Prof. Dr. José Alexandre M. Demattê

Professor of Soil Science Department
University of São Paulo, ESALQ/USP
Av. Pádua Dias 11, PO Box 9
Piracicaba (SP), Brazil. CEP 13418-900

13/12/2021
Dear Dr. Juan A. Añel,
Editor of Geoscientific Model Development (GMD)

We would like to thank you and the reviewers for the suggestions concerning the manuscript "**A new methodological framework by for geophysical sensors combinations associated with machine learning algorithms to understand soil attributes**". We made all corrections following the pointed issues by reviewers, and their corrections or clarifications can be found in the new manuscript provided. Not only the material and methods were better explained but also the entire text was improved in order to fulfil reviewers' requests. In addition, we sent the research for a general review of English (American English) to a specialized company, where a geoscience specialist also reviewed the entire manuscript (Proofreading service). A certificate attesting to the new revision of the manuscript was inserted in the "supplementary material" field.

The following updated version was prepared following the previous instructions:

- New version of the manuscript incorporating the issues raised by the reviewers. The alterations are highlighted in green colour in the manuscript.
- Detailed comments about reviewers' remarks, mainly concerning the methodology.
- The manuscript was adapted according to the code & data policy, and the reviewer suggestion.
- We would like to highlight that many questions were asked in a different way, but the essence of the question was the same. Therefore, the answers were similar for some questions.

Please do not hesitate to contact us for further clarification.

Thank you for your attention in this matter.

Sincerely,

Prof. José A. M. Demattê

Dear authors,

After checking your manuscript, it has come to our attention that it does not comply with our Code and Data Policy.

A: The manuscript was adapted according to the code and data policy.

You do not include in your manuscript the necessary information in the Code and Data Availability section. I strongly recommend you to check some of the recent papers published in Geosc. Mod. Dev. to get an example of what kind of information is necessary to include in it.

A: The manuscript was adapted according to the code and data policy and, following examples from other articles published in GMD journal.

For example, you use several R packages, but you do not list them in this section; you only do it in the main text of the manuscript. Moreover, you do not include the version number for such packages, and you must do it.

A: All packages used in "R software", as well as their respective versions are listed in the database and codes available in the section *list_of_packages_and_versions* in the indicated repository: https://zenodo.org/record/5733366#.YaTXa9DMKUk (DOI: 10.5281/zenodo.5733366).

Also, in your work where machine learning techniques are applied, it is of the utmost importance that you store the input and output data used in a permanent and open repository (recheck our guidelines for a list of suitable repositories). In this way, you must include the modified 'Code and Data Availability' section and the corresponding DOI of the datasets in a potential reviewed version of your manuscript.

A: The manuscript was adapted according to the code and data policy and the reviewer suggestion: https://zenodo.org/record/5733366#.YaTXa9DMKUk (DOI: 10.5281/zenodo.5733366).

In the meantime, you should reply to this comment as soon as possible with the requested information. In this way, it will be available for the review process, as it must be.

A: the questions were answered in time

Juan A. Añel

Geosc. Mod. Dev. Exec. Editor

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.

A: We improved the writing on the entire manuscript and sent the work to a company that specializes in proofreading the English language, as suggested by the reviewers and editor. A certificate attesting to a review and editing of the English will be place in the "supplementary material". In addition, we have corrected all the points indicated by the reviewer in the document that he sent us separately with his comments.

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 parameters tuning and feature selection. There is not a true independent evaluation of model performance in this study.

A: The reviewer stated that Nested-LOOCV is not a rigorous method of validating the true performance of the modeling. We disagree with the reviewer's opinion.
Our training and test separation process was repeated 75 times using the nested leave one out cross-validation (“Nested-LOOCV”) method (Clevers et al., 2007; Honeyborne et al., 2016; Rytky et al., 2020). The “Nested-LOOCV” method is indicated as a set for small data sets, which other methods of evaluation of test samples would not be

viable/appropriate due to the low number of samples in a test samples (Ferreira et al., 2021), being more used in the field of medicine in human experiments or where the number of samples is limited, providing an unbiased estimate of the true error (Chen et al., 2017; Li et al., 2018; Xing et al., 2011; Xu et al., 2020). The explanation of how the Nested-LOOCV operates is detailed explained below:

"The nested-LOOCV method is a double loop process, where in the internal loop, the model is trained with a data set of size n-1, using the LOOCV for the optimization of the final model. On the other hand, the external loop corresponds to the test. In this loop, the remaining sample is predicted using the final model calculated in the inner loop. This prediction result is stored with the observed value of the remaining sample and later used to calculate the algorithm's performance (Jung et al., 2020; Neogi and Dauwels, 2019). *The two loops are run n times (n = total number of samples, in our case 75). All samples are inserted into the outer loop, where the values predicted by the final model of each algorithm are calculated with the predicted and observed values of each sample. Then, the final result of the machine learning algorithm's performance will be obtained by predicted and observed values stored in the external loop. This is a robust method to evaluate the algorithm's performance and detects possible samples with problems in the collections or outliers. The training set generated in each loop went through the process.*

Other validation methods for our experimental conditions (dataset with small number of samples) such as Hold-out validation or Repeated hold-out validation, would not be suitable due to the test sample size (which would be very small).

If the reviewer still disagrees after our explanations, we kindly ask the reviewer to send us some updated works and/or references that affirm and demonstrate that Nested-LOOCV is not a rigorous method for the same experimental conditions of our work.

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.

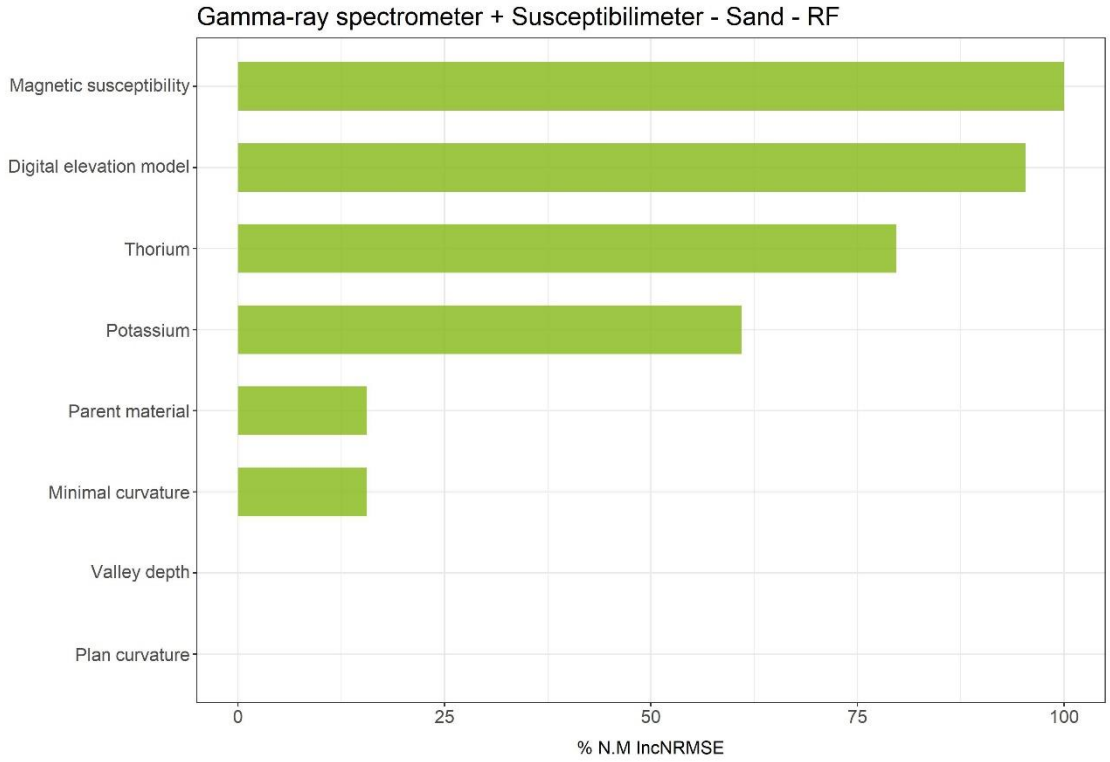
A: We thank the reviewer for that point. However, it is noteworthy that the method used was not the LOOCV, but the "Nested-LOOCV", which already performs exactly the

