Comment on gmd-2021-153
Anonymous Referee #1

Referee comment on "A new methodological framework by geophysical sensors combinations associated with machine learning algorithms to understand soil attributes" by Danilo César de Mello et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-153-RC1, 2021

Overall

This manuscript needs further proofreading and editing to due to spelling and grammatical errors and awkward phrasing in English. Detailed grammatical edits have not been specified in this review as a complete editing of the manuscript is required. There are also undefined acronyms.

While I understand the challenges the authors faced with sample collection. I do not think nested LOOCV is sufficiently rigorous. Especially since the feature selection was done using LOOCV, and in effect then the entire data set, there is no truly independent test set. LOOCV is an insufficiently rigorous test method, particularly when it is used for both hyper parameter tuning and feature selection. There is not a true independent evaluation of model performance in this study.

The authors should be using LOOCV on a separate training data set for feature selection and hyperparameter optimization, with the test data set withheld for final validation only. The data should either be split into relatively small sample sets given only 79 samples are available, or more data should be acquired and added to the study.

The R² values using the LOOCV are low enough that a true independent validation might have no relationship at all between the sensors and the parameters of interest. The results with an independent test set might also be very similar. There is no way to tell given the analysis approach which will have introduced an unknown amount of positive bias into the validation results. The author’s conclusions that it is possible to model soil attributes satisfactorily is not warranted given the analysis approach.

Scientific Questions

Line 24: Make clear how validation was done in the abstract. Cross validation? Independent test dataset?

Line 57: Why did you choose to focus on gamma-spectrometry, magnetic susceptibility
and apparent electrical conductivity. The authors need to add a justification. The explain
the value of these approaches, by not why they selected them vs other methods.

Line 147: What correction factor was used for the Walkley-Black method? Was a soil
specific correction factor available and used? Elemental analysis by dry combustion would
be a better analysis option, however I understand due to cost is may not always be
possible. However, the under consumption of organic matter during Walkley black and
associated correction factor should be discussed.

Line 228: While I understand the challenges the authors faced with sample collection. I do
not think nested LOOCV is sufficiently rigorous.

**Technical Corrections**

Line 25: Please state actual best r2 values not just greater than 0.2. That is unclear

Line 91: While I agree that the best models are those that use the smallest number of
variables, the authors need to explain why and not just state it is better as a self-evident
truth.

Line 198: Table 1 is introduced much earlier than it is discussed.

Line 422: Elaborate on what the relationships are between the VisNIR and XRF sensors.

Line 427: Add more specifics about what pedogenesis and soil attributes are being
accounted for with these sensors

Line 433: State performance metrics not just satisfactory performance. Satisfactory is
contextual and subjective. Just stating not greater than 0.5 is still vague.

Line 451: It can justify the low $R^2$ values obtained is very unclear.

Line 452: Further justification and explanation needs to be added that 0.2 to 0.5 is
satisfactory. I don’t necessarily disagree, but a more rigorous argument needs to be made
rather than just stating a convention. Why does this present informative results?

Line 454: It is unclear what point the authors are trying to make with the statement about
standardized laboratory conditions

Line 461: Add more explanation why, not just that Cracknell and Reading (2003) states it.

Line 465: Explain why this is satisfactory not just that it is

Line 471: The discussion about NULL_RMSE and NULL_MAE is valuable. This needs to be
explained earlier and used to justify the results as satisfactory.

Line 532: Based on the nested LOOCV approach, the results are not truly unbiased
estimates

Line 544: The authors cannot truly conclude that modelling results was satisfactory with
the validation approach they used. Soil attributes is very general and they should be
specific about which attributes they could actual model.

Figure 3: Spelling mistakes
Figure 4 - 8: Define acronyms in caption so figure is stand alone. Write out x axis label to make it easier to read.

Figure 9: Spelling mistakes