Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-437-RC2, 2022

Comment on essd-2021-437

Anonymous Referee #2

Referee comment on "Colombian soil texture: building a spatial ensemble model" by Viviana Marcela Varón-Ramírez et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-437-RC2, 2022

This is a very interesting study and I enjoyed reading it. The authors have applied some very novel ensemble machine learning algorithms to predict soil particle size fractions across the entire country of Colombia. The idea of also incorporating global predictions like SoilGrids with national model predictions is another novel application that can help to improve national digital soil maps in areas where training data is limited. The authors also treat the soil particle size fraction data as compositional data, using ALR transformed variables in their ensemble models. I think that this study has a lot of potential but there is one major issue that needs to be addressed.

Based on my understanding of the spatial ensemble approach presented here, this approach has some fundamental issues that need to be corrected. As correctly stated by the authors, soil particle size fraction data is compositional data. The authors appropriately used the ALR transformation for predicting the different PSFs. However, once the data is back transformed to sand, silt, and clay fractions, the authors then treat each fraction independently, thus ignoring their compositional nature. In their spatial ensemble, at a given pixel they may take the predicted clay from SoilGrids, sand from MACHISPLIN, and silt from Landmap, thus ignoring the interdependence of the predictions within a given model. From visually evaluating figures 4-8, it appears that this can result in pixels where modeled sand, silt, and clay percentages far exceed 100 percent. This observation was later confirmed in the manuscript on lns 385-387 and lns 410-411. Spatial ensembling of compositional data requires the preservation of the compositional structure of the data and the authors' current approach violates this. I see this as a fundamental flaw in the current approach that can be corrected.

I can see two possible solutions to this problem:

- Each individual pixel would only be populated with a single model for all PSFs, e.g., all SoilGrids or all Landmap. Model selection would be based on the lowest model error

averaged across the three PSFs.
- Perform the spatial ensemble on the ALR transformed values (Tans_1 and Trans_2) for MACHISPLIN, Landmap, and SoilGrids. I'm not familiar with all of the processing steps in the MACHISPLIN algorithm but applying a similar approach for interpolating error and selecting the best model.

Another issue with the current spatial ensemble approach is that it produces lots of spatial artifacts (e.g., circle and blob patterns) which is likely due to the patchwork of models used to populate each pixel. Applying some type of model weighting, like the MACHISPLIN algorithm does, might improve these results. Another potential reason for the spatial artifacts seen in Figs 4-8 (e.g., vertical striping running north-south) is likely due to the ML algorithms. Documentation for the MACHISPLIN algorithm states that Boosted Regressive Trees and Random Forests models can produce blocky outputs, thus MACHISPLIN provides the option to exclude those two models from the ensemble using the model parameter 'smooth.outputs.only'.

There are many more 'minor' issues that the authors should address which I list below:

Specific comments:

Introduction: The introduction is long and overly general. The authors should focus on topics directly relevant to this study, including ensemble modeling, spatial cross validation, and modeling compositional data. The section starting of line 59 extending to the sentence ending on line 96 does not add to the manuscript and can be replace with more relevant background text. For example this section provides a very general discussion of geostatistics which isn't direclty relevant to this study and discusses current research questions in DSM that are not addressed in this study.

Lns 60-61. Are there examples of unsupervised statistical learning for soil PSFs or texture classes? If not remove this statement.

Lns 100-101. Which accuracy indicators? The ones previously stated? If so, then say it rather than generally listing indicators and then not stating which ones you used.

Ln 123. From Fig. 1 it appears that not all sample locations were sampled at all depths. I assume this was due to the presence of shallow soils? If so please state this. Also, please provide a breakdown of the samples (training and validation) represented for each modeled depth. This could be easily included in Fig. 2.

Ln 124. 'Soil particle-size fractions (PFS)' -- acronym not consistent with abstract, i.e.,

SPF, soil particle fraction

Ln 145. Equations 1 and 2 need to be better defined, i.e., all equation parameters need to be explicitly defined. It is not clear in this particular application how Trans_1 and Trans_2 were calculated. You state that clay was used as the denominator variable, so does that make the denominator in equation 1 (zeta-D) equal to the clay fraction. Yet on line 141 you state that D =3, and i=3 (D) represents the silt fraction. Please explicitly define how Trans_1 and Trans_2 are calculated.

Table 1. Information on the spatial resolution of covariate data is missing for several sources, e.g., soil index, sand and clay minerology, landsat. It would also be helpful to provide an approximate grid cell resolution equivalent to the 1: 100,000 map scale. Also, many references in table are missing.

