Can AI replace field survey?

Systematic investigation of smoothing of machine learning to predict soil texture

Madlene Nussbaum[1,2], Stefan Vogel[2], Stefan Oechslin[2], Simon Tanner[2], Stéphane Burgos[2]

[1]Utrecht University, Physical Geography, Computational Geography, [2] Bern University of Applied Sciences, Agronomy

Soil texture needed for irrigation

- Potatoes irrigation network for farms in **Switzerland** (https://bewaesserungsnetz.ch)
- Goal: Improve water efficiency and crop quality
- Installation of soil moisture probe per potato parcel
- Calibration needs soil texture in 2 soil depth (USDA texture classes)
- So far done by manual auger and haptic in-situ estimation of soil texture

Investigate smoothing in predictions: 4 arable land study areas

→ total 16 responses: sand + clay in 15 and 50 cm (silt as difference to 100 %)
→ figures show computations on validation sets

	area (km²)	calibration samples	samples per km²	validation samples	validation strategy	R^2 random forest Clay in 50 cm
Zurich arable land of Canton Zurich	710	3500	**4.9**	290	data splitting	0.24
Berne pilot Mineral soils	10	1590	**159**	190	new stratified random sample	0.40
Rhine Valley Cultivated Histosols	38	1980	**52**	140	new stratified random sample	0.78
Seeland Cultivated Gley- and Histosols	60	1620	**27**	120	data splitting	0.44

Can we replace texture sampling with machine learning?

$$y_i \sim f(X_i)$$

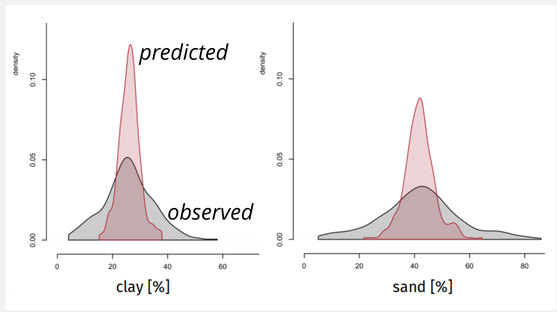
y_i: clay or sand content [%] in 15 or 50 cm soil depth at location i
X_i: soil forming factors represented by terrain attributes, climate maps, geology map, overview soil maps, etc.
$f()$: prediction function found by statistical learning algorithm

Function f(X) allows to predict at new locations where only the maps of soil forming factors X are known.

First results with random forest: Predictions not satisfactory!

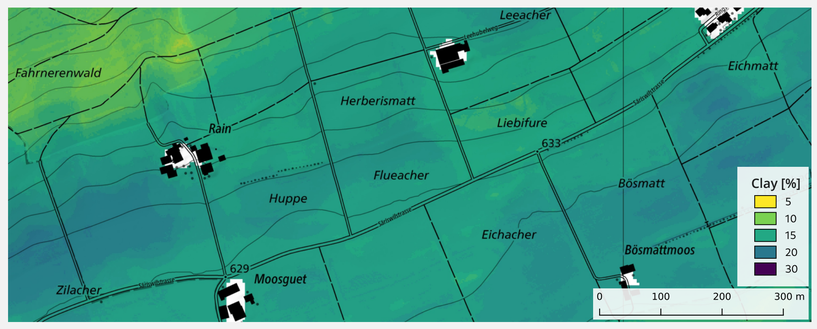
Main problem:

- smoothing of original distribution
- no/reduced prediction of sandy or clay soils
- poor skill to predict USDA texture classes



