Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-437-RC1, 2022

Comment on essd-2021-437

Anonymous Referee #1

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-RC1, 2022

General comments:

The manuscript "Colombian soil texture: Building a spatial ensemble model" by Varón-Ramírez et al. presents soil texture maps (clay, sand, and silt) for Colombia for different depth intervals by using and comparing different machine learning techniques. The authors compare their predicted maps with the global SoilGrid product for Colombia and provide maps that are based on the best model for each pixel. The soil data used to derive the maps and the final maps are provided as independent datasets/raster files and are easily accessible and usable. Unfortunately, the authors do not provide the code to reproduce their maps.

The manuscript and the related maps are unique and useful for the scientific community since, as stated by the authors, they provide the first Colombian soil texture maps obtained by spatial ensemble of national and global soil products. The methods are explained in detail (sometimes maybe too detailed for a general reader). However, I have quite a few comments regarding the manuscript and data presentation which I will outline below. Additionally, I think the language needs to be improved – sometimes the grammar and structure of sentences is not correct which makes it a little bit difficult to always know what the authors are trying to say. The usage of many abbreviations (which are mostly explained at the beginning of the manuscript) makes it also sometimes hard do follow the argumentation. Personally, I think it would help if you could just write out the words or use more intuitive acronyms.

Specific comments:

Abstract: Overall, I think the abstract is a little bit too long and detailed and should focus more on the novelty of the data products provided by this work and how they can be used. It is good that you describe what you did, however, it is probably not necessary to

provide all the details about the model performance and comparison in the abstract.

Line 5: How are the depth intervals exactly defined? I assume 0–5, 5–15, 15–30, 30–60, and 60–100 cm.

Line 6: What do you mean by ‘stack’ in this context. Try to avoid to use too technical language in the abstract.

Line 6: Maybe better: “the most important” instead of “top”

Line 10: Maybe better: “smallest” instead of “fewest”

Line 15: What do you mean by “compared to other algorithms”? Aren’t all the methods you used spatial ensemble? Your abstract should really be understandable to a reader that is not that familiar with all the methods you used.

Line 18: Should be “SPF” instead of “PSF”.

Line 19: Without geographic context this information is difficult to follow. I think you can be less specific in the abstract and just say that the ensemble machine learning algorithms usually performed better, but in some regions the SoilGrid product also resulted in reliable predictions.

Introduction: The introduction gives a nice and comprehensive overview about the idea and state of the art of digital soil mapping in general and provides details about the methods used in this work. However, it can probably also be shortening a little bit. For example, the first two sentences (line: 29-31) are probably not needed.

Line 57: Maybe better: “Digital soil maps are derived from soil datasets that represent the continuous nature of soil variability.”

Line 88: Missing brackets around references.

Line 90: Not sure what you mean with “what are the best big-data management strategies

for generating high-spatial resolution maps across large areas?". Your manuscript is not really about data management, but predicting soil maps.

Line 101: Is your objective really to develop a digital soil texture dataset? As far as I understand, the soil data already exists and you are applying machine learning techniques to create digital soil maps. So, your objective should be more about the maps than the soil data. Maybe I am misunderstanding something here and you just have to be more clear what you did for this work and what is based on previous work.

Line 102: Are these soil data already part of any international soil databases (e.g. ISRIC, ISCN) so that they can also be used easy by other researches? I really encourage you to put your dataset in one of these international soil databases if you haven't done it yet.

Line 117: Awkward phrasing: What do you mean with "positive implications"

Methodology: This section provides a good overview about the applied methods. In general, I think that this section can be improved by focusing more on the actual methods rather than the R packages and functions that were used. If you provide the code to reproduce your maps this information can all be presented in the R script. Please also make sure that you always correctly cite the R packages that you are using – the reference is quite often missing.

Line 120: Awkward phrasing: "A total of five major steps". Maybe better: "Our work flow contains five major steps, including … which will be discussed in detail below".

Line 124: Not needed: "Soil particle-size fractions (PSF) such as clay, sand, and silt were collected, including geographical coordinates (EPGS: 4326)." You can just write in the sentence before: "A total of 4,203 georeferenced (EPGS: 4326) soil profiles were collected from … that all contained information about particle-size fractions (clay, sand and silt)."

Line 125: Did you create these geographic regions or are there somewhere defined. If so, please provide the reference.

Line 130: Awkward phrasing of first sentence. Maybe better: "Dataset quality was ensured by i) sum of particle-size data equals 100 % and ii) no overlapping sampling depth". By definition, two soil horizons cannot be overlapping. Also, what about number of samples for each soil profile? Your Figure 1 shows that some profiles only have data down to 5 cm. Did you exclude any profiles that contained, e.g. less than 3 measurements or that did not reach a certain depth? If not, how did you treated these samples.

