SOIL Discuss., referee comment RC1
https://doi.org/10.5194/soil-2020-102-RC1, 2021

Comment on soil-2020-102

Anonymous Referee #1

Referee comment on "On the benefits of clustering approaches in digital soil mapping: an application example concerning soil texture regionalization" by István Dunkl and Mareike Ließ, SOIL Discuss., https://doi.org/10.5194/soil-2020-102-RC1, 2021

This manuscript is about using clustering in a number of applications for digital soil mapping. The first application is feature selection of the mapping model. The second application is for dealing with imbalance training data. The third application is to build folds, to be used in model tuning and model prediction validation (cross-validation).

Overall I found that in each of the cases, the rationale for clustering was very unclear. I could not find real justification for these approaches. Besides, the description of what has exactly been done is missing. I read some parts several times, but I stopped at the end of the Methods section because I think this study in fundamentally flawed in its objectives.

- Clustering for the feature selection: it is not clear what the authors did, but from their explanation I dot not see the rationale for doing this. The description made in Section 2.2.3 makes no sense to me. None of these tests are needed. Why clustering the soil texture data? How is feature selection made, for each cluster? The authors claim that feature selection is not needed (Introduction), but then make cluster for feature selection. How has clustering anything to do with feature selection?
- Clustering for imbalance learning problem. Not sure why this is actually a problem. The introduction seems to mix several concepts, such as bias, representativity, predictive performance etc. All papers cited are about categorical mapping, where class imbalance is indeed a problem. I have never heard that imbalance data in a regression setting is a problem. How is imbalance defined in a regression perspective, where classes do not exist? Is a cluster a class? The authors in Section 2.2.4 discuss about strata, not clusters, the two are different.
- Clustering for building robust models. I would argue that the approach used by the authors is not correct. I have seen papers dealing with building geographical strata to define the CV folds, but the authors used the response variable for this. It does not make sense to me. The risk with building clusters on the training data is that large part of the feature space covered by the data within the cluster will be missing to the model when the model is fitted, so that there is a serious risk of extrapolation when predicting in that fold. Random k-fold is preferred for this, because each fold with contain values that are likely to be different.

In addition, the authors are constantly mixing the terms between clusters and strata. They are different terms with different meaning.

This manuscript is about clustering, but the clustering is poorly described and often not done in the correct way. For example, the correlation among variables is not accounted for. Another example is the number of clusters is determined from a sample of only 2,000 points. Even if it is repeated, this is simply not sufficient for determining the optimal number of clusters.

Other comments on Introduction and Methods.

Line 32: Do the authors mean linear regression? The other models cited after are also models for regression, (e.g. RF).

Line 37: citation to Blanco et al. (2018) does not really support the claim of this sentence, look at review papers.

Line 37: citation to Møller et al. (2019) does not really support the claim of this sentence, look at review papers.

Lines 39-40: this is a very wrong statement. RF is of course prone to overfitting, like all regression models. It has been well described in the literature. For a recent example, see Makungwe et al. (2021): https://doi.org/10.1016/j.geoderma.2021.115079, but there are many more. Further, how does overfitting relate to the presence of outliers or "parameters". What do you mean by parameters?

Lines 41-45: I do not understand this part. Feature selection do not reduce the noise cause by uninformative predictors. Feature selection would also not omit predictors of importance because these would damage the prediction accuracy, which machine learning models are concerned with.

Lines 51-52: Already mentioned at L. 48-49.

Lines 51-53: So in one sentence the problem is the bias and in the next one the problem is the prediction performance of the model. I think the problem is that with oversampling (clustered) data, we might be overoptimistic about the predictive performance of the model. It has nothing to do with getting biased estimates of the property.

Lines 53-55: very unclear what the link between the two sentences is.

Lines 58-59: to this point I still have no good justification to why the problem of having an imbalanced data is a problem for regression. All the cited papers at L. 58-59 are for classification!

Lines 61-63: What does it mean model robustness? The cross validation is only here to estimate error needed to estimate the validation statistics, but the "model robustness" is not evaluated. What does this sentence mean: "the outcome can still be compromised by an uneven distribution of classes between the data subsets", I do not think this paper is about classes?

Line 65: the authors should systematically differentiate between dependent and independent values. Do they do the clustering on the independent or independent variables? It makes a lot of difference, but both options are possible.

Lines 94-96: Removing a soil sample from the modelling because the sand content is too high is not a good reason. This is clearly not an outlier.

Lines 97-99: a clustering was done on what? On the predictors or the values of the soil properties? It should be systematically written.

Lines 130-131: this is not an accurate description of what random forest is. Also, RF is based on RT, but the authors do not mention the differences.

Line 134: rectangular region of the predictor space?

Line 135: I would argue this is not correct. Of course RT are sensitive to the training data, because it is a data-driven model. The text should really be made more precise.

Line 138: bootstrap sample of the data: training data or predictors?

Line 142: the authors should be consistent in the writing: feature, but at other places I see variable, covariate, predictor, predictor variable, predictor data.

Lines 159-160: what are these traits?

Lines 161-166: applying clustering on a mix of categorical and continuous variables is a very complex problem. More information on what has exactly been done is necessary. How are these categorical variables transformed? How can a PCA be made on categorical variables? Also, you should use systematic random sampling, not simple random sampling for collecting a subset. No need to include the minimum and maximum value, and no need to add one sample per class. If the sample is sufficiently large, all classes should be included.

Lines 160-170: Is correlation included in the clustering? The correlation should be accounted for in the clustering process. This can be done by rescaling the predictors by the inverse of Cholesky transformation of the variance-covariance matrix.

Line 174: so the number of cluster is actually decided based on 2,000, selected randomly out of the 33 millions of points?

Section 2.2.3. None of these normality tests are needed, what is the rationale for doing this.

Line 218: the response data?? Why not the predictor variables??