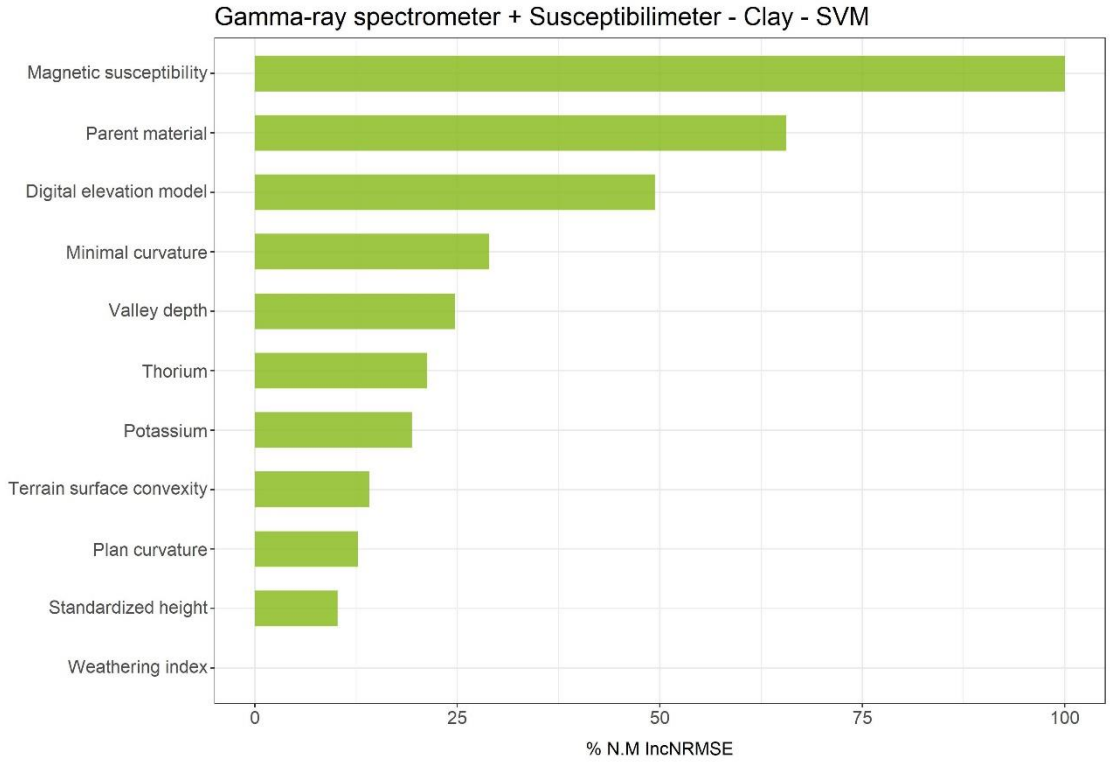
procedure described and requested by the reviewer. It is even explained in the manuscript methodology:

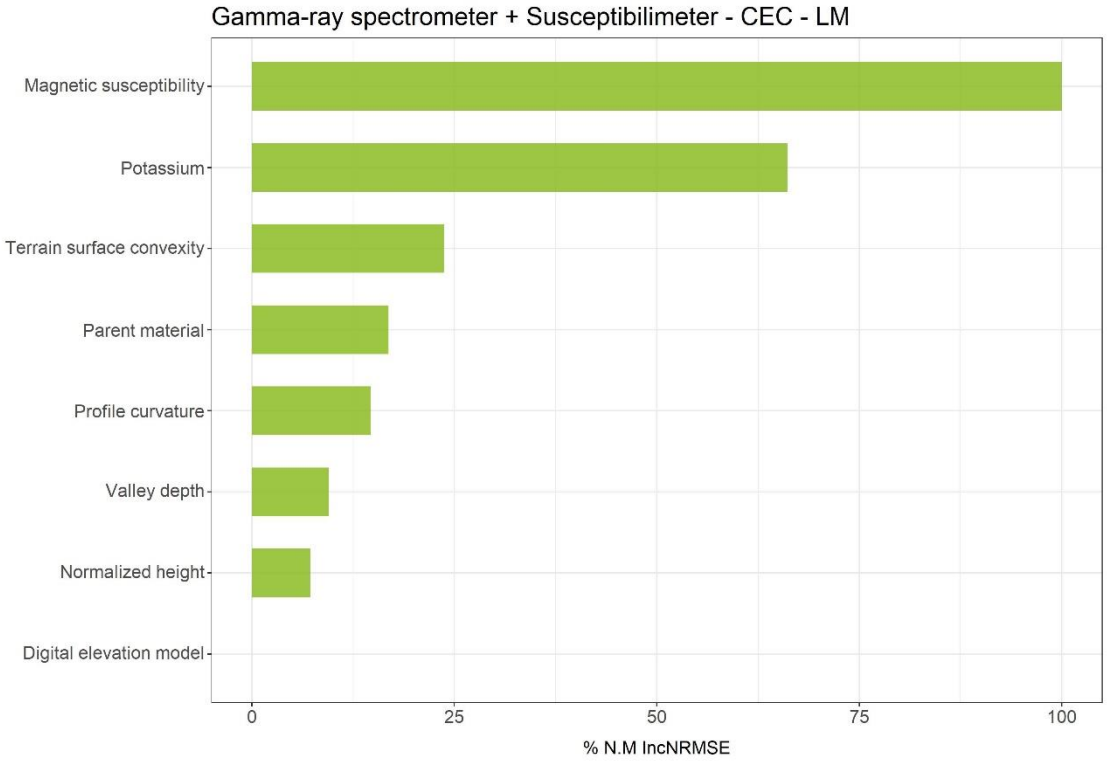
"*…The separation of training and test was performed using the "Nested leave one out" ("Nested LOOCV") method (Clevers et al., 2007; Honeyborne et al., 2016; Rytky et al., 2020). It is important to highlight that our number of soil samples and readings with geophysical sensors is small (75), due to several difficulties encountered in the field in data collection (high sugar cane size, sloping terrain, dense forest, etc.). In this sense, the "Nested LOOCV" method is indicated for small sample sets (values near 100 samples) to which other validation/test methods (as holdout validation) would not be viable due to the low sample set in the test and /or training group (Ferreira et al., 2021). This is one of the main innovations of this research.*

