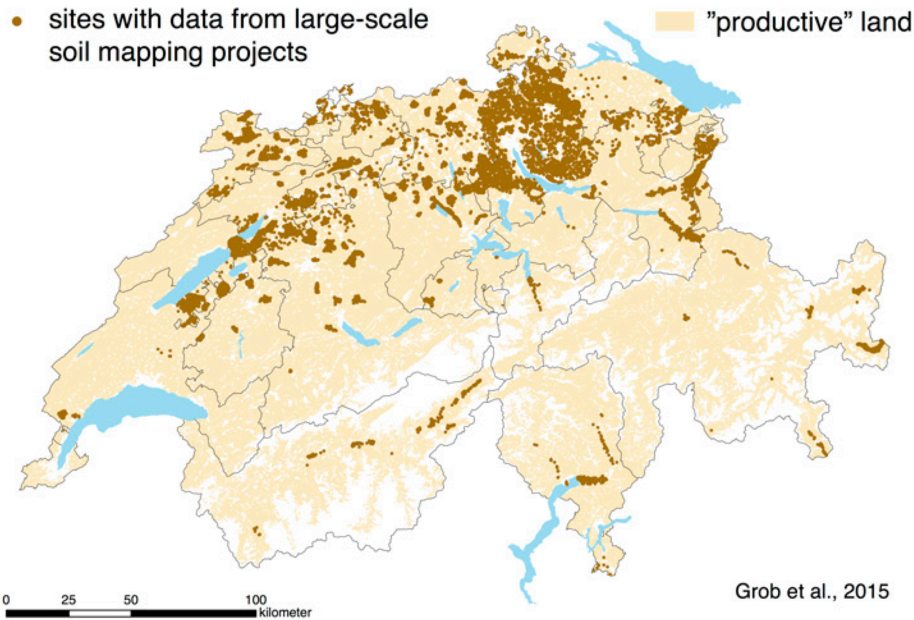
Mapping soil properties at regional scale by robust external drift kriging

Andreas Papritz[1], Madlene Nussbaum[1], Kay Spiess[1];
Marielle Fraefel[2], Andri Baltensweiler[2], Lorenz Walthert[2];
Armin Keller[3], Urs Grob[3]; Sanne Diek[4]

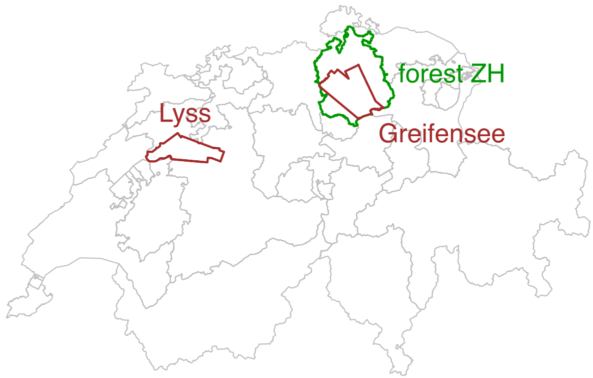
[1]Department of Environmental Systems Science, ETH Zurich;
[2]Swiss Federal Institute for Forest, Snow and Landscape Research WSL;
[3]Swiss Soil Monitoring NABO, Agroscope;
[4]Remote Sensing Laboratories, University of Zurich;
Switzerland

papritz@env.ethz.ch

context: spatial soil information in Switzerland



NRP project PMSoil: DSM pilot study at regional scale



- forest ZH — focus: soil acidification
  $\Rightarrow$ CEC, pH, base saturation, Al_{ex}
- Greifensee and Lyss — focus: agricultural production
  $\Rightarrow$ SOC, pH, silt, clay, stone content, water logging, soil depth
- 2.5D-approach: mapping soil properties for several depth layers

review of statistical approaches for DSM

geostatistics: external-drift kriging

- 😐 only linear relations between response $Y(s)$ and covariates $x(s)$
- ☹ model building difficult for numerous $x(s)$
- ☺ model structure easily interpretable
- ☺ modelling prediction uncertainty

review of statistical approaches for DSM

geostatistics: external-drift kriging

- only linear relations between response $Y(s)$ and covariates $x(s)$
- model building difficult for numerous $x(s)$
- model structure easily interpretable
- modelling prediction uncertainty

tree-based methods: random forest, quantile regression forest

- complex relations between $Y(s)$ and $x(s)$
- model building straightforward
- model structure not easily interpretable
- modelling prediction uncertainty (quantregForest)

review of statistical approaches for DSM

boosted geo-additive models (Hothorn *et al.*, 2011):

- complex relations between $Y(\boldsymbol{s})$ and $\boldsymbol{x}(\boldsymbol{s})$
- model building straightforward (boosting)
- model structure easily interpretable
- modelling prediction uncertainty

$\Rightarrow$ ***main approach***

$\Rightarrow$ ***comparison with external-drift kriging***

model building strategy for external drift kriging

1. selection of initial set of covariates by LASSO

model building strategy for external drift kriging

1. selection of initial set of covariates by LASSO
2. stepwise backward/forward covariate selection based on AIC (REML fit)

model building strategy for external drift kriging

1. selection of initial set of covariates by LASSO

2. stepwise backward/forward covariate selection based on AIC (REML fit)

