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

Anonymous Referee #1

Received and published: 19 October 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".

We have clarified it in our revised paper as: The important wavelengths were selected based on the variables that were used within the model either as predictors (blue lines) or conditions (pink lines).

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

Anonymous Referee #2

Received and published: 16 March 2020

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

The manuscript tackles with an important and interesting topic; however, the presentation was really poor, not easy to follow. The most important issue is that the manuscript lacks the Discussion section! Actually the manuscript is not ready to be submitted to a journal.

We will elaborate the discussion section.

- Apart from the Abstract and Introduction sections, the other sections were totally mixed in a way that in some parts you could not get, which section you are reading. For example, Lns. 176-196 are method but have been presented in the Results sections. This is a critical issue in a paper that needs to be solved.

We will move it into Method section, and keep the results within Result section.

- The authors have compared CNN with PLSR and Cubist, as two common machine learning techniques, although Cubist have not been very common in soil spectroscopy so far compared to RF and SVM. It would be fine if these algorithms also be taken into account.

We’ll remove the word “common method” when referring to Cubist model. Several of the studies below had shown that Cubist > RF and SVM. Hence, the inclusion of RF and SVM would not be necessary.

> Sorenson, P. T., Small, C., Tappert, M. C., Quideau, S. A., Drozdowski, B., Underwood, A., and Janz, A. 2017. Monitoring organic carbon, total nitrogen, and pH for reclaimed soils using field reflectance spectroscopy. Canadian Journal of Soil Science, 97(2), 241-248.
>
> Sharififar, A., Singh, K., Jones, E., Ginting, F. I., and Minasny, B. 2019. Evaluating a low-cost portable NIR spectrometer for the prediction of soil organic and total carbon using different calibration models. Soil Use and Management, 35(4), 607-616.
>
> Silva, E. B., Giasson, É., Dotto, A. C., Caten, A. t., Demattê, J. A. M., Bacic, I. L. Z., and Veiga, M. d. 2019. A Regional Legacy Soil Dataset for Prediction of Sand and Clay Content with Vis-Nir-Swir, in Southern Brazil. Revista Brasileira de Ciência do Solo, 43.

- Some parts repeating the same thing several times. For instance, the section 4.3. generally repeats the same contents in Lns. 158-163 and Lns. 168-173 that should be avoided.

We understand that it seems that we are repeating the same thing several times. However, R^2 itself is not enough in chemometrics. Thus, RMSE are also implemented to evaluate model performance.

In our example, although it seems that Cubist and PLSR performed better than the CNN model in terms of R^2 for smaller sample size (see Figure 5); there are larger variances in the RMSE of Cubist model in comparison to CNN model. Based on the R^2 itself, both PLSR and Cubist seemed to also perform similar. However, when we compare the model performance in terms of RMSE ratios, we can see that there are less variances using the PLSR model.

- In presenting the comparison between PLSR and Cubist has been missed. Please compare them as well. In general, the Results sections should be more detailed furnished with more obtained values and comparison of them.

Yes, we shall include the comparison between the PLSR and Cubist model within the script as well (see updated figure 6)

- Surprisingly, the manuscript does not have the Discussion section, which is one of the most important parts of each paper. There are only some lines in the Result section whitin authors have presented the results of other similar studies (e.g. Lns. 148-151, Lns. 198-207, Lns. 212-215), which cannot be considered as the discussion of the results of the current work. Please separate the section of Results from the Discussion with detailed and informative discussion of your works' outputs.

We'll elaborate the discussion as we have mentioned above..

All to all, I reject the manuscript at this step but highly recommend its resubmission after the corrections done.

We have revised the paper, and added discussion of our findings.