Line 131: “to” instead of “at”

Line 133: Citation for R package aqp is missing

Line 136: Great that you are transforming the compositional data. However, I think you need to provide some more background information why this is necessary and why you decided to use the additive log-ratio transformation (especially for non-experts in this field). It would also be helpful if you could provide some details on how to interpret the transformed values.

Line 142: Earlier you wrote, you applied the additive log-ratio transformation, now you are mentioning inverse additive log-ratio transformation. Please provide some more information what this is and why you are doing this. Also, in the published dataset it is not clear to me which of these transformations refers to transformation_1 and transformation_2. You need to provide this information in the metadata file and in the manuscript and maybe also think about a more self-explanatory naming of these two variables.

Line 146: Could you elaborate a little bit on why clay is used in the denominator? Just because it is used by other studies is not necessarily sufficient as an explanation.

Line 147: Missing citation for the R package Compositional.

Line 153: Table 1: Number of covariates does not match with the description of the covariates, e.g. soil has 28 covariates but only 5 are mentioned. A detailed list with precise data sources should be provided in the supplement. Also, the acronym GSI is not explained. What are the land categories you used? Are the extracted years matching with the year of sampling?

Line 154: Table ?? – missing cross-reference

Line 154: How did you do the adjustment to 1 square kilometer? Could you provide some more details here, including uncertainties?

Line 155: Again, explain what you mean by “stack”

Line 156: Missing citation for the R package carat

Line 156 ff: The idea/method behind the recursive feature elimination is not clear to me. If I understand it correctly, you first built a model with all covariates and then selected the most important predictors (based on what?). How is it possible that you only then extract the values for the covariates at the profile level? Maybe, I am missing something here. So, if this a common approach it is probably fine, but I cannot really evaluate this.

Line 166: Did you also consider different training and test datasets? In your introduction you mention spatial cross-validation which is probably a good approach for your data given its clustered nature. Could you please provide some more details about the bootstrapping technique since the splitting of the data is crucial for the validation of your predictions later.

Line 165: Space missing

Line 167: Space missing

Line 168–191: After reading this section I am not quite sure if I fully understand your methods. The description of the two R packages is quite technical and detailed, however, I am missing a little bit a more general description of the methods that are not necessarily restricted to the two R packages. Also, at the end of the section you mention spatial cross-validation, but earlier, you talked about bootstrapping the data. I think I am having a little bit difficulty to follow what you did at each step and why. Maybe you can emphasize this a little bit better. A flow chart might also help to guide the reader. Yet, I also have to admit that I am not really an expert in this field of digital soil mapping and other reviewers may be able to evaluate it better.

Line 194: As mentioned by Fuat Kaya, why did you resample the SoilGrid data and not just extracted if for the sampling locations of your soil profiles?

Line 195: Maybe better: “Next” instead of “After”

Line 195: Did you do the validation for the SoilGrid product only or, as I assume, also for your own predictions? This is not really clear based on the sentence.

Line 196–203: You do not provide an explanation/definition for the concordance correlation coefficient, but for all other measures you mentioned here. Additionally, the

whole section (line: 194–203) is not that clear written: first you talk about comparing your results to the SoilGrid products and then you talk about the validation measures you used for your own predictions and the SoilGrid product. Again, try to streamline the description of your different steps so that it becomes more clear what you did when and why. I really think a flow chart-type figure would help here.

Line 206: You did not introduce the term "independent residual" yet. Why did you not just used the prediction error terms? The interpolation was probably done for the final map – if so, say so here. Can you provide some more details about the kriging? There are a lot of different ways doing it.

Line 210–115: Could you please provide some references for this section. Maybe you can also mention more clear that the spatial ensemble was based on your two models and the SoilGrid product.

Results: Overall this section is quite descriptive (which is ok for a result section), however, I think the section could be improved by providing some more context and by focusing on the main results.

Line 225: Table ?? – missing cross-reference; not clear which transformation refers to which

Line 225: Awkward phrasing: "the minimum contents were 1% or less for 5 standard depths" Maybe better: "For all depths, the particle-size fractions ranged from circa 0 to more than 90 %, except for silt, which only ranged from circa 0 to 80%". However, I think you can also just say that the particle-size fractions are covering more or less the entire range, which is to be expected for continental-scale analysis.

Line 234–236: This is repetition from the method section and probably not needed in the result section; Table ?? – cross-reference missing

Line 238: You have not defined the acronyms TEM, RH and PPT also units are not provided for the covariates. Did you scaled the covariates before using them? As I wrote earlier, you probably need to provide a table with all the covariates and description of them, including units and sources.