3. stepwise backward selection based on AIC (REML fit) from model expanded by first-order interactions between covariates

model building strategy for external drift kriging

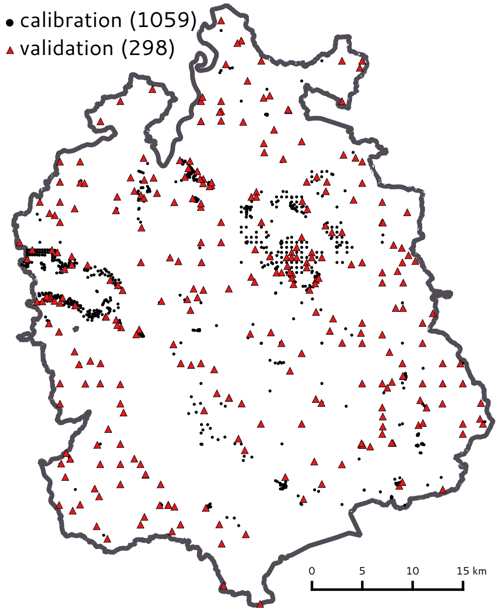
1. selection of initial set of covariates by LASSO
2. stepwise backward/forward covariate selection based on AIC (REML fit)
3. stepwise backward selection based on AIC (REML fit) from model expanded by first-order interactions between covariates
4. manual backward covariate selection by cross-validation

model building strategy for external drift kriging

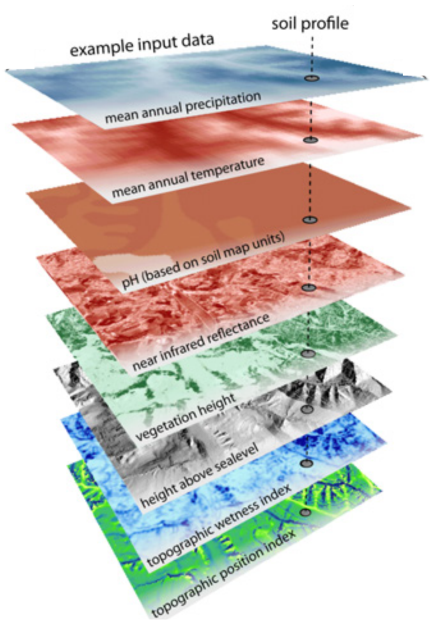
1. selection of initial set of covariates by LASSO
2. stepwise backward/forward covariate selection based on AIC (REML fit)
3. stepwise backward selection based on AIC (REML fit) from model expanded by first-order interactions between covariates
4. manual backward covariate selection by cross-validation
5. manual merging of levels of categorical covariates