~~Estimation~~The influence of ~~effective calibration~~training sample size ~~using visible near infrared spectroscopy:~~on the accuracy of deep learning ~~vs machine learning~~models for the prediction of soil properties with NIR spectroscopy data

Wartini Ng[1], Budiman Minasny[1], Wanderson de Sousa Mendes[2], José A.M.Demattê [2],

[1] School of Life and Environmental Sciences & Sydney Institute of Agriculture, The University of Sydney, NSW, Australia

[2] Department of Soil Science, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Av. Pádua Dias 11, Portal Box 9, Piracicaba, São Paulo state Code 13418-900, Brazil

Correspondence to: Wartini Ng (wartini.ng@sydney.edu.au)

Abstract

The number of samples used in the calibration dataset affects the quality of the generated predictive models using visible, near and shortwave infrared (VIS-NIR-SWIR) spectroscopy for soil attributes. Recently, the convolutional neural network (CNN) is regarded as a highly accurate model for predicting soil properties on a large database~~, however~~. However, it has not been ascertained yet how large the sample size should be for CNN model to be effective. ~~This paper~~This paper investigates the effect of training sample size on the accuracy of deep learning and machine learning models. It aims at providing an estimate of how much calibration samples are needed to improve the model performance of soil properties predictions with CNN~~. It is hypothesized that the larger the amount of data, the more accurate is the CNN model. The performances of two commonly used~~ as compared to conventional machine learning models. In addition, this paper also looks at a way to interpret the CNN models, which are commonly labelled as black box. It is hypothesized that the performance of machine learning models will increase with an increasing number of training samples, but it will plateau when it reached a certain number, while the performance of CNN will keep improving. The performances of two machine learning models (Partial least squares regression (PLSR) and Cubist) are compared against the CNN model. A VIS-NIR-SWIR spectral library from Brazil containing 4251 unique sites, with averages of 2-3 samples per depth (a total of 12,044 samples), was divided into calibration (3188 sites) and validation (1063 sites) sets. A subset of the calibration dataset was then created to represent smaller calibration dataset ranging from 125, 300, 500, 1000, 1500, 2000, 2500 and 2700 unique sites, or equivalent to sample size approximately 350, 840, 1400, 2800, 4200, 5600, 7000, and 7650. All three models (PLSR, Cubist, and CNN models) were generated for each sample size of

Style Definition: Normal

Style Definition: Heading 2: Space After: 8 pt, Don't add space between paragraphs of the same style, Line spacing: Double, Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0 cm + Indent at: 0.63 cm, Don't keep with next

Style Definition: Heading 3: Indent: Hanging: 1.4 cm, Space After: 8 pt, Don't add space between paragraphs of the same style, Line spacing: Double, Outline numbered + Level: 2 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.63 cm + Indent at: 1.4 cm, Don't keep with next

Style Definition: Heading 4: Font: Bold, Indent: Left: 0 cm, Hanging: 1.25 cm, Space Before: 12 pt, After: 8 pt, Don't add space between paragraphs of the same style, Line spacing: Double, Outline numbered + Level: 3 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 1.27 cm + Indent at: 2.16 cm, Don't keep with next

Style Definition: Bullets: Indent: Left: 0 cm, First line: 0 cm, Bulleted + Level: 1 + Aligned at: 0.63 cm + Indent at: 1.27 cm, Tab stops: 0.63 cm, List tab

Style Definition: No Spacing

Formatted: English (Australia)

Formatted: English (Australia)

Formatted: Line spacing: 1.5 lines

Formatted: English (Australia)

Formatted: English (Australia)

the unique sites for the prediction of five different soil properties, i.e. cation exchange capacity, organic matter, sand, silt and clay content. These calibration subset sampling processes and modelling were repeated ten times to provide a better representation of the model performances. ~~Similar results were observed when~~Learning curves showed that the ~~performances~~accuracy increased with an increasing number of ~~both~~training sample. At a lower number of samples (<1000), PLSR and Cubist ~~model were compared to the~~performed better than CNN ~~model where the~~. The performance of CNN outweighed the PLSR and Cubist model at a sample size of 1500 and 1800 respectively. It can be recommended that deep learning is most efficient for spectral modelling for sample size above 2000. The accuracy of the PLSR and Cubist model ~~seemed~~seems to reach a plateau above sample size of 4200 and 5000, respectively~~.~~, while the accuracy of CNN has not plateaued. A sensitivity analysis ~~was performed on~~of the CNN model demonstrated the ability to determine important wavelengths region that affected the predictions of various soil attributes.