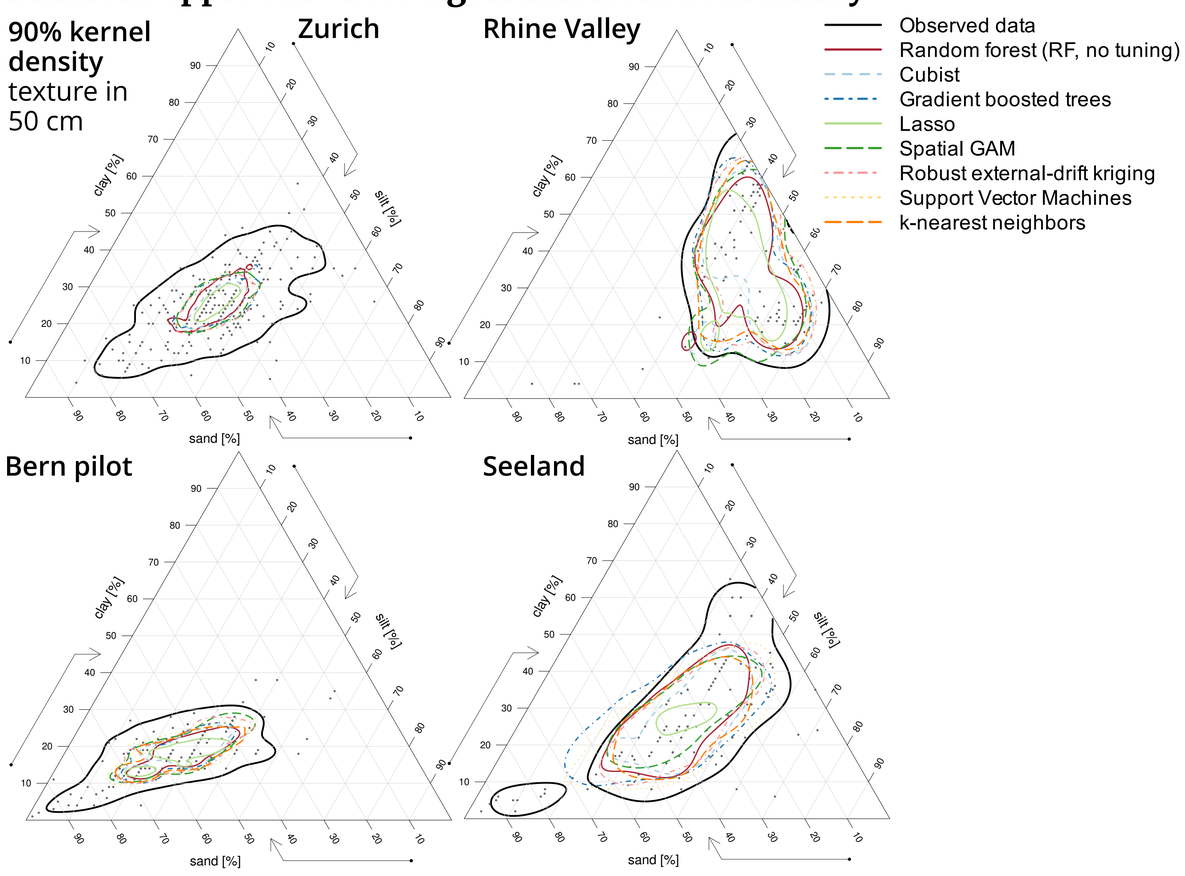
Smoothing also affects map predictions.

Clay content in 0-20 cm soil depth. Section of pilot area north of Berne, Switzerland

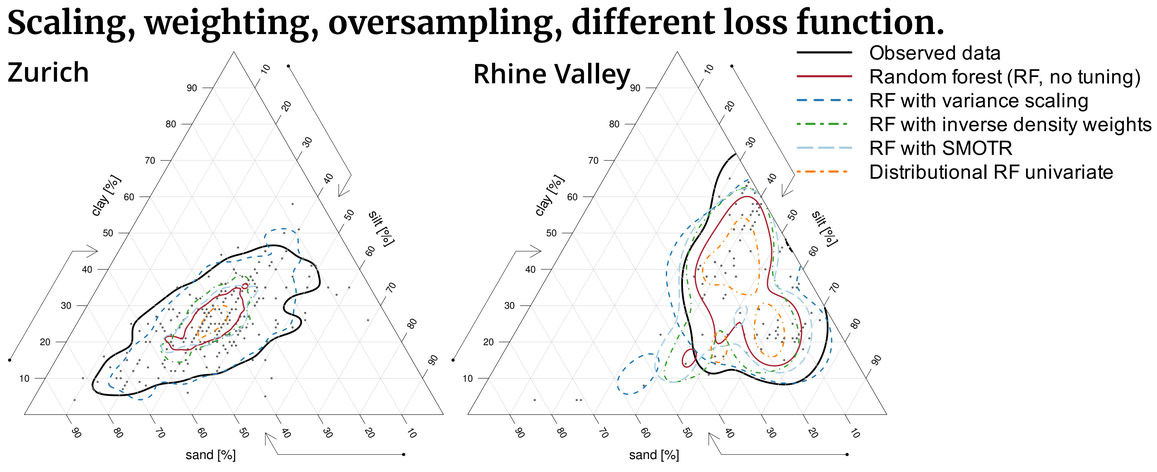


Different approach? Test algorithm from each family.