$\Rightarrow$ 😰 😰 😰 😰

forest soil CEC_{eff} 0–20 cm legacy data

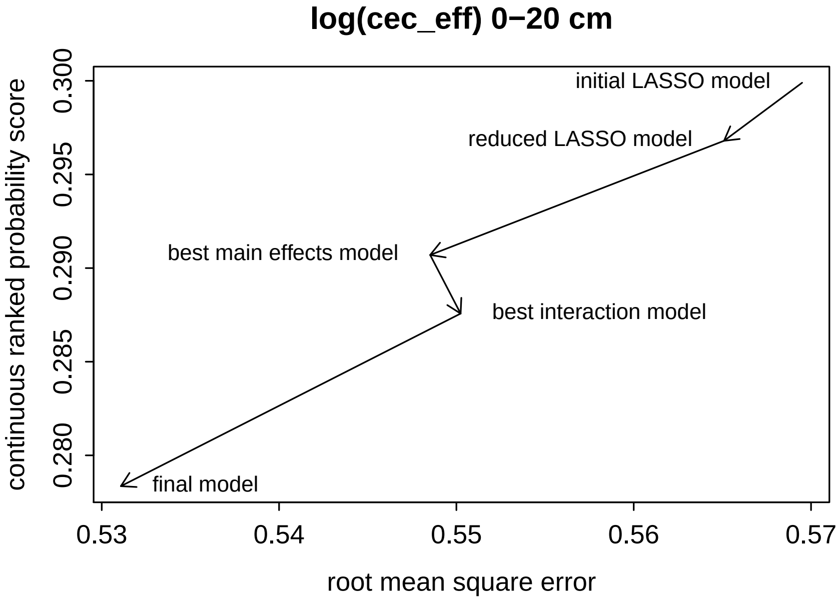


forest soil CEC_{eff} 0–20 cm: > 300 covariates

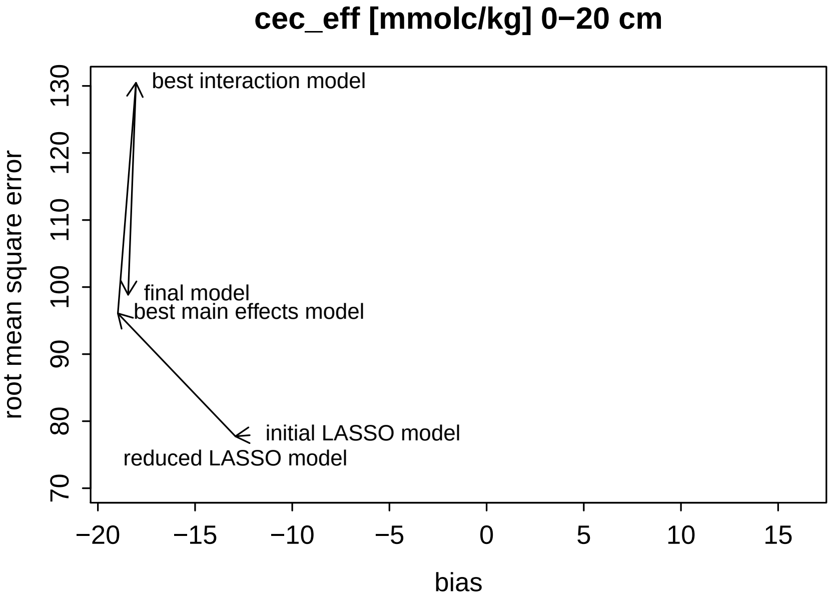


	resolution
climate	
annual rainfall, temperature, …	
period 1961-1990	*25 m*
soil, parent material	
soil map 1:200‘000	
geological map 1:50‘000	
vegetation	
forest type map 1:5'000	
percentage of coniferous trees	*25 m*
SPOT5 mosaic	*10 m*
UK-DCM2 image	*22 m*
digital surface model (lidar)	*2 m*
terrain	
digital elevation model	*25 m*
digital terrain model (lidar)	*2 m*