Line 242: Table ??: missing cross-reference

Line 254: Table ??: missing cross-reference

Line 262: Fig. 4–8 instead of listing all figures

Line 265–274: It is not always easy to follow which model was best where, which is partly also due to wrong grammar. I encourage you to have someone checking the language and grammar throughout your manuscript. For example, what do you mean with "MACHISPLIN had representation in all natural regions, and in the deepest layers"?

Line 268: It should read MACHISPLIN instead of MACHISPLIS

Line 275: Maybe better: "In terms of SG" instead of "Concerning SG"

Line 278: Table ??: missing cross-reference; missing space

Line 279: It should read "On the other hand"

Discussion: This section provides some context for the results and also discusses limitations of the data and methods. However, it sometimes also repeats things from the introduction and result sections, which is probably unnecessary.

Line 286: "Soil texture is a key property required for many applications in environmental sciences" – This sentence is not adding anything new and the statement was already made in the introduction.

Line 288: Delete the second "previous"; what do you mean with the word "detail" in this context?

Table 4: It should read "Root mean square error"

Table 5: Not clear what is meant by "adjusted parameters"; the table is showing the validation terms for clay, sand and silt for the five depth intervals. The acronyms are also not defined.

Line 290: Missing reference. Without any references it is difficult to follow your argumentation her. Also, this is probably material for the introduction and not really for the discussion.

Line 290–296: This is a description of your methods and could probably be moved to the beginning of your result section to summarize what you did before presenting the results.

Line 297: Does the soil diversity really changes with depth?

Line 297–312: This seems to be more part of the result section than the discussion section.

Line 327: "with" instead of "whit"

Line 330: You are not providing any details why Araujo-Carrillo et al. 2021 provides the best example. I think in this section (line 323–343) you don't need to describe the methods of the other studies rather focus on comparing your results with their results and discussing which improvements you achieved and why a direct comparison is not that easy. You can then just briefly talk about differences and similarities at broad spatial scale.

Line 343–355: I think this section could also be part of the introduction (and is partly already in it) as a justification for the methods you used in this work.

Line 350: "than" instead of "that"

Line 356–369: Again, you provide a lot of details about the methods from other studies and yet, I don't think that this is necessary. In general, I like the idea of comparing error terms between the different studies, but I am not convinced that it needs to be done in such detail. Also, you have not mentioned this method (the error term comparison) anywhere else in the manuscript and it comes a little bit unexpected. So, I think you either need to set it up better earlier in the manuscript or just provide a general discussion which study/method had better/worse error terms and what the reason for this is.

Line 372: I think a word like "only" is missing between "but" and "in"

Line 374–376: Could you be more precise here? What do you mean with “in general terms” and “with good quantitative statistics”

Line 383: Maybe better: “for the entire country of Colombia”

Line 384: Sentence structure: “However, the differential factor included maps that represent the best model (EML or SG) in each area of the country at different depths, called in this work spatial ensembled.”

Line 386: Unit % is missing; could you maybe discuss here what approach could overcome some of these limitations?

Line 389: What do you mean by “new and great challenges”

Line 391: Do you really think adding covariates will improve the predictions? You already tested 83 covariates. If you think you are missing important covariates, you should elaborate on why they are important and why they are missing.

Line 392: “with” instead of “whit”

Line 396: What do you mean by “homosoil”?

Conclusions: The conclusion section many repeats statements from the result and discussion sections. Maybe think of some new aspects that could conclude the manuscript and only give a short summary of the main results and how they related to previous work.

Line 403: Is your map "just" better or does it reveal new patterns of soil texture in Columbia? Maybe you can add 1-2 sentence here (or earlier in the result/discussion section).

Data availability: The provided links to the dataset and maps are working and everything can be downloaded easily. Please consider also to publish your code in order to fulfill the FAIR principles.

Line 421: You have not clarified what trans_1 and trans_2 are and this information is also missing in the metadata of your published data.

References: For some of the references the doi link is not correct, e.g. line 442

Figures:

Figure 1: Maybe think of using a different color scheme since red and green are not distinguishable for many people. www.colorbrewer2.org provides a nice tool for picking colorblind-friendly color schemes for maps.

Figure 2: Could you provide this figure in a better resolution?

Figure 3: Again, think of using a colorblind-friendly color scheme. Why are you not providing the maps for each model and each depth? It is not clear from the figure legend (which should stand for its own) why landmap is used for 5 and 30 cm and MACHISPLIN for the other depths in the same figure.

Figure 4–8: Colorblind-friendly scheme; The scale for clay, sand and silt should always have the same range, otherwise it is impossible to compare the maps which each other. Also, the contrast for the color range is not ideal, differences are difficult to see in the maps.