SOIL Discuss., referee comment RC1
https://doi.org/10.5194/soil-2021-107-RC1, 2021

Comment on soil-2021-107

Anonymous Referee #1

Referee comment on "Spatial prediction of organic carbon in German agricultural topsoil using machine learning algorithms" by Ali Sakhaee et al., SOIL Discuss., https://doi.org/10.5194/soil-2021-107-RC1, 2021

Overall this is a very interesting paper in which a fair comparison as regards different DSM approaches has been made across Germany, including the effect of a data-size extension (after combining 2 databases) and whether mineral and organic soils should be treated separately (by creating two different models).

In addition, the paper is well structured and writing, though some minor spellings and grammar improvements are possible (please note that I only focused on the language in the first couple of pages, but I'm convinced that the entire paper could benefit from some slight language polishing)

Nevertheless, I believe that this paper may require some major revisions based on following comments:

I. Main Suggestions/Remarks

I.1 This research considers agricultural soils, including both grassland and cropland, and as such I have some serious concerns as regards the presented (0-20 to 0-30) depth interpolation approach in order to match both databases (P. 3), which seems to be based on a (first order?) linear function depending on the soil class. However, in my opinion this analysis should be carried out per land use – soil type combination, because the depth distribution in cropland topsoil is remarkably different to that in grassland topsoil (i.e. more or less a Cte value versus exponential decline, respectively). Hence, I would like to

ask the authors to carry-out this analysis again per land use soil – type. Moreover, only a (general / average?) slope parameter value has been given (in L115 – 116), and as such I would like to ask the authors to provide the readers with a much more detailed picture on the different slope parameter values obtained depending on the land uses (and soil types) setting. This can be done in a tabular format (in annex) by presenting the slope parameter (+/- the associated SE) for each land use and soil type combination - or - in a graphical format showing the distribution of slope parameter values per land use type.

I.2. From section 2.2 I can see that a wide range of covariates has been considered. However, I was wondering whether the authors did carry-out any multi-collinearity analysis in order to identify those who may be too strongly correlated (e.g. $r > 0.9$). Subsequently, I was wondering what they have done to solve this potential issue?

I.3 The model performance evaluation indicators (section 2.6) are all quite similar and have a particular focus on "random error". Hence, I would like to suggest to include some others that could provide the readers with some information as regards the (%)bias. In addition, within 'the spirit' of SVR I think that including also a model performance evaluation indicator that also takes into account the concept of 'model complexity' could be an interesting add-on here. (I know that in the context of this kind of model this can be interpreted quite widely and may include a penalization term that depends on the number of parameters (like AIC and BIC) or the complexity of the trees / nodes, ect….)

I.4 I think that the "Results and Discussion" section requires some clarifications as regards the structure. In essence, I would like to suggest to add a short intro-paragraph explaining briefly the logic behind the structure (and clarifying as such the meaning of AP1, AP1L, AP2 and AP2L). Moreover, in the (bold) headings of the separate sub-section you could add the corresponding abbreviation in brackets to it at the end as well as give a short statement at the start of the section which case you're going to consider (actually, similarly to what have been done in L318). In addition, I believe that in some cases a bigger effort could be made to discuss the regional differences (in results obtained by applying the different approaches). In that respect I would like to suggest the authors to provide the readers with a relative residual map with annotation of + or – in order to be able to interpret the under / over predictions patterns in a spatial explicit way (I think this will have more value that the maps in fig. 3 and 6).

I.5 I believe that 'the main message' should be highlighted more, i.e. the fact that creating 2 separate models (one for mineral soils and one for organic soils) is much more important than the choice of the type of model (at least those considered here) and/or the suggested data-size extension. Please make sure that this is highlighted in the discussion and the conclusions sections. In that respect I think that some small additional analysis could be useful, for example a table / figure showing the potential model improvement (e.g. average RMSE decrease – or any other model performance indicator – see comment I.3.) due to this 3 factors (i.e. model separations (org vs. mineral), type of model, data extension). I think that this can be calculated rather easily from the information given in figure 2.

I.6 As I understood (from reading section 2.2) that all the covariates are represented by a spatially continuous map, I was wondering whether it could also be an option to provide the readers with one spatially continuous predicted SOC map, for example created by applying 'the best model' (AP2L?) on the various covariate maps. I think this could be useful in order to obtain a more detailed interpretation of the results taking into account regional differences depending on various environmental settings.

II. Specific Suggestion/Remarks:

L 9-10: "to influence climate change and mitigation" is a somewhat strange formulation. (I guess this should have been "to influence and mitigate climate change"?) Please rephrase.

L 15: define topsoil, (e.g. add '(0-30cm)')

L 32 – 37: you make several references to Meersmans et al 2012 but in your reference list there is 2012a and 2012b, so please specify "a" of "b" here.

L 46: "at a different scale" is a somewhat strange formulation. (I guess this should have been "at different scales' or 'across different scales") Please rephrase or delete.

L 54: I suggest replacing "SOC inventory" by "SOC monitoring" because you make reference to the periodic character of it.

L57: "with a sampling depth down to 100 cm" is a somewhat strange formulation. A more common way to say this could be "considering a sampling depth of 1m" or "considering a reference depth of 1m".

L61: What do you mean with "complete"? Is this a good spatial distribution? Please clarify.

L73: add a space between "(2017)" and "concluded".

L 81: What do you mean with disparity? (Do you mean "sample design"? Or "spatial

distribution"?) Please clarify.

L 120-126: Why didn't you just use just the best quality product? Are all covariates resampled to a resolution of 100m? And if yes, why not use the any higher level of detail / precision if you have been provided with it anyway? Was this done in order to deal with some computation intensity issues?

L 137: What is the (initial) resolution of this DEM? Was this layers also resampled to 100m (see previous comment)? And if so, was this done before or after deriving the related co-variates (such as slope, curvature ect…). Please be more clear / specific about the exact methodological approach followed here.

L145-147: please make a reference to the source of this map.

L 164: What kind erosion map has been considered? Is it a map highlighting water erosion and/or tillage erosion? Hence, please specify what kind of model has been considered to generate this map (e.g. Is this map based on RUSLE or WatemSedem)? I'm also wondering whether it was really required to add this map, because you have already a lot of topographical related input variables which may provide you with similar info. (In that respect I like to reiterate my main comment I.2 – see above)

L 175: What kind of interaction depth did you consider?

L 188: Are you sure this needs to be "maximum error"? To me it sounds more logic to go for a model with "minimum error" but still with a limited model complexity.

L237: Can you clarify what you exactly mean with "shuffled 10 times". I guess this is a kind of random perturbation? (following a normal distribution?) Is it similar to what one will do in Monte Carlo?

L 265 / Figure 2: Please add subplot labels to fig2 (a1, a2, a3, ….. b1, b2,…) and make always reference to the specific subplots in the text so the reader know immediately which subplots needs to be considered / compared (and which one he / she can ignore).

Figure 2: Besides adding subplot labels (see comment just above this one), I think there is an error in the x-ax labeling, because in all cased it is either "AP2" or "AP2L", so there is no "AP1" or "AP1L" present, whereas I think that all the plots on the left-hand side of the figure (which are making reference to "one model approach") should have the labels "AP1"

or “AP1L” (and not “AP2” or “AP2L” is currently the case). Right?

Figure 5: Please add a regression line though these clouds of dots so one can evaluate a potential bias and /or over- /underprediction. (please note that this suggestion is related to my main comment I.3)