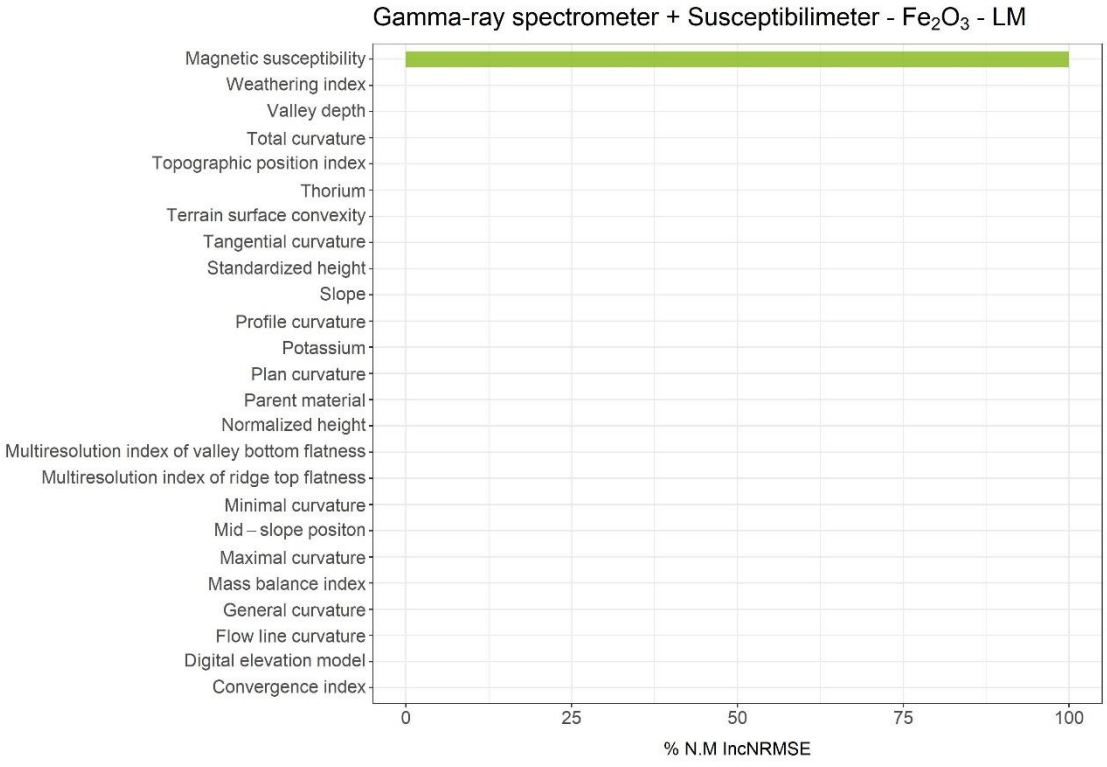
The "Nested LOOCV" method is a double loop process, where in the first loop the model is trained with a data set of size n-1, and the test is done in the second loop with the missing sample and used to validate the training performance (Jung et al., 2020; Neogi and Dauwels, 2019). The final result of the performance of the machine learning algorithm will be the mean performance indicators for all points (Training / test). This is a robust method to evaluate the performance of the algorithm and detects possible samples with problems in the collections or outliers. The training set generated in each loop went through the process of selecting covariates for importance and subsequent training..."

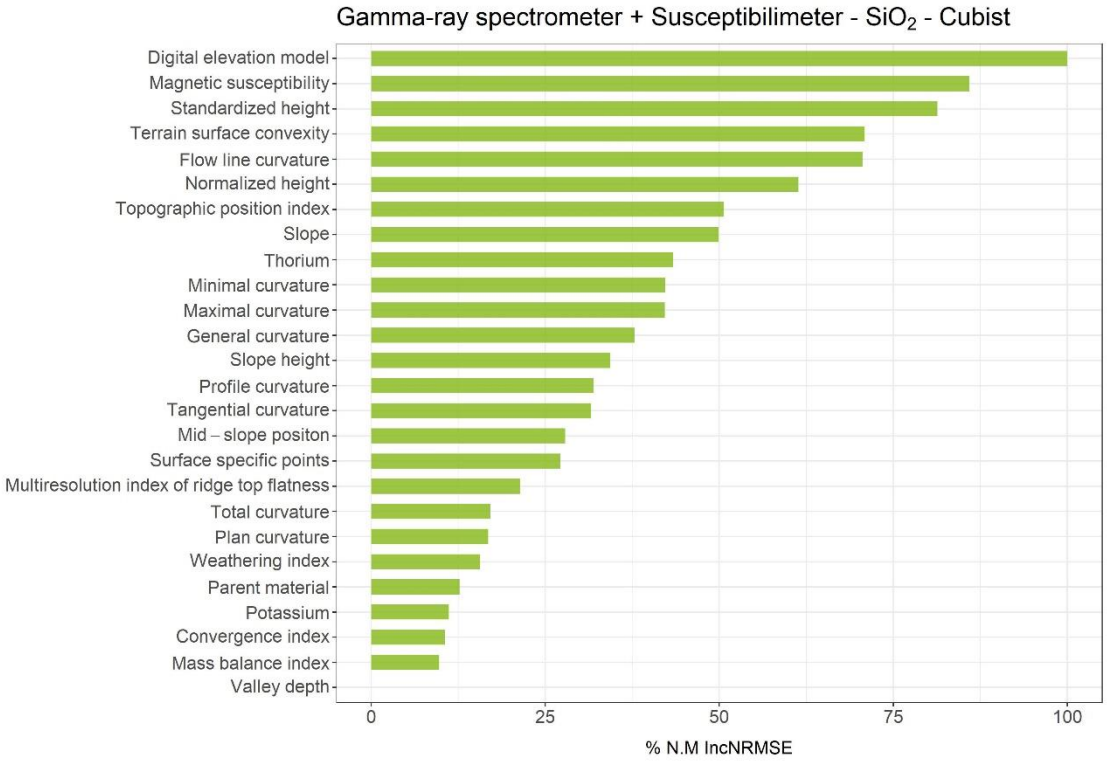
Regarding the division of data into small sample sets, we performed a test by creating 3 random covariates (var_1, var_2 and var_3). These variables were not chosen in any case (Figures below), showing that the models are capable of detecting and adequately separating the variables used.

