SOIL Discuss.,
https://doi.org/10.5194/soil-2019-57-AC2, 2019

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: 28 November 2019

Thanks for your feedback. Regarding your comments:

I am afraid that the keyword "Machine Learning" is too restrictive for capturing the targeted papers. Due to the limited number of keywords that we are allowed to select, a lot of authors (including me) select as a keyword the precise name of the algorithm, e.g. random forest, neural network,..., rather than selecting "Machine learning"

We did not perform a search by keyword. As we mention in Page 3, Line 9, we performed a full-text search. Of course, not every article uses the "Machine Learning" term within the text, but that is the criteria that we set for this review.

... As an example, it seems that you missed the paper from Nussbaum et al (2018) that was published in this journal (Soil).

That paper was not included because we did not include any paper from SOIL. Table A1 has the list of all the journals that had matches for our criteria. We excluded various journal, for different reasons, including the ones mentioned in Page 3, Line 10 (our institution having access to full-text articles, and that they provide text-mining permission). SOIL was excluded because, to our knowledge, does not provide an API to query their database. We could have written custom code to download and process their publications, but the number of machine learning articles in SOIL is low so we decided not to do it. The results of the review would not be different from what we present in the current article.

The selection of the number of topics is a critical operation. By selecting 12 topics from figure 8, you privileged a local (slight) maximum whereas you could have selected 6 topics by considering the number of topics at which the coherence indicator reaches a plateau. It would have been interesting to check whether this more parsimonious choice could provide a clearer classification with more identifiable and less correlated topics or not.

We did explore other values for the number of topics but we do not present the results mainly because 12 topics, besides providing the highest coherence score, also presents fewer overlaps between topics (graphically represented in Figure 9). Of course, we could select fewer topics to group the articles into more general groups, but that degree of granularity is not interesting for a review. The point was to show the diversity of topics. We agree that it could be interesting to explore the hierarchy of topics, but we think it is out of the scope of this review.

An important difference among ML algorithms is their ability/inability to predict the uncertainty of their predictions. For example, Quantile Random Forest (Meinshauzen et al, 2006) has this functionality that was successfully applied for map-

ping soil properties (e.g. Vaysse Lagacherie, 2017). This is of paramount importance in the subdomain of soil science called Digital Soil Mapping. I guess it can be of interest for other domains. Therefore, I think that this aspect should be examined and discussed as it is done for the interpretability of the models (and not in the few words in the section "commercial ML application").

We completely agree with this point. Thanks for the reminder. We moved the uncertainty paragraph from Section 4.3 to its own Section, showing some extra results (see attached Figure) and references.

I agree that the selection of the hyper-parameters of a ML algorithm is very important. You should mention that this aspect has been worked also for the random Forest algorithm (Probst et al, 2019).

We added the reference to that paper and to more general reading on hyper-parameter tuning.

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

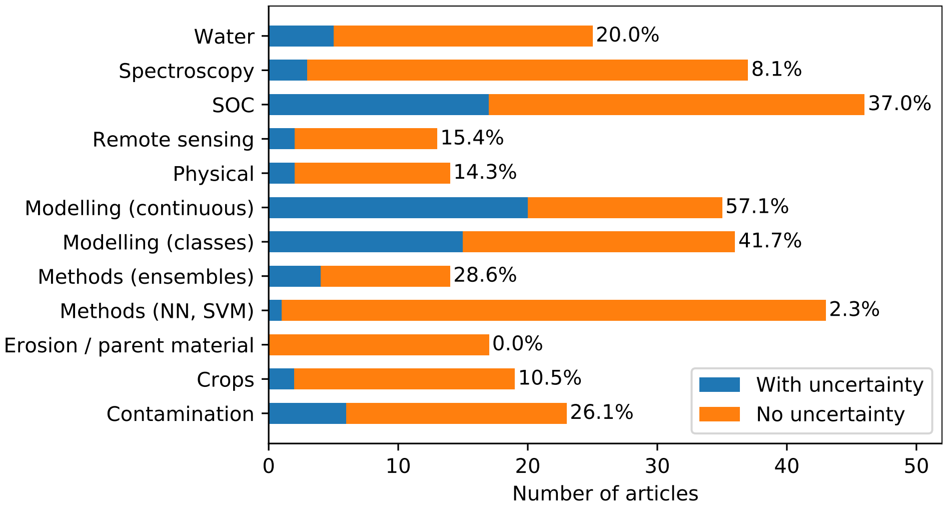

Fig. 1.