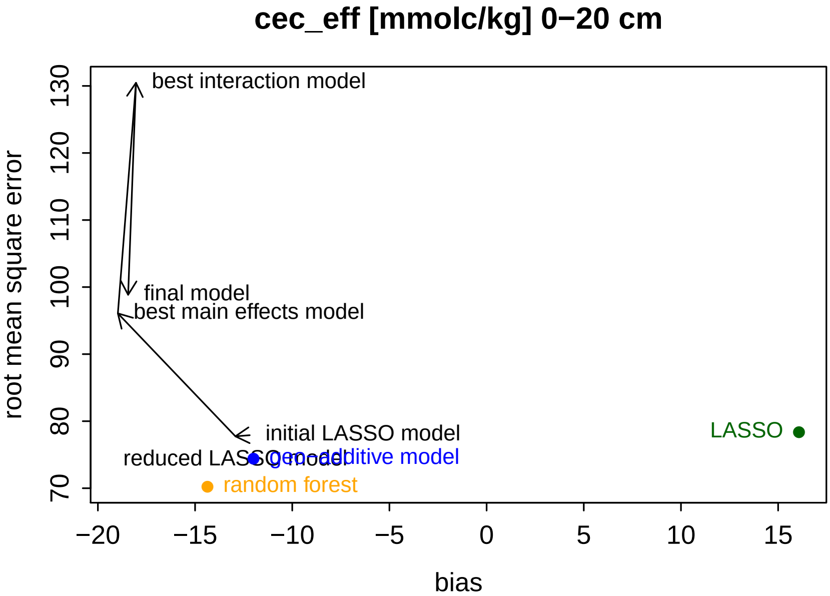
cross-validation results model building



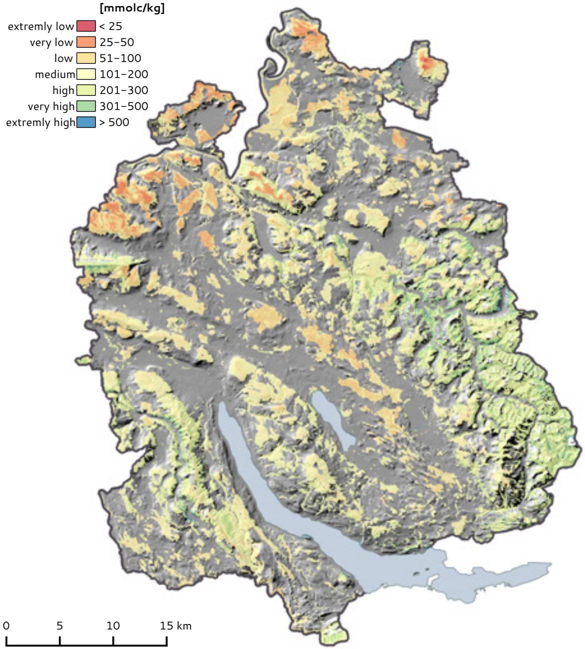
external validation results



external validation results

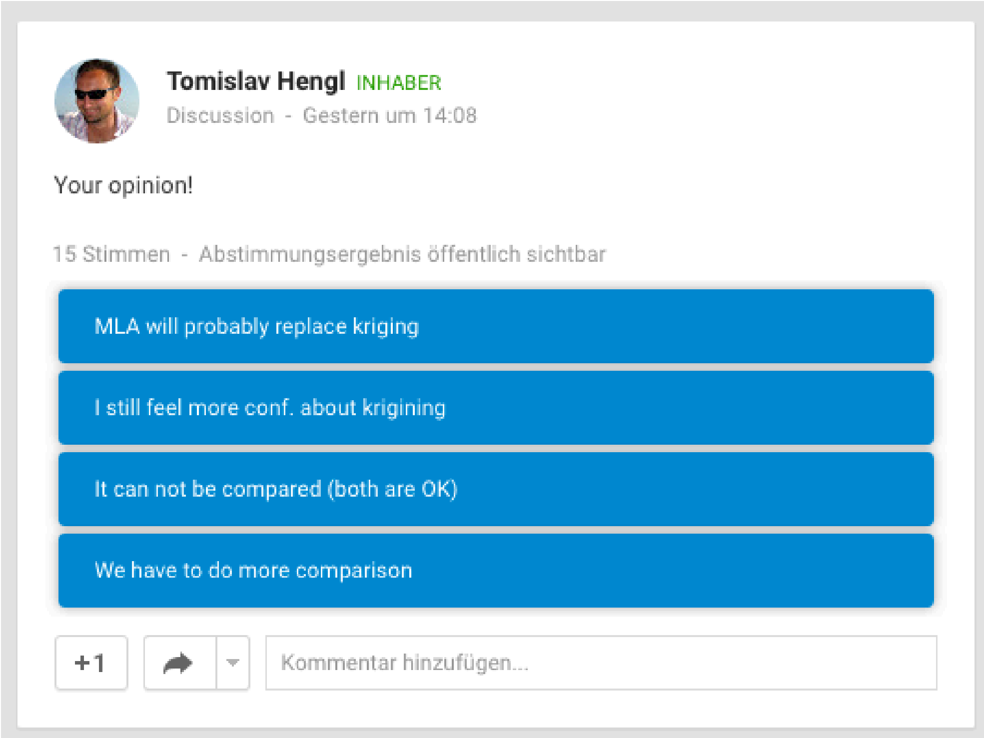


CEC_{eff} predictions by geo-additive model



conclusions

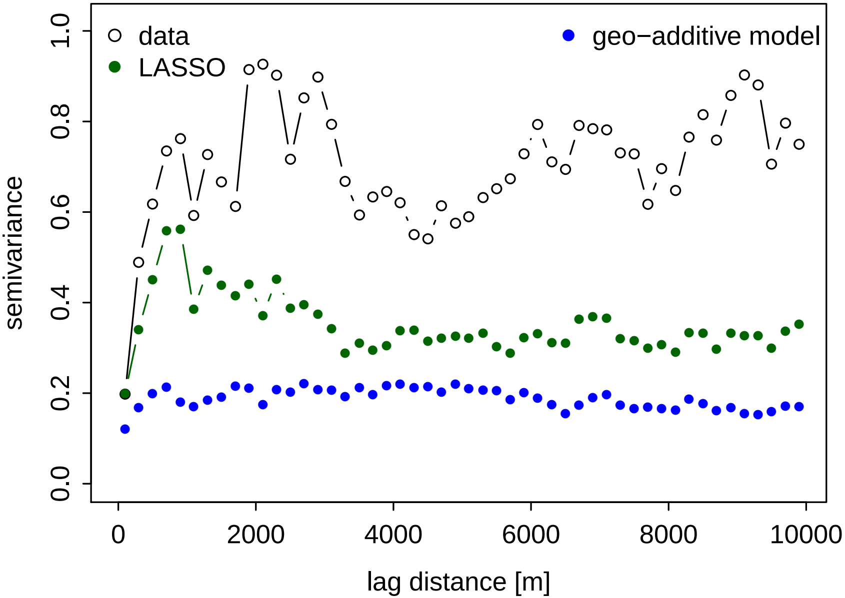
- external-drift kriging tuned for optimal AIC overfits
- boosted geo-additive models and quantile regression forest predict more precisely than external drift-kriging
- external-drift kriging not better for modelling prediction uncertainty?
- LASSO under-rated prediction approach?


Tomislav Hengl INHABER
Discussion - Gestern um 14:08
Your opinion!
15 Stimmen - Abstimmungsergebnis öffentlich sichtbar
MLA will probably replace kriging
I still feel more conf. about krigining
It can not be compared (both are OK)
We have to do more comparison
+1
Kommentar hinzufügen...

acknowledgements

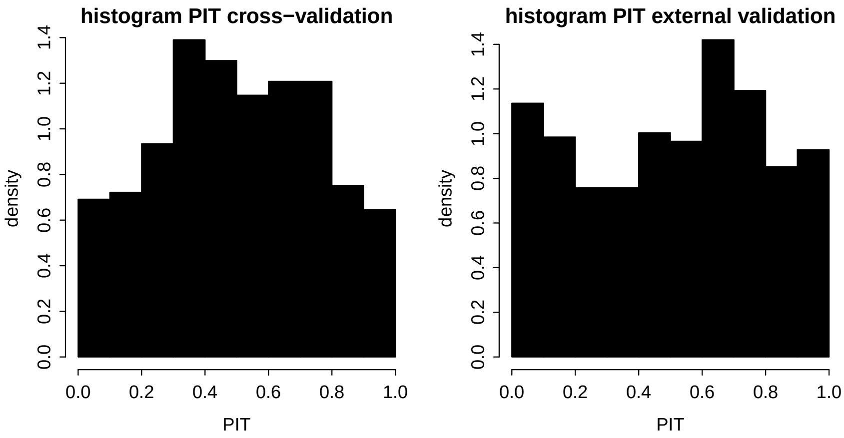
part of this works was supported by the Swiss National Science Foundations under grant no 406840_143096

(residual) variogram log(CEC_{eff} 0–20 cm



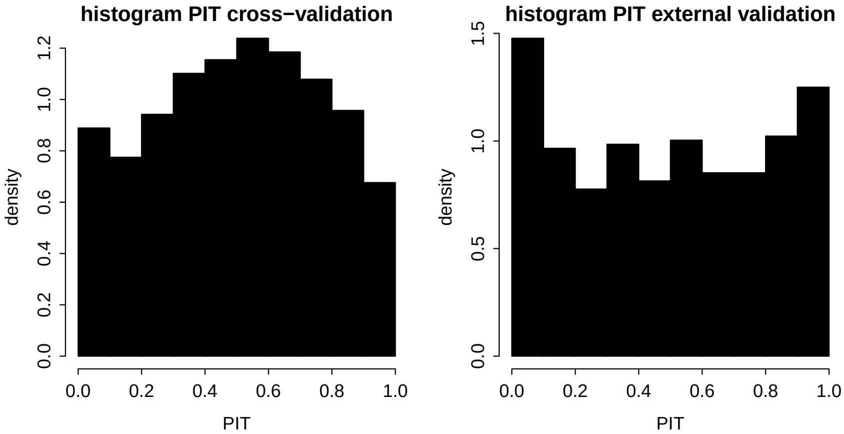
modelling prediction uncertainty: PIT histograms

