SOIL Discuss.,
https://doi.org/10.5194/soil-2017-14-AC2, 2017

Interactive comment on "Evaluation of digital soil mapping approaches with large sets of environmental covariates" *by* Madlene Nussbaum et al.

Madlene Nussbaum et al.

madlene.nussbaum@env.ethz.ch

Received and published: 2 September 2017

Many thanks you for your detailed and very helpful feedback. We comment on your review in the subsequent text (P: page, L: line). Our answers to minor, more specific comments were directly added to the supplement to this document with our suggestions for changes of the manuscript.

Benefit of large sets of covariates (comment P2L21)

P2 L21: You state that you presume DSM will benefit from a large number of covariates. Why then do you aim to find models which eliminate covariates? This presumption seems at odds with the entire premise of this paper. For modelling datasets of this size, the advantage to feature elimination is to avoid over-fitting but above all to enable model interpretation. The time advantages tend to be small (i.e. irrelevant) given the computing power available in desktop computers these days.

Numerous studies cited in the introduction of the manuscript demonstrated the benefit of using large sets of covariates for DSM. But there are still strong reasons for favouring parsimonious models: We obviously agree that ease of interpretation of modelled effects and avoiding of over-fitting are arguments in favour of such models. However, we do not fully agree on your opinion that computational gains are irrelevant. This is likely the case for fitting models, but it is clearly not true for pre-processing covariates prior to computing predictions. The respective effort depends linearly on the number of covariates in the models. It makes a difference if the data of 300 or only 20 covariates must be pre-processed. Hence, ease of interpretation, avoiding over-fitting and optimization of computational efforts for computing predictions are the main incentives for building parsimonious models from large sets of covariates. Selecting covariates a priori based on expert knowledge has been shown to be inferior to model-based covariate selection (Brungard et al. 2015). Hence, a challenge for DSM remains to select the relevant covariates for each soil property, soil depth and study region.

Time aspect in soil legacy data (P7L15)

P7 L15: which statistical models did you use to harmonise your data temporally? Your description implies that you fit your models to all data from all years, using the year *as a categorical predictor variable in the respective model, but predicted only for recent years. However, this will not account for temporal changes in a response variable. To do this, you would have to fit models to predict current response based on past measurements (y(t=current) f(y(t=past)) and then predict the current y at each location as a function of past collected data. Obviously, this can only be down where re-sampling of the same sites between sampling campaigns has taken place. This harmonised data set can then be used to train models of the current data for DSM purposes. Merely including a year in the code for a model will not do this, because you will still train the model on outdated data, which may be spatio-temporally distinct (i.e. models will train on one location sampled in 1996 adjacent to another location sampled in 2006, which will affect the spatial predictions but not account for temporal differences between sampling campaigns). If you do not have enough re-sampled sites, you could create separate models for each sampling campaign and then look at correlation between model predictions between years to derive your correction models and harmonise your dataset.*

Our wording “temporal changes” was admittedly somewhat misleading (we will mend this in the revised manuscript): We did not attempt to model the temporal evolution of soil properties by time series models because we did either not have repeated samples at the same locations at all (Berne region), or we did not have enough data to calibrate such models in a meaningful way (Greifensee, ZH forests). We do not believe that a time series modelling approach would improve the correction for time. The sites were chosen by purposive sampling (see comment below) and do therefore not give unbiased estimates for each sampling period. Moreover, the maps would be calibrated on rather small data sets for each sampling period.

We added — as you correctly describe in your comment — therefore a categorical covariate that grouped the data by sampling year and type of observation (field estimates, measurements, predictions by pedotransfer function). We agree that this approach is far from satisfactory and does not allow us to distinguish between real temporal

changes of soil properties and analytical artifacts caused by changes in procedures, fluctuations between laboratories, between sample batches, etc. Our correction accounts for the lumped effect of all these causes. Our objective was to adjust the data to a common level, irrespective of the cause of the discrepancies. In doing so, we implicitly assume that all the fluctuations are analytical artifacts, and soil properties do not change over time. This assumption is likely reasonable for physical properties like soil texture, gravel content and soil depth, but it might be questionable for SOM and pH. But also for these chemical properties, fluctuations by analytical artifacts are by no means negligible (unpublished study of data gathered for the soil monitoring programme of the Canton of Zurich). If we would not have adjusted the data — as it is usually done when legacy data are used for DSM — then the modelled spatial patterns would be partly induced by temporal effects, either caused by real changes or analytical artifacts. We do not know how large these effects would be. However, we know that the legacy data correction was for many responses an important covariate.

Soil texture – separate model for *sand*

P10 L4ff: Why not model sand individually, or compute either clay or silt to sum up to 100 %. This modelling approach seems arbitrary to me.

