Reply to comments by Dr. Alexandre Wadoux on our manuscript “Oblique geographic coordinates as covariates for digital soil mapping”

Anders Bjørn Møller[1], Amélie Marie Beucher[1], Nastaran Pouladi[1], Mogens Humlekrog Greve[1]

[1]Department of Agroecology, Aarhus University, Tjele, 8830, Denmark

Correspondence to: Anders Bjørn Møller (anbm@agro.au.dk)

General reply

We thank Dr. Alexandre Wadoux for the insightful comments on our manuscript “Oblique geographic coordinates as covariates for digital soil mapping” (Møller et al., 2019, Wadoux, 2019). We have found the comments very helpful in improving the manuscript, and we would like to give our replies to the comments.

We will start with a general reply to the commenter’s use of “pseudocovariates” as a label for oblique geographic coordinates. We see this label as misplaced. We believe the term “pseudocovariates” is only appropriate for covariates, which are clearly unsuited for the purpose, and this is not the case for oblique geographic coordinates.

Notable examples of pseudocovariates in the statistical literature have included randomly generated covariates for testing variable selection (Wu et al., 2007, Sandri and Zuccolotto, 2008, Sandri and Zuccolotto, 2009, Ghosal et al., 2019). In the mapping literature, recent studies have used pictures projected in geographic space as cautionary tales (Fourcade et al., 2018, Wadoux et al., 2019). The commenter correctly asserts that pseudocovariates with a spatial pattern can predict properties in geographic space with moderate success. However, we do not believe this to mean that researchers should disregard covariates that explicitly account for spatial position.

In fact, the digital soil mapping literature has a rich number of studies, which have included spatial position as a covariate. The *scorpan* approach to digital soil mapping presented by McBratney et al. (2003) explicitly includes spatial position as a component. Although most studies in the review include spatial position through kriging or regression-kriging, the authors are open to the use of covariates to account for spatial position. We quote:

“As was discussed in Section 2, soil can be predicted from spatial coordinates alone. […] This may indeed reflect some other environmental variable such as climate, and because of this it can be argued that n is not really a factor, but simply putting the coordinates is a simple way to ensure that spatial trends not included in the other environmental variables are not missed. Therefore, n could also be described by some linear or nonlinear (nonaffine) transformation of the original spatial coordinates," (McBratney et al., 2003).

Oblique geographic coordinates represent such a transformation of the spatial coordinates. As one may expect from the previous reference, several studies have included x- and y-coordinates as covariates (Poggio and Gimona, 2014, Nussbaum et al., 2018, Koch et al., 2019, Lagacherie et al., 2019). Other studies have included spatial position in the form of distance-based covariates, for example using distances to the coastline (Holmes et al., 2015) or rivers (Rudiyanto et al., 2018).

Recently, studies have included additional distance-based covariates, including distances to the corners and middle of the study area (Behrens et al., 2018b), and distances to observations (Hengl et al., 2018). We hope therefore to have demonstrated that the use of covariates to account for spatial position is a theoretically sound, well-established practice, which does not warrant the label "pseudocovariates". Using covariates to include spatial position in machine learning models is in itself not new. Oblique geographic coordinates are simply a new method for doing this, with some advantages over previous methods.

In addition to this general reply, we would like to address the specific comments in the following.

Specific replies

We structure our replies by first showing the comment in question, then our reply to the comment.

COMMENT

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.

REPLY

In our manuscript, we mainly aim to introduce oblique geographic coordinates as a concept and to demonstrate the method on a dataset. We find that in this case it yields good accuracies and meaningful results. We agree that it would be advantageous to test the method on several datasets in order to conclude more generally. However, it would also dilute the focus of the manuscript, as we would not be able to report the results in as much detail as we do.

We will add that it is usual to use only a single dataset for introducing a new method, as several studies have used this approach (Grimm et al., 2008, Odgers et al., 2014, Padarian et al., 2019). We admit that there are notable exceptions (Behrens et al., 2018a, Behrens et al., 2018b, Hengl et al., 2018), but we still assert that our chosen approach is not problematic.

We agree that it is important to test oblique geographic coordinates on additional datasets, and we plan to do this in the future. This is also part of our rationale to share our code, as this will allow other researchers to test the method on their own datasets. This, we hope, will allow a more thorough assessment of the capabilities of the method.

COMMENT

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).