initial LASSO model

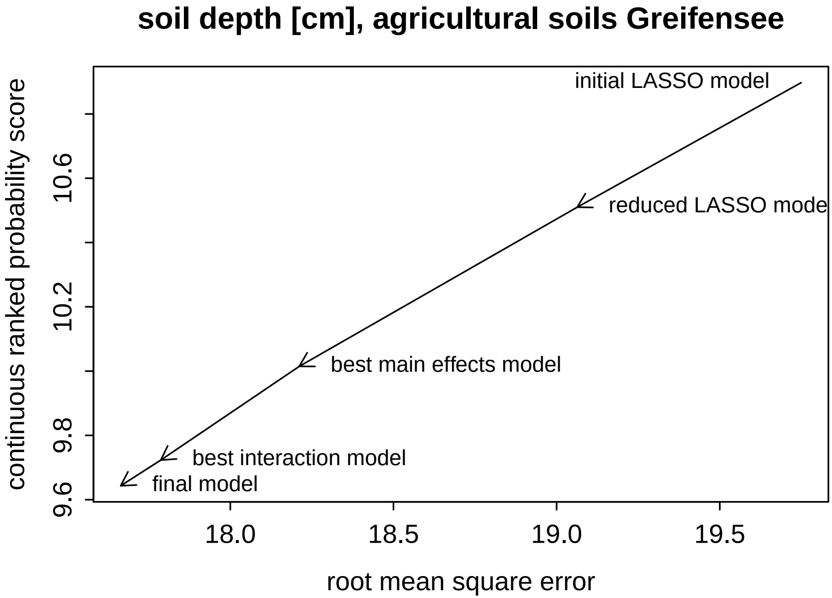


modelling prediction uncertainty: PIT histograms

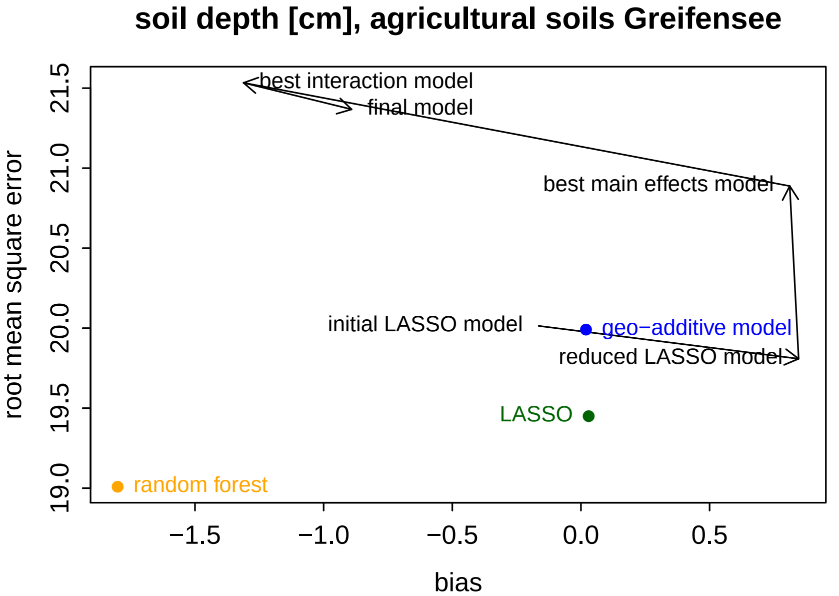
final geostatistical model



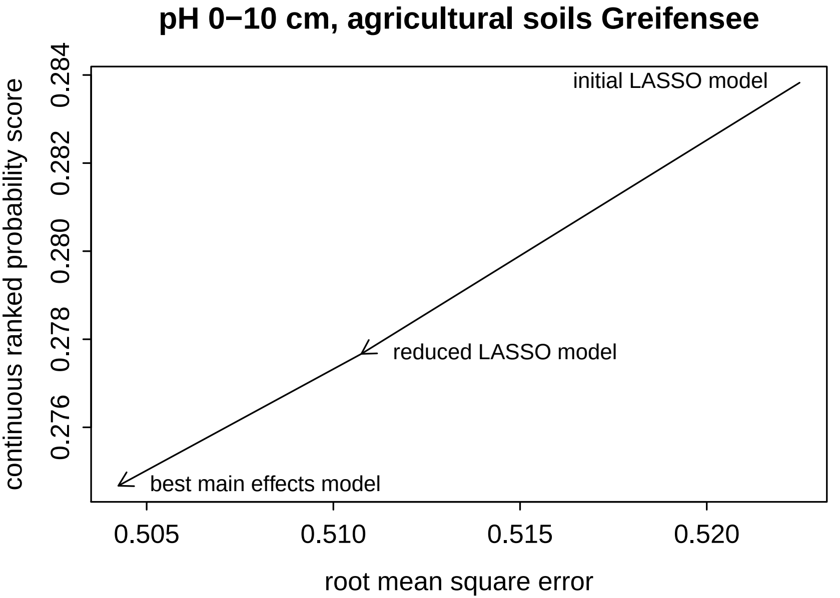
cross-validation results model building



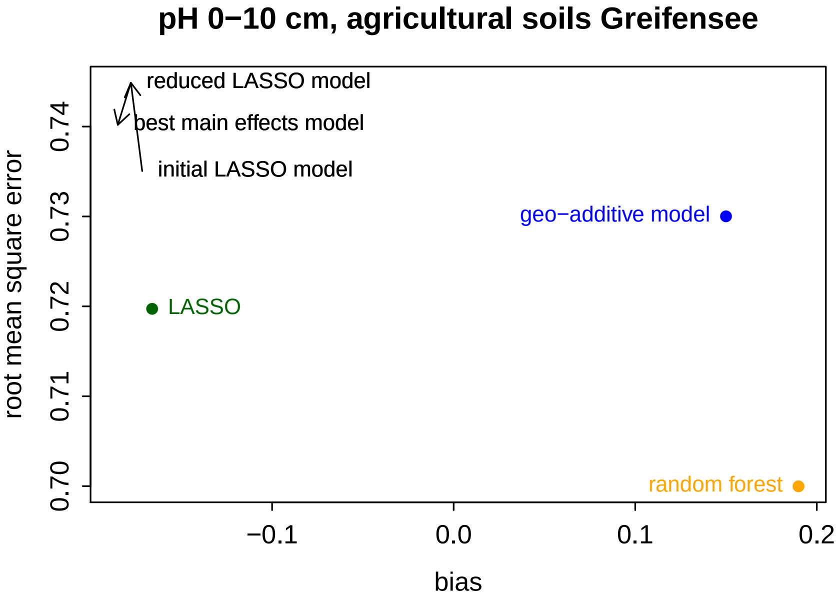
external validation results



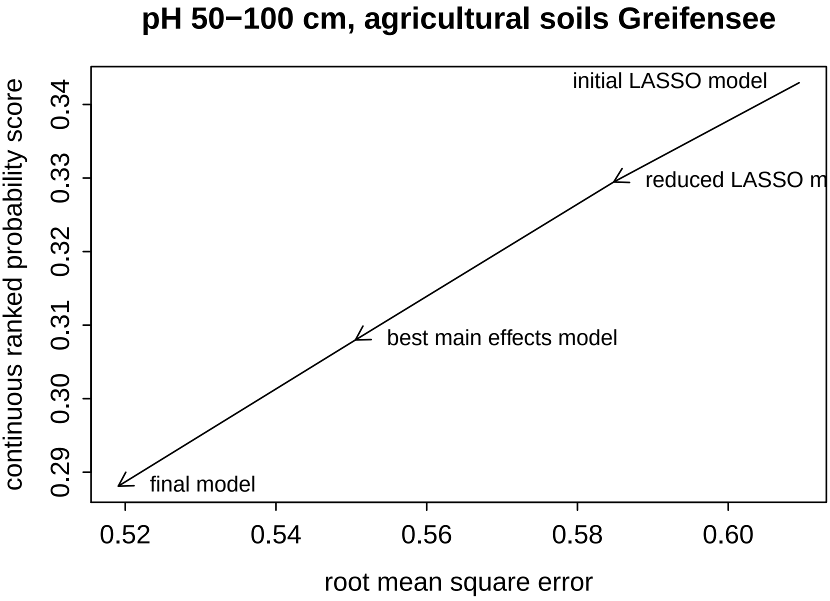
cross-validation results model building



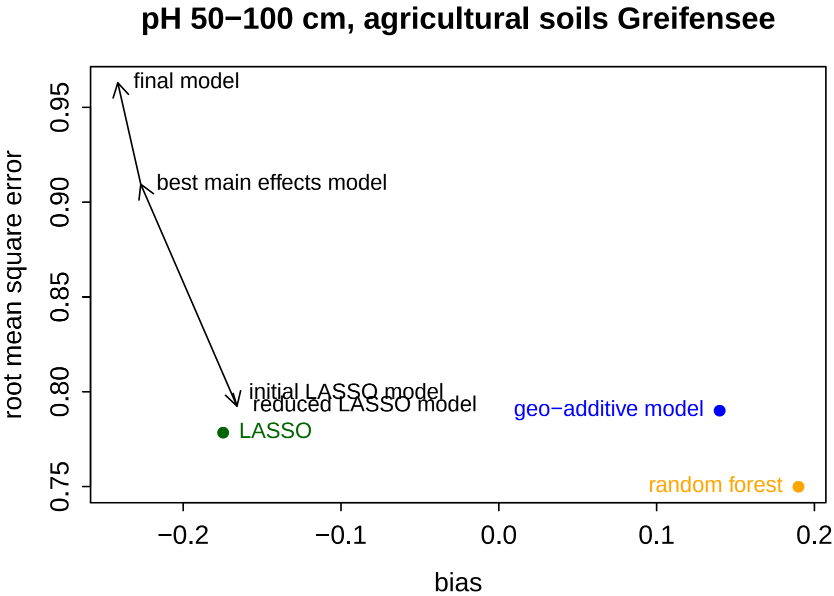
external validation results



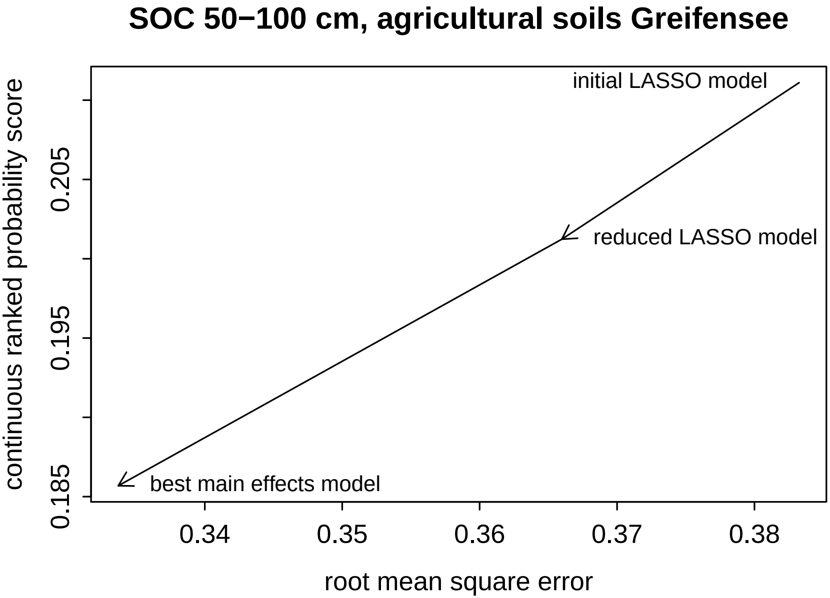