Ln 154 and throughout manuscript. Table citations are missing table numbers.

Ln 154. ‘adjusted to 1 square kilometer’ Which upscaling method was used, e.g., nearest neighbor? bilinear?

Lns 160-162. Please provide additional detials about this bootstrapping technique. Does it account for ranges in covariate space when splitting the samples? Was spatial autocorrelation accounted for when creating the training/testing split? Based on your statement on lns 189-190, without accounting for spatial autocorrelation, your training and testing datasets are not independent. on the other hand, there are arguments against the use of spatial cross validation, see https://doi.org/10.1016/j.ecolmodel.2021.109692

Also, for reference, please provide the number of samples in training and validation sets, i.e., (75%, n=???)

Lns 169-171. It would be good to provide additional details comparing the two ensemble modeling techniques. For example, the Landmap algorithm applies a stacking ensemble approach using 5 base learners and a 'meta model' or super learner to produce an ensembled prediction. What type of super learner was used? Also, how does the stacking ensemble compare to the weighting approach applied in the MACHISPLIN algorithm. How were the model weights calculated? These types of details are more relevent to this study than the very general discussion of machine learning vs geostatics presented in the introduction.

Lns 171-174. This statement is not clear. Is the residual error interpolated for each model? Are these error surfaces used to determine the model weighting in the final

ensemble? The details of how this is done needs to be made more clear.

Ln 187. How is the cross validation used to determine the meta-learner?

Lns 189-191. The authors have done a nice job of citing recent work relevant to this study. In regards to the use of spatial cross validation, there has been recent debate as to its appropriateness for map validation. It might be good to reference this here.

https://doi.org/10.1038/s41467-020-18321-y

https://doi.org/10.1016/j.ecolmodel.2021.109692

Lns 194-195. So resampled from 250m to 1km? It is helpful to state this, as well as the resampling method.

Lns 206-207. Kriging assumes some spatial autocorreation among the errors. Was this the case? Might be helpful to provide the semivariograms. Did you consider using a thin-plate spline approach similar to the MACHISPLIN method?

Lns 212-214. The accuracy of this approach depends on the accuracy of your kriged error maps. It seems like applying a model weighting approach similar to MACHISPLIN might provide a better result rather than select the model with the lowest error at each pixel.

Ln 242. 'Boundary adjustment parameters'? I'm not sure what you mean by this. It is not referenced anywhere else. Do you mean Accuracy metrics or indices?

Tables 4 and 5. 'Adjustment parameters'? Why are these model accuracy metrics referenced as adjustment parameters? Also, it should be stated here that these accuracy statistics are based on the validation dataset.

Table 5. Why was CC for clay at 5cm lower than either MACHISPLIN or Landmap for that depth and fraction? I would have thought the spatial ensemble would select the most accurate model for each site and therefor produce more accurate results relative to the other models. There are other instances of this among the depths and fractions. Could this be a result of combining PSFs from different models?

Ln 292. Is this a reference to the predicted map uncertainty for SoilGrids? Was this uncertainty evaluated? It would be interesting to see uncertainty maps of SoilGrids PSFs for this area. This is also an important aspect of digital soil mapping not addressed in this paper. Since SoilGrids quantifies model uncertainty this would be an interesting point of comparison to national model results.

Lns 299-301. Fig 2. is presented in black and white and as a low resolution figure. making it difficult to interpret.

Lns 385-387. Was the data normalized to 100 after the spatial ensemble? This might explain why some of the PSF at certain depths had lower performance relative to the EML models. I see this as a major problem with your spatial ensemble approach.

Lns 402-403. This was seen in Table 4 with the validation statistics. What is missing is a visual comparison of the two EML algorithms. In Fig. 3 you show either Landmap or MACHISPLIN but not both. I would like to see a figure similar to Fig. 3 but showing Landmap, MACHISPLIN, SoilGrids, and the spatial ensemble at one or two depths.

Lns 410-411. Evaluating model accuracy in these areas is tricky because there is limited data to accurately model the spatial distribution of model error. Using ordinary kriging won't do a great job in these data sparse regions.

Lns 414-416. It is good to see that the authors recognize this issue with the spatial ensemble approach. However, I see this as a major flaw that diminishes or even removes the prior efforts to account for the compositional nature of the data. This could have been avoided using one of the alternative approach I outlined above.