soil map 1:200'000

- from 1980
- broad classes
- hardly any observations in forests

e.g. water storage capacity



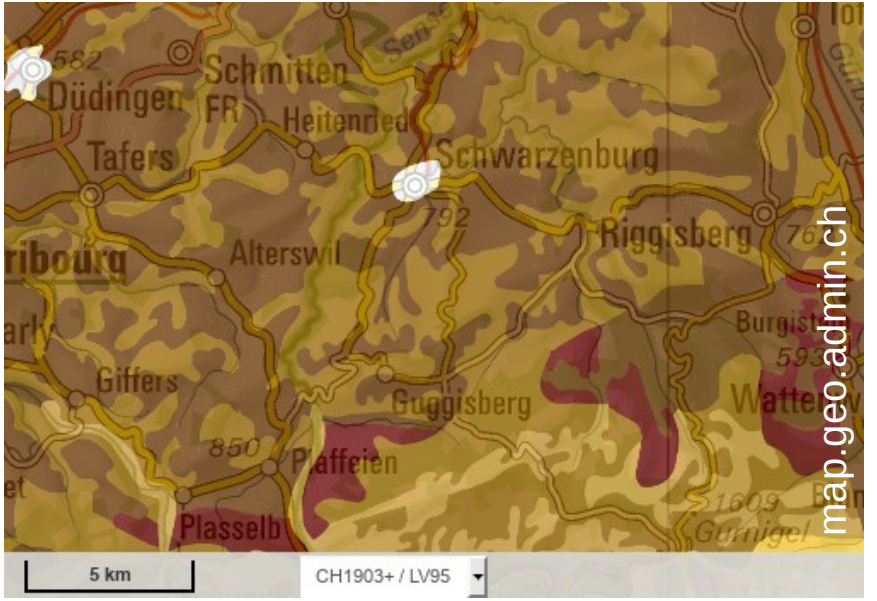
Soil map in these models:

Best soil map available:

Overview soil map 1:200'000

- from 1980
- broad classes
- hardly any observations in forests

e.g. water storage capacity



Available, but not suitable:

Soil organic carbon maps

0–30 cm and 0–100 cm (resolution 100 m), done for greenhouse gas inventory (Nussbaum et al. 2014)

Available soil data

Available soil data

- **Soil database**
  - Swiss Federal Institute for Snow, Forest and Landscape Research
  - sampling since 30 years
  - homogeneous dataset

Available soil data

- **Soil database**
  - Swiss Federal Institute for Snow, Forest and Landscape Research
  - sampling since 30 years
  - homogeneous dataset

- **2'071 soil profiles**
  - validation 382
  - calibration 1689

Available soil data

- **Soil database**
  - Swiss Federal Institute for Snow, Forest and Landscape Research
  - sampling since 30 years
  - homogeneous dataset

- **2'071 soil profiles**
  - validation 382
  - calibration 1689

clay, sand, gravel
density of the fine soil fraction <= 2 mm
soil organic carbon, pH
(soil depth)

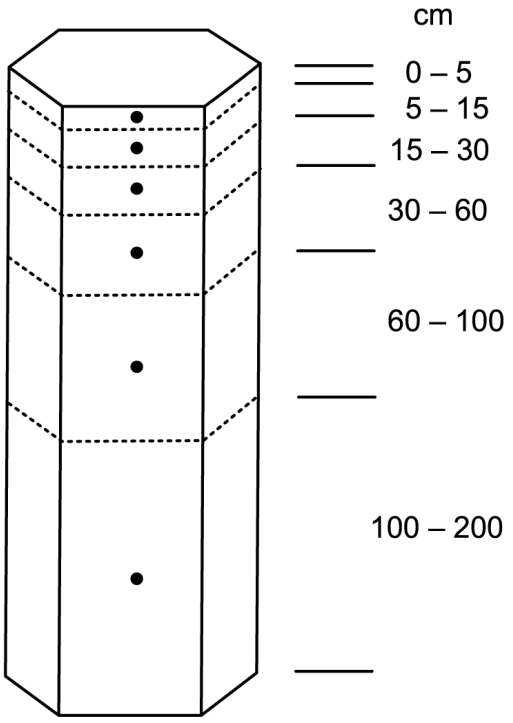
Available soil data

- **Soil database**
  - Swiss Federal Institute for Snow, Forest and Landscape Research
  - sampling since 30 years
  - homogeneous dataset

- **2'071 soil profiles**
  - validation 382
  - calibration 1689

clay, sand, gravel
density of the fine soil fraction <= 2 mm
soil organic carbon, pH
(soil depth)

GlobalSoilMap Specs
(Arrouays et al. 2014)