This was remarked by referee 1 as well, hence we repeat our arguments here:

You suggested to model *sand* content separately instead of just computing it as the remainder of the sum of *clay* and *silt* content to 100 %. We agree that it would be nice to predict *sand* content with meaningful estimates of prediction uncertainty. Nevertheless, we refrained from separately modelling *sand* content because a substantial part of the soil texture data were field estimates by soil surveyors. For field estimates, sand content is computed as the remainder of the sum of the estimated clay and silt content to 100 % (Brunner et al. 1997, Jaeggli et al. 1998). Furthermore, soil function assessment (Greiner et al. 2017) relies on clay and silt content as input. Hence, uncertainty assessment for soil functions using soil texture data does not depend on predictive distributions for *sand* content.

Covariate importance

P13 L8 and L12: Variable importance $>$ 0 does not necessarily mean that the variable is relevant to the model. A variable is relevant if variable importance is greater than expected variable importance for a random model (cf relative variable importance, Hobley et al. 2015, Plant and Soil).

We agree that covariate importance of > 0 does not mean that a covariate is indeed important. To evaluate how effectively RT and BRT reduce the covariate set, we had to count how many covariates had been used at least once for a fitted model. This number is by definition equal to the number of covariates with importance > 0. But this criterion was not used to select any covariates during model building by RF or BRT. And by the way: Selecting non-relevant covariates by the relative covariate importance ($\frac{1}{p}$ for p covariates, Hobley et al. 2015) seems arbitrary to us. We prefer to use recursive backward elimination (Brungard et al. 2015) for covariate selection for RF and BRT.

Evaluation of overfitting

P18 L8ff: The general definition of overfitting is not that cross-validation error differs from external validation error, but that during model construction, the fit improves with increasing model complexity (R2-fit evaluated on the training dataset) despite a lack of improvement of predictive performance (R2-prediction evaluated on the test dataset)

cf. Hasties et al. As such, I am neither surprised nor concerned by difference between cross-validation RMSE and external validation RMSE and I would attribute this to differences between evaluation and fit datasets, not to overfitting. Looking at the statistics in Tables S3-S7, there were several validation datasets whose range lay outside that of the training dataset (e.g. gravel below 10 cm depth, SOM in top 10 cm, pH below 30 cm depth). This will detrimentally affect the tree based methods, because that cannot interpolate to data outside their range, whereas the other models can (though as you point out, they can interpolate to nonsensical data e.g. negative textural proportions). Thus, it is highly likely that these results reflect differences in the training and test data. In fact, you did not assess overfitting because you do not report the improvement in R2-prediction as a function of R2-fit. This section should be cut.

We acknowledge that our use of the term “over-fitting” is not in accordance with its strict definition by Hastie et al. (2009). However, we do not agree with your view that the differences in $\mathsf{SS}_{\mathsf{mse}}$, observed between cross-validation and independent validation, must be attributed solely to random variation between calibration and validation sets. Figure 7 of our manuscripts shows that for the very large majority of the evaluated cases, cross-validation $\mathsf{SS}_{\mathsf{mse}}$ were smaller than validation $\mathsf{SS}_{\mathsf{mse}}$. As we split the data sets by weighted random sampling one would expect a more even distribution of positive and negative differences of $\mathsf{SS}_{\mathsf{mse}}$. Furthermore, we believe that cross-validation MSE does not provide an unbiased estimate of the expected test error Err (Hastie et al. 2009, eq. 7.16) for the calibration sets because we used cross-validation on the same data to tune all the parameters of the various methods or to select covariates.

We will therefore not drop this section but we will carefully revise in the spirit of the above thoughts, thereby acknowledging that our definition of over-fitting is more general than Hastie et al.’s.

Revision of introduction and methodology

I found the Introduction somewhat poorly constructed. The introduction you frame the state of knowledge and identify the gaps, lastly stating how you try to fill the knowledge gap(s). Some of the paragraphs are without clear structure or connections between the sentences. A paragraph should start with an introductory sentence to the topic, provide supporting information, and finish with a concluding sentence. They should not be lists of information. Paragraphs should clearly fit within the introductory purpose of framing the knowledge gaps or methods of the current paper. See e.g. P3 L18-24 (connection between sentences, is there a point to this information?) and P3 L25-29 (why is this relevant? You use neither svm nor ann!) for examples or poor paragraph construction. Furthermore, there is cut-over between the Introduction and the Model descriptions (Section 3). I suggest you re-read these sections, cut out repetition and irrelevant information and tighten the Introduction to frame the importance of your study within the field.

The descriptions of the models (Methodology section 3.1-3.6) are at times vague and generally appear better suited as introductory remarks rather than methods. Consider restructuring.

We agree on your comments on the introduction. We will revise the introduction accordingly and we will consider restructuring the methods section before resubmission. Moreover, we will revise the language of the manuscript.

Answer to further minor comments

Purposive sampling (P6L15-17)

P6 L17: "The sites for WSL were chosen purposively according to the aims of the project". Vague and borderline tautology. I would hope that all sites for scientific studies are chosen on purpose!