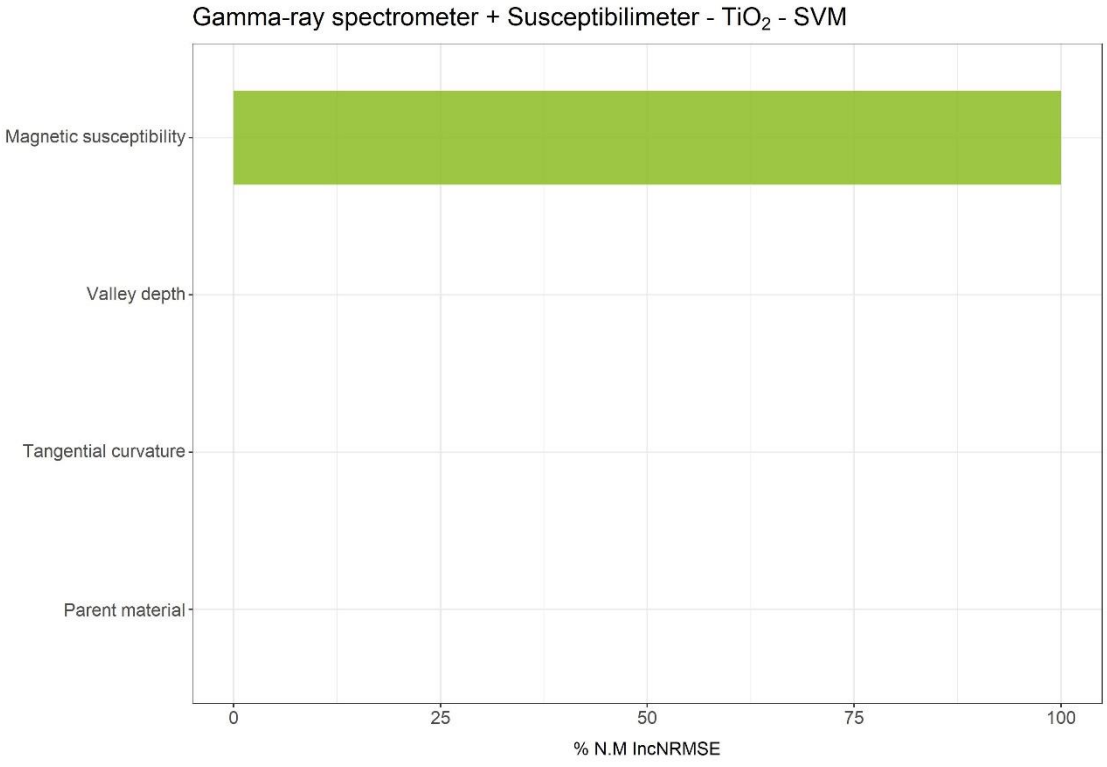

Gamma-ray spectrometer + Susceptibilimeter - Sand - RF
Magnetic susceptibility
Digital elevation model
Thorium
Potassium
Parent material
Minimal curvature
Valley depth
Plan curvature
0
25
50
75
100
% N.M IncNRMSE


Gamma-ray spectrometer + Susceptibilimeter - Clay - SVM
Magnetic susceptibility
Parent material
Digital elevation model
Minimal curvature
Valley depth
Thorium
Potassium
Terrain surface convexity
Plan curvature
Standardized height
Weathering index
0
25
50
75
100
% N.M IncNRMSE


Gamma-ray spectrometer + Susceptibilimeter - CEC - LM
Magnetic susceptibility
Potassium
Terrain surface convexity
Parent material
Profile curvature
Valley depth
Normalized height
Digital elevation model
0
25
50
75
100
% N.M IncNRMSE


Gamma-ray spectrometer + Susceptibilimeter - Fe2O3 - LM
Magnetic susceptibility
Weathering index
Valley depth
Total curvature
Topographic position index
Thorium
Terrain surface convexity
Tangential curvature
Standardized height
Slope
Profile curvature
Potassium
Plan curvature
Parent material
Normalized height
Multiresolution index of valley bottom flatness
Multiresolution index of ridge top flatness
Minimal curvature
Mid – slope positon
Maximal curvature
Mass balance index
General curvature
Flow line curvature
Digital elevation model
Convergence index
0
25
50
75
100
% N.M IncNRMSE





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.

A: Dividing into relatively small sample sets requested by the reviewer is already the standard operating procedure performed by Nested-LOOCV. In addition, one of the

objectives of the work is just to work with a database with a small number of samples. This is because this is a challenging and quite common situation in soil science and geosciences. It is not always possible, for several reasons, to get a large/ideal number of samples for the most varied analyses. Obtaining more samples at this time, in addition to not being possible any further field incursions, is beyond the scope of our work.

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.

A: Analyzing table 5 of the manuscript (our best result of combination of sensors) it is possible to observe that the NULL_MAE values are superior to the MAE values for all the variables addressed and discussed in the manuscript (except BS and OM). In other words, MAE values are below NULL_MAE. This means that the error is smaller compared to the NULL model and that our model presents gain and predicts better than when using the mean values for prediction.

Table 5. Model performance for the combined use of the gamma-ray spectrometer and the susceptibilimeter, for all soil attributes, based on R^2, RMSE, MAE, and NULL_RMSE.

Gamma-ray spectrometer + Susceptibilimeter	***R^2***				
	Random Forest	***Cubist***	***SVM***	***LM***	***-***
Clay	0.465	0.441	0.494	0.366	-
Sand	0.422	0.152	0.367	0.233	-
Fe2O3	0.36	0.426	0.096	0.47	-
TiO2	0.308	0.282	0.284	0.328	-
SiO2	0.159	0.207	0.169	0.167	-
CEC	0.147	0.152	0.296	0.303	-
BS	0.169	0.082	0.112	0.002	-
OM	0.046	0.033	0.028	0.034	-
Gamma-ray spectrometer + Susceptibilimeter	***RMSE***				
	Random Forest	***Cubist***	***SVM***	***LM***	***NULL_RMSE***
Clay	127.149	132.977	123.84	148.11	140.885
Sand	165.624	244.635	175.35	202.104	176.521
Fe2O3	53.418	52.737	67.759	48.513	53.341
TiO2	10.724	11.37	10.846	10.659	10.239
SiO2	40.898	40.244	42.207	42.993	35.45
CEC	41.902	44.296	38.723	37.645	36.139
BS	19.294	21.318	20.856	1024.32	17.142
OM	7.8	7.842	7.81	8.131	6.158
Gamma-ray spectrometer + Susceptibilimeter	***MAE***				
	Random Forest	***Cubist***	***SVM***	***LM***	***NULL_MAE***
Clay	102.229	105.123	97.173	117.097	119.751
Sand	134.525	168.957	140.318	166.083	153.803
Fe2O3	33.284	32.411	42.282	33.124	41.578
TiO2	6.548	6.573	6.447	7.049	8.074
SiO2	30.394	29.691	30.396	32.951	29.534
CEC	28.977	30.945	25.376	25.815	27.187
BS	15.597	17.321	16.96	137.422	14.425
OM	5.805	5.836	5.966	6.262	4.813

Clay and sand content in $g.kg^{-1}$; Fe_2O_3, TiO_2 and SiO_2 in $g.kg^{-1}$ CEC in $mmol_c\ dm^{-3}$; abbreviations: CEC: Cation Exchange Capacity; OM $g.dm^{-3}$; BS: mmolc dm^{-3}. Clay and sand content in $g.kg^{-1}$; Fe_2O_3, TiO_2 and SiO_2 in $g.kg^{-1}$ CEC in $mmol_c\ dm^{-3}$; abbreviations: CEC: Cation Exchange Capacity; OM $g.dm^{-3}$; BS: mmolc dm^{-3}. Support Vector Machines (SVM); Linear Models (LM).