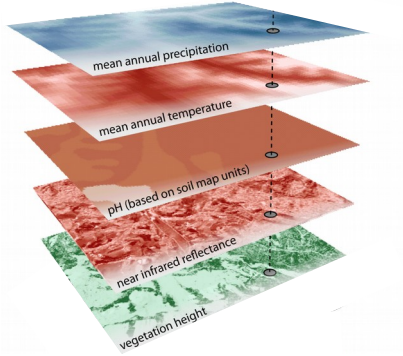
Digital soil mapping task

Swiss forest area – 11 800 km²

clay
sand,
density
gravel
pH
SOC

~1700
locations with
soil properties in

6 depth intervals



180
environmental
covariates

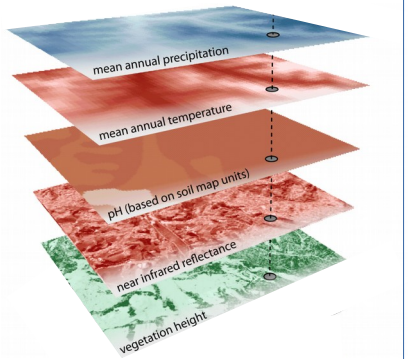
Digital soil mapping task

Swiss forest area – 11 800 km²

clay
sand,
density
gravel
pH
SOC

~1700
locations with soil properties in

6 depth intervals



180
environmental covariates

36 statistical models

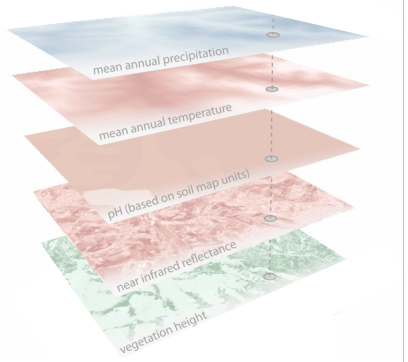
Digital soil mapping task

Swiss forest area – 11 800 km²

clay
sand,
density
gravel
pH
SOC

~1700
locations with
soil properties in

6 depth intervals

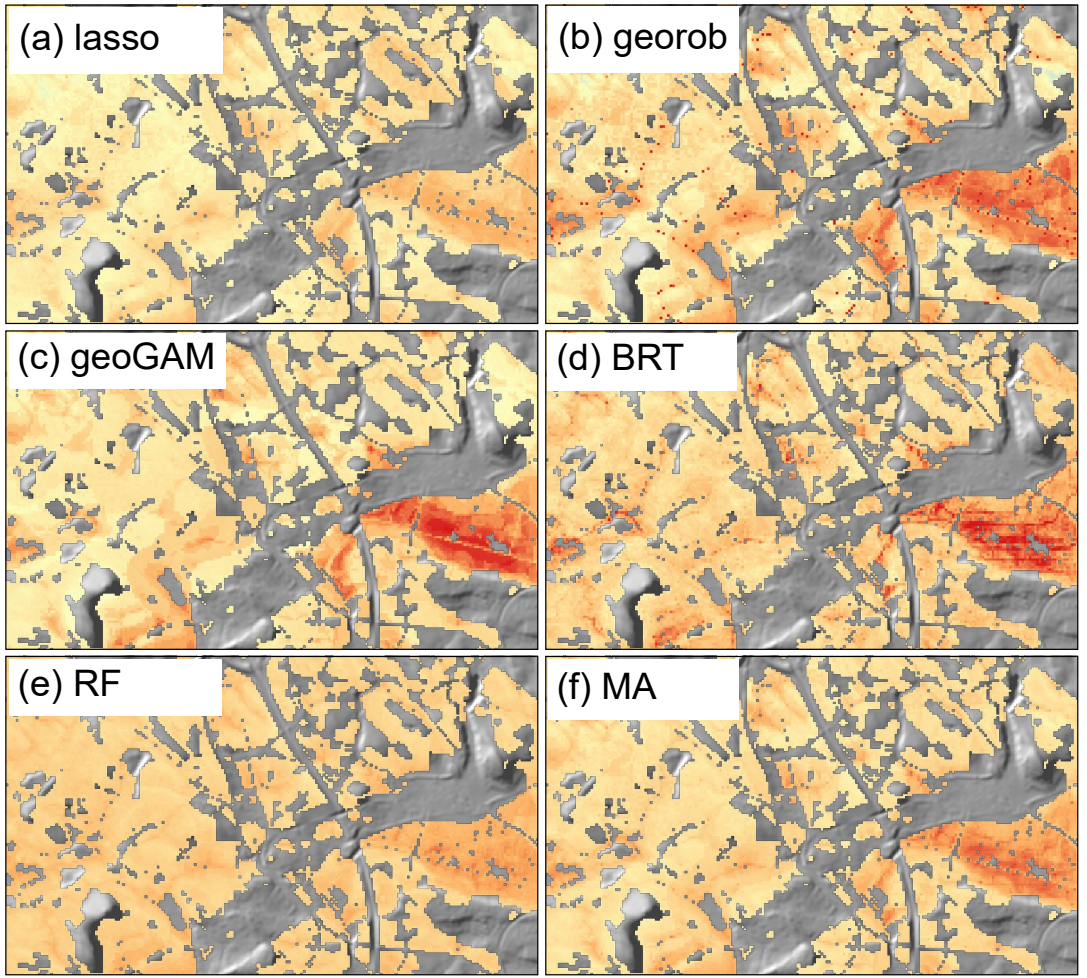


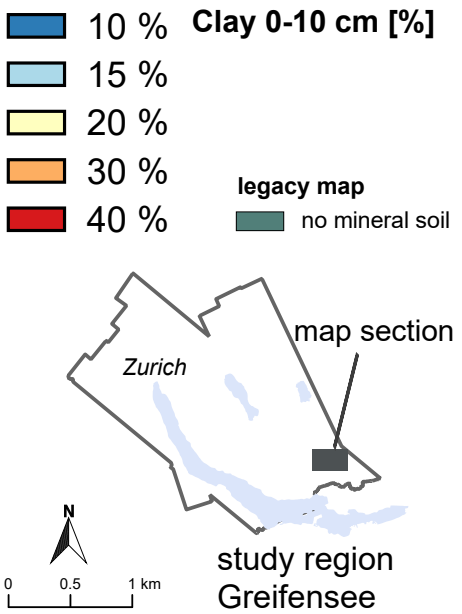
180
environmental
covariates

36 statistical models

Geodata sets	*n*
• **Climate data** (25, 100, 250 m)	52
• **Soil map** (1:200'000) **Geological maps** (1:200:000, 1:500'000), Glaciation	14
• **Vegetation** landscapes, biogeographic regions, historic peatlands, main watersheds, Sentinel 2 indices (10 m) vegetation height (2 m)	18
• Lidar **terrain modell** (2 m), derivations at different radii and with different algorithms	88
• **Coordinates** (rotated)	6
• **Sampling year** (meta data)	1

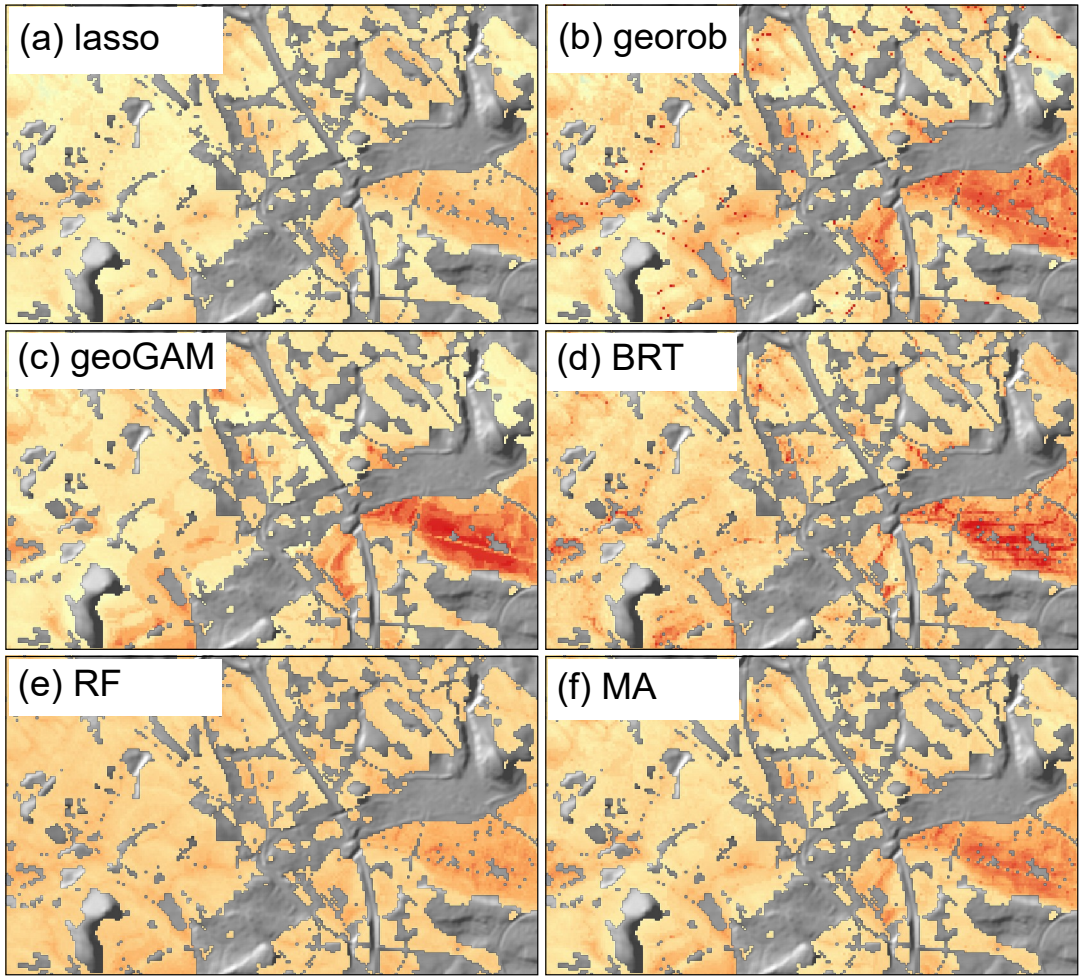
Method: model averaging (MA)

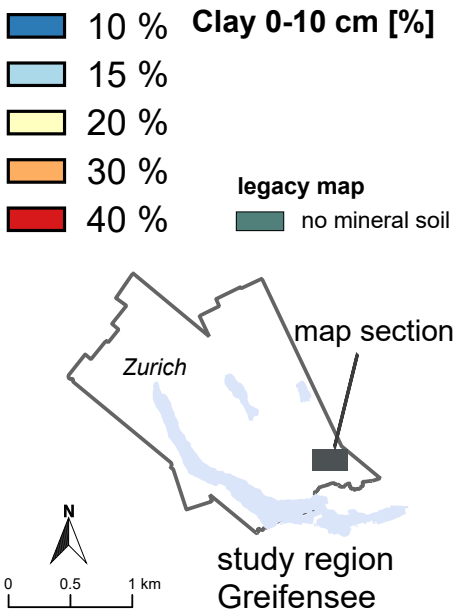




(Nussbaum et al. 2018)

Method: model averaging (MA)





(Nussbaum et al. 2018)