cross-validation results model building



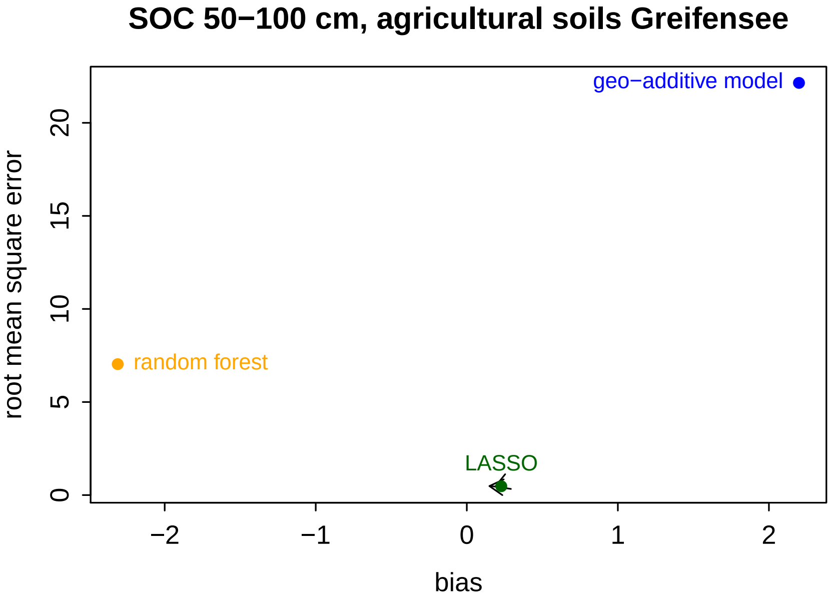
external validation results



cross-validation results model building



external validation results



review of statistical approaches for DSM

boosted geo-additive models (Hothorn *et al.*, 2011):

- complex relations between $Y(\boldsymbol{s})$ and $\boldsymbol{x}(\boldsymbol{s})$

$$\mathsf{E}\left[Y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\right] = \underset{\text{global}}{\sum_k