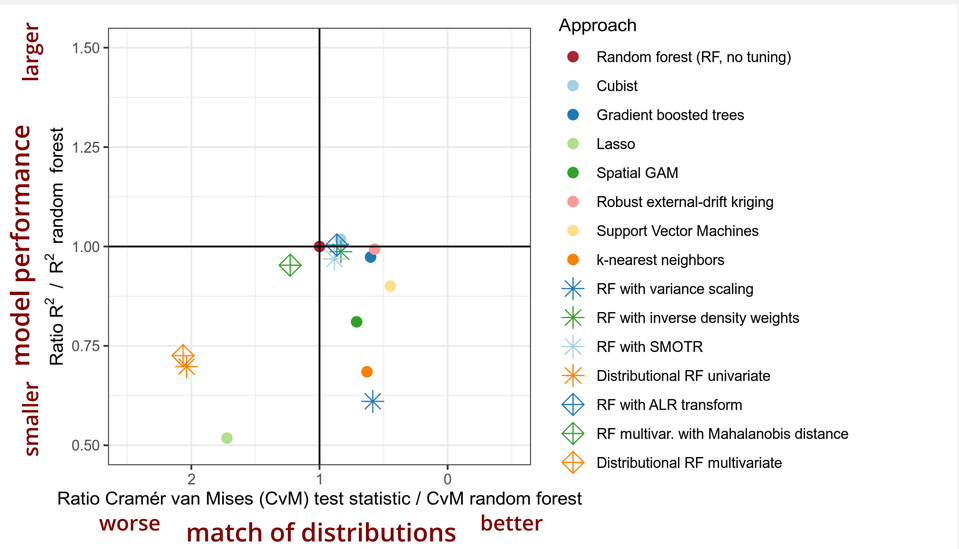
90% kernel density texture in 50 cm



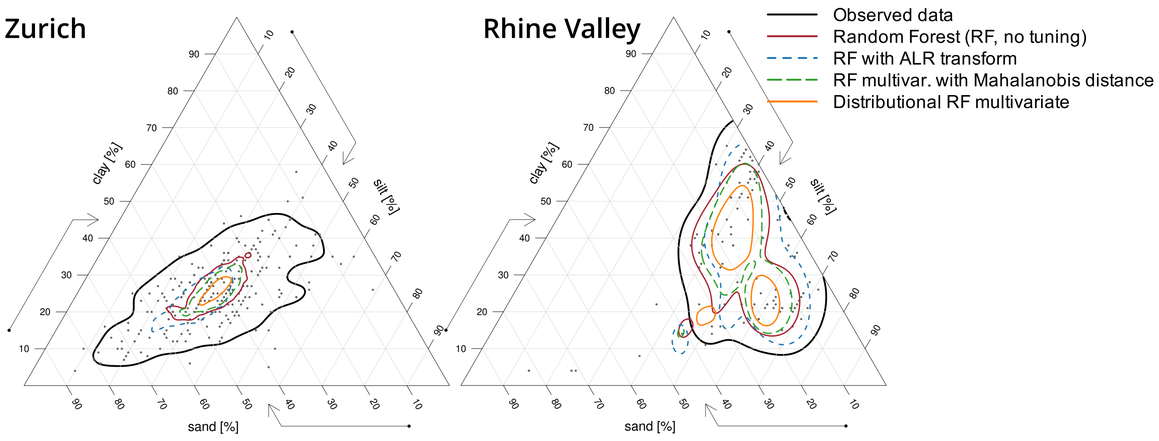
Custom fix for random forest? Scaling, weighting, oversampling, different loss function.



Improvement over random forest



Should we consider clay and sand in one model? Multivariate response.



Conclusions

- All methods failed to well predict clay and sandy soils for all datasets.
- Less smoothing with better model performance.
- No approach outperforms RF clearly. Small advantages of Cubist, SMOTR, inverse weights, ALR.
- Geostatistics performs equally well and multivariate RF disappointing.