EGUsphere, author comment AC4
https://doi.org/10.5194/egusphere-2022-301-AC4, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Reply on CC4

Ashenafi Ali et al.

Author comment on "Reference soil groups map of Ethiopia based on legacy data and machine learning-technique: EthioSoilGrids 1.0" by Ashenafi Ali et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-301-AC4, 2022

We thank Fuat Kaya for having an interest in the work and voluntary community review. We respond to the key issues raised as indicated below:

Dear Associate Editor,

I have carefully read the study As the voluntary "commentor" of the article "Reference Soil Groups Map of Ethiopia Based on Legacy Data and Machine Learning Technique: EthioSoilGrids 1.0".

Since I am not an official referee, my comments are sincere.

The authors should be commended for their work in Ethiopia, feeling sincerely about the data sharing process.

Response 1: We are grateful for the positive comments

However, the authors have edited this article to produce only one output. I have concerns about research questions. There are many challenges to address in digital soil mapping. And these challenges are voiced by the DSM community. Here's an example: Ten challenges for the future of pedometrics (https://www.sciencedirect.com/science/article/pii/S0016706121002354).
Response 2: Thank you for bringing this to our attention, we are aware of the publication you indicated and found it helpful.

In this regard, I invite the author, who does the modeling in this valuable team, to model the events globally with two more accepted algorithms in SoilGrids 1.0 and SoilGrids 2.0.

https://soil.copernicus.org/articles/7/217/2021/--SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty------Used

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105992---SoilGrids1km — Global Soil Information Based on Automated Mapping

Response 3: This work considered the SoilGrids 250m (2017) as a base which succeeded the development of the SoilGrids 1km (https://www.isric.org/explore/soilgrids/faq-soilgrids-2017). As indicated in

the Soil Grids2.0 (https://soil.copernicus.org/articles/7/217/2021/), the numeric soil variables were only modelled and mapped (but not the soil reference groups/soil types). We understand that SoilGrids250m (2017) is the framework in which soil type/class modelling and mapping are done using Random Forest (RF), and as shown in lines 178 to 188 of this manuscript, RF was used for EthioGrid 1.0.

Specific comments:

Line 1:

As far as We know, This map not "conventional", well this map "digital" map.

I think "digital" must added to title.

Response 4: It is possible to qualify the map by adding "Digital" to the title. However, digital maps can be generated either based on a predictive/digital soil mapping framework or digitalised conventional maps. Therefore to avoid confusion, we prefer to qualify the map as it is generated based on the legacy soil data and machine learning techniques which explicitly indicate that the digital soil mapping approach was followed.

Line 35:

Really, honestly, "awesome" work for this team to collaboratively extract and collate the data.

But, We (DSM community and public) know, Soilgrids 1.0 and 2.0 versions have been released.

Publishing by running a single algorithm here is just to produce an output. There is a need for an approach to address current DSM issues. We know that there is something "Unknown" in Big data. And we will discover the unknown in Data with machine learning algorithms. So why one algorithm. Comparative results are necessary for this study to make accurate inferences for regional results.

multinomial logistic regression for Soilgrids 1.0 and quantile random forests for Soilgrids 2.0.

If reference soil groups are estimated in the field with these algorithms, their outputs will be appreciated by the DSM community at the international level.

Response 5: Yes, the data extraction and compilation process is something that we are proud of. Regarding the algorithm used as explained under response 3, the scope of the work is not to compare algorithms, but to develop SoilGrid1.0 using a selected algorithm.

Line 70:

the last part of the introduction, the authors define a brief research purpose/question. In the last paragraph of the Introduction chapter, the Authors wrote that ... objectives of this study. In this part of the article, I rather expected a clearly formulated research goal. I suggest that in the article it is precisely stated what the purpose of the research is, using the example statement: "The goal of the study / research was ...". When formulating the research goal (s), it would be worth writing what was the cognitive (scientific) goal and what was the utilitarian (useful) goal. Before stating the purpose of the study, it would be

worth formulating the research problem. The research problem may constitute a premise to indicate a gap in the current state of knowledge. It is worth writing what the current gaps in knowledge the Authors would like to fill in on the basis of planned and conducted research.

Response 6: Thank you for this specific comment, we will revisit and clear up confusing statements.

Line 178:

Is it just "model accuracy" ?

How do we evaluate uncertainty?

To evaluate classification-based algorithms that produce probabilistic predictions, D.G. I recommend Rossiter's valuable work.

https://www.sciencedirect.com/science/article/pii/S0016706116303901#bb0110

Please control "confusion index" released by Burroug et al. (1997 --https://www.sciencedirect.com/science/article/pii/S0016706197000189)

And the other 2 sources applied quantify in different regions, large and small areas.

https://www.sciencedirect.com/science/article/pii/S0016706116304864

https://www.tandfonline.com/doi/full/10.1080/02571862.2022.2059115

Response 7: The accuracy assessment (overall, user's and producer's accuracy) method and uncertainty are indicated in lines 361 to 365. Among the reviewed techniques, we have used the most commonly used cross-validation technique and accordingly the 95% confidence interval is indicated (lines 362 and 363). These are in line with the approach followed by global/regional soil grid development frameworks. However, as you indicated, there are various accuracy assessment techniques or issues that need to be considered in selecting an accuracy assessment of modelling soil classes e.g. accounting for taxonomy distance (which has also different sub-techniques), spatial cross-validation which is presumed to have limitations, dealing with clustered samples for assessing map accuracy by cross-validation, and dealing with imbalanced data in categorical mapping which might lead to issues on the accuracy of majority and minority classes. We recommend future studies to consider these issues in line 441 to 444.

Line 263:

What "reference" soil group did the models predict in areas with these classes? Is there a taxonomic relationship here? Please read this title paper: Accounting for taxonomic distance in accuracy assessment of soil class predictions

Response 8: Thank you for the recommendation. The reference soil groups indicated in line 263 were excluded from the modelling and hence comparison was not made. However, we now get insights to include some RSGs left unmapped and improve the accuracy of this beta version. As indicated in the confusion matrix even those soil groups modelled and mapped have depicted different accuracy values and we noticed that some reference soil groups are mapped at the expense of others which enables to interpret taxonomic

relationships.

Line 305:

Climate, Organism and topgrapy. If it is related to them, how would it be to compile it with a sentence?

Response 9: It indicates the relative importance of the predictor variables in determining the spatial distribution of reference soil groups across the landscapes of Ethiopia. It is an effort to go beyond prediction and incorporate model interpretations i.e. extract information on the relationships among variables found by the models. However, as is clearly indicated in various kinds of literature, model interpretations are not straightforward/simple in complex/ensemble models e.g. Wadoux et al. (2022): Beyond prediction: methods for interpreting complex models of soil variation, https://www.sciencedirect.com/science/article/abs/pii/S0016706122002609?via%3Dihub

Line 420, Fgure 7:

Very nice map. Most probable class maps, I think, for True phrase

Response 10: We are grateful for the appreciation.