f_{\text{env},k}(x_k(\boldsymbol{s}))} + \underset{\text{spatial}}{f_{\text{s}}(\boldsymbol{s})} + \underset{\text{local}}{\sum_l f^*_{\text{env},l}(x_l(\boldsymbol{s})) \cdot f^*_{\text{s},l}(\boldsymbol{s})}$$

- model building straightforward (boosting)
- model structure easily interpretable
- modelling prediction uncertainty

$\Rightarrow$ ***main approach***

$\Rightarrow$ ***comparison with external-drift kriging and random forest***

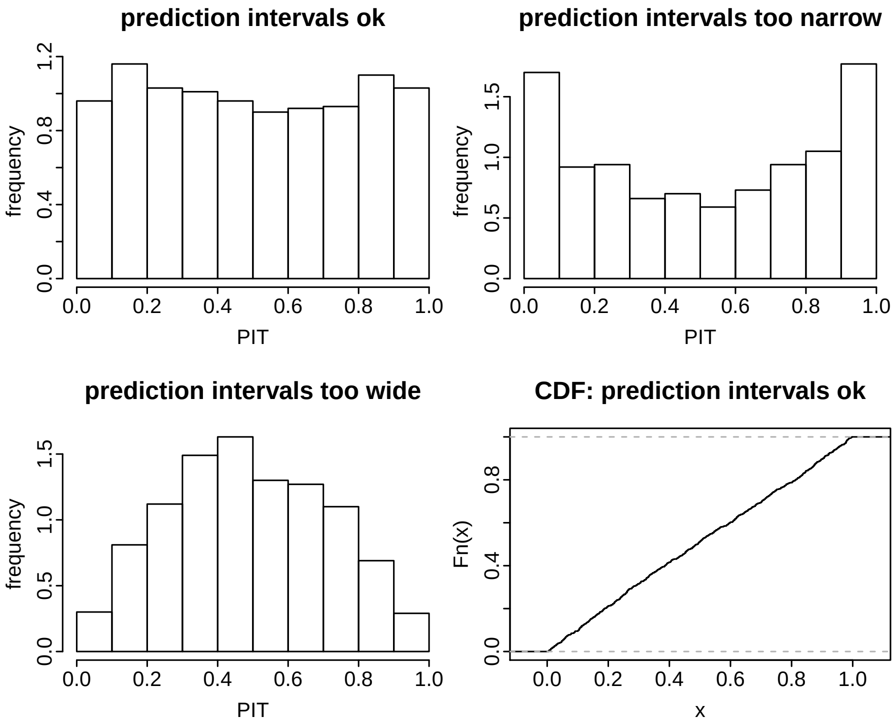
criteria to assess probabilistic predictions

