SOIL Discuss.,
https://doi.org/10.5194/soil-2017-13-RC1, 2017

SOIL Discussions Open Access EGU

Interactive comment on "Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models" *by* Madlene Nussbaum et al.

Anonymous Referee #1

Received and published: 3 July 2017

I. General comments This study makes an important methodological contribution to the Digital Soil Mapping research community by evaluating in depth a novel modelling technique (i.e. boosted geoadditive model), which clearly aims to find a good balance between the model predictive performance and its interpretability (as driven by the level of complexity). This is in particular useful, when a (very) large set of co-variates are / can be considered, and given the recent remarkable increase in data-resources availability, due to for example improved remote sensing techniques, the present model, will most probably have a large potential to be used in future research in soil science. Hence, I'm in favour of accepting this paper for publication in 'SOIL' journal (after minor revisions). Nevertheless, I think that some further clarifications are needed in terms of

the methodological approach. More precisely, it would be good to explain why both a K-fold cross validation as well as a validation based on an independent set of data were needed. Furthermore, I believe that the results needs to be discussed much more in detail, by comparing this study's output with other models' performances when predicting & mapping the considered variables in the literature. Please, consider as well my specific comments below (which will be helpful to tackle these general comments).

II. Specific comments P.2 L 20-30: You refer quite a lot to McBratney et al. 2003's overview paper on DSM. Although, I'm convinced that this is a very important paper and you certainly need to mention it, I believe that it would be much better to integrate as well much more specific examples when refereeing to specific modelling techniques as well as more recent publications. P. 2 L 45-46: Add in specific references? P. 2 L 50-52: You say "Lately" but subsequently refers to McBratney et al. 2003, which is a fairly old reference by now, so please, add more recent and specific references. P.4 L 80 – 90: It's unclear to me when K-fold cross validation will be used and when an independent set of data will be considered for validation. Moreover, why a "10-fold" has been considered (and not something different than 10?) P. 6 L 12-13 You state that the accuracy of the coordinates in about 25m. Is this something you interpret that way (because records have been made in the field on topographical maps) or has this info been documented somewhere? P. 6 L 35-39: Why did you consider this additional 5% of data for which you needed a PTF to estimate ECEC? Furthermore, did you test the effect of having included this data on the overall outcome / model performance / uncertainty ect. . .? P. 6 L 40 So that's 21.7% of the data used for Validation. Why 21.7%? P. 6 L 67 So that's 20.6% of the data used for Validation. Why 20.6%? P. 7-8 Section 3.3.3 It's fairly hard for me to understand when a certain statistical measure (to test the model predictive performance ect. . .) have been used on (i) the calibration data set or (ii) in the context of the K-fold cross-validation or (iii) considering the independent set of validation data-points. (See related general comment). Anyway, I guess it would be good to clarify this further in the MS and probably improve as well the structure of this section of the text in the light of this comment. P. 9 L 31-37: It

would be good to compare these results with results from other models as presented in the literature (See related general comment). I'm aware this need to be done in "Section 5.3" (see below / ultimate comment) P. 9 L 50-55: By saying that "the model explained about 40% of the variance of the log-transformed and 37% of the variance of the original data", I'm wondering if you did make this statement by comparing your RMSE-value with the STDEV value in the validation dataset? If yes, was it done in a (K-fold) cross validation context and/or based on the independent validation dataset? Can you please clarify what you did to come up with this conclusion? P. 10 Table 2: you refer to "SD slope" and "SD elevation" ect... as being the standard deviation in local neighbourhood. Can you please clarify this a little bit more? Does I got it right that you did calculate the standard deviation based on an X.X raster window in order to obtain alternative measures for terrain/topographical complexity?? P. 10 Table 2: Can you specify the 'r' (raster-resolution, i.e. 2m or 25 m) value of the maps of the considered topographical variables? P.13 L 10 – 25 & L 48 – 63. In the interpretation of the results as regards the influence of topography on the modelled soil characteristics it will be crucial to mention the raster-resolution, i.e. 2m or 25 m, because different levels of topographical detail will represents different process, i.e. small scales irregularity within a field (e.g. reflecting local surface irregularities related to agricultural practices) visible at 2 m resolution versus larger scale topographical general slope signature (e.g. reflecting variability induced by soil erosion processes) visible at 25m resolution. p. 15 Section 5.3. I believe that the discussion of the predictive performance of the fitted models can be worked out much more in detail by comparing this study's output with other models' performances when predicting & mapping the considered variables in the literature. (See related general comment)

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2017-13, 2017.