Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-584-RC1, 2019

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data" *by* Mo Zhang and Wenjiao Shi

Anonymous Referee #1

Received and published: 15 March 2019

General comments The article presents a comparison among five machine-learning methods (k-nearest neighbor, multilayer neural network, random forest, support vector machines and extreme gradient boosting), when applied to the spatial analysis of soil particle-size fractions collected in the Heihe River Basin (China), together with environmental covariates (topographic, remote sensing, climate, soil physicochemical attributes and categorical maps). The performances of the methods are tested when data are considered both on the original scale, and on a log-ratio basis. In the latter case, the authors consider three transformations widely used in compositional data

analysis, i.e., additive log-ratio, centered log-ratio, isometric log-ratio. The comparison among methods is quantitatively performed through a Monte-Carlo procedure (30 repetitions of subsampling) on the basis of objective performance indicators (AUC, AUPRC for classification, R2, RMSE, MAE, AD, STRESS for regression). In my opinion, the paper is overall clear and the methods used fairly detailed. The models and methods chosen are overall appropriate. However, it is not always explained how the methods were applied to the soil fraction vectors, i.e., whether they were applied jointly to the fractions or component-wise. This indeed makes a relevant difference, the former being definitely more meaningful than the latter. In general, I found interesting the thorough comparison of those machine learning methods that are nowadays widely used, and particularly the comparison of the two views of the Euclidean and the Compositional geometry. However, I have two main concerns – reported in the next section – on the approach used by the authors for the investigation, related with two points that, in my view, would be relevant for the topic of the paper but are not considered by the authors.

Specific comments I have two major concerns on the study, that regard two topics that are relevant in my view but not developed by the authors. Uncertainty. The authors do not address the key topic of uncertainty, neither in the results of classification/regression, nor in the performance indicators. In fact, it would be key to understand the degree of uncertainty associated with the results, and if the used method can indeed provide a clear indication of the variability of the estimates and not only the estimates themselves. In a Monte Carlo study, one should also verify (i) if the estimators' variability has a reasonable order of magnitude with respect to the values of the estimates and (ii) if it is representative of the actual error that one makes on an independent test set. In fact, the results provided by different methods and compared in the paper may be even indistinguishable if their variability is high. I do believe that point estimates are relevant, but their uncertainty does provide a meaningful information that in my view cannot left out of this kind of comparisons. In addition, the authors should indicate the standard deviation of the indicators of performance (e.g., those in Table 2), to appreciate the stability of the results across the repetitions with different sampling

points. Generality of the results. The work provides a very broad comparison among the (classification or prediction) results obtained with different methods. However, it is not clear to me how general these results indeed are and thus how usable they will be for other scientists, that more likely work in other context than the field study here considered. In fact, even if I see the value of the Monte Carlo study and the quantitative indices it provides, I'm less convinced on the evaluation of the methods in terms of classification and prediction power for regions were no data is available (section 3.2.2, 3.3.2) – i.e., where it is hard to say which result is actually better that the others. It would be much easier (and convincing) to evaluate the method performances on a large scale simulated case rather than on this field case, at least for what concerns the classification and prediction in areas far from the data – and this would also provide a more general indications to other scientists.

Technical corrections - If I understood correctly, the authors widely use the term "interpolation" to refer to the fit of the models. However, I'm not sure that all the models used are indeed interpolating the data

- P. 9 line 10: The fact that one variable would be omitted without loss of information does not provide a convincing explanation of why the Euclidean geometry is not appropriate to treat compositional data. I suggest to better explain the point. Further, the log-ratio approach provides a geometrical structure to the space of compositions, but it is not formally correct to say that the approach consist of the transformations alr, ilr, clr. Instead, it is more correct to say that the transformations ilr and clr can be used to operate within the log-ratio approach by simply using the Euclidean geometry on the transformed data. One should also note that for a number of method (among which averaging and regression) ilr and clr provide equivalent results, whereas alr may provide different results.

- P. 9 last line: it is not true that CLR is inapplicable – in fact, it is widely used in multivariate analyses.

- P. 10 line 2: It would be more appropriate to refer to Egozcue et al 2003.

- P. 10 line 15: the back-transformations are well known, the author should also refer to classical references appeared before their recent work.

- P. 12 line 15 and P. 16 line 15: if the ROC curve is not appropriate – as the author state – it should not be used for comparison.

- P. 14 line 5 to 10: since the data are multivariate, multivariate notions of median (e.g., based on depth measures) should be used. Component-wise medians and quantiles should be avoided. Similarly, indices computed on the single proportions have a limited meaning because of the constraint to 100% – joint indices should be used instead.

- P. 19 line 15: the methods on the original scale are designed in the Euclidean geometry, so there is no surprise in that they outperform methods developed to optimize other criteria (log-ratio geometry). The authors should better highlight the conceptual difference of working on the original scale or on the transformed scale.

- Table 2: I'm not sure it is meaningful to provide information on single part if the analysis – as I understood – was performed as to ensure that the total is 100%. I think it would be more meaningful and appropriate to display the overall RMSE, MAE and R2 (sum of the element-wise numbers) – in the same way as AD is just one value for the Aitchison distance between compositional vectors.

- P. 30: the authors discuss their results in comparison with previous ones. They should however discuss whether these differences are due to the particular case study considered or if they can be considered of more general validity.

- There are several typos and sentences to be revised in terms of English wording; I suggest a careful revision.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584, 2019.