REPLY

We concur that it would probably be possible to obtain the accuracies obtained with OGC + AUX with other (but not just any) sets of covariates. For example, RFsp + AUX achieves similar accuracies, although with a larger number of covariates. However, we will also remind the commenter that OGC do not simply have spatial structure – they have only spatial structure and nothing more. As we have already stated in our general reply, using covariates to account for spatial position is a well-established practice. OGC account for spatial position in a clear and systematic way, which is useful for decision tree algorithms and easily yields to interpretation.

COMMENT

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.

REPLY

We agree that it is important to use pedologically relevant covariates in machine learning models when mapping soil properties. We do not intend OGC to be used on their own, but in combination with auxiliary data of this form. As we hope to have demonstrated in our general reply, several studies have used spatially explicit covariates in combination with the other six components of the scorpan concept for digital soil mapping. Other studies have accounted for spatial autocorrelation in the residuals by means of regression-kriging, another well-established practice.

The commenter’s dedication to purely pedologically relevant covariates has merit. However, due to the complexity of soil-forming processes, the hunt for a set of covariates that perfectly explain spatial variation in soil properties, is in many cases likely to be fruitless.

COMMENT

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 pseudocovariates 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 pseudocovariates do not interfere with the pedologically relevant covariates.

REPLY

Redundancy is generally not a risk for decision tree models, as they simply choose the optimal covariate in each split (Breiman, 2001). See also our reply to the next comment.

Furthermore, we doubt if the approach, which the commenter suggests, would be useful. We are not sure how the commenter would create a covariate based on the residuals. However, the attempt would create a serious risk of circular logic, which could invalidate model fitting and the assessment of model accuracy. Models should be based on covariates, not vice versa.

COMMENT

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.

REPLY

Firstly, we refer to our general reply. Secondly, we will state that we see the commenter’s allegation of “harmfulness” as a misunderstanding. We see the integration of spatial and environmental covariates as one of the strengths of using oblique geographic coordinates. Firstly, it allows the machine learnings model to map complex processes characterized by spatial dependence as well as environmental effects (Behrens et al., 2018b). This has an advantage over regression-kriging, where separate, mostly incomparable models treat environmental and spatial effects.

The commenter fears a scenario where a coordinate raster gains a higher importance than environmental covariates in a model. If this were the case, it would indeed be a cause of worry, but not for the reasons stated by the commenter. If a coordinate raster gains a higher importance than an environmental covariate, it suggests that the pedological process represented by the environmental covariate is probably not highly relevant for the soil property in this specific area. Therefore, if all environmental covariates turn out to be less important than coordinate rasters, it would show that the environmental covariates did not adequately account for spatial variation in the soil property.

In our case, the most important coordinate raster was the 12th most important covariate. OGC only became the second most important covariate, when we summed their importance. This shows that spatial effects have a large influence on SOM in the study area. However, it also shows that the model did not discard environmental covariates when we included OGC. Instead, it successfully integrated the two sets of covariates and their combined effects.

COMMENT

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.

REPLY

We never claim in the manuscript that oblique geographic coordinates fully account for spatial autocorrelation in the data. This comment would be more helpful if the commenter provided the lines where we allegedly state this.

We once refer to Hengl et al. (2018), who found that RFsp fully accounted for spatial autocorrelation in the data, but it is quite clear from the sentence that we refer to results in another study, not our own results. Our own results contrast with this earlier finding, and we will include a comment on this in the final paper.

Furthermore, the commenter appears to reverse the interpretation of Figure 9. We intend soil mappers to use OGC as an addition to environmental covariates, not on their own. The figure shows that the addition of OGC greatly reduces spatial autocorrelation in the residuals relative to the model relying only on environmental covariates. We mainly include OGC, EDF and RFsp on their own to demonstrate more clearly the effects of these sets of covariates. We do not recommend that researchers use them on their own.

COMMENT

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.

REPLY

We believe that it is quite an overstatement to say that any Random Forest model with spatially autocorrelated residuals is potentially “flawed”. Such a conclusion would most likely invalidate a very large portion of Random Forest models used in digital soil mapping. However, we agree that it is not an optimal situation, and that it might be useful to add more environmental covariates.

As per the commenter’s request, we have calculated bias as mean error (ME) for each method. We have based this calculation on residuals from models using all observations:

Method	**ME**
Kriging	-0.011
AUX	0.040
EDF	0.041
EDF + AUX	0.042
RFsp	0.011
RFsp + AUX	0.028
OGC	0.029
OGC + AUX	0.036

The values show that kriging has lower bias than the other methods except RFsp, but all methods have low bias.