Keywords: convolutional neural network, deep learning, machine learning, infrared spectroscopy, soil properties, soil analysis

1. Introduction

There has been an increasing demand for a rapid and cost-effective method as an alternative for conventional laboratory soil analysis. Visible, near and shortwave infrared (VIS-NIR-SWIR) spectroscopy has been proposed to be used as an alternative tool for soil analysis for the last few decades (Bendor and Banin, 1995;Shepherd and Walsh, 2002;Stenberg et al., 2010). This method enables the simultaneous prediction of various properties and has non-destructive characteristics.

Various machine learning models, such as Partial Least Squares Regression (PLSR), Cubist, random forest, and support vector machines had been utilized to model spectroscopy data. However, the performances of these regression models are dependent on the spectra pre-processing methods (Rinnan et al., 2009), as well as the size ~~of calibration dataset~~ and ~~its~~ representativeness of the calibration samples (Kuang and Mouazen, 2012;Ng et al., 2018). Different orders and combinations of the spectra pre-processing methods, ~~which are~~ developed to remove artefact in the spectral signal, will result in ~~different~~various model performances. Furthermore, the spectra pre-processing techniques developed for a particular dataset might not work for a different dataset. Better generalization can be made by training the model in a larger dataset. ~~However, reduced or plateau performance on the machine learning model was found as the sample size increased to several thousands (Ng et al., 2018)~~However, several studies demonstrated that the performance of the machine learning model did not increase significantly or even plateaued as the calibration sample size increased (Figueroa et al., 2012;Ramirez-Lopez et al., 2014;Ng et al., 2018).

Advances in ~~the~~ artificial intelligence, such as deep learning enable the possibility of extracting features from data without hand-engineered features (LeCun et al., 2015), such as pre-processing. Various deep learning convolutional neural network (CNN) model (AlexNet, VGGnet, GoogLeNet, ResNet~~)~~), had been developed and trained on large volumes of data, which included over 10 million image data (Krizhevsky et al., 2012;Simonyan and Zisserman, 2014;Szegedy et al., 2015;He et al., 2016).

Although CNN often deals with images as input data, it has recently been successfully applied to vibrational ~~spectroscopy.~~and reflectance spectroscopy (Acquarelli et al., 2017;Cui and Fearn, 2018;Liu et al., 2018;Ng et al., 2019;Padarian et al., 2019). Acquarelli et al. (2017) found that the CNN based model outperformed other models (Partial Least Square – Least Discriminant Analysis, logistic regression and k-nearest neighbour) for the classification of various vibrational spectroscopy data. CNN also has recently been successfully utilized for regression modelling using reflectance spectroscopy data (Cui and Fearn, 2018;Liu et al., 2018;Ng et al., 2019;Padarian et al., 2019)~~. In particular, recent studies~~ (Ng et al., 2019;Padarian et al., 2019) ~~had shown that CNN model had the capability to outperform PLSR and Cubist model. However, the CNN model usually requires a large number of calibration samples~~. Cui and Fearn (2018) compared the performance of CNN and PLSR to predict protein and ash content of wheat kernels and wheat flour from the NIR-SWIR spectra data with calibration sample size ranging from 415 – 6,987. Liu et al. (2018) developed one-dimensional CNN model using VIS-NIR-SWIR spectra data to predict soil clay content with a calibration sample size of 16,000. Other studies had shown that CNN model had the capability to outperform PLSR and Cubist model for the prediction of various soil properties using VIS-NIR-SWIR (Ng et al., 2019;Padarian et al., 2019), mid-

Formatted: Line spacing: 1.5 lines

Formatted: Space After: 12 pt

Formatted: Space After: 12 pt

infrared (MIR) and combined VIS-NIR-SWIR with MIR spectra (Ng et al., 2019) with a calibration sample size greater than 10,000.

~~The~~Deep learning such as CNN was developed to handle a large amount of data (millions of images), and clearly soil spectra libraries these days are not that large yet. For example, a recent study used deep learning on 135 soil samples (Chen et al., 2018). The advantage of using CNN on such a small number of samples is uncertain. A recent review on spectroscopy showed that there were several studies where deep learning was used with a small calibration sample size (Yang et al., 2019). The review indicated that increased calibration samples should further improve the calibration performance, however, there is no guideline how much improvement can be expected and what is the minimum number of samples for it to be effective.

