# Replies to Referee 1's comments on "Insights into the prediction uncertainty of machine learning-based digital soil mapping through a local attribution approach" (egusphere-2024-323)

We would like to thank the Editor and Referee 1 for giving us the opportunity to correct our manuscript based on the new comments. We agree with most of the suggestions and, therefore, we have modified the manuscript to take on board their comments. Marked changes are indicated in green. The line numbers are those of the manuscript with tracked changes.

## Editor:

*R1 has noted improvement in the revised manuscript but also considered that some more changes are needed before we can accept the manuscript. I agree with this report and hope that authors can answer the remaining comments.*
We thank the Editor for this new round of review. We have paid a special care to address the problems raised by Referee 1 by clarifying multiple aspects; in particular the problem of spatial clustering, and of spatial extrapolation, as well some confusing terms.

In the acknowledgement section, we also indicate that we are grateful to the two anonymous referees for their comments and recommendations that led to the improvements of the manuscript.

## Referee 1:

*I commend the authors on an improved manuscript. The methodology section now reads better, and the various steps are also clearer. I am also pleased with the improved methodology in terms of performing the screening analysis within the cross-validation. However, even with the improvements, certain sections are still a bit cryptic. See additional concerns below (some are minor and other are more major). I based my second review on the track changes document.*

*Line 260. The sentence starting with "Overall, the RF …" reads strange.*
The term "Overall" was removed.

*Line 280. maybe rather: "Shapley values, as defined in Sect. 3.2, … ". Avoid to overly depend on brackets.*
This is now corrected.

*Line 289-291. Try rewriting in more than one sentence. It is currently hard to follow. In addition, I am unsure what the authors mean by "… having uniquely high and low values…".*
We agree that this sentence is unclear. We now more clearly indicate the problem of representativeness of the training dataset as follows: "These differences in importance may be related to the scarcity of soil samples in both zones (see Fig. 2). This means that the training data are not representative of both zones".

*Line 299-300. First part of the sentence does not make sense. The part with "to estimate the conditional mean, which is used as the best estimate of the prediction,…". Rewrite, because this is technically wrong. How can the estimate of the conditional mean be the estimate of the prediction?*

The sentence is now rewritten as follows: "We use the conditional mean as the best estimate of the prediction, and the interquartile width *IQW* as the uncertainty estimate, with the 25$^{th}$ and 75$^{th}$ quantiles computed using a qRF model."

*Line 301. I must admit I am getting lost with this part. Maybe other readers will as well. The authors mention the difficulty of related to the clustering (i.e., verb) of the observations. Because this term is also used in this manuscript to refer to the clustering algorithm, this is a bad choice for meaning how the points are distributed. Could the authors clarify what they mean here. Are they referring to how the points are spatially distributed?*

We thank Referee 1 for noticing this problem. We confirm that Referee 1's interpretation is correct. The problem is related to how the points are spatially distributed.

*Line 301: I am also confused as to why this is a problem? Given that my understanding of the above point is right.*

*Lines 302-308. Is all of this necessary? Was this discuss in the methodology section? So, to make sure I understand all of this. Since the points are spatially clustered, that is, the points are not well dispersed over the region, the authors define weights which must then be used when observations are sampled when the bootstrap samples (i.e. trees) are drawn. If my understanding is correct, then this seems all a bit unnecessary. Could the authors elaborate why this is necessary?*

*In addition, why would you bring additional methodology that was not discussed in the previous sections?*

*Also, what if the weights do not address the feature space well?*

*Another question, is this step necessary when you include covariates that used to address the spatial aspect of the data? I mean, you included covariates such as the coordinates and various distances.*

*Can the authors highlight DSM studies where this has been done? Again, I am just trying to understand the motivation behind this methodology in these lines.*

We reply below to this series of comments.

In order to clarify why this is a problem, we provide more details in Sect. 4.2 (page 14, lines 300-304) as follows: "In our case, one additional difficulty is related to how the points are spatially distributed. Figure 3a shows that the points are spatially clustered as they overrepresent some regions while underrepresent, or even miss, others. This situation might lead to biased predictions, because the same weight is given to every point and thus regions with high sampling density are overweighted."