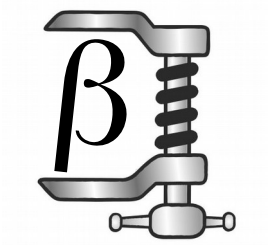
Experience with MA

- Good performance
- Balanced artefacts

Used methods

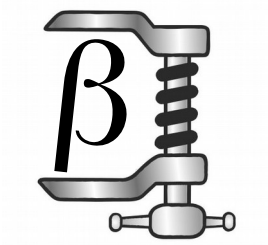
lasso

group lasso

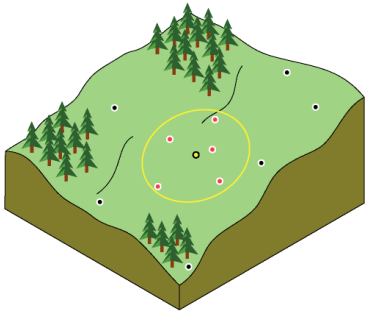


Used methods

lasso
group lasso

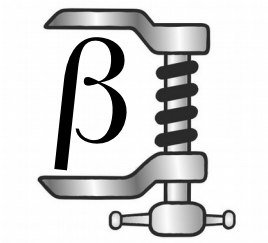


georob
robust external drift kriging

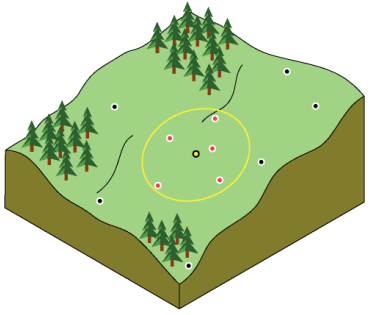


Used methods

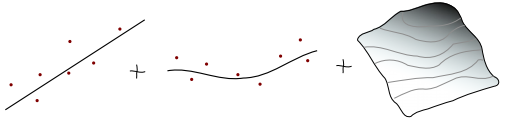
lasso
group lasso



georob
robust external drift kriging

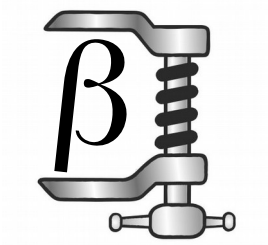


geoGAM
geoadditive model

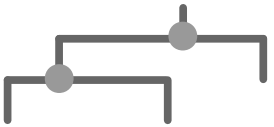


Used methods

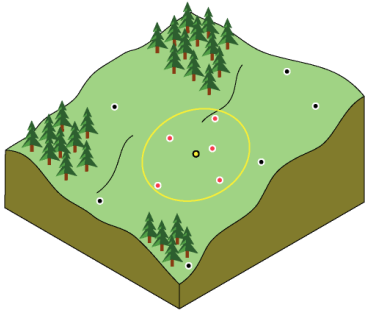
lasso
group lasso



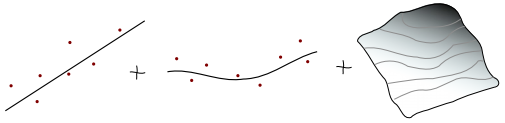
Cubist
rule based regression



georob
robust external drift kriging

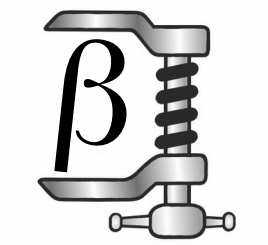


geoGAM
geoadditive model

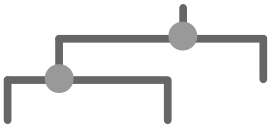


Used methods

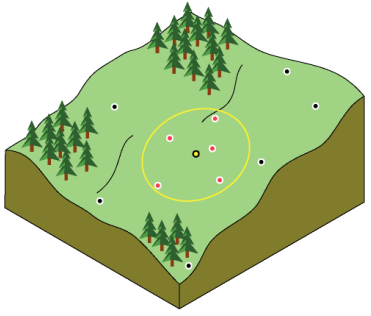
lasso
group lasso



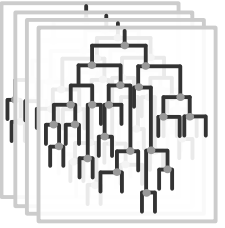
Cubist
rule based regression



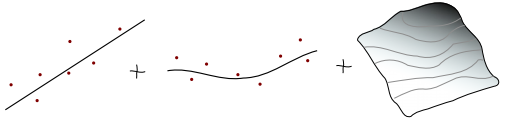
georob
robust external drift kriging



RF
random forest

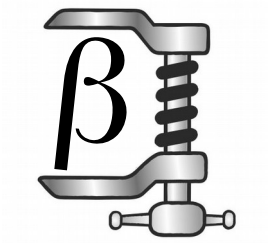


geoGAM
geoadditive model

