SOIL Discuss., editor comment EC1
https://doi.org/10.5194/soil-2022-17-EC1, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on soil-2022-17

Nicolas P.A. Saby (Editor)

Editor comment on "Weathering intensities in tropical soils evaluated by machine learning, clusterization, and geophysical sensors" by Danilo César de Mello et al., SOIL Discuss., https://doi.org/10.5194/soil-2022-17-EC1, 2022

I made an in depth reading of the paper and I have several major concerns about the work and the paper.

The abstract is quite vague and not very informative. Please explain the overall goal, the assumption, the approach implemented and give precise results and findings.

The English does not seem to be ok.

The work is not enough clearly presented. A lot of questions arise when reading the actual version.

The overall goal of this study is not enough clearly presented in the introduction. First (i), modelling with ML of weathering index (WI) is quite vague. But for what ? Mapping, monitoring, statistical application? In (ii), it is said that importance of the covariates will be discussed but what is these covariates. This is not clear if these are the one for the digital soil mapping approach or the observed ones using sensor data. It is also a general comment about this paper where a confusion is made between these 2 types of information. The description of the proximal and remote sensing data should be better explained. The first ones are not spatially explicit.

The way geophysical data are used is rather difficult to understand and does not seem at all appropriate. It is quite difficult to understand why these data are mixed with the WI in the PCA step. This means that Geophysical data are not used as covariates but as a covariable. The PCA is not a method to evaluate (as it written in the abstract) but a multivariate analysis to explain the multivariate relations between variables. Moreover, the results of the PCA are not presented. It is not then possible to understand the signification of the principal components.

Why did you use a clustering approach ? We could expect a more **quantitative** digital soil mapping approach of the raw value of WI instead. Why are you creating clusters of WI and sensor data to map them after? Again, as the results of the PCA and the clustering are not presented, it is difficult to understand the signification of the cluster. It could happen that no correlation occurs between these data and the clusters represents only one of 2 variables. The results table 3 seem to validate this assumption as the discrimination between the different cluster are not very significant. However, an extensive discussion is provided about the signification of these clusters based on an

interpretation of the dsm maps and by expertise. However, all the maps are wrong. Maybe a point map of the clusters would be better to be interpreted first. Finally, it is difficult to understand the adding value of the geophysical data.

In the 2.1, two sets of data are presented (16 and 79 points) but only the 79 points are used for the study. Could you explain ? finally, it is topsoil map of WI ?

The equation 1 is not explained enough. What is SIO2 and TIO2 ?

In 2.7, two approaches are presented for the clustering, eg scott lethod and k-means. Which one did you use? Do you mean that a kmeans approach was implemented and the optimal number of clusters was selected using a scott approach ?

The 2.8.2 should be titled "validation of the map".

The explanation of the nested approach is quite difficult to follow. Do you mean that there is a tuning of the ML algorithm at each step of the LOOCV? This is not indicated in the fig 3. How this tuning is done ? Bye cross validation?

283 should be titled validation not training

2.8.3 Why are you explaining that you are using a LOOCV now? This is very confusing as it is not explained in the fig 3. All the indicators explained in this section are based the result of the confusion matrix where the 4 kinds of results are computed, eg FP TP, FN and FN.

It is not explained how the different maps produced during the LOOCV are combined at the end of the process. Do you compute the dominant value? Are using a probability approach?

The "pls" is usually a regression approach and is not adapted to a classification exercise. Could you detail which algorithm you used?

The size of the dataset is quite not adapted to the use of ML algorithm like random forest. You may consider more classical algorithm like the multinomial algorithm. As you use a LOOCV, we could not exclude an overfitting of the data. A k-fold cross validation would be more adapted. The hyper parameter of the different algorithms are not described so it is difficult to understand the quality of the models.

The sampling design is also not all adapted to a digital soil mapping approach. The locations were selected by expert judgement and are not very well spread over the area. There is extensive discussion about how to collect data in the recently published paper. This not really discussed in this paper. It is only acknowledged that the sample are not enough even but the spread of the data into the covariate space should be checked to discuss the validity of the map produced.

I think this article needs to be thoroughly revised before publication. Th use of the clustering step and the geophyical data should be better justified.