With respect to relatively low R^2 values, to understand the relationship that exists between geophysical variables, soil, morphometry and geology by itself it's already complicated. The low accuracy results can be attributed to the difficulty of modeling soils and their attributes. This is related to the high complexity of soils: high spatial variability in surface and depth; occurrence of geomorphic processes, weathering and pedogenesis; performance of different soil formation factors etc. In addition, it does not exist a

minimum value of R^2 and there are several researches published which presented low performance (R^2 values) of models for soil/attributes prediction due to the high pedoenvironmental complexity:

- ✓ **Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains.** (Andrade et al., 2020a)
- ✓ **Modelling and mapping soil organic carbon stocks in Brazil.** (Gomes et al., 2019)
- ✓ **Satellite data integration for soil clay content modelling at a national scale.** (Loiseau et al., 2019)
- ✓ **The effectiveness of digital soil mapping to predict soil properties over low-relief areas.** (Mosleh et al., 2016)
- ✓ **Evaluation of digital soil mapping approaches with large sets of environmental covariates.** (Nussbaum et al., 2018)
- ✓ **Validation of digital maps derived from spatial disaggregation of legacy soil maps.** (Bargaoui et al., 2019)

- ✓ **Is it possible to map subsurface soil attributes by satellite spectral transfer models?** (Mendes et al., 2019)
- ✓ **Pedology and soil class mapping from proximal and remote sensed data** (Poppiel et al., 2019)

There are many research papers published in high quality scientific journals which already demonstrates the relationship at all between the geophysical sensors and the parameters of interest (soil attributes). However, with different approaches to our research.

Scientific Questions

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

A: We adjusted the text as the reviewer suggested: *"The validation of the results was performed using the method "Nested leave one-out cross validation"."*

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.

A: We chose to work with these 3 geophysical sensors for three reasons: First, we and our research partners only had these three sensors available for the work. Second, these three geophysical sensors are easy to operate, quickly and accurately in acquire field data and provide geophysical variable information, with high correlation with various soil attributes as well as their formation factors. Finally, the EM38 (conductivimeter) and RS 230 (gamma-ray spectrometer) provide information at satisfactory depth, where the most of pedogenetic processes occur. In addition, the information of EM38 and RS 230 associate with KT10 (susceptibilimeter) on soil surface provide additional information about some soil attributes related to soil subsurface horizons, which is also related to the other geophysical variable used (gamma-ray and apparent electrical conductivity).

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.

A: We used the conventional factor of 1.724 known as “van Bemmelen factor” (Van Bemmelen, 1890).

“Elemental analysis by dry combustion would be a better analysis option, however I understand due to cost is may not always be possible. We agree with the reviewer statement. However, we deem it unnecessary to discuss this factor as it is only a detail in the determination of organic matter and it already, we cited the correct methodology for such analysis. In addition, we did not address at any point in the work (results and discussion) the variable organic matter and organic carbon, due to the very low performance of the models in predicting this attribute.

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

A: The reviewer stated that Nested-LOOCV is not a rigorous method of validating the true performance of the modeling.

Our training and test separation process was repeated 75 times using the nested leave one out cross-validation ("Nested-LOOCV") method (Clevers et al., 2007; Honeyborne et al., 2016; Rytky et al., 2020). The "Nested-LOOCV" method is indicated as a set for small data sets, which other methods of evaluation of test samples would not be viable/appropriate due to the low number of samples in a test samples (Ferreira et al., 2021), being more used in the field of medicine in human experiments or where the number of samples is limited, providing an unbiased estimate of the true error (Chen et al., 2017; Li et al., 2018; Xing et al., 2011; Xu et al., 2020). The explanation of how the Nested-LOOCV operates is detailed explained below:

"The nested-LOOCV method is a double loop process, where in the internal loop, the model is trained with a data set of size n-1, using the LOOCV for the optimization of the final model. On the other hand, the external loop corresponds to the test. In this loop, the remaining sample is predicted using the final model calculated in the inner loop. This prediction result is stored with the observed value of the remaining sample and later used to calculate the algorithm's performance (Jung et al., 2020; Neogi and Dauwels, 2019). *The two loops are run n times (n = total number of samples, in our case 75). All samples are inserted into the outer loop, where the values predicted by the final model of each algorithm are calculated with the predicted and observed values of each sample. Then, the final result of the machine learning algorithm's performance will be obtained by predicted and observed values stored in the external loop. This is a robust method to evaluate the algorithm's performance and detects possible samples with problems in the collections or outliers. The training set generated in each loop went through the process*

Other validation methods for our experimental conditions (dataset with small number of samples) such as Hold-out validation or Repeated hold-out validation, would not be suitable due to the test sample size (which would be very small).

If the reviewer still disagrees after our explanations, we kindly ask the reviewer to send us some updated works and/or references that affirm and demonstrate that Nested-LOOCV is not a rigorous method for the same experimental conditions of our work.

Technical Corrections

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

A: The text was adjusted following the reviewers' recommendations.

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.

A: Models that use fewer variables usually optimize the modelling process and give better results. Also, it makes easier to explain the variables influence in modelling process. Furthermore, information generated from a large number of variables can generate ambiguous information, which makes it difficult to interpret the results generated by the models.

We adjusted the text as the reviewer suggested

"…Models that use fewer variables usually optimize the modelling process, makes easier to explain the variables influence in modelling process and give better results easier to interpret. In addition, that facilitates the understanding and the faster computer processing (Brungard et al., 2015)…"

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

A: We fully agree with the reviewer. The table was properly relocated in the new version of the manuscript.

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

A: We added the follow statement highlighted in the text "…*The VisNIR spectroscopy acts on targets with low energy levels, showing the ability to identify soil mineral species, strongly linked to soil attributes (Coblinski et al., 2021). In addition, pXRF spectroscopy allows the identification of total elementary contents by acting with high levels of ionizing energy, which is not identified by Vis-NIR, and is strongly correlated with minerals and*

soil attributes (Silvero et al., 2020). Therefore, the addition of pXRF with Vis-NIR data for obtaining information about soil constituents is highly efficient for modeling soil attributes.…"

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

A: The sentence was adjusted following the reviewer's recommendations.

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

A: Comparing our results with those obtained in other works of prediction of soil attributes, our results fall within the range of R^2 considered satisfactory (approximately between 0.2 - 0.5) (Dharumarajan et al., 2017; Henderson et al., 2005; Khaledian and Miller, 2020; Mansuy et al., 2014; Mosleh et al., 2016; Poggio et al., 2016).

