SOIL Discuss.,
https://doi.org/10.5194/soil-2019-83-RC2, 2020

Interactive comment on "Oblique geographic coordinates as covariates for digital soil mapping" *by* Anders Bjørn Møller et al.

Anonymous Referee #2

Received and published: 31 January 2020

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:

- 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). - 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. Certain aspects need to be better clarified:

- The main advantage of OGC+AUX over using only XY+AUX is the high number of co-

ordinates, 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.

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. - 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. - 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. - 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". - 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?

Further comments:

- Please delete equations (1) – (3). This is simple trigonometry. 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. - lines 135-136. Please add the tested mtry values - line 136. Please explain how extratrees allows for sub-optimal splits - Does the approach work on any type of coordinate system? I suppose coordinates have to be projected?

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2019-83, 2019.