Reply to Referee #2 on our manuscript "Oblique geographic coordinates as covariates for digital soil mapping"

We thank the referee for the qualified and insightful comments on our manuscript. In the following, we will address the referee's comments and describe the changes that we have made to the manuscript because of the comments.

COMMENT

The manuscript "Oblique geographic coordinates as covariates for digital soil mapping" from Møller et al. presents a valuable contribution to integrate predictor information on spatial position into machine learning approaches for digital soil mapping. It, thereby, seeks to overcome the known problem of orthogonal artefacts sometimes introduced by the usage of xy-coordinates as covariates in recursive partitioning algorithms. While commonly applied covariates usually relate to site characteristics that approximate the soil forming factors, the inclusion of coordinates provides a chance to reflect further spatial patterns we are not necessarily aware of. The authors show that the usage of a multitude of oblique spatial coordinates reflects spatial anisotropy. Major spatial axes identified through predictor importance measures may then give a hint on the geographic direction of the underlying processes as the authors demonstrate. The article compares the new approach (OGC) to existing approaches such as Euclidean distance fields (EDF) and spatial Random Forest (RFsp). The article is written using adequate language and it follows a clear structure. The figures are well prepared. Furthermore, it is a rare, but highly welcome choice of the authors to provide the R code of their approach. While the authors very clearly demonstrate the power of their approach particularly due to the clear figures and the comparison to similar approaches, certain aspects would require reconsideration:

REPLY

We thank the referee for the support for our manuscript. Furthermore, we are happy that the referee appreciates our choice to share the code for our study. We will consider the issues that the referee raises in the following.

COMMENT

- I do not understand why the OGC+AUX approach is not directly compared to regression kriging, but to ordinary kriging. Ordinary kriging would require a stationary mean which is not given in this particular research setting. Accordingly, a regression model would first have to be fitted to model the trend from covariate data, while then spatial autocorrelation in the residuals will be accounted for by ordinary kriging of the residuals. While the regression model is fitted by random forest, this would also allow for direct comparability. The authors provide rather vague arguments against regression kriging (lines 27-32).

REPLY

Our study focuses on one-step methods, as one of the goals in developing OGC is to create a feasible one-step method. Two-step approaches such as regression-kriging require researchers to interpret two models at once, which can confound analyses of uncertainty and the processes that govern the spatial distribution of soil properties. We believe that this is a relevant consideration, but it is not our main reason for omitting regression-kriging. Our first reason for this choice is that a previous study carried out in the same area showed that kriging predicted SOM more accurately than regression-kriging using both Cubist and Random Forest models (Pouladi et al., 2019). When a relatively simple method outperforms complex approaches, we believe that it is right to consider the complex approaches as redundant. Without this previous finding, we believe that it would have been relevant to include regression-kriging in the comparison.

CHANGES

We see that the manuscript does not clearly state our reasons for omitting regression-kriging. We will therefore add the following paragraph to section 2.3:

"A previous study in the same area showed that kriging predicted SOM more accurately than regression-kriging (Pouladi et al., 2019). We therefore omitted regression-kriging from the analysis, although, without this previous finding, it would have been relevant to include it."

COMMENT

The data in this study display spatial autocorrelation. Specifically, a range of 139 m is mentioned. This is not surprising due to the high spatial data density. Furthermore, the authors mention a couple of processes that may have caused this spatial dependency. However, this aspect is not accounted for in the evaluation approach. 100 random splits 75/25 (training/ test set) make it very likely that spatially autocorrelated sampling points will end up in the test and training set for the majority of the 100 splits. As a consequence, the test sets are not independent of the training sets and will lead to overly optimistic error values. This aspect at least needs to be mentioned. Particularly, in the context of Figure 7. The argumentation line of the introduction requires some improvement.

REPLY

Our main priority in the study is to compare the accuracies of several methods, not to assess their accuracies in absolute terms. Furthermore, we do not consider the issue of spatial autocorrelation to be as grievous as to warrant attention in the manuscript. Firstly, geostatistical approaches such as kriging would be useless if there was no spatial autocorrelation in the data. Secondly, the sample distribution in the field is very even, and as a result, only very few areas in the field are more than 20 meters from the nearest sample, and all areas are within the range of spatial autocorrelation. Therefore, having training and test samples within the range of spatial autocorrelation actually represents the general conditions in the field quite well. We therefore do not believe that our accuracy metrics are very much overly optimistic. If we were to extrapolate our results to a larger area, spatial autocorrelation would be an issue to consider, but this is not the goal of our study.