Strategy to select adequate calibration set in terms of representativeness and size is vital in obtaining a model with good generalization ability. Although various sampling algorithms (Kennard-Stone, conditioned Latin Hypercube sampling, k-means clustering) to select representative samples have been explored (Ramirez-Lopez et al., 2014;Ng et al., 2018), the question of how much samples are needed for the CNN model to perform better than ~~the~~ machine learning ~~model using the~~models for spectroscopy data has yet to be determined. It is commonly depicted and hypothesized in a learning curve that as more data are available, CNN will perform better compared to traditional machine learning models ~~which will reach a plateau with an increasing amount of data~~ (Mahapatra, 2018) (see Figure 1). Machine learning models tend to reach a plateau or show marginal improvement with an increasing amount of data as the model has limited complexity to deal with an increasing amount of data (Zhu et al., 2016).

Thus, the purpose of this study is to assess the amount of calibration data needed for the CNN model to perform better than machine learning models. PLSR and Cubist are chosen as the representatives of the ~~regression and machine learning models which has been commonly used to develop predictive models based on soil spectra data.~~machine learning models which had been found to perform well in soil spectra data (e.g., Dangal et al. (2019)). In addition, to be able to predict soil properties accurately, we need to understand and interpret how a CNN model can predict soil properties from spectra. ~~The sensitivity analysis of the VIS-NIR-SWIR region used in the CNN model is performed to uncover the CNN black box.~~Specifically, this paper presents the following specific contributions:

- testing the idea that common machine learning models with reach a plateau in accuracy with an increasing number of calibration samples,
- establishing the number of calibration samples required for deep learning to be effective for VIS-NIR-SWIR spectra data,
- establishing how much improvement in accuracy is achieved the number of calibration sample for deep learning and machine learning models is increased, and
- demonstrating how to interpret deep learning model using a sensitivity analysis.

Formatted: Space After: 12 pt

2. Materials and Methods

2.1. Dataset and chemical analysis

This dataset comprises of 12,044 soil samples from 4,251 unique sites. The soil samples, collected from several regions of Brazil, i.e., states of Sao Paulo, Minas Gerais, Goias, and Mato Grosso do Sul. This dataset is part of The Brazilian Soil Spectral Library and extracted from Terra et al. (2018) and Bellinaso et al. (2010). The soils were derived mostly from basalt (volcanic rock) and sedimentary ones (sandstone). Each site has up to seven samples measurements from the surface up to 1 m depth.

The measured properties include soil texture (sand, silt, and clay), organic matter (OM) content and cation exchange capacity (CEC). ~~The hydrometer and sedimentation method was used to obtain soil particle size (Bouyoucos, 1927). Organic matter (OM) was determined using~~The soil particle size was quantified by the pipette method, as described in Donagema et al. (2011). The method consists of using a 0.1 M NaOH solution as dispersing agent under high-speed mechanical stirring during 10 min. Then, the sand fraction was separated by sieving and the clay portion by sedimentation. The silt was quantified based on pre- and post-difference. Organic carbon (OC) was determined by the Walkley and Black method (Walkley and Black, 1934)~~. CEC was determined as the sum of exchangeable cations using the method described by Raij et al. (2001).~~, in which OC was oxidised using $K_2Cr_2O_7$ in a wet environment and then measured by titration with 0.1 M ammonium iron sulphate. After that, the organic matter (OM) was calculated by multiplying the OC quantified per the Van Bemmelen factor of 1.724. As described in Donagemma et al. (2011), a 1 M KCl solution was used to extracted aluminium, exchangeable calcium and magnesium. The atomic absorption spectrophotometry was used to quantify Ca and Mg concentrations. Aluminium concentration was determined by titrating with 0.025 M NaOH. Potassium and phosphorus contents were extracted using Mehlich-1 (0.05 M HCl with 0.0125 M H_2SO_4) solution. The concentration of P was quantified by colourimetry and the K concentration by flame photometry. Afterwards, CEC was determined as the sum of exchangeable cations. The descriptive statistics of the soil properties measured are included in Table 1.

