Interactive comment on "Machine learning and soil sciences: A review aided by machine learning tools" *by* José Padarian et al.

José Padarian et al.

jose.padarian@sydney.edu.au

Received and published: 30 November 2019

We like the open discussion format because we can get extra interesting feedback. Thanks. Regarding your comments:

The term "advanced ML" is mentioned several times. What do you mean by this? Any criteria? And please offer a list of the included ML algorithms by advanced ML.

We discussed that point at the moment of writing, without reaching a conclusion, mainly because it is hard to draw a clear limit when a model becomes complex or advanced. We try to capture that discussion in the first paragraph of Page 3. We understand it is vague but we think it illustrates the gradient between simple and complex and that any

modelling practitioner can understand it.

It will be useful if the author can offer some details about the performance of ML and simpler approaches (at least the best and/or popular ones) with validation statistics such as R2, RMSE and ME, which are reported by most studies.

We tried to present the performance differences in a more quantitative way but we quickly realised that the variation between studies made the comparison very difficult and not very informative. That is why we opted for a natural language approach, looking for significant differences between methods, reported by the authors of the articles, looking for sentences such as "significant improvement", "significant improvement", etc. That is what is reported in the first paragraph of Section 3.4.

Authors reviewed both the best performance (though no details offered) ML and the most used ML. Could you compare these two? There may be a gap between them and need attention for researchers to choose their ML wisely. As discussed by the author, performance and interpretability may both affect the choice of ML or other methods. Suggestion and insights may be offered by reviewing the most used, performance and interpretability (even not well defined, you may still classify them such as low, medium, high, potential low...).

This is a very good point. In general, we prefer to limit our recommendations to simple ones, which can "resist the passage of time". Other articles have classified algorithms according to their complexity and interpretability (e.g. Brungard *et al.*, 2015) and we think that it is not necessary to do it again, especially because, at this point of time, there are many ML researchers focusing on this research topic and probably the landscape will change quickly.

In terms of the gap that you mention, we see a natural tendency to leaving behind the methods that do not perform that well. For instance, Partial Least Squares Regression (PLSR) is very popular and has been used since the 80-90s but, when used in the studies included in this review (mostly published post 2000s), very few studies use

PLSR as their main algorithm and it is generally used in comparative studies where it is outperformed by more advanced models. We will add this example to Section 3.4, not as a recommendations to stop using simpler models, of course, since they are a core component of science (for a good reason).

We find the topic of model selection very fascinating since also has an important human/social component. Most scientists/groups have their favourite models, there are traditional methods, and also fashionable models. Probably a topic for another research.

Consider using the cited number of a paper in addition to the number of paper only. When defining the most used paper, you may use the cited number as a weight to each published paper. In this way, we may see a different pattern from Figure 7.

We think that mixing usage with "popularity" is not ideal. Probably we would need to do some correction by "time since publication" and, at the moment, it is not very clear to us how to account for biases such as "publications by famous authors" or "publications in famous journals". We believe the simplest measure of usage of a method is usage by itself.

Some short names have no full names, e.g. PLSR, PCR, kNN. And there are so many short names in the paper. Consider make a list of them as appendix. If it only appears one or two times, do not use a short name.

Thanks for the suggestion. We will add an appendix with a list of acronyms.

———————————————

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2019-57, 2019.