SOIL Discuss., community comment CC4
https://doi.org/10.5194/soil-2022-17-CC4, 2022

Comment on soil-2022-17

Rafael Siqueira

Community 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-CC4, 2022

I've read the pre-print version of this manuscript with great interest, due to the novelty associated to the application of modern technologies to investigate soil weathering and pedogenetic development, a very welcome initiative in Soil Science. The authors successfully combine two powerful approaches or tools constantly applied in the Pedometrics field: computational statistical techniques (machine learning and multivariate analysis) and geophysics proximal sensing data, aiming to understand in a quantitative manner how the soils develop in a tropical landscape. Nonetheless, I would like to write down some comments as reflections and suggestions:

I am not a native speaker, but a new look at the English writing for corrections would be important to improve the quality of the text. Furthermore, some parts of the text are a bit clumsy and a review would be welcome.

Abstract

Line 17 - I am not quite sure if you could call machine learning as a "geotechnology", for the reason it was not originally developed for geosciences.

Line 19 – Satellite imagering is not a proximal sensing technique, but a remote sensing technique.

Line 21 – You must explain in the Abstract why you used the PCA and clusters analysis. At the same time, I recommend to cite what information have you used to create the clusters, and moreover, what they really mean, in the Abstract.

Line 22 – Change "we determine and used the ideal number of clusters" for "we used the ideal number of clusters". You've already cited you determined the number of clusters before.

Line 27 – "The Nitisol over East diabase presented greater weathering intensity than".

The last sentence of the Abstract is clumsy.

Introduction

The Introduction is very well written, making a brief literature review about weathering and the geophysical sensors.

Line 94 – Would not be better if you used just "model weathering intensity using combined data from geophysical sensors" than "model weathering index using combined data from geophysical sensors"?

Line 97 – I suggest to remove the objective 4, since you did not explore it in your results.

Line 99 – I would remove the citation about satellite here.

Material and methods

Line 109 - I guess the name of the municipality cited in the text is with just one "f".

Line 135 - You could standardize the Embrapa citation. Sometimes you cite Embrapa, 2011 and sometimes Embrapa, 2017.

Line 135 – In the laboratory analysis topic, you make a complete description about physical and chemical analysis of soils according to Embrapa. However, you do not use any of them, unless the oxides contents. I suggest to remove all analyses which you did not apply in this paper.

Line 146 - I think it'd be productive you write one or two more lines explaining better how the oxides contents, the only analysis you really used, are obtained. Moreover, you cite Fe_2O_3 content but you also do not use it, only the SiO_2 and TiO_2 contents.

Line 150 - If you consider that index (WI) used in the text as sufficient for your goals, I suggest you to better characterize this index in particular and its meaning about weathering. Moreover, I see in many parts of the text you using "weathering indexes" when in this paper you use just one index.

Line 197 - You are very correct in saying that the use of the IGC data allowed the creation of a more detailed topographic data than orbital sources, such as SRTM, could do. But just a hint for future works: with the scale of IGC database you claimed (1:10.000) you'd be able to obtain an MDE with a way finer resolution than you obtained, which would be more appropriate to the goals that you wanted to achieve.

Table 1 – very nice table explaining each morphometric variable. It will be a good reference for other authors.

Line 217 - It will be interesting you cite in an only place the "7 parameters derived from geophysical sensors data" which you used. According to your descriptions in the "Geophysical data collection" and the Table 3, you used 5 parameters from the geophysical sensors, not 7.

Line 228 - "argiluviation or ferralitization indexes". Again, why two indexes if you supposedly used just one? Furthermore, it is the SiO2/TiO2 index that you described previously one of them? If so, you should have to indicate that before.

Line 229 - "These groups characterize themselves by present similar values within the groups, but different values between one group to another". You should refine the description about the statistical clustering.

Line 275 - Very good insight about the combined use of the PCA and k-means analysis as well as the use of the Nested LOOCV for handling few samples.

Line 233 - “This result was used to extract the values of covariates (morphometric and geological data)”. Not only these covariates, but also the Synthetic Soil image (SYSI).

Line 236 – “base database”?

Line 250 – “increases the performance of machine processing algorithms”. Not necessarily.

Line 260 – the remotion by correlation is not only associated to “reduce computational time” but also to minimize the problem of multicollinearity.

Line 272 – I think it is “repeatecv”.

Line 350 – You cite here that you used the Kruskal Wallis test to choose the best model. But in your results, you only showed the Kruskal Wallis being used to differentiate the weathering clusters. You need to synchronize your material and methods with your results in this regard.

Line 359 – You don’t need to cite RF twice in the sentence.

Line 360 – Before “The RF algorithm presented equivalent performances than other algorithms”, insert “In other studies,”.

Line 381 – “performance precision”? Redundancy.

Line 381 – Paragraph shows some clumsy sentences.

Figure 5 - I’ve seen that the authors explain for the first time what the clusters really mean (weathering intensity) in a clear way only on the Figure 5. I suggest the authors to anticipate this important explanation. At the same time, I suggest the authors to label the clusters with the information of weathering, at last, the numbers 1,2,3 that the authors arbitrarily chose are just categorical, not expressing by themselves the weathering degree.

Line 451 – According to the rest of your text, the cluster 3 has higher weathering, followed by cluster 1 and then the cluster with lesser weathering, cluster 2. In this sentence, you have been confounded the order of the clusters.

Line 458 – “West Rodic Nitisol” – I think the authors confounded the terms. Here, the correct would be “East Rodic Nitisol”.

Line 465 – “For the weathering index”.

Table 3 – interesting table. Could explain in a paragraph which of the parameters explain better the weathering, according with your data? For me, the most important were the W1, plus magnetic susceptibility and ECa.

Line 494 – East diabase

Line 495 – West diabase and West Rhodic Nitisol

Line 514 – free drainage

Line 523 – but there is higher K40 in the supposedly more weathered Lixisols (cluster 1)

Line 526 – decreasing or increasing ECa? You must explain that.

Line 539 – you do not model with the nested LOOCV, but you evaluated it.

My suggestions aim to improve the quality of this excellent paper. I congratulate the authors for their work and hope to read more on this very interesting topic.