The R^2 value (approximately between 0.2 - 0.5) results can be attributed to the difficulty of modeling soils and their attributes. This is related to the high complexity of soils: high spatial variability in surface and depth; occurrence of geomorphic processes, weathering and pedogenesis; performance of different soil formation factors etc. In addition, there are several researches published which presented similar performance with our results to models for soil/attributes prediction:

- ✓ **Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains.** (Andrade et al., 2020)
- ✓ **Modelling and mapping soil organic carbon stocks in Brazil.** (Gomes et al., 2019)
- ✓ **Satellite data integration for soil clay content modelling at a national scale.** (Loiseau et al., 2019)
- ✓ **The effectiveness of digital soil mapping to predict soil properties over low-relief areas.** (Mosleh et al., 2016)
- ✓ **Evaluation of digital soil mapping approaches with large sets of environmental covariates.** (Nussbaum et al., 2018)
- ✓ **Validation of digital maps derived from spatial disaggregation of legacy soil maps.** (Bargaoui et al., 2019)

✓ **Is it possible to map subsurface soil attributes by satellite spectral transfer models?** (Mendes et al., 2019)

✓ **Pedology and soil class mapping from proximal and remote sensed data** (Poppiel et al., 2019)

Line 451: It can justify the low R2 values obtained is very unclear.

A: We explain in detail in the text the factors that may account for the relatively low values of R2.

"…The relatively low R^2 value (approximately between 0.2 - 0.5) results can be attributed to the difficulty of modeling soils and their attributes. This is related to the high complexity of soils: high spatial variability in surface and depth; occurrence of geomorphic processes, weathering and pedogenesis; performance of different soil formation factors etc…".

In addition, there are several researches published which presented similar performance with our results to models for soil/attributes prediction:

✓ **Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains.** (Andrade et al., 2020)

✓ **Modelling and mapping soil organic carbon stocks in Brazil.** (Gomes et al., 2019)

✓ **Satellite data integration for soil clay content modelling at a national scale.** (Loiseau et al., 2019)

✓ **The effectiveness of digital soil mapping to predict soil properties over low-relief areas.** (Mosleh et al., 2016)

✓ **Evaluation of digital soil mapping approaches with large sets of environmental covariates.** (Nussbaum et al., 2018)

✓ **Validation of digital maps derived from spatial disaggregation of legacy soil maps.** (Bargaoui et al., 2019)

✓ **Is it possible to map subsurface soil attributes by satellite spectral transfer models?** (Mendes et al., 2019)

✓ **Pedology and soil class mapping from proximal and remote sensed data** (Poppiel et al., 2019)**.**

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?

A: The low R^2 value (approximately between 0.2 - 0.5) results can be attributed to the difficulty of modeling soils and their attributes. This is related to the high complexity of soils: high spatial variability in surface and depth; occurrence of geomorphic processes, weathering and pedogenesis; performance of different soil formation factors etc. Also, there is no common census among researchers regarding these R^2 values. Other researchers used and argued. We just use it. If reviewers want, we can remove that part of the text.

In addition, there are several researches published which presented similar performance with our results to models for soil/attributes prediction:

- ✓ **Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains.** (Andrade et al., 2020)
- ✓ **Modelling and mapping soil organic carbon stocks in Brazil.** (Gomes et al., 2019)
- ✓ **Satellite data integration for soil clay content modelling at a national scale.** (Loiseau et al., 2019)
- ✓ **The effectiveness of digital soil mapping to predict soil properties over low-relief areas.** (Mosleh et al., 2016)
- ✓ **Evaluation of digital soil mapping approaches with large sets of environmental covariates.** (Nussbaum et al., 2018)
- ✓ **Validation of digital maps derived from spatial disaggregation of legacy soil maps.** (Bargaoui et al., 2019)
- ✓ **Is it possible to map subsurface soil attributes by satellite spectral transfer models?** (Mendes et al., 2019)
- ✓ **Pedology and soil class mapping from proximal and remote sensed data** (Poppiel et al., 2019)

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

A: We try to emphasize that in the laboratory the analyzes are conducted under standardized experimental conditions in terms of dosage, temperature, humidity, substance concentrations, and other variables that interfere with the analysis results (in this case they tend to optimize). We adjusted the sentence following the reviewer suggestion.

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

A: We agree with the reviewer. We chose to remove the final sentence, since the cited author used a classification, while our case is a regression.

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

A: The sentence was adjusted following the reviewer's recommendations. However, the justification for this statement was given in previous paragraphs. Repeating it here would make the text repetitive and redundant information.

"For example, the low R^2 value (approximately between 0.2 - 0.5) results can be attributed to the difficulty of modeling soils and their attributes. This is related to the high complexity of soils, such as: high spatial variability in surface and depth; occurrence of geomorphic processes, weathering and pedogenesis; performance of different soil formation factors etc. It can justify the low R^2 values obtained. For soil mineralogical attributes predicted by machine learning algorithms, results can be classified as satisfactory from 0.2 to 0.5, as for preliminary evaluation, since these values present more informative results (Beckett, 1971; Dobos, 2003; Malone et al., 2009)."

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.

A: The Null model (NULL_RMSE and NULL_MAE) emulates other model building functions, but returns the simplest model possible given a training set: a single mean for numeric outcomes. The percentage of the training set samples with the most prevalent class is returned when class probabilities are requested. The Null model can be considered

the simplest model that can be adjusted and that serves as a reference. Models that presents similar or worst perform to the Null model should be discarded. The best models had lower RMSE and MAE results than those found for NULL_MAE and NULL_RMSE. This shows that the final model is better than using the mean values, which also demonstrates better quality in creating the models.

$$NULL\_RMSE = \left[\frac{1}{N}\sum_{i=1}^{N}(\overline{Qtrain_i} - Qobs_i)^2\right]^{\frac{1}{2}} \text{(Eq.4)}$$

$$NULL\_MAE = \frac{1}{n}\times\sum|\overline{Qtrain_i} - Qobs_i| \quad \text{(Eq.5)}$$

Where:

Qtrain = the mean of the training samples

Qobsi = the validation sample

N =number of samples (loop).

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

A: *"An accuracy estimate obtained using LOOCV is known to be almost unbiased, but it has high variance, leading to unreliable estimates (Efron, 1983). It is still widely used when the available data are very rare, especially in bioinformatics where only dozens of data samples are available"* (Liu and Özsu, 2009).

Cross-Validation, Table 1 Pros and cons of different validation methods

Validation method	Pros	Cons
Resubstitution validation	Simple	Over-fitting
Hold-out validation	Independent training and test	Reduced data for training and testing, large variance
k-fold cross-validation	Accurate performance estimation	Small samples of performance estimation, overlapped training data, elevated type I error for comparison, underestimated performance variance or overestimated degree of freedom for comparison
Leave-one-out cross-validation	Unbiased performance estimation	Very large variance
Repeated *k*-fold cross-validation	Large number of performance estimates	Overlapped training and test data between each round, underestimated performance variance or overestimated degree of freedom for comparison

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.

