SOIL Discuss.,
https://doi.org/10.5194/soil-2019-83-SC1, 2019

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

Alexandre Wadoux

alexandre.wadoux@sydney.edu.au

Received and published: 19 November 2019

This study tries to account for residuals spatial autocorrelation of a machine learning model by adding a set of pseudo-covariates. I have a few comments on the paper. I hope the authors find them useful and that it helps them to improve their manuscript. Overall, the study would benefit from a test of the method on several case studies, using different scales, different calibration sampling designs. A single case study at local scale and predicting a single soil property is in my opinion not enough to draw general conclusions.

About the methodology: 1) Any set of covariates with spatial pattern added to the original set of covariates may result in higher accuracy with a ML algorithm. This

is because ML algorithms can find relevant patterns even when the covariates are meaningless and not related to any soil forming process. The increase of accuracy that the authors obtain with the RF OGC + AUX model may well be obtained by adding any set of covariates with a spatial structure (see Fourcade et al., 2018).

2) Spatial autocorrelation in the raw data is not a problem per se and one should rather focus on remaining spatial autocorrelation on the residuals. I am strongly in favor of using only pedologically relevant covariates in a RF model. If the residuals of a model built using pedologically relevant covariates present autocorrelation, then one should consider making a map of the residuals because he may see a clear pattern of why this happens. The authors might then see that they are missing an important spatial process not included in the analysis. In this case one can add additional pedologically relevant covariates that could explain this pattern, and refit the model.

3) In case one made the previous step and admits that there is unexplained residual variation, one could consider using additional pseudo-covariates because there is no better proxy to explain the soil spatial variation. I stress here that these pseudo-covariates should not correlate with the pedological covariates because there would be redundancy (see next comment). In this case the pseudo-covariates should be covariates computed based on the remaining residuals. This would effectively tackle the problem of the residual autocorrelation and the authors would ensure that the pseudo-covariates do not interfere with the pedologically relevant covariates.

4) In this study, the authors include the set of pseudo-covariates with the set of pedologically relevant covariates. This is in my opinion very harmful because they can have pseudo-covariates which integrate over several of the pedologically relevant covariates, making them in some cases even better predictors. This is unrealistic and undesirable. This also makes the model less interpretable in terms of variable importance.

5) It is concluded that adding a set of pseudo-covariates effectively accounts for spatial autocorrelation in the data. This is clearly not the case as shown in Fig. 9 and admitted

by the authors at line 315 'the models built exclusively on spatial relationships had the most autocorrelated residuals.' The reason for this is that the covariates have a spatial pattern but are not related to the raw data and either to the residuals of the prediction made by a RF model. When the authors compared the sample variograms of kriging and RF residuals, it is visible that kriging do much better. The method would work if the sample variogram of RF OGC would be close to that of kriging. We can also see in Fig. 9 that the model with OGC covariates only have strong residual autocorrelation. The reduction in terms of residual autocorrelation in the OGC + AUX model is obtained by adding the pedologically relevant covariates. This is also a contradiction with the conclusion that OGC covariates account for the spatial autocorrelation.

6) Fig. 9 shows that there is still autocorrelation in the residuals of the RF model. This violates the assumption made in RF modelling, i.e. independence between the data points. Since this assumption is not satisfied, the calibrated RF model is potentially flawed. The authors have potentially missed important soil processes which could be added to the model as covariates. I would be interested to see a measure of the bias in the prediction.

Other considerations: Nugget to sill ratio should not be used to compare sample variograms, see Section 3.3. in https://doi.org/10.1016/j.catena.2013.09.006

Very surprised to read at Line 297 that the advantage of ML algorithms is their interpretability. I think the authors refer to the variable importance of the RF algorithm for the interpretability of the ML models. There is in my opinion a misunderstanding of the difference between ML and geo-statistical methods such as kriging. In ML you do not do inference and so you should not directly interpret the fitted model, or at least with caution. In geostatistics you can interpret because you make inference on the process that generated the data.

ML are also mostly black boxes. For example, is it impossible to interpret all the trees in a RF model, or all the neurons in a neural network model. This is in consequence

not justified to claim that ML algorithms have the advantage to be interpretable.

L 305: I would disagree with this conclusion; this would need to be justified by the literature or comparison between different case studies.

L 313: It is quite high accuracy a minimum CCC = 0.83.

L. 315. The authors have contradictory statements in the last paragraph of the Discussion. The last sentence is not very clear. Dealing with spatial data, which are auto correlated, a spatial methods is always needed otherwise you miss an important process and the fitted model is probably flawed because of the i.i.d assumption of the errors.

How did the authors compute the R2? A R2 can either indicate the closeness of the predicted values to the fitted regression line or the proportion of variance explained by the predictors. Authors should check that the R-square was computed against the 1:1 line and not against the fitted linear regression between observed and predicted, see https://doi.org/10.5194/soil-4-1-2018, Section 3.8 where the authors called it a skill score.

Impact of the sampling design is not considered. A spatial coverage design is very poor for random forest, while it is very efficient for kriging (assuming the variogram parameters are known). You should also consider that the sampling designs affect greatly the way the sample variograms are computed.

How did the authors compute the sample variograms? The authors gave no information about it.

It seems that the sample variogram for ordinary kriging is not at the same scale. It is either a much better model or the authors did not back-transformed the log-transformed observations. The authors mentioned that they log-transformed the observations prior to variogram fitting, it is not clear whether they also did it for the RF model.