Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-584-RC3, 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 #3

Received and published: 25 April 2019

A review of "Systematic comparison of five machine-learning methods. . ." by Mo Zhang and Wenjiao Shi. The manuscripts describes a comparison between five machine learning methods for soil classification and interpolation of soil particle size fractions. It explores different transformed data. There are a few major problems with the manuscript:

1. The five machine-learning methods come falling from the air. That is to be regretted, as there is much more information available in the literature. The manuscript now provides only some technical aspects, and for instance not prior assumptions, restrictions

in their use or their general comparability. The manuscript would largely benefit from a short mathematical introduction to the five techniques, from where it would become clear whether in fact comparable methods are compared, or that there is a comparison between apples and oranges.

2. Section 2.5 discusses log-ratio transformation methods.

a. For one, here the term 'methods' is used (which I like) and at other places the term 'approaches' (which should be avoided throughout).

b. But more importantly is the lack of mathematical rigor throughout. Line 13 gives a condition ïĂćj = 1,. . .,j-1,j+1,. . .,D. I do not at all understand this restriction. I think it is wrong.

c. In equation (4) it is unclear why on the left hand side there is a term yj mentioned: is that exceptional? I do not think so: it has to be removed.

d. In equation (9) there seems to be the D-ith root: is that correct? What is the rationale behind this?

e. Again in equation (9), the z is only defined for all except for the last component. Why is that the case?

In all, this section needs a very careful revision by a professional mathematician.

3. The results section is not at all convincing in its current status. I could understand that a selection is made for the best of machine-learning method x transformation combination for the particular study area. That requires some space, but it then should be followed up by only one outcome, namely the best. As a scientist I am not interested in maps of an inferior quality. Hence, in figure 6, four of the five maps are redundant. The authors could use small sections of the maps, though, to show where the other techniques are critically sub-optimal, but not more than that.

4. Still in the results section, I am not convinced about the repetition on page 20 of what

is already in the table. I want to know why one method and one transformation is the best, and in particular whether that is due to the specific case study or to an inherently superior combination of the transformation with the methods. What stands out is that RF_ORI is the bet. Fine, point taken. But then in figure 7, I am only interested in the RF_ORI map, and all the other maps should be avoided. Similar remarks apply to table 3 and figures 8 and 9. The authors should make more out of the data that they had to their disposal, than just creating sub-optimal maps! Figure 10, by the way, is interesting and may lead to a sub-optimal map (in terms of the RMSE). XGB is not so bad in terms of RMSE and much faster, hence if speed is an issue (when would that be?) then a researcher may opt for XGB. It should then be clear what s/he essentially loses in terms of representations on maps.

5. The discussion section has some interesting elements, but I could easily imagine an improvement when better focusing upon what has been achieved and how it should be interpreted, also in terms of the soils and the particle sizes. In particular, a transition of the methods to other areas should be discussed: how should we take it? Maybe do it hierarchically, i.e. first a quick and imprecise method, followed by a precise method?

6. Finally, the abstract should be improved: the opening statements are too wordy, as a single sentence justification for the study is enough. The term 'notable consequences' is too vague for an abstract. The final main concluding sentence '. . . helps to elucidate the processing and selection of compositional data in spatial simulation' is not justified from the manuscript. We only see that one method x transformation combination is doing best, another combination is fast at the expense of a loss in precision. That seems to be a good conclusion, and in that sense the study is valuable as an interesting case study on soil analysis in a rather large area.

Some minor comments:

1. There is little need to describe the vegetation in the study area, please remove. Also rainfall patterns are not so interesting when it comes to soil particle size fractions, but I

may be mistaken here.

2. The English needs a careful check: in general the manuscript is well readable, but some fine-tuning may further improve the accessibility.

3. What are 'close correlations' (p. 2, l. 10)?

4. Whence the sentence 'kriging methods (so-called geostatistics)' (p.3, l. 18)? Kriging is a geostatistical interpolation method, in fact it is a whole collection of those.

5. I object to the use of the term 'attribute' for a variable (p. 6). It is a GIS term and not a scientific term, a much better term is 'variable'. Further down in the paragraph we suddenly read about evapotranspiration data. The manuscript would benefit from a more careful separation between variable and data.

6. Notation is inconsistent for k-nearest neighbor (p. 6 ff.). It should either be capitalized throughout, or not at all.

7. On page 9 there is the requirement stated that the sum of the components is equal to 1. Fine, but how is that guaranteed? Is a correction being made if it is not the case at a prediction location?

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