A: The reviewer questioned the performance of the models with the validation method used, claiming that the performance values obtained were not satisfactory. This implies that Nested-LOOCV is not a rigorous method of validating the true performance of the modeling. We disagree with the reviewer's opinion. We would like, if possible, the reviewer to send us some updated works and/or references that affirm and/or demonstrate that Nested-LOOCV is not a rigorous method for the same experimental conditions of our work.

Our training and test separation process was repeated 75 times using the nested leave one out cross-validation ("Nested-LOOCV") method (Clevers et al., 2007; Honeyborne et al., 2016; Rytky et al., 2020). The "Nested-LOOCV" method is indicated as a set for small data sets, which other methods of evaluation of test samples would not be viable/appropriate due to the low number of samples in a test samples (Ferreira et al., 2021), being more used in the field of medicine in human experiments or where the number of samples is limited, providing an unbiased estimate of the true error (Chen et al., 2017; Li et al., 2018; Xing et al., 2011; Xu et al., 2020). The explanation of how the Nested-LOOCV operates is detailed explained below:

"The nested-LOOCV method is a double loop process, where in the internal loop, the model is trained with a data set of size n-1, using the LOOCV for the optimization of the final model. On the other hand, the external loop corresponds to the test. In this loop, the remaining sample is predicted using the final model calculated in the inner loop. This prediction result is stored with the observed value of the remaining sample and later used to calculate the algorithm's performance (Jung et al., 2020; Neogi and Dauwels, 2019). *The two loops are run n times (n = total number of samples, in our case 75). All samples are inserted into the outer loop, where the values predicted by the final model of each algorithm are calculated with the predicted and observed values of each sample. Then, the final result of the machine learning algorithm's performance will be obtained by predicted and observed values stored in the external loop. This is a robust method to evaluate the algorithm's performance and detects possible samples with problems in the collections or outliers. The training set generated in each loop went through the process*

Other validation methods for our experimental conditions (dataset with small number of samples) such as Hold-out validation or Repeated hold-out validation, would not be suitable due to the test sample size (which would be very small).

If the reviewer still disagrees after our explanations, we kindly ask the reviewer to send us some updated works and/or references that affirm and demonstrate that Nested-LOOCV is not a rigorous method for the same experimental conditions of our work.

Figure 3: Spelling mistakes

A: The spelling mistakes was adjusted by proofreading service.

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

A: The figure was adjusted following the reviewer's recommendations.

Figure 9: Spelling mistakes

A: The spelling mistakes was adjusted by proofreading service.

The manuscript aims to provide a novel methodological framework for combining terrain attributes and data from geophysical sensors with machine learning algorithms in order to understand the pedosphere system and model soil attributes. An analysis of the importance of pedoenvironmental variables in predictive modelling is also presented.

A: Exactly, this was the main proposal of this work.

Although the study has scientific significance to it, it is undermined by the writing issues present in the manuscript. The manuscript contains numerous grammatical and spelling errors (Eg - misspelt hyperparameters in 244 etc.). Furthermore, some ideas are not well fleshed out which hinders the understanding. For example, the

discussion about the null model could be further elaborated while discussing the results.

A: We improved the writing on the entire manuscript and sent the manuscript to a company that specializes in proofreading the English language, as suggested by the reviewers and editor. With respect to the discussion section about null model we improve this section by conceptualize and further explain how this model works.

"The Null model is a simple model (naive) that expresses the value of the mean of the Y (variable to be predicted or target variable). The RMSE and MAE values are calculated for the Null model. This value is further compared with MAE and RMSE calculated to other models. If the RMSE and MAE values from other models present similar or worse performance than the Null model, the model compared it is not an informative model. In these cases, it is better to choose to use a simple mean as a predictor, rather than using a more complex model to explain a given phenomenon. The null model sets a minimum performance threshold to be reached by models. There are little studies using NULL_RMSE and NULL_MAE as parameters for model evaluation and decision making."

Additionally, I agree with the comments that the LOOCV is not a rigorous method of validating the true performance of the modelling since it lacks a true test set to evaluate the model against. It is recommended that the authors use a separate test set for validation purposes.

A: Thanks for the contribution. However, it is noteworthy that we do not use LOOCV to evaluate our results. We use the "Nested-LOOCV" method. The method is described in detail below.

"The nested-LOOCV method is a double loop process, where in the internal loop, the model is trained with a data set of size n-1, using the LOOCV for the optimization of the final model. On the other hand, the external loop corresponds to the test. In this loop, the remaining sample is predicted using the final model calculated in the inner loop. This prediction result is stored with the observed value of the remaining sample and later used to calculate the algorithm's performance (Jung et al., 2020; Neogi and Dauwels, 2019). *The two loops are run n times (n = total number of samples, in our case 75). All samples are inserted into the outer loop, where the values predicted by the final model of each algorithm are calculated with the predicted and observed values of each sample. Then,*

the final result of the machine learning algorithm's performance will be obtained by predicted and observed values stored in the external loop. This is a robust method to evaluate the algorithm's performance and detects possible samples with problems in the collections or outliers. The training set generated in each loop went through the process

The reviewer stated that Nested-LOOCV is not a rigorous method of validating the true performance of the modeling. We disagree with the reviewer's opinion. We would like the reviewer to send us some updated works and/or references that affirm and demonstrate that Nested-LOOCV is not a rigorous method for the same experimental conditions of our work.

Other validation methods for our experimental conditions (dataset with small number of samples) such as Hold-out validation or Repeated hold-out validation, would not be suitable due to the test sample size (which would be very small).

Some of the specific issues are highlighted below:

Grammatical errors and spelling mistakes need to be fixed.

A: We improved the writing on the entire manuscript and sent the work to a company that specializes in proofreading the English language, as suggested by the reviewers and editor.

The methodological flowchart - In Pearson's test 95% threshold branching, at least one branch must have 'yes'.

A: The flowchart was redone and we performed the adjustment recommended by the reviewer.

The mathematical notations in the paper need to be more consistent. For Eg, in eq 4, change RMSE_NULL to NULL_RMSE.

A: The mathematical notations was adjusted to be more consistent as the reviewer suggested.

Many of the symbols used have not been clearly explained earlier. For example, it is not specified what the values Om and Oi mean in Eq 4 (I believe these must be Qm and Qobs_i, respectively). Similarly, in eq 5, it is not clear what Qtrain means.

A: All the symbols used were clearly explained as the reviewer suggested.

Qpred = predicted samples

Qobs = observed samples

n = the number of samples

Overall, the presentation of the manuscript needs to be improved significantly in further revisions, along with a diligent model validation approach.

A: The presentation of the manuscript underwent a review by the authors of the work and it was improved. The validation method was done using the Nested-leave-one out method. This method is rigorous and careful, due to the fact that the "Nested-leave-one-out" ("Nested-LOOCV") has an external set of validation (test). This external validation sample differentiates this method from LOOCV (Leave-one-out-cross validation).

References

Andrade, R., Faria, W. M., Silva, S. H. G., Chakraborty, S., Weindorf, D. C., Mesquita, L. F., Guilherme, L. R. G. and Curi, N.: Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains, Geoderma, 357(September 2019), 113960, doi:10.1016/j.geoderma.2019.113960, 2020.