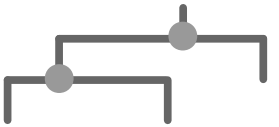


Used methods

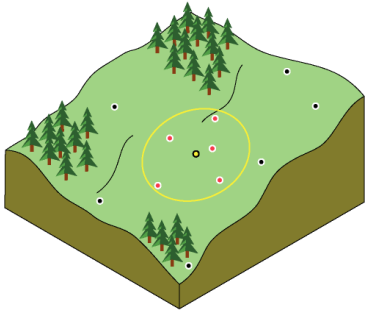
lasso
group lasso



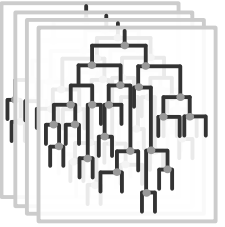
Cubist
rule based regression



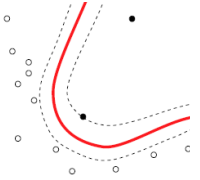
georob
robust external drift kriging



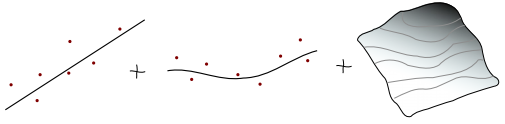
RF
random forest



SVM
support vector machines

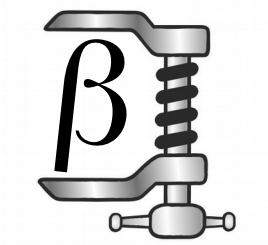


geoGAM
geoadditive model

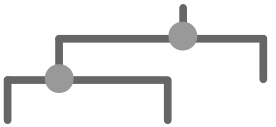


Used methods

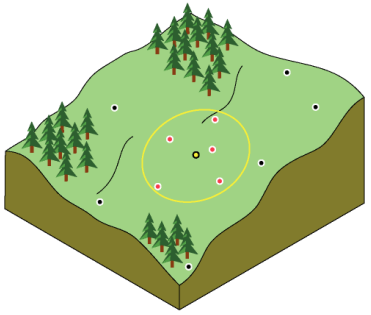
lasso
group lasso



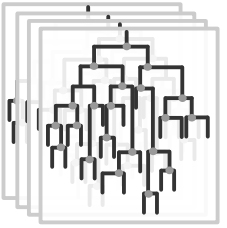
Cubist
rule based regression



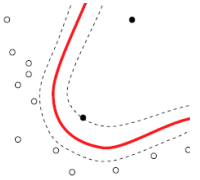
georob
robust external drift kriging



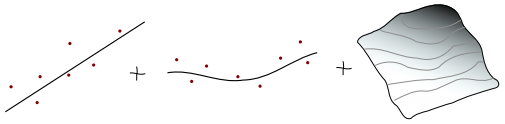
RF
random forest



SVM
support vector machines



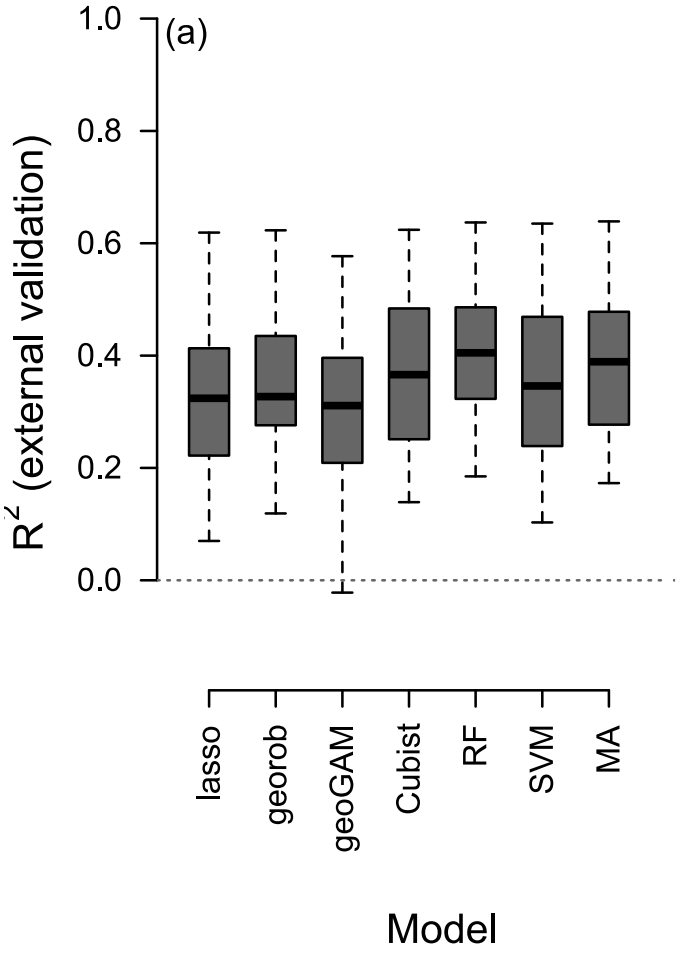
geoGAM
geoadditive model

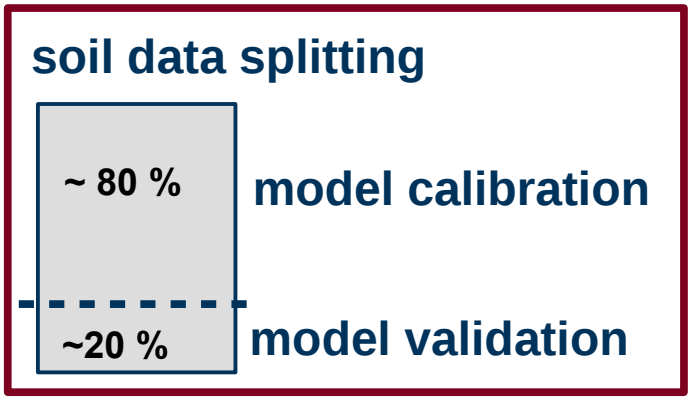


MA
weighted model average

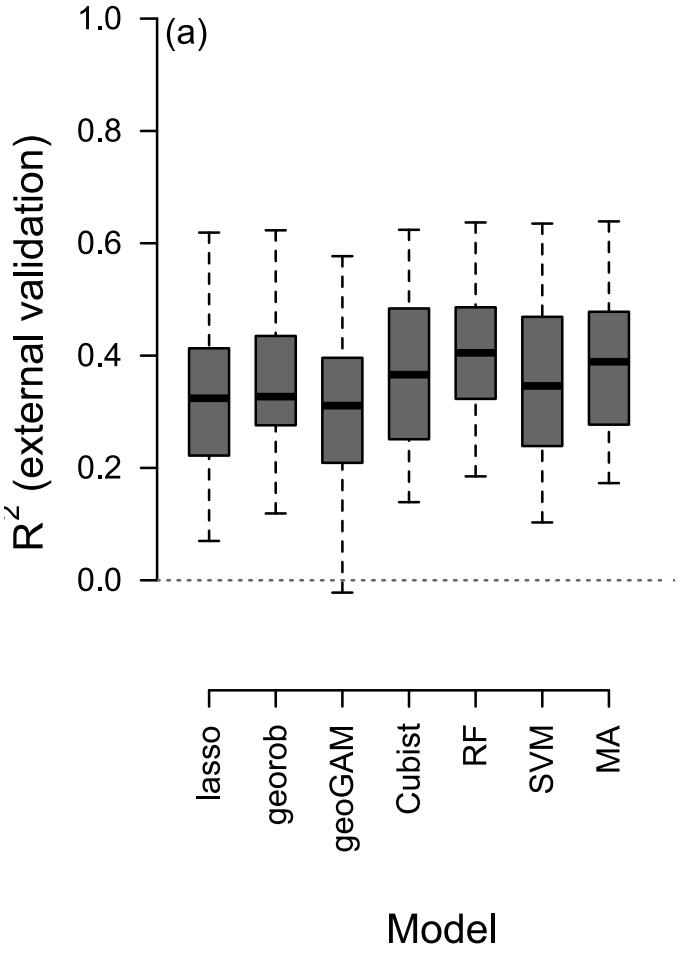
$$\frac{1}{MSE}$$

Results – Validation



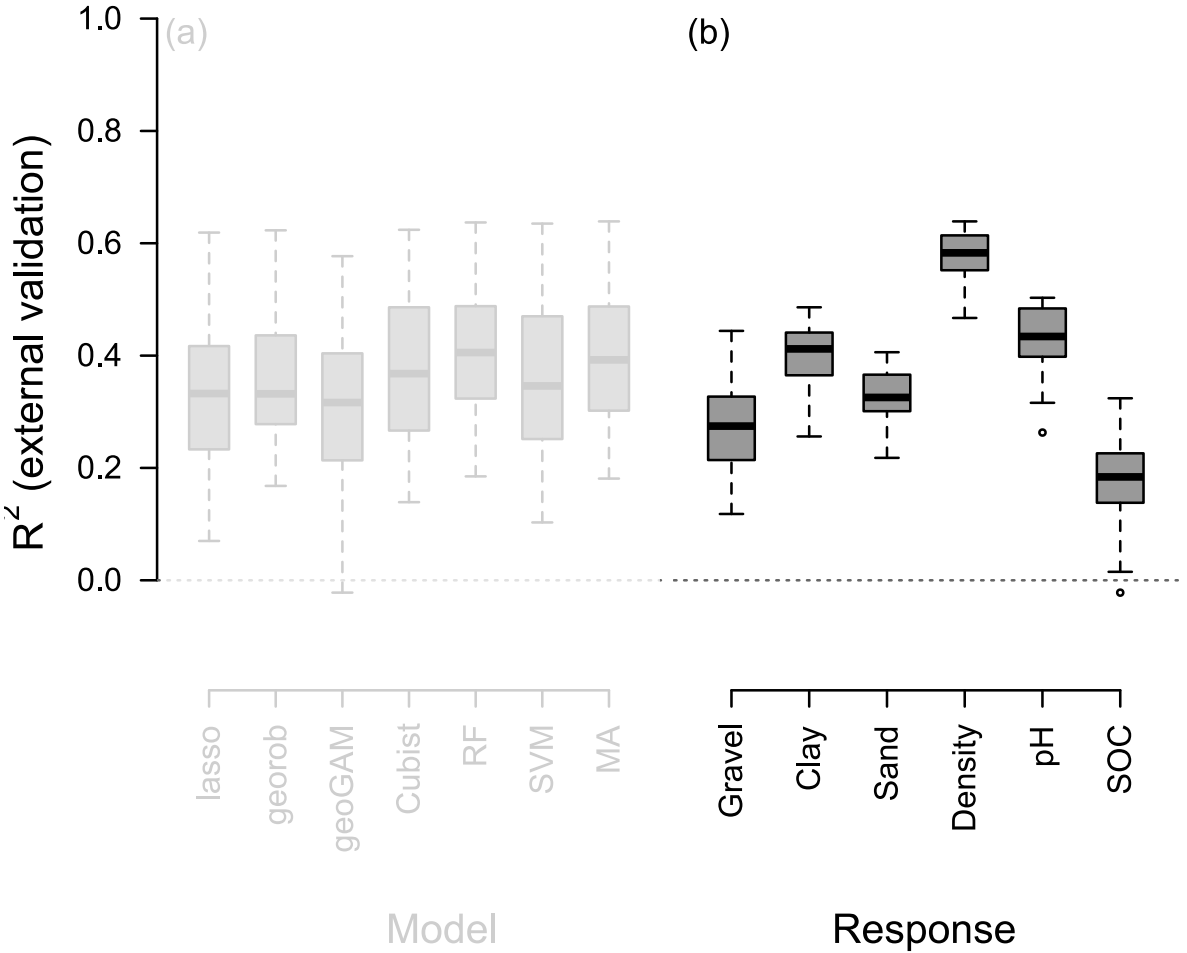


Results – Validation

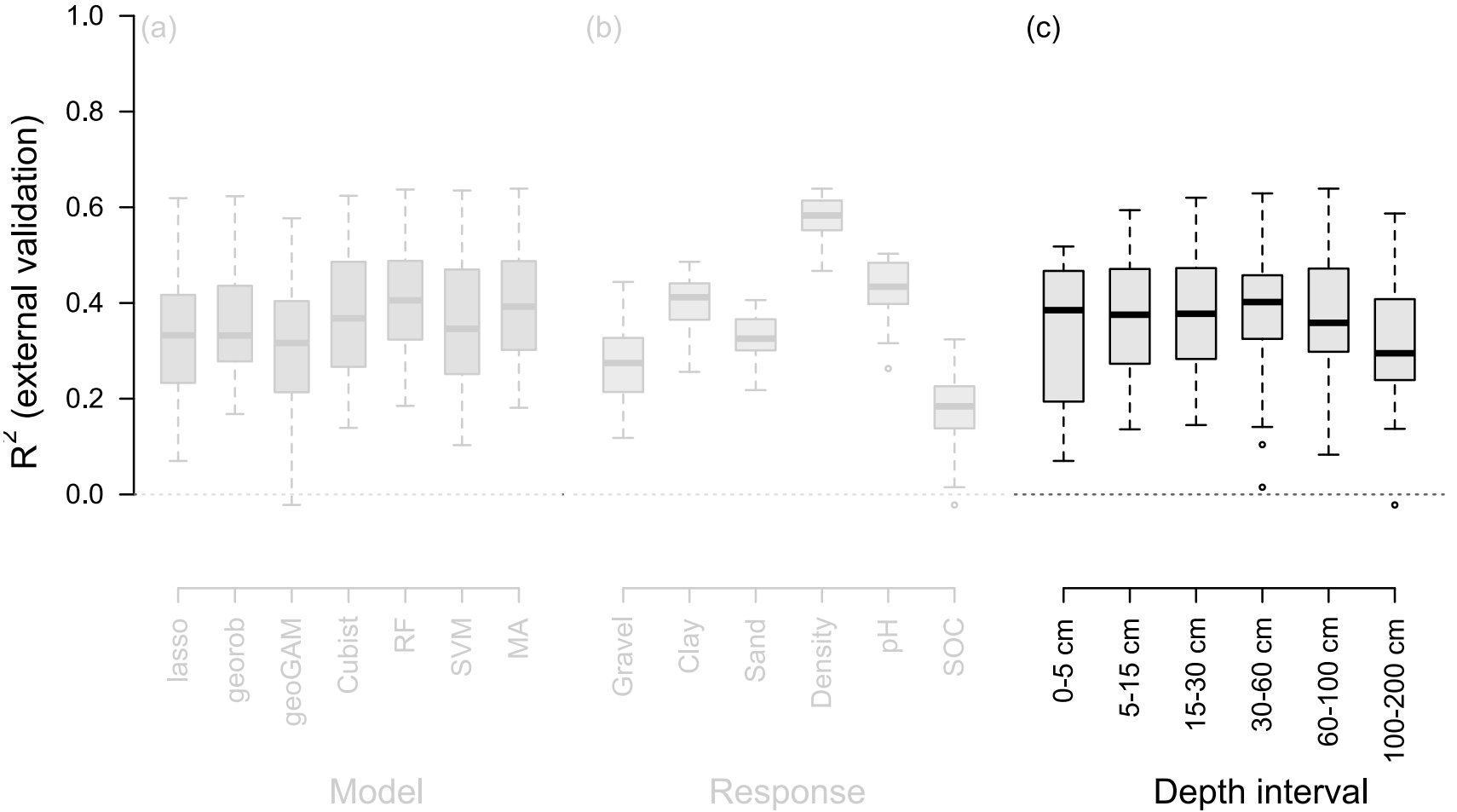


- 36 models basically very similar
- as single model RF is best for 22
- MA best for 7 (6x outperformed RF)