COMMENT

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

REPLY

We see the error. In the final paper, we will remove mentions of the nugget-to-sill ratio when comparing the variograms.

COMMENT

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.

REPLY

This comment is confusing. The commenter appears to assert that (1) machine learning models are not interpretable, but that, on the other hand, (2) geostatistical models are interpretable. The commenter seems to conflate interpretation and inference, but we believe that one should understand these two as separate terms.

Furthermore, it gives the impression of a contradiction when the commenter states that machine learning models should not include spatial relationships, but also states that geostatistical models are interpretable. Likewise, the statement that machine learning models are not interpretable contrasts with the commenter’s insistence that they should only contain covariates that represent pedological processes. If spatial position matters, even to the point where a geostatistical model is exclusively interpretable, why should we not use it in a model? Moreover, if we cannot interpret a machine learning model, then why does it matter what sort of covariates we use?

In themselves, geostatistical models only inform us on the spatial structure of the data. We agree that this can be useful, but any sort of interpretation would rely almost exclusively on the user’s knowledge of the target variable and the processes that affect it. On the other hand, machine learning models are potentially far more informative.

Researchers should exert caution when interpreting any form of statistical model, but we agree with the commenter that it is especially relevant for machine learning models. Machine learning models are more complex than geostatistical models, and their interpretation is therefore also more complex and requires a higher level of abstraction. Tools to interpret machine learning models include covariate importance, which we use, but other tools exist, for example partial dependency plots (Friedman, 2001). Irrespective of the tools that researchers use, it is important that they critically use their knowledge of soils and the study area as well as the machine learning algorithm.

We can see that our statement that geostatistical models and machine learning models differ in interpretability is misleading. In the final paper, we will change the phrasing to state that the difference lies in the information content provided by the models.

COMMENT

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

REPLY

We cannot see why the commenter would outright disagree with this conclusion, as the commenter also states that spatial coverage sampling favors kriging. However, we do see the need for justification from the literature. Several studies have shown that machine learning models using environmental covariates are more accurate than geostatistical models for large, less densely sampled areas, including Zhang et al. (2008), Greve et al. (2010) and Keskin et al. (2019). We will include these references in the final paper.

COMMENT

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

REPLY

We agree. In the final paper we will rephrase this sentence: “as EDF, RFsp and OGC all yielded lower accuracies without auxiliary data”.

COMMENT

L. 315. The authors have contradictory statements in the last paragraph of the Discussion.

REPLY

We do not see the contradiction, but we agree that the sentences are not quite clear enough. In the final paper, we will rephrase the last two sentences: “The results suggest that these methods should be used in combination with auxiliary data, but not on their own. If no auxiliary data are available, kriging is a better option.”

COMMENT

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.

REPLY

We agree on the lack of clarity. Please see our reply to the previous comment.

COMMENT

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.

REPLY

We used Peason’s R2, this is, closeness to a fitted regression line. We see that we did not include this information in the manuscript, and we will make sure to include it in the final paper.

We will not change the way we calculate R2, as Pearson’s R2 indicates if the predictions have the same trend as the observations, which we believe is relevant in itself. We rely on several accuracy metrics, including also RMSE and CCC. CCC gives information on closeness to a 1:1 line, which the commenter requests. Furthermore, the skill score, to which the commenter refers, uses on the mean square error (MSE) of the predictions, and the variance in the dataset. It is very useful for comparing accuracies across different regression problems. However, for any single regression problem, as in our study, the variance in the dataset will be constant, and variation in the skill score will depend only on variation in MSE. As we already provide RMSE, this information would be redundant.

COMMENT

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.

REPLY

We agree that the sampling design favors kriging. In fact, we already state in the manuscript that an earlier study in the same area (Pouladi et al., 2019) found that kriging yielded higher accuracies than machine learning models. It is therefore quite remarkable that OGC + AUX and RFsp + AUX allow Random Forest models to achieve accuracies on par with kriging.