The wording "... sites were selected purposively ... " and "... were chosen purposively ..." is used in textbooks and research articles (Brus et al. 2011, Webster and Lark 2013; Webster and Oliver 2007, p. 45). In purposive sampling, locations for soil investigations are selected by an experienced soil surveyor where he or she thinks that the site best represents the local conditions. It is — opposed to random sampling — a non-probability sampling strategy and sometimes called targeted sampling. We suggest to slightly change the text and to add a reference.

Soil function assessment (Table 1)

P4 L17 and Table 1: I do not know how you know that these properties are required for assessing regulation, habitat and production functions. This is also for the caption to Table 1 (Basic soil properties needed). How did you assess the requirement to include these properties? SD is not defined in the Caption of Table 1.

One work-package of the PMSoil project — in which frame the work for our manuscript was done — reviewed methods for a thematic broad assessment of soil functions (Greiner et al. 2017). Based on this review we chose a comprehensive set of soil functions that was (spatially) assessed in the 3 study regions. A subset of the soil properties needed in this assessment are listed in Table 1 of our manuscript.

Covariate interpretation (section 4.3.2)

Section 4.3.2: You only give one example of covariate interpretation. I realise that given the number of models and covariates you cannot interpret everything, but for me the main advantage to covariate reduction is to enable model interpretation (for a dataset you size the time factor between different model runs is insignificant). I would either expand this section to include more examples or cut it completely.

We agree on your comment that discussing two covariates of one model does not allow insight in the structure of the 48 models, each containing numerous covariates. But detailed interpretation of covariate effects for all models would clearly be beyond the scope of this manuscript. By discussing the two covariates we intended to demonstrate the possible options of covariate interpretation (partial residual and dependence plots for a continuous and a categorical covariate). To our knowledge partial dependence plots were rarely presented for tree based methods in digital soil mapping studies (an exception is Behrens et al. 2014). Hence, we prefer to leave this section unchanged as an illustration of suitable tools for inspection of covariate effects.

Predictive skill (Figure 6)

Figure 6: The y-axis is not labelled correctly (what do the numbers mean?). You imply in your methods that $\mathrm{SS}_{\mathrm{MSE}}$ is less than or equal to 1, but the values reported here are all above 1 (reported as %). "Covariate topic" is unclear.

Thank you for pointing out the missing y-axis labels. The units do not represent the $\mathrm{SS}_{\mathrm{mse}}$, but the "predictive skill" of a covariate. Computational details are given in the caption of the figure. Covariate topic or theme refers to the content of the covariate (climate, vegetation, soil etc.). We refer in the figure caption to Table 3 where the datasets are listed under the same thematic names.

References

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, 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, 10.1016/j.geoderma.2014.09.019, 2015.

Brunner, J., Jäggli, F., Nievergelt, J., and Peyer, K.: Kartieren und Beurteilen von Landwirtschaftsböden, FAL Schriftenreihe 24, Eidgenössische Forschungsanstalt für Agrarökologie und Landbau, Zürich-Reckenholz (FAL), 1997.

Brus, D. J., Kempen, B., and Heuvelink, G. B. M.: Sampling for validation of digital soil maps, European Journal of Soil Science, 62, 394–407, 10.1111/j.1365-2389.2011.01364.x, 2011.

Greiner, L., Keller, A., Grêt-Regamey, A., and Papritz, A.: Soil function assessment methods for quantifying the contributions of soils to ecosystems services, Land Use Policy, 68, dx.doi.org/10.1016/j.landusepol.2017.06.025, 2017.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning; Data Mining, Inference and Prediction, Springer, New York, 2 edn., 2009.

Hobley, E., Wilson, B., Wilkie, A., Gray, J., and Koen, T.: Drivers of soil organic carbon storage and vertical distribution in Eastern Australia, Plant and Soil, 390, 111–127, 10.1007/s11104-015-2380-1, https://doi.org/10.1007/s11104-015-2380-1, 2015.

Jäggli, F., Peyer, K., Pazeller, A., and Schwab, P.: Grundlagenbericht zur Bodenkartierung des Kantons Zürich, Tech. rep., Volkswirtschaftsdirektion des Kantons Zürich und Eidg. Forschungsanstalt für Agrarökologie und Landbau Zürich Reckenholz FAL, 1998.

Webster, R. and Lark, R.: Field Sampling for Environmental Science and Management, Environmental science/statistics, Routledge, https://books.google.ch/books?id=7Xz5u05QC4AC, 2013.

Webster, R. and Oliver, M. A.: Geostatistics for Environmental Scientists, John Wiley & Sons, New York, 2 edn., 2007.

Please also note the supplement to this comment:
https://www.soil-discuss.net/soil-2017-14/soil-2017-14-AC2-supplement.pdf

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2017-14, 2017.