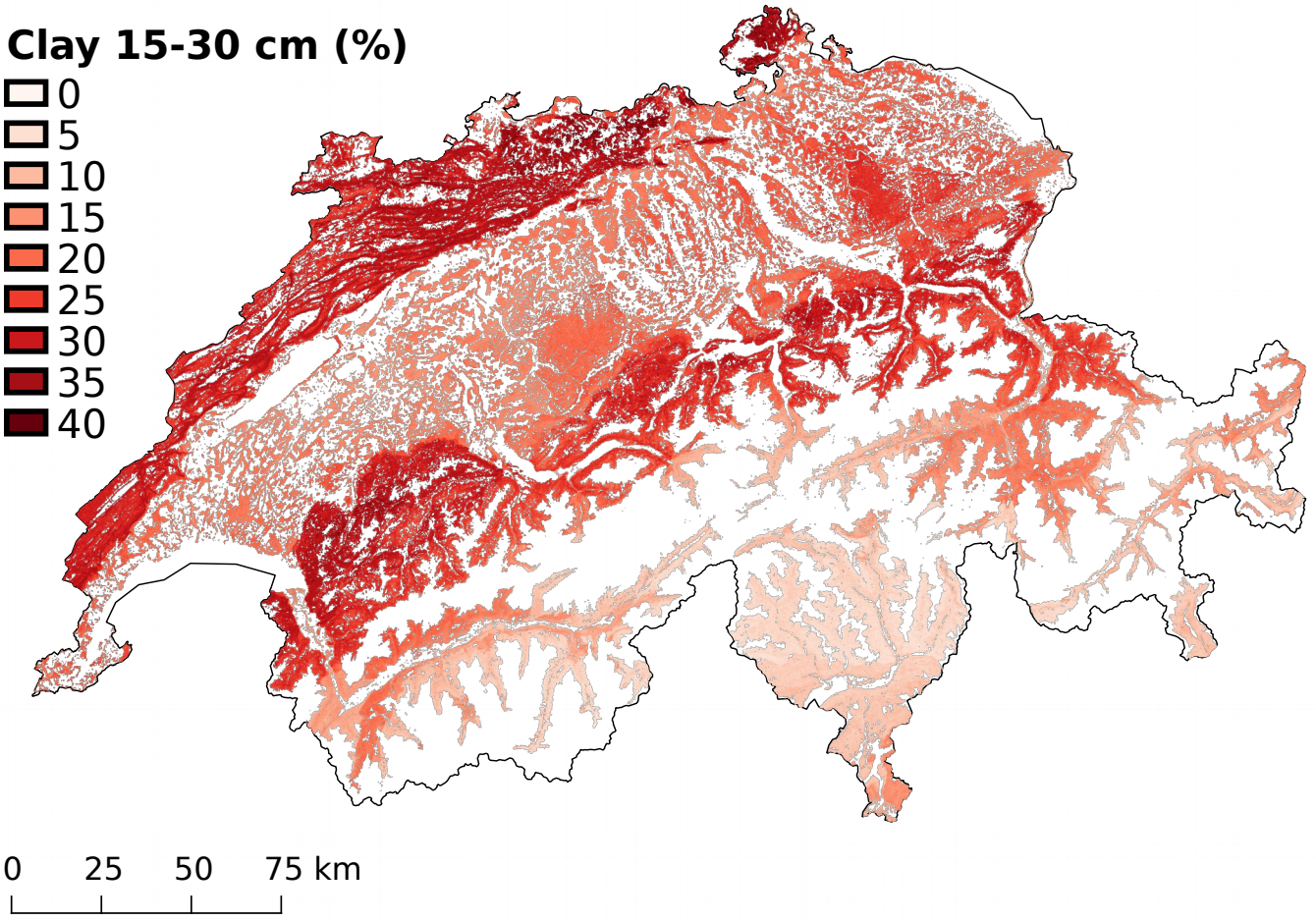
Results – Validation



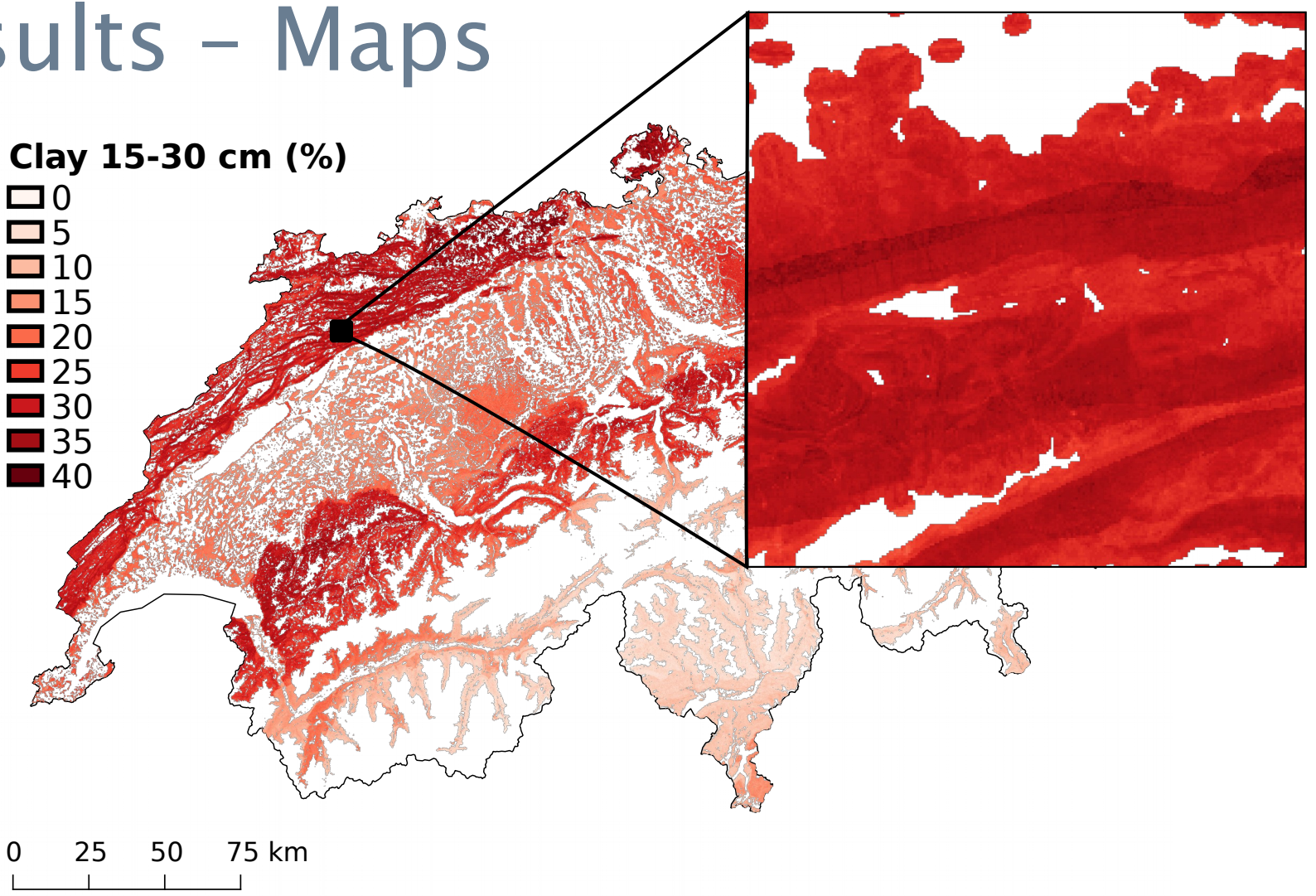
Results – Validation



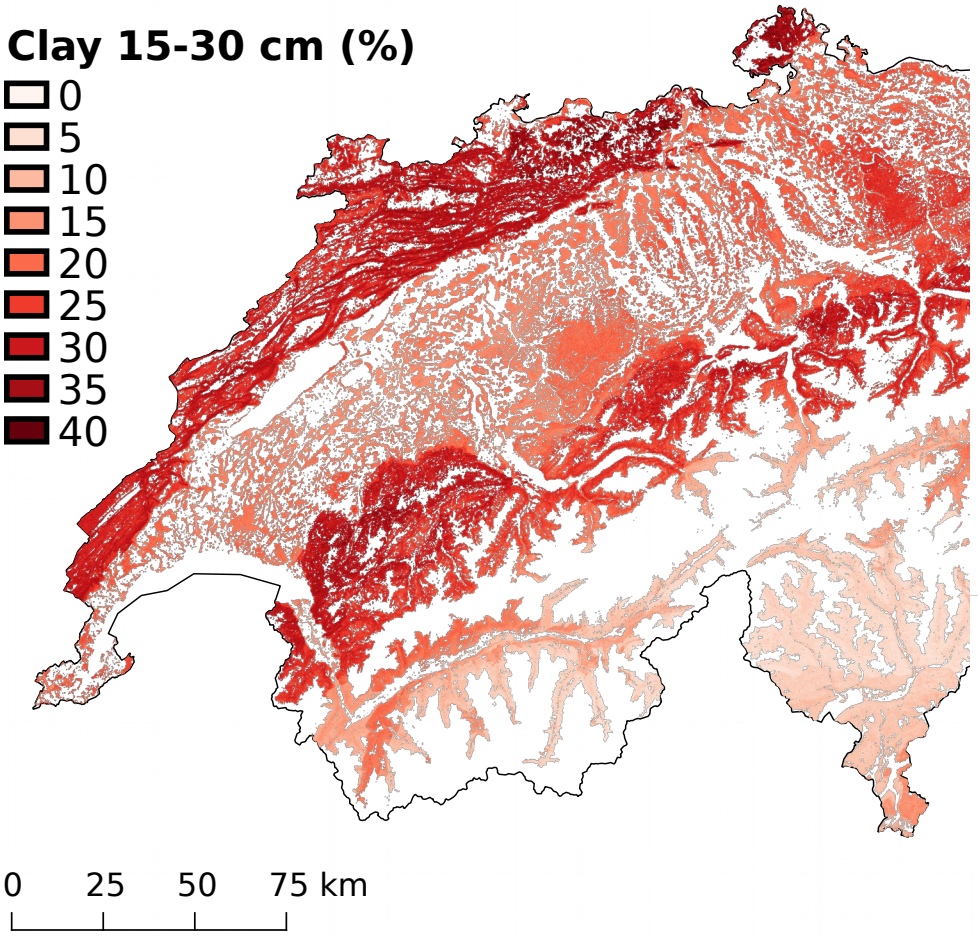
Results – Maps

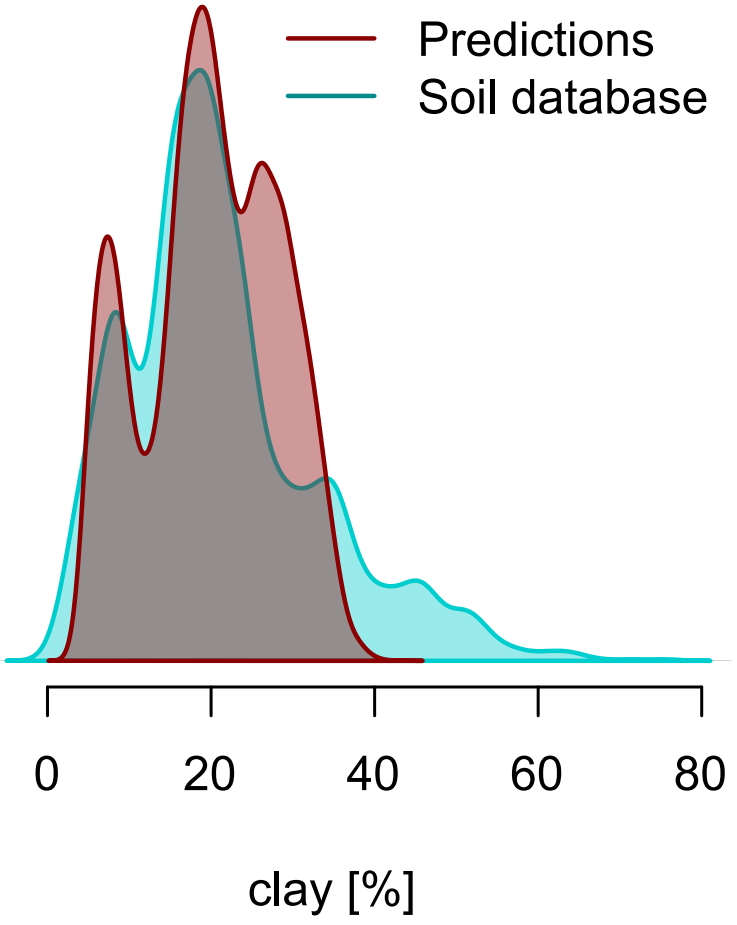


Results – Maps