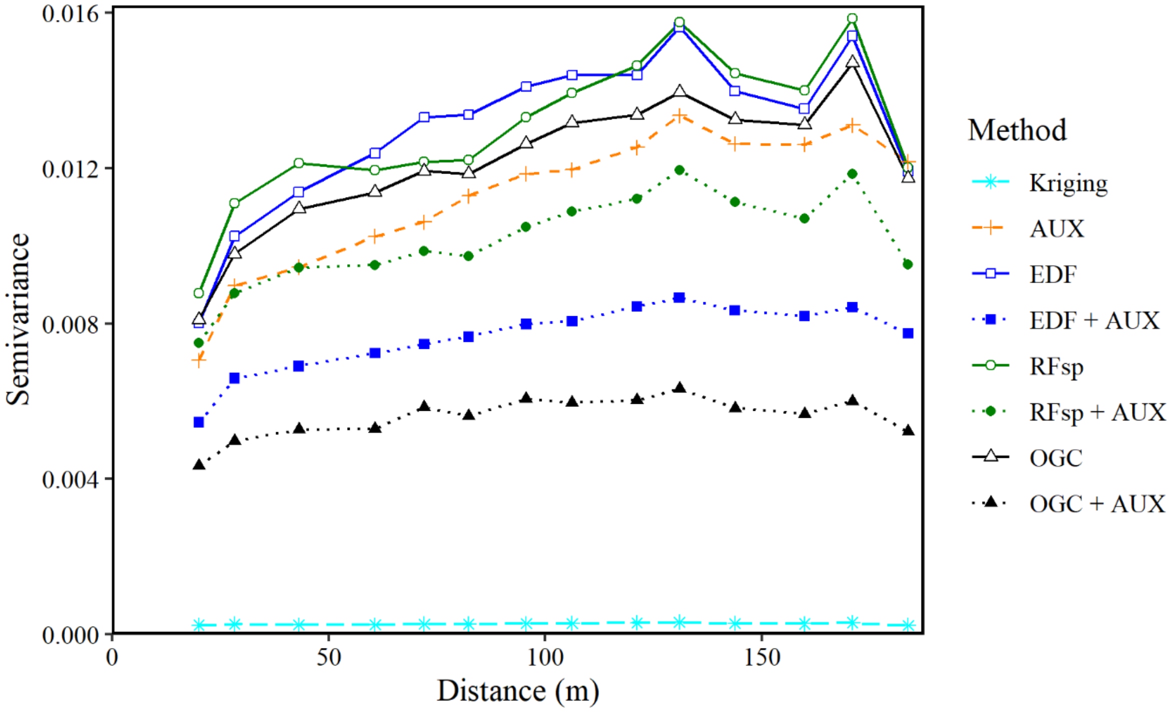
COMMENT

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

REPLY

Firstly, we produced maps with each method using all observations. Secondly, we converted both observations and predictions to natural logarithmic scale. We then subtracted the predictions from the observations and calculated variograms for these residuals. For this purpose, we used the function ‘variogram’ from the R package ‘gstat’ with its default parameters. We will include this information in the final paper.

Furthermore, we have discovered an error in our code, which caused us to use only 75% of the observations when calculating the variograms. We have therefore recalculated the variograms using all observations and produced a new version of Figure 9. We will include this updated figure in the final paper:



COMMENT

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.

REPLY

The variograms are all on the same scale. Kriging has smaller residuals than the other methods, as the variogram had a very small nugget, but we do not believe that this shows it to be a “better” model. For example, with inverse distance weighting interpolation, the residuals would be zero, but it would not necessarily by a very good model. We will also point out to the commenter that the residuals for OGC + AUX show nearly no trend. So the residuals are larger, but they have very little spatial autocorrelation.

References

Behrens, T., Schmidt, K., MacMillan, R.A. and Rossel, R.A.V. Multi-scale digital soil mapping with deep learning. Sci. Rep. 8(1), 15244. http://dx.doi.org/10.1038/s41598-018-33516-6, 2018a.

Behrens, T., Schmidt, K., Viscarra Rossel, R., Gries, P., Scholten, T. and MacMillan, R. Spatial modelling with Euclidean distance fields and machine learning. Eur. J. Soil Sci. 69(5), 757-770. http://dx.doi.org/10.1111/ejss.12687, 2018b.

Breiman, L. Statistical modeling: The two cultures. Stat. Sci. 16(3), 199-215. http://dx.doi.org/10.1214/ss/1009213726, 2001.

Fourcade, Y., Besnard, A.G. and Secondi, J. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Glob. Ecol. Biogeogr. 27(2), 245-256. http://dx.doi.org/10.1111/geb.12684, 2018.

Friedman, J.H. Greedy function approximation: A gradient boosting machine. Ann. Stat. 29(5), 1189-1232. http://dx.doi.org/10.1214/aos/1013203451, 2001.