COMMENT

Certain aspects need to be better clarified:

- The main advantage of OGC+AUX over using only XY+AUX is the high number of coordinates, as the usage of only two oblique coordinates would lead to similar artefacts as demonstrated in the results. - The usage of coordinates as predictors in a regression model differs from fitting a geostatistical model to the residuals of a regression model. The approach closest to fitting a semivariogram is RFsp, as it accounts for the distance between points. However, it comes at the cost of introducing a high number of covariates as the authors state, correctly. It is important that the authors also compare their approach to RFsp, but the difference in calculating a different set of coordinates and taking the distance between points into account should be explicitly mentioned. In contrast, OGC+AUX and EDF+AUX really follow a similar approach in calculating a set of different coordinates. OGC+AUX is demonstrated to be superior to EDF+AUX. - Overall, whether it is worse to make the effort of calculating a high number of oblique coordinates could only be decided while being compared to regression kriging.

REPLY

We agree with the referee, and we see the need for further clarification. We will add several statements to the final version of the manuscript for this purpose.

CHANGES

We will add the following statements:

L44: “The main advantage of this approach [RFsp] is that it incorporates distances between observations in a similar manner to geostatistical models”.

L301: “Of the previous approaches, OGC is most similar to EDF, as it used the x- and y-coordinates, and the distances to the corners of the study area resemble coordinates. On the other hand, RFsp is more similar to geostatistical models, as it relies on distances between observations. However, this similarity comes at the cost of calculating a large number of distance rasters.”

L319: “The method [OGC] eliminated the orthogonal artefacts that arise from use of x- and y-coordinates and also achieved higher accuracies than maps created with only two coordinate rasters.”

COMMENT

There are a couple of statements that are problematic. Please consider rephrasing:

- lines 19-21 "…decision tree algorithms…are immune to correlated and redundant covariates". There are a couple of publications that show the contrary.

REPLY

Our experience has shown that decision trees are less vulnerable to correlated and redundant covariates than other model types, such as artificial neural networks. However, we admit that this does not constitute a full immunity.

CHANGES

We see that our statement is not correct, and we will therefore remove it from the final version of the manuscript.

COMMENT

- line 29 "By kriging the residuals…soil mappers have been able to reduce or remove spatial bias". We usually fit a geostatistical model to explain spatial autocorrelation not to remove spatial bias. Please also correct throughout the manuscript, e.g. lines 45/46.

REPLY

We agree with the referee that our phrasing is incorrect, and we will therefore change it (see below). However, the phrasing in lines 45 – 46 is in line with the study to which we refer. We quote the authors: "Further analysis shows that in both cases there is no remaining spatial autocorrelation in the residuals […]. Hence, both methods have fully accounted for the spatial structure in the data" (Hengl et al., 2018). The authors of this study refer to a figure, which shows a pure nugget variogram for the residuals of their model.

CHANGES

We will rephrase the sentence in question:

"By kriging the residuals of the predictive model and adding the kriged residuals to the prediction surface, soil mappers have been able to explain spatial autocorrelation and achieve higher accuracies."

COMMENT

- line 47 "...methods are able to integrate spatial relationships…" I am not convinced that by the mere consideration of coordinates we account for spatial relationships, leave alone spatial autocorrelation. Please explain or rephrase.

REPLY

We agree that our use of the term "spatial relationships" is inaccurate.

CHANGES

We will replace the term “spatial relationships” with the term “spatial trends” throughout the manuscript.

COMMENT

- lines 51-56 “Another shortcoming relating to EDF and RFsp is that…” As EDF and RFsp did not intend to keep the number of coordinate covariates variable I would suggest “reduced flexibility” instead of “shortcoming”.

REPLY

We agree with the referee, and we will rephrase as requested.

CHANGES

We will rephrase L51:

“EDF and RFsp also have limited flexibility as both methods specify the number of geographic data layers a priori.”