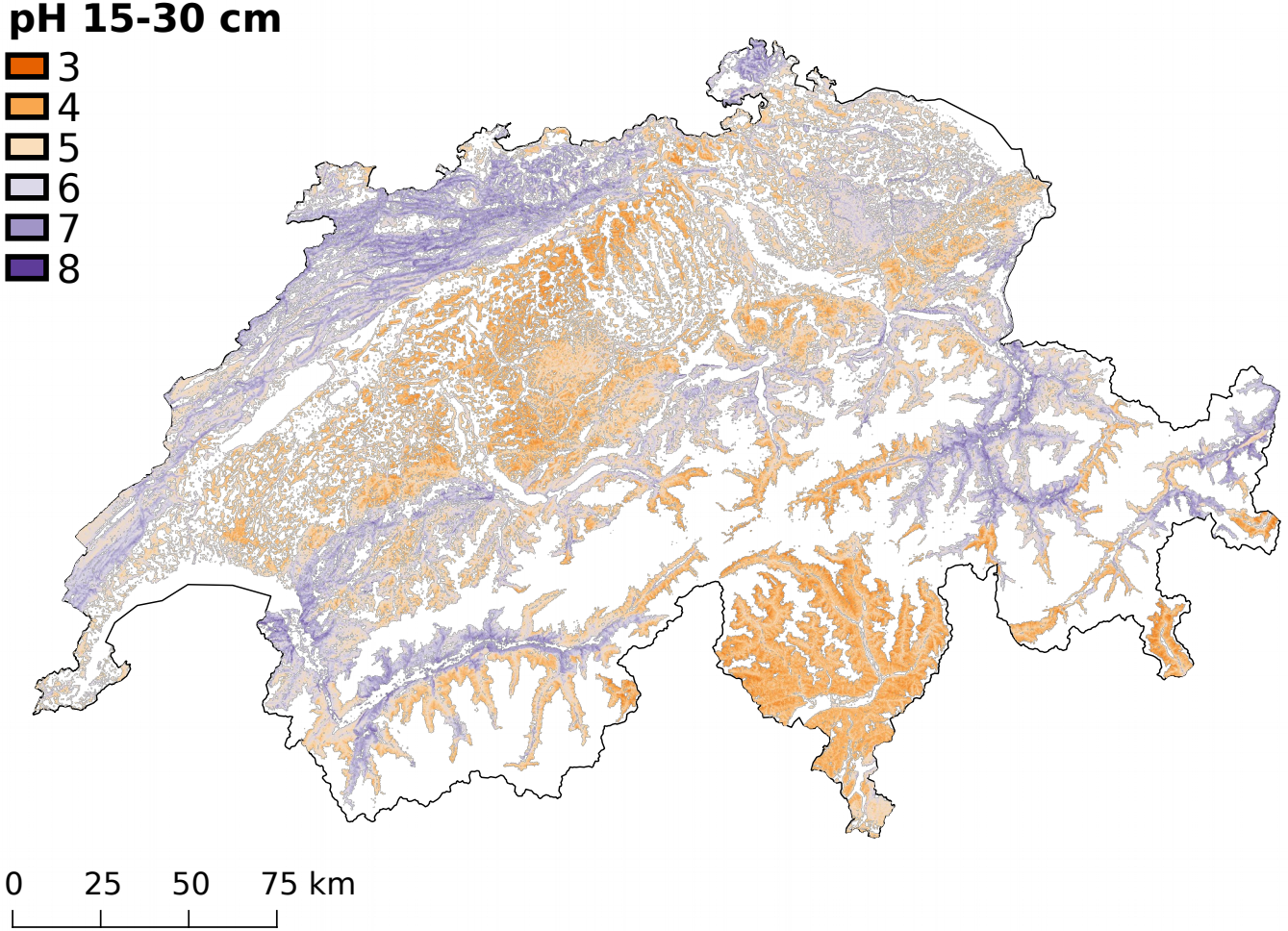


Results – Maps

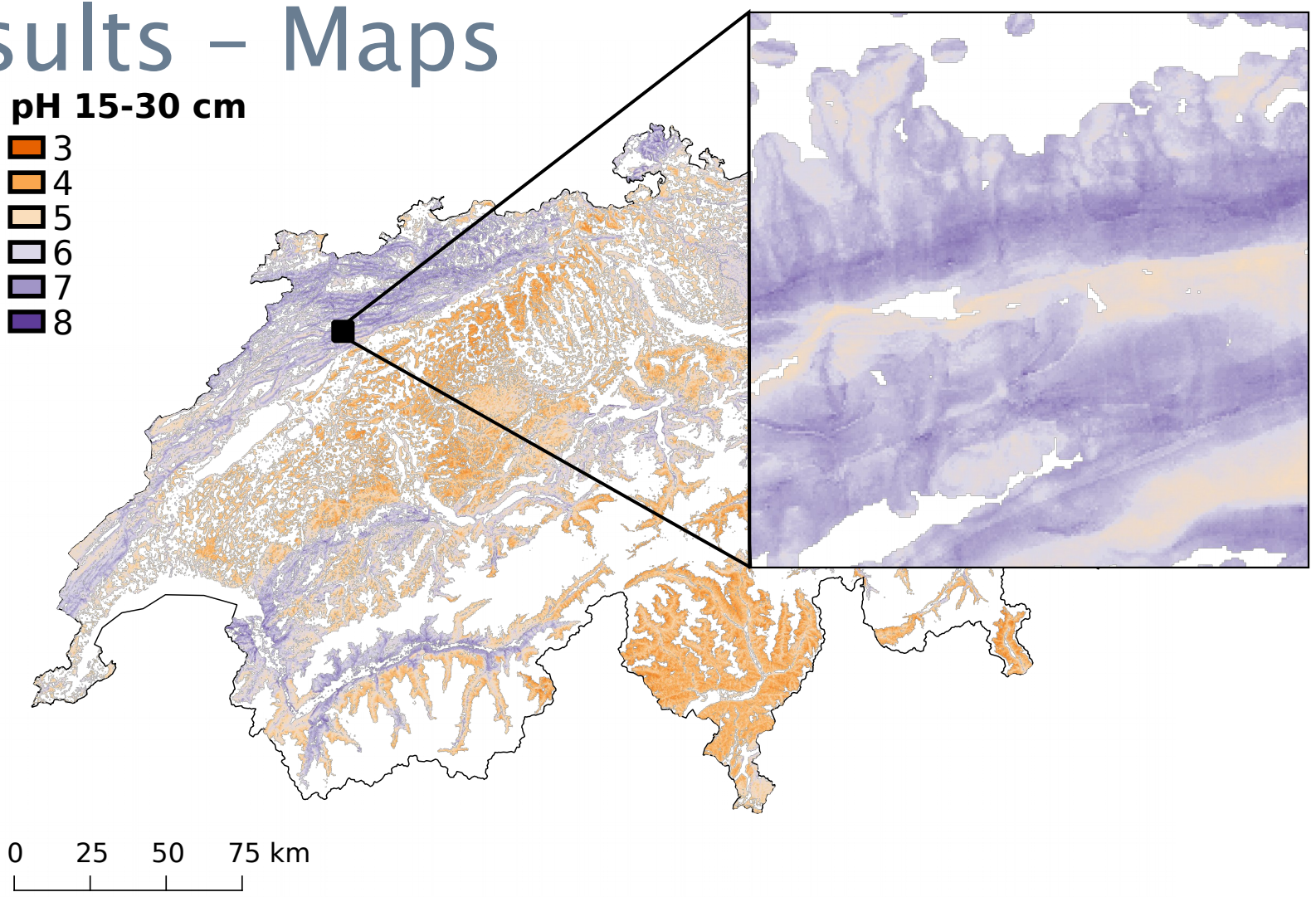




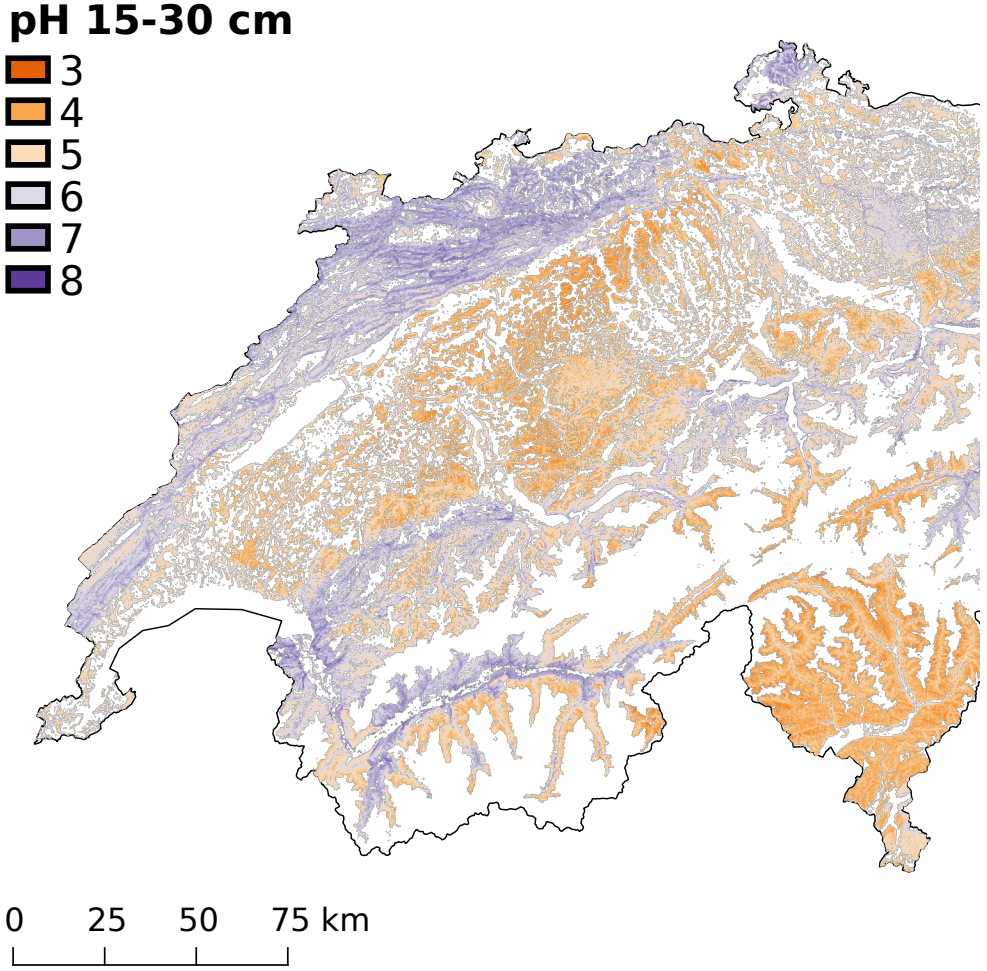
Results – Maps

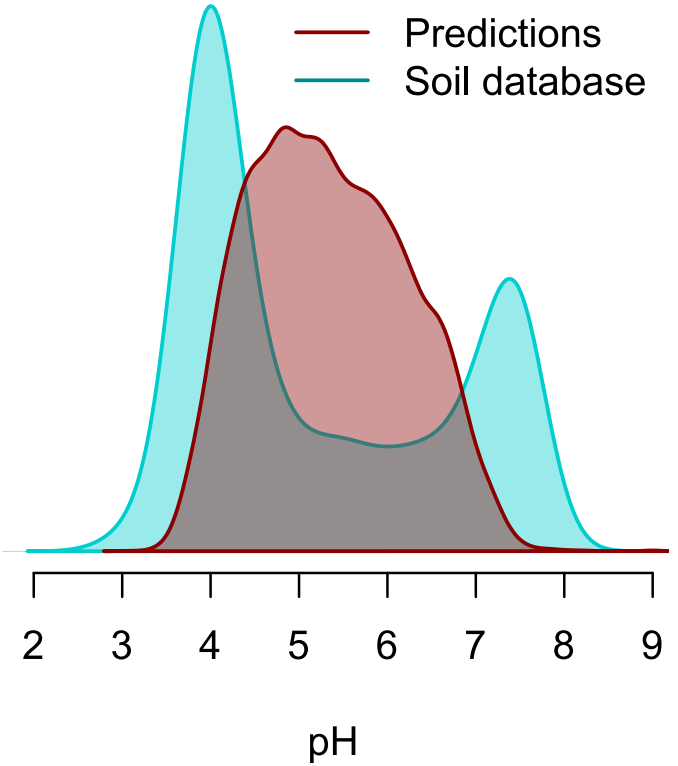


Results – Maps



Results – Maps

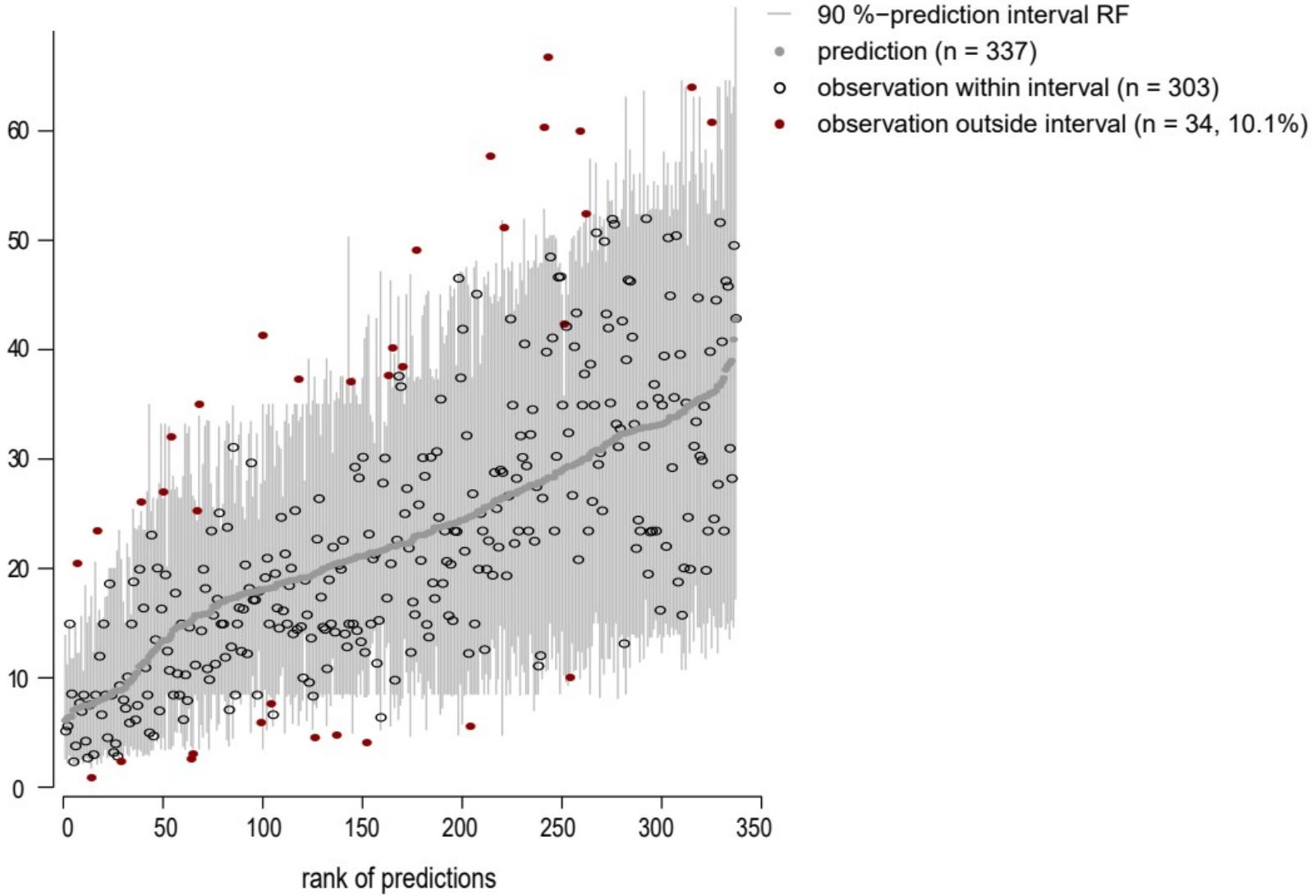