It is important to note that in the original version of the study, we overlooked this problem. At the first submission stage, the Editor rightly pointed out the need to remedy it. We have therefore adopted a weighting procedure adapted from Bel et al. (2009) and Xu et al. (2016) that implies weighting the training data by the inverse sampling intensity.

We preferably describe the approach in section 4.2, because this problem is specific to our real case and because we do not claim to provide a new approach to solving it (see our last answer below).

Adopting the weighting procedure had two implications:

- a clear decrease of the prediction uncertainty, and an intensification of the prediction uncertainty in the zones where the initial RF (without weighting) was not “confident”;
- the emergence of some clearer spatial patterns.

As rightly pointing out by Referee 1, the proposed approach using sampling intensity can be improved in different manners. Estimating the weights using the feature space is certainly a valuable suggestion. Some preliminary investigations have been tested in a conference (see e.g., https://geostat23.sciencesconf.org/489353). Alternative options have also been proposed in the literature; see e.g., the reference to de Bruin et al. (2022), which provides a recent example of DSM study having addressed this spatial clustering problem.

It is important to note that our objective in this study goes beyond improving the predictive capability of the RF model. Given the level of prediction uncertainty obtained using “classical methods”, we aim to investigate what are the main drivers of this uncertainty and whether they differ from the ones driving the best estimate of the prediction. Therefore we have highlighted in the concluding remarks (page 24, lines 512-516) some lines of improvements on that particular topic as well on others as follows: “Third, we focused on one type of machine learning model, i.e., the quantile RF model. Alternative approaches should be considered in future research: different types of machine learning models, such as deep learning techniques, which have shown promising results (see, e.g., Kirkwood et al. (2022)), and improved approaches, in particular, to address complex sample distributions such as clustering (see, e.g., de Bruin et al. (2022) and references therein)”.

**Reference:**

De Bruin, S., Brus, D. J., Heuvelink, G. B., van Ebbenhorst Tengbergen, T., and Wadoux, A. M. C.: Dealing with clustered samples for assessing map accuracy by cross-validation. Ecological Informatics, 69, 101665, 2022.

*Line 337: “…covariates are retained in the construction of the RF model.” But the RF was already constructed if the cross-validation was performed. So why are covariates retained? What does this mean?*
*Line: 361. Oh, I see retained for the group based shap. Is this what the authors meant at Line 337? If so, then make it clearer. If not, please explain.*
Referee 1 is correct. Thank you for noticing this confusing statement. We now specifically specify in Sect. 4.2 (page 15, line 340) that: “These covariates are retained in the construction of the final RF model that is used for the application of the group-based SHAP approach.”

*Line 406: models, plural?*
This typo is now corrected.

*Line 406: This is also a very strange sentence, because the RF model cannot extrapolate. See this post for example that explains it (https://stats.stackexchange.com/questions/235189/random-forest-regression-not-predicting-higher-than-training-data#:~:text=Decision%20Trees%20%2F%20Random%20Forrest%20cannot,outside%20of%20the%20observed%20range. ). So again, all of this is a bit cryptic, and I am cautious to what the authors mean (Lines 405-412). The authors referenced here the paper by Takoutsing*

*and Heuvelink. Note the paragraph right above section 3.5 that also notes that RF cannot extrapolate beyond training data.*
*L411: What limitations?*
We reply to the two comments (L406 and L411). We now clarify the term "extrapolation" as follows: "This indicates that the RF model is being used here beyond the area from which the training data were taken. This is a situation of spatial extrapolations, where tree-based methods such as RF can fail completely; see a recent study highlighting the limitations by Takoutsing and Heuvelink, (2022)".

We also clarify line 396 in Sect. 3.5 as follows: "[…] resulting in an "optimal" prediction situation in which the RF model is used to predict cases that are relatively similar to those used for its training".

*Lines 438-443: rewrite to include the long line-in reference in the quotes.*
Both sentences are now grouped in a unique one as follows: "The SHAP results are expected to improve the framing of the prediction results together with the associated uncertainty as illustrated with the synthetic test case described in the introduction as follows: […]".

*Line 456: extrapolation mode: odd way of stating that RF is used to make spatial extrapolations. See also in Line 411.*
To avoid confusion, we now use the term "spatial extrapolation" as suggested by Referee 1.

Orleans,
August 12th, 2024
J. Rohmer[1] on behalf of the co-authors

[1] BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 – France