COMMENT

- line 65-66. “…it should be possible to optimise it” Please be specific: is it possible or not? Does it make sense to optimise it? Why did the authors then merely test all numbers of coordinate covariates?

REPLY

We see that the sentence is not very clear. We will therefore rephrase it.

CHANGES

We will rephrase lines 65 – 66:

“Furthermore, the number of oblique angles is adjustable, and soil mappers can therefore choose a number that suits their purpose. Some mapping tasks may require a higher number of oblique angles than others, and soil mappers can therefore increase the number as necessary. Alternatively, if a small number of oblique angles suffices, soil mappers can reduce their number and thereby shorten computation times.”

COMMENT

Further comments:

- Please delete equations (1) – (3). This is simple trigonometry.

REPLY

We agree with the referee.

CHANGES

We will delete equations 1 – 3.

COMMENT

Please also consider adapting the symbology: b2 is the knew oblique coordinate that replaces b1 (=x) and a1 (=y) not only b1 as somehow suggested by naming it b2.

REPLY

Our reason for naming b_2 is that it forms one of the sides of the right triangle a_2b_2c. We will therefore not rename it, as it would obscure interpretation of Figure 2. However, we see that the equations and the figure do not sufficiently stress the fact that the length of b_2 is equal to the new oblique coordinate.

CHANGES

We will add "OGC" to equation 4, to stress that OGC is equal to the length of b_2:

$$OGC = b_2 = \sqrt{{a_1}^2 + {b_1}^2} * \cos\left(\theta - \tan^{-1}\frac{a_1}{b_1}\right)$$

COMMENT

- lines 135-136. Please add the tested mtry values

REPLY

In each model, we tested five *mtry* values at even intervals between 2 and *NC*, where *NC* is the total number of covariates (counting both auxiliary data and spatially explicit covariates). The tested *mtry* values therefore depended on the method, and the number of covariates differed between methods.

CHANGES

We will add this explanation to the paragraph, starting at line 137:

"We tested *mtry* values at even intervals between 2 and the total number of covariates, including auxiliary data and spatially explicit covariates. The tested *mtry* values therefore varied depending on the number of covariates."

COMMENT

- line 136. Please explain how extratrees allows for suboptimal splits

REPLY

We will rephrase the sentence to better clarify how *extratrees* works.

CHANGES

We will rephrase the sentence as follows:

"The *extratrees* splitting rule generates random splits, as opposed to the *variance* splitting rule, which chooses optimal splits. Per default, *extratrees* generates one random split for each covariate and then chooses the random split that gives the largest variance reduction (Geurts et al., 2006). It therefore leads to a greater degree of randomization."

COMMENT

- Does the approach work on any type of coordinate system? I suppose coordinates have to be projected?

REPLY

This is a very interesting question, which we have given some though, although we have not included these thoughts in the first version of the manuscript. In the study, we use UTM coordinates, which have the advantage that the x- and y-coordinates have the same unit. Furthermore, it is reasonable to treat relatively small study areas as two-dimensional planes. In practical terms, OGC may also work reasonably well for larger areas with other coordinate systems, such as latitude/longitude systems. However, interpretation would not be as straightforward as in this study.

Using OGC at a global extent would probably require changes to the method. Because longitude is circular, points located on different sides of 180° L would have drastically different coordinates, even if the actual distances between them were short. One solution to this problem could be to replace the present version of OGC with latitudes rotated at various angles around a pair of equatorial axes. However, the implementation and testing of such an approach is far outside the scope of this study.

Due to the interest of this question, we will shortly address it in the conclusions section of the revised manuscript.

CHANGES

We will add the following statement to the conclusions section:

"One should note that we carried out this study for a small area using UTM coordinates as input. Using OGC for larger areas and other coordinate systems may require alterations to the method."

References

Geurts, P., Ernst, D. and Wehenkel, L. Extremely randomized trees. Mach. Learn. 63(1), 3-42. http://dx.doi.org/10.1007/s10994-006-6226-1, 2006.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B. and Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518. http://dx.doi.org/10.7717/peerj.5518, 2018.

Pouladi, N., Møller, A.B., Tabatabai, S. and Greve, M.H. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85-92. http://dx.doi.org/10.1016/j.geoderma.2019.02.019, 2019.