Ghosal, R., Maity, A., Clark, T. and Longo, S.B. Variable Selection in Functional Linear Concurrent Regression. arXiv preprint arXiv:1904.08507. 2019.

Greve, M.H., Greve, M.B., Kheir, R.B., Bøcher, P.K., Larsen, R. and McCloy, K. Comparing Decision Tree Modeling and Indicator Kriging for Mapping the Extent of Organic Soils in Denmark, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E. and Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer Netherlands, Dordrecht, 267-280, 2010.

Grimm, R., Behrens, T., Märker, M. and Elsenbeer, H. Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. Geoderma 146(1-2), 102-113. http://dx.doi.org/10.1016/j.geoderma.2008.05.008, 2008.

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.

Holmes, K.W., Griffin, E.A. and Odgers, N.P. Large-area spatial disaggregation of a mosaic of conventional soil maps: Evaluation over Western Australia. Soil Res. 53(8), 865. http://dx.doi.org/10.1071/sr14270, 2015.

Keskin, H., Grunwald, S. and Harris, W.G. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40-58. http://dx.doi.org/10.1016/j.geoderma.2018.12.037, 2019.

Koch, J., Berger, H., Henriksen, H.J. and Sonnenborg, T.O. Modelling of the shallow water table at high spatial resolution using random forests. Hydrol. Earth Syst. Sci. 23(11), 4603-4619. http://dx.doi.org/10.5194/hess-23-4603-2019, 2019.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M. and Saby, N.P.A. How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma 337, 1320-1328. http://dx.doi.org/10.1016/j.geoderma.2018.08.024, 2019.

McBratney, A.B., Mendonça Santos, M.L. and Minasny, B. On digital soil mapping. Geoderma 117(1-2), 3-52. http://dx.doi.org/10.1016/s0016-7061(03)00223-4, 2003.

Møller, A.B., Beucher, A.M., Pouladi, N. and Greve, M.H. Oblique geographic coordinates as covariates for digital soil mapping. SOIL Discuss. http://dx.doi.org/10.5194/soil-2019-83, 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. http://dx.doi.org/10.5194/soil-4-1-2018, 2018.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B. and Clifford, D. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214-215, 91-100. http://dx.doi.org/10.1016/j.geoderma.2013.09.024, 2014.

Padarian, J., Minasny, B. and McBratney, A.B. Using deep learning for digital soil mapping. Soil 5(1), 79-89. http://dx.doi.org/10.5194/soil-5-79-2019, 2019.

Poggio, L. and Gimona, A. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation - An example from Scotland. Geoderma 232-234, 284-299. http://dx.doi.org/10.1016/j.geoderma.2014.05.004, 2014.

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.

Rudiyanto, Minasny, B., Setiawan, B.I., Saptomo, S.K. and McBratney, A.B. Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma 313, 25-40. http://dx.doi.org/10.1016/j.geoderma.2017.10.018, 2018.

Sandri, M. and Zuccolotto, P. A bias correction algorithm for the Gini variable importance measure in classification trees. J. Comput. Graph. Stat. 17(3), 611-628. http://dx.doi.org/10.1198/106186008x344522, 2008.

Sandri, M. and Zuccolotto, P. Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms. Stat. Comput. 20(4), 393-407. http://dx.doi.org/10.1007/s11222-009-9132-0, 2009.

Wadoux, A. Interactive comment on “Oblique geographic coordinates as covariates for digital soil mapping” by Anders Bjørn Møller et al. SOIL Discuss. http://dx.doi.org/10.5194/soil-2019-83-SC1, 2019.

Wadoux, A.M.J.C., Samuel-Rosa, A., Poggio, L. and Mulder, V.L. A note on knowledge discovery and machine learning in digital soil mapping. Eur. J. Soil Sci. http://dx.doi.org/10.1111/ejss.12909, 2019.

Wu, Y.J., Boos, D.D. and Stefanski, L.A. Controlling variable selection by the addition of pseudovariables. J. Am. Stat. Assoc. 102(477), 235-243. http://dx.doi.org/10.1198/016214506000000843, 2007.

Zhang, X., Lin, F., Jiang, Y., Wang, K. and Wong, M.T. Assessing soil Cu content and anthropogenic influences using decision tree analysis. Environ. Pollut. 156(3), 1260-1267. http://dx.doi.org/10.1016/j.envpol.2008.03.009, 2008.