SOIL Discuss.,
https://doi.org/10.5194/soil-2019-48-AC1, 2019

Interactive comment on "Estimation of effective calibration sample size using visible near infrared spectroscopy: deep learning vs machine learning" *by* Wartini Ng et al.

Wartini Ng et al.

wartini.ng@sydney.edu.au

Received and published: 1 November 2019

Thank you for taking the time to review our manuscript. We will address the comments and revise the paper accordingly. Our detailed responses are as follow:

The paper presents a case study of prediction performance comparisons between a couple of standard machine learning methods commonly used in soil spectroscopy (PLS and Cubist) and a deep learning algorithm (convolutional neural networks, CNN). These algorithms are tested in a large soil spectral library. The paper clearly shows that CNN outperforms PLS and Cubist when the

number of calibration observations is large. All the algorithms tend to perform poorly when they are used in small (calibration) sample sets. In my opinion, the manuscript does not present a clear contribution to soil science. I just have some comments that I hope help the authors to improve their manuscript.

We agree that the comparison on the performance of deep learning vs machine learning (Cubist and PLS) using a large NIR spectra library to predict soil properties has been published by our group (Padarian et al and Ng et al., 2019). Deep learning application in soil spectroscopy, and even in spectroscopy is still new (Yang et al. 2019)

Currently, there is still no guideline on how many samples would we need to effectively use deep learning methods. Deep learning was developed to handle a large amount of data (millions of images), and clearly soil spectra data are not that large. For example, a recent study used deep learning on 135 soil samples (Chen et al., 2018). Clearly the advantage of using deep learning on such small number of samples is questionable. A recent review on spectroscopy showed that there are a large number of studies where deep learning was used with small sample size (Yang et al. 2019). The review indicated that increased training samples could further improve the calibration performance, however there is no guideline how much improvement can be expected and what is the minimum number of samples.

In addition, there is a hypothesis in machine learning literature that common regression methods will reach a plateau with increasing sample size, while the performance of deep learning will still increase. We tested this hypothesis in soil NIR modelling,

Hence, the contribution to soil science is:

- Establishing the number of samples required for deep learning to be effective
- To test the hypothesis that common machine learning models with reach a plateau in accuracy with an increasing number of samples.
- Establishing how much improvement in accuracy when we increase the number of calibration sample
- Demonstrating how to interpret deep learning models

Chen, H., Liu, Z., Gu, J., Ai, W., Wen, J. and Cai, K., 2018. Quantitative analysis of soil nutrition based on FT-NIR spectroscopy integrated with BP neural deep learning. Analytical methods, 10(41), pp.5004-5013.
Yang, Jie, Jinfan Xu, Xiaolei Zhang, Chiyu Wu, Tao Lin, and Yibin Ying. "Deep learning for vibrational spectral analysis: Recent progress and a practical guide." Analytica Chimica Acta (2019).

General comments: - The method used by the authors to estimate the effective calibration sample size is entirely based on prediction performance indicators (e.g. root mean square error) which requires prior knowledge of the response variables of the samples used as candidates for calibration. Therefore, the method is rather unrealistic/impractical.
The objective of this paper is not on calibration sampling or estimating the effective number of sample size for an area as done by Ramirez-Lopez et al., 2014. We recognise the title would need a revision, which we will revise. The aim of the application of spectroscopy in soil science field is to provide rapid prediction of soil properties. In order to do it, prior knowledge of the response variables is indeed needed. There is no other way of estimating the effective calibration size rather than using empirical data.

- Since the main objective of the paper is related to calibration sampling for soil spectroscopy, I encourage the authors to review the literature available on this topic. This might help to clearly identify research needs and also to identify already available methods to optimize the number of samples used in calibration (see Esbensen et al., 2014; Ramirez-Lopez et al., 2014; De Gruijter et al., 2006 ; Petersen et al., 2005; Minkkinen, 2004).
The objective of this paper is not on calibration sampling or estimating the effective number of sample size for an area. This paper aims to provide a guide on the number of samples to be used within the calibration model for both deep learning and machine learning model in a large and diverse dataset. In our previous study, we

have found that the machine learning model reaches plateau performance when the number of sample size reach several thousand. We would like to understand a comparison of performance of machine learning vs. deep learning with number of samples.

- The effective size of the calibration set for a given spectral dataset largely depends on the variability or complexity embedded in such dataset. For example, a small area where a large number of soil spectra is available (as in the case of on-the-go soil spectroscopy), the optimal size of the calibration set would be rather small. Furthermore, in such non-complex scenario, the use of CNN would be arguable, as the conventional methods would be expected to perform well (as it has been proven). In this respect, the authors seem to focus only on the size of the calibration sets disregarding very important aspects of the theory of sampling (see Minkkinen, 2004) and draw general conclusions from a single experimental dataset.
We agreed that for a smaller area, the use of conventional machine learning would be suitable. We did not disregard the theory of the sampling and as described above this paper is not about establishing sampling for calibration. The sample selected for the calibration model is based on stratified sampling scheme, where the samples from the same sites are grouped together.

- The conclusions are not clear and despite their original research question (how many samples are required to get CNN performing better than PLS and Cubist) is answered for their particular dataset, there is no useful procedure or method presented by the authors to reproduce or extrapolate this to other cases in a useful way.
Although the conclusion we drew applied only for this dataset, it would provide the readers a guide of when to and not to use the CNN model within their own dataset. Clearly CNN requires a large dataset. It would not be possible to analyse all combination of spectra library. We believe the 2000 samples are applicable as a general guide.

Specific comments: - Section 3 (chemometrics model): the authors need to provide information on model optimization and references. For example, why do they choose a learning rate of 0.001 and adam optmizer, what does it mean? Is there any reference the readers can be referred to?
We will provide this information in the revised paper. We need to ensure that the learning rate is not too high or too low. Adam is one of the learning optimizer that is used when training neural network (aside from RMSprop, SGD, Adadelta, etc.) For more information regarding this optimizer, refer to Kingma and Ba (2014).
Kingma, D. P., and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

How many PLS components and committees were tested in PLS and cubist respectively? Optimization of the algorithms play a key role in their performance.
We could not agree more that the optimization is important. We had included this within the paper. For the PLS model, we selected number of components that resulted in the lowest RMSE based on cross validation approach. The committees for the Cubist model was set as 1.

- Section 4.4 (sensitivity analysis): this whole section does not seem to bring any significant contribution to the objectives of the paper.
We still think this is an important part of modelling. This section is not related to sampling size for both machine learning and deep learning model, however interpretation is as much important as getting a highly accurate prediction. We would like to demonstrate that deep learning does not necessarily be a black box. We show that the sensitivity analysis can relate the model back to the basic knowledge. The sensitivity of certain region can be related to the presence of certain molecules that affect the prediction of certain soil properties.

- Section 4.4 (sensitivity analysis): The estimations of the importance of variables for modeling for different modeling algorithms are based on different methods, therefore the comparisons between the results carried out in the paper are not appropriate.
We couldn't agree more. We recognise the differences and that each method is unique, and it is mentioned in the paper. For example, PLS regression parameters method cannot be applied to Cubist and CNN. We just want to mention that there are different methods to interpret the model.

- Section 4.4 (sensitivity analysis): the authors need to be more clear with their statement: "the wavelengths used Cubist were derived based on model usage".
The important wavelengths were selected based on the variables that were used within the model either as predictors (blue lines) or conditions (pink lines). We will clarify it in our revised paper.

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2019-48, 2019.