Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2022-291-RC2, 2022

Comment on essd-2022-291

José Lucas Safanelli (Referee)

Referee comment on "Improving the Latin America and Caribbean Soil Information System (SISLAC) database enhances its usability and scalability" by Sergio Díaz-Guadarrama et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-291-RC2, 2022

General comments:

The paper “Improving Latin American Soil Information Database for Digital Soil Mapping enhances its usability and scalability” describes the effort of gathering and harmonizing Latin America soil data from historical surveys, which was promoted by FAO's South American Soil Partnership and involved several collaborators across from region. The authors presented a quality assessment analysis, described a new improved version of the dataset, and demonstrated the potential of SISLAC for generating new soil information through digital soil mapping. This type of work is important in order to document soil data integration efforts and document the best practices for harmonizing heterogeneous soil datasets. In addition, it makes clear that avoiding removing a lot of data that can be simply adjusted has an enormous impact on the final number of samples and potentially the spatial representation across a region. Overall, the authors did a great job in describing their quality analysis, but I was not convinced by the results from digital soil mapping. I think the authors could rather explore the dataset with a denser descriptive analysis, avoiding a predictive approach (which was very simple and suboptimal). Therefore, I don’t have any major objection to its publication. However, I think that a moderate revision of the second goal is required before reaching a final decision. Finally, I congratulate the authors for making available the improved SISLAC dataset on a public persistent repository (Zenodo) with an open-access license.

Specific comments:

Although the first introduction paragraphs describe what soil is and how they form, the current structure seems a bit overloaded to me. For example, the first three sentences have a lot of information that is hard to grasp at first moment. I would suggest starting

from line 72 and relocating those first sentences after explaining the soil importance, bringing the definitions after a gentler introduction.

The data are well described. I was able to access their online website (http://54.229.242.119/sislac/es) and check some soil profiles. However, I had some issues with signing up to the portal (could not confirm my email address to log in). The public access does not have any download button, but it seems the user can copy and paste single profile tabular data. They do not mention any application programming interface (API) in this data section, which is a characteristic of modern web 2.0 platforms (https://en.wikipedia.org/wiki/Web_2.0). I would suggest at least discussing data distribution through APIs and explaining in the manuscript if this feature is planned as a potential improvement of future SISLAC versions.

It is not clear in the manuscript if the SISLAC from their website is the older or the improved version.

When navigating their website, I found that many samples come from the WoSIS snapshot of 2016. There are other datasets, such as the SISINTA. I just wonder if the authors could provide an overview of the original sources (WoSIS, SISINTA, etc.) similarly to what they did with country numbers. This new table could be placed as supplementary material to help readers quickly evaluate the difference between SISLAC and other available public datasets, such as WoSIS.

How do the authors expect to update SISLAC when newer versions of the original sources are released? Have they automated the quality analysis keeping in mind new updates or has this current work involved a workforce for manual inspection?

Why the authors defined 150 cm as the bottom limit instead of 200 cm? 200 cm is an arbitrary convention from pedology but at least is the standard limit of GlobalSoilMap. A simple justification would be enough in my view, as reprocessing the data would be very expensive.

Both good-of-fitness equations have minor mistakes, although the result will not be impacted as the difference between observed and predicted are squared. However, the sum of squared residuals should be observed-predicted in both RMSE and R2 numerator.

The authors did a good job of describing and reporting their quality assessment analysis. I wonder if they used some published guidelines or proposed those based on the issues they faced in the project development. I think this data description paper and methods can help many other efforts for soil data integration and harmonization.

I only have serious concerns about the results from the data usability section. The authors provided reasonable summary statistics and visualizations. However, the cross-validation statistics are very intriguing, at least from the current scatterplot visualization. In my view, it is impossible to get moderate to good R2 from the scatter distribution they plotted, especially for the third panel where they reached an R2 of 0.83. All the fitted lines are almost flat, with a narrower predicted variance compared to the original values. In addition, when many data points are overlapped, it is common to present a scatterplot with point density, making possible the evaluation of the linear trend around the fitted line. The bias of these models is really high, so other performance metrics like Lin's correlation concordance coefficient (CCC) would indicate a potential unsatisfactory performance. Therefore, I'm not convinced with the results from this data usability section and even question the authors if they are willing to keep these results in their manuscript. Instead of presenting these questionable results from digital soil mapping or another predictive approach, I think the authors could rather crunch the dataset with a denser exploratory data analysis with summary statistics, multivariate data analysis using PCA in combination with grouping factors (coloring by color, biome, or any other physical information), some spatial statistics (like Moran's index, or even screening variograms for the whole region), etc. In my opinion, those results would be a greater fit for the manuscript type, which is a data description paper. If they follow this suggestion, I think they should adjust the paper title.

The discussion is well developed; however, I would only suggest adjusting it if the digital soil mapping results are revised.

Technical corrections:

Overall, the paper is clear and well-structured. I'm not an English native speaker, but I think the readers would benefit from a proofread version of the paper.

In line 214, I think the authors should define ordinary kriging as an interpolation method rather than a method to estimate SOC, e.g.: "On the other hand, ordinary kriging (OK) was used for horizontal variability assessment, a method frequently used to spatially predict SOC …"