Bargaoui, Y. E., Walter, C., Michot, D., Saby, N. P. A., Vincent, S. and Lemercier, B.: Validation of digital maps derived from spatial disaggregation of legacy soil maps, Geoderma, 356, 113907, 2019.

Beckett, P. H. T.: Soil variability: a review, Soils Fertil., 34(1), 1–15, 1971.

Van Bemmelen, J. M.: Über die Bestimmung des Wassers, des Humus, des Schwefels, der in den colloïdalen Silikaten gebundenen Kieselsäure, des Mangans usw im Ackerboden, Die Landwirthschaftlichen Versuchs-Stationen, 37(279), e290, 1890.

Chen, X., Zhang, H., Lee, S.-W. and Shen, D.: Hierarchical high-order functional connectivity networks and selective feature fusion for MCI classification, Neuroinformatics, 15(3), 271–284, 2017.

Clevers, J. G. P. W., Van Der Heijden, G. W. A. M., Verzakov, S. and Schaepman, M. E.: Estimating grassland biomass using SVM band shaving of hyperspectral data, Photogramm. Eng. Remote Sensing, 73(10), 1141–1148, doi:10.14358/PERS.73.10.1141, 2007.

Dharumarajan, S., Hegde, R. and Singh, S. K.: Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India, Geoderma Reg., 10, 154–162, doi:https://doi.org/10.1016/j.geodrs.2017.07.005, 2017.

Dobos, E.: The application of remote sensing and terrain modeling to soil characterization, Innov. Soil-Plant Syst. Sustain. Agric. Pract., 328–348, 2003.

Efron, B.: Estimating the error rate of a prediction rule: improvement on cross-validation, J. Am. Stat. Assoc., 78(382), 316–331, 1983.

Ferreira, R. G., da Silva, D. D., Elesbon, A. A. A., Fernandes-Filho, E. I., Veloso, G. V., de Souza Fraga, M. and Ferreira, L. B.: Machine learning models for streamflow regionalization in a tropical watershed, J. Environ. Manage., 280, 111713, 2021.

Gomes, L. C., Faria, R. M., Souza, E. De, Veloso, G. V., Ernesto, C., Schaefer, G. R., Inácio, E. and Filho, F.: Modelling and mapping soil organic carbon stocks in Brazil, Geoderma, 340, 337–350, doi:10.1016/j.geoderma.2019.01.007, 2019.

Henderson, B. L., Bui, E. N., Moran, C. J. and Simon, D. A. P.: Australia-wide predictions of soil properties using decision trees, Geoderma, 124(3), 383–398, doi:https://doi.org/10.1016/j.geoderma.2004.06.007, 2005.

Honeyborne, I., McHugh, T. D., Kuittinen, I., Cichonska, A., Evangelopoulos, D., Ronacher, K., van Helden, P. D., Gillespie, S. H., Fernandez-Reyes, D., Walzl, G., Rousu, J., Butcher, P. D. and Waddell, S. J.: Profiling persistent tubercule bacilli from patient sputa during therapy predicts early drug efficacy, BMC Med., 14(1), 1–13, doi:10.1186/s12916-016-0609-3, 2016.

Jung, Y., Lee, J., Lee, M., Kang, N. and Lee, I.: Probabilistic analytical target cascading

using kernel density estimation for accurate uncertainty propagation, Struct. Multidiscip. Optim., 1–19, 2020.

Khaledian, Y. and Miller, B. A.: Selecting appropriate machine learning methods for digital soil mapping, Appl. Math. Model., 81, 401–418, 2020.

Li, Y., Liu, J., Huang, J., Li, Z. and Liang, P.: Learning brain connectivity sub-networks by group-constrained sparse inverse covariance estimation for Alzheimer's disease classification, Front. Neuroinform., 12, 58, 2018.

Liu, L. and Özsu, M. T.: Encyclopedia of database systems, Springer New York, NY, USA:, 2009.

Loiseau, T., Chen, S., Mulder, V. L., Dobarco, M. R., Richer-de-Forges, A. C., Lehmann, S., Bourennane, H., Saby, N. P. A., Martin, M. P. and Vaudour, E.: Satellite data integration for soil clay content modelling at a national scale, Int. J. Appl. Earth Obs. Geoinf., 82, 101905, 2019.

Malone, B. P., McBratney, A. B., Minasny, B. and Laslett, G. M.: Mapping continuous depth functions of soil carbon storage and available water capacity, Geoderma, 154(1–2), 138–152, doi:10.1016/j.geoderma.2009.10.007, 2009.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V. and Beaudoin, A.: Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method, Geoderma, 235–236, 59–73, doi:https://doi.org/10.1016/j.geoderma.2014.06.032, 2014.

Mendes, W. D. S., Medeiros Neto, L. G., Demattê, J. A. M., Gallo, B. C., Rizzo, R., Safanelli, J. L. and Fongaro, C. T.: Is it possible to map subsurface soil attributes by satellite spectral transfer models?, Geoderma, 343(January), 269–279, doi:10.1016/j.geoderma.2019.01.025, 2019.

Mosleh, Z., Salehi, M. H., Jafari, A., Borujeni, I. E. and Mehnatkesh, A.: The effectiveness of digital soil mapping to predict soil properties over low-relief areas, Environ. Monit. Assess., 188(3), 195, 2016.

Neogi, S. and Dauwels, J.: Factored Latent-Dynamic Conditional Random Fields for Single and Multi-label Sequence Modeling, arXiv Prepr. arXiv1911.03667, 2019.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L.,

Schaepman, M. E. and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, Soil, 4(1), 1–22, 2018.

Poggio, L., Gimona, A., Spezia, L. and Brewer, M. J.: Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA, Geoderma, 277, 69–82, doi:https://doi.org/10.1016/j.geoderma.2016.04.026, 2016.

Poppiel, R. R., Lacerda, M. P. C., Demattê, J. A. M., Oliveira, M. P., Gallo, B. C. and Safanelli, J. L.: Pedology and soil class mapping from proximal and remote sensed data, Geoderma, 348(October 2018), 189–206, doi:10.1016/j.geoderma.2019.04.028, 2019.

Rytky, S. J. O., Tiulpin, A., Frondelius, T., Finnilä, M. A. J., Karhula, S. S., Leino, J., Pritzker, K. P. H., Valkealahti, M., Lehenkari, P., Joukainen, A., Kröger, H., Nieminen, H. J. and Saarakkala, S.: Automating three-dimensional osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography, Osteoarthr. Cartil., 28(8), 1133–1144, doi:10.1016/j.joca.2020.05.002, 2020.

Xing, J., Wang, S. X., Jang, C., Zhu, Y. and Hao, J. M.: Nonlinear response of ozone to precursor emission changes in China: a modeling study using response surface methodology, Atmos. Chem. Phys., 11(10), 5027–5044, 2011.

Xu, X., Li, W., Tao, M., Xie, Z., Gao, X., Yue, L. and Wang, P.: Effective and Accurate Diagnosis of Subjective Cognitive Decline Based on Functional Connection and Graph Theory View, Front. Neurosci., 14, 2020.