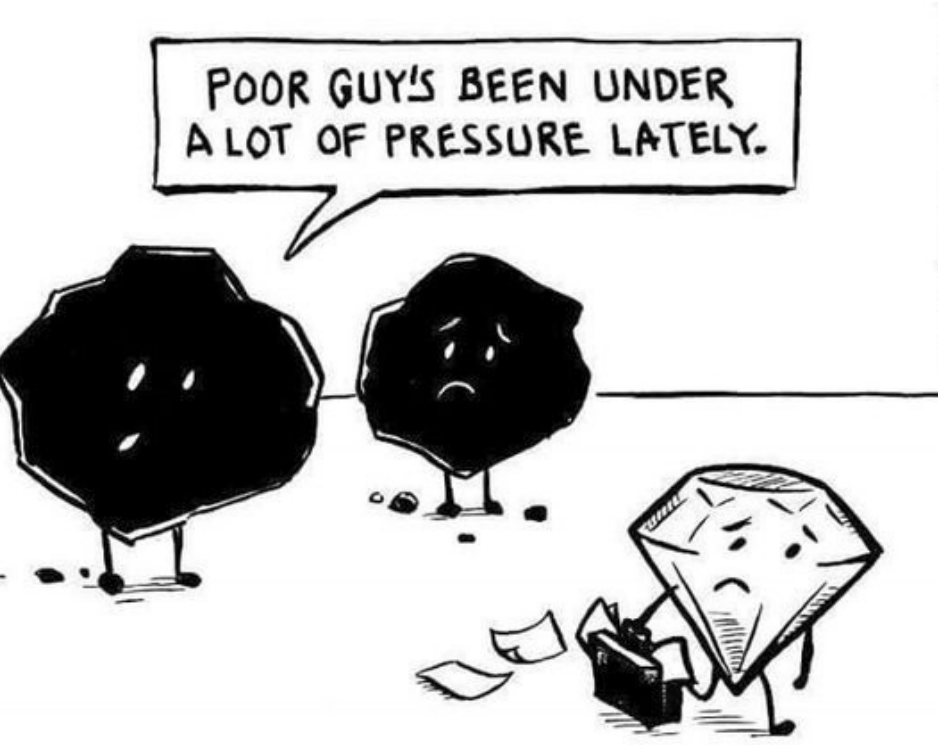
Prediction intervals

Quantile regression forest

Clay 15-30 cm



Conclusions



Pressure to produce for end users, we will publish although:

- Model **performance** low to medium
- Prediction intervals: roughly correct, but **very wide**
- **Artefacts** from covariates at low scale

Bern University
of Applied Sciences

Many thanks to ...

Lorenz Walthert for soil sampling and maintaining the soil data base and

Andri Baltensweiler for preparing the geodata!

And for the funding by WSL