- for Gaussian stochastic processes kriging provides estimates of mean and variance of conditional distribution of target $Y(\boldsymbol{x}'_j)$ given the data $\boldsymbol{Y}$ (predictive distribution)
- denote cdf of predictive distribution by $\widehat{F}_{Y(\boldsymbol{x}'_j)|\boldsymbol{Y}}(y)$
- probability integral transform PIT (Gneiting *et al.*, 2007)

$$\mathsf{PIT}_j = \widehat{F}_{Y(\boldsymbol{x}'_j)|\boldsymbol{Y}}(y_j)$$

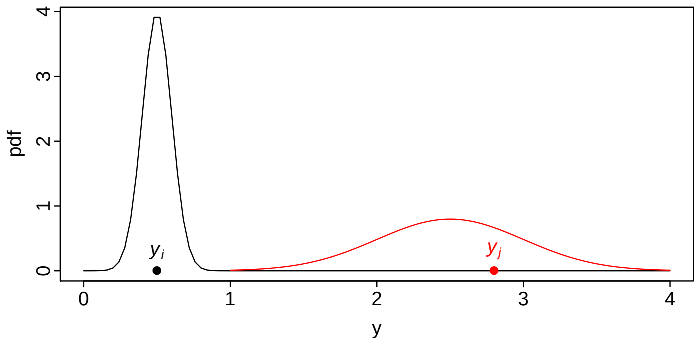
- PIT has a uniform distribution on interval $[0, 1]$ if predictive distribution is ok

$\Rightarrow$ histogram of PIT_j should be flat


prediction intervals ok
frequency
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0
PIT
prediction intervals too narrow
frequency
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
PIT
prediction intervals too wide
frequency
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
PIT
CDF: prediction intervals ok
Fn(x)
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
x

criteria to assess "sharpness" of $\widehat{F}_{Y(\boldsymbol{x}')|\boldsymbol{Y}}(y)$

- overall criterion to assess quality of probabilistic predictions
- predictive distribution is "sharp" if it is narrow (small variance) and is centred on true value (no bias)

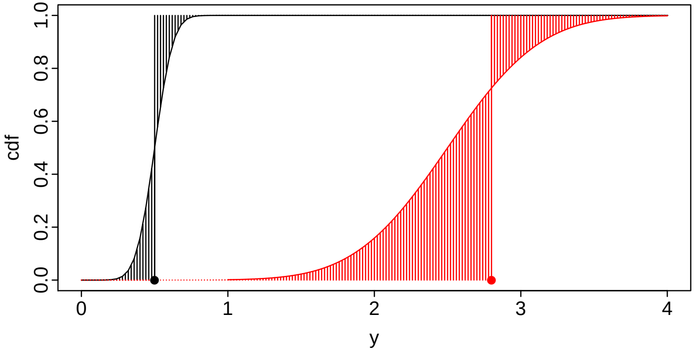


criteria to assess "sharpness" of $\widehat{F}_{Y(\boldsymbol{x}')|\mathbf{Y}}(y)$

- measure for sharpness of predictive distribution for single prediction site $\boldsymbol{x}'_j$

$$\int \{\widehat{F}_{Y(\boldsymbol{x}'_j)|\mathbf{Y}}(y) - I(y_j \leq y)\}^2 dy$$

where $I(A)$ is indicator function with value equal to 1 if A is true and zero otherwise



continuous ranked probability score

- continuous ranked probability score (CRPS) measures average sharpness of predictive distributions for all sites of a data set

$$\text{CRPS} = \frac{1}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty}\{\widehat{F}_{Y(\boldsymbol{x}'_j)|\boldsymbol{Y}}(y) - I(y_j \leq y)\}^2 dy$$

- CRPS equal to integral over Brier score (BS = averaged MSEP for predicting that observations y_j do not exceed cutoff y)

$$BS(y) = \frac{1}{n}\sum_{j=1}^{n}\{\widehat{F}_{Y(\boldsymbol{x}'_j)|\boldsymbol{Y}}(y) - I(y_j \leq y)\}^2$$

⇒ CRPS criterion of choice for assessing quality of probabilistic predictions (strictly proper scoring rule, cf. Gneiting *et al.*, 2007)

References

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, **69**(2), 243–268.

Grob, U., Ruef, A., Zihlmann, U., Klauser, L., and Keller, A. (2015). Agroscope-Bodendatenarchiv : Bodendaten aus Bodenkartierungen 1953–1996. Agroscope Science 14, Agroscope Reckenholz, Zürich. http://www.agroscope.ch/publikationen/07703/07705/index.html?lang=de.

Hothorn, T., Müller, J., Schröder, B., Kneib, T., and Brandl, R. (2011). Decomposing environmental, spatial, and spatiotemporal components of species distributions. *Ecological Monographs*, **81**(2), 329–347.