SOIL Discuss.,
https://doi.org/10.5194/soil-2019-57-RC2, 2019

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

Anonymous Referee #2

Received and published: 17 October 2019

General comments The paper presents a review of the applications of Machine Learning (ML)in Soil science. This is undoubtedly an interesting topic since ML arouse more and more interest in many scientific domains and, to my knowledge, there has not been any review focusing on soil science. Furthermore, the paper conveys an innovative methodology for doing a literature review, using an automatic reading of the papers followed by a topic modelling algorithm (the latent Dirichlet Allocation) for identifying the topics within a given scientific domain and allocate the paper to these topics. This however a very nice paper for SOIL. However, some details dealing with the application of the methodology and with the discussion of the results are debatable and are discussed below.

Specific comments âĂć I am afraid that the keyword "Machine Learning" is too restric-

tive 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". Consequently, such author's papers could not belong to the set of 322 papers that were analyzed in this review. As an example, it seems that you missed the paper from Nussbaum et al (2018) that was published in this journal (Soil). âĂć 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. âĂć 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 mapping 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"). âĂć I agree that the selection of the hyperparameters 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)

References Meinshausen, N., 2006. Quantile Regression Forests. J. of Machine Learning Res. 7, 983–999. Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4, 1–22. Probst, P., Wright, M., Boulesteix, A., 2018. Hyperparameters and Tuning Strategies for Random Forest. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1–19. Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.

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