2.2. Spectral measurements

The VIS-NIR-SWIR spectra of the soil samples were obtained with FieldSpec3 spectroradiometer (Analytical Spectral Devices, Boulder, Colorado) with a spectral range of visible to shortwave infrared (350 – 2500 nm) and spectral resolution of 1 nm from 350 to 700 nm, 3 nm from 700 to 1400 nm, and 10 nm from 1400 to 2500 nm. The sensor scanned an area of approximately 2 cm^2, and a light source was provided by two external 50-W halogen lamps. These lamps were positioned at a distance of 35 cm from the sample (non-collimated rays and a zenithal angle of 30°) with an angle of 90° between them. A Spectralon (Labsphere Inc., North Sutton, NH) standard white plate was scanned every 20 min during calibration. The samples were oven-dried at 45°C for 48 hours before being ground and sieved ≤ 2 mm. The sample was distributed homogeneously in

Formatted: Space After: 12 pt, Line spacing: 1.5 lines

Formatted: Space After: 12 pt

Formatted: Space Before: 0 pt, After: 12 pt

Formatted: Space After: 12 pt, Line spacing: 1.5 lines

Formatted: Space After: 12 pt

~~petri~~Petri dishes for spectra measurement. Three replicates (involving a 180° turn of the Petri dish) were obtained for each sample. Each spectrum was averaged from 100 readings over 10 s.

2.3. Training and validation

To better represent the soil distribution, we split and subset the data based on sites. The dataset is first randomly split into 75% calibration (3188 sites) and 25% validation (1063 sites) based on the unique sites.

From the calibration dataset, ~~we created~~ smaller sample sizes ranging from 125, 300, 500, 1000, 1500, 2000, 2500 and 2700 unique sites were created, which is equivalent to a sample size of approximately 350, 840, 1400, 2800, 4200, 5600, 7000, and 7650. Better representations of model performances were provided by ten replicates of these sizes. Each sampling for the same number of sites could generate a slightly different number of samples since the number of measurements varied from one site to another. However, the model performance was evaluated on the common validation dataset using a total of 1063 sites (sample size N = 3017). Thus, we create a learning curve of the accuracy of the models of the validation dataset as a function of the number of calibration samples.

2.4. Chemometrics model

Prior to the development of machine learning models (PLSR and Cubist), the spectra data were subjected to some pre-processing methods: (i) conversion to absorbance followed by (ii) Savitzky - Golay smoothing filter with window size of 11 and second-order polynomial (Savitzky and Golay, 1964), (iii) spectral trimming to discard region that has low signal to noise ratio (<500 nm and between 2450 – 2500 nm) and (iv) standard-normal-variate (SNV) transformation (Barnes et al., 1989). For the ~~deep learning~~CNN model, the spectra were only normalized with SNV ~~prior to~~before being fed into the model. Our previous research (Ng et al., 2019) found that CNN has its own filtering algorithm that made pre-processing not necessary. This filtering approach will be discussed in the results section.

2.4.1. PLSR model

PLSR is one of the standard and most commonly used models with the spectroscopy data. It is a linear chemometric regression model that projects spectra data into latent variables that explain the variances within the spectra data and the response variables (Wold et al., 1983). The optimal number of latent variables used in the PLSR regression that resulted in the smallest root mean square error (RMSE) using the cross-validation approach was used to create the models.

~~1.1.1.~~ PLSR was~~Cubist model~~

Formatted: Space After: 12 pt, Line spacing: 1.5 lines

Formatted: Space After: 12 pt

Formatted: Heading 3, Space After: 12 pt, Line spacing: 1.5 lines

Formatted: Space After: 12 pt

Formatted: Heading 4, Space After: 12 pt, Line spacing: 1.5 lines

Formatted: Heading 4, Space After: 12 pt, Line spacing: 1.5 lines

~~Cubist is a rule-based data-mining model, which is an extension of the M5 model tree by Quinlan (1993). The model creates one or more rules, in which if the rules are met, a certain linear model can be utilized to predict the target task.~~

~~These machine learning models were~~ implemented in the R statistical software (R Core Team, 2019) using the "pls" package (Mevik et al., 2018).

2.4.2. Cubist model

Formatted: Heading 4, Space After: 12 pt, Line spacing: 1.5 lines

Cubist is a rule-based data mining model, which is an extension of the M5 model tree by Quinlan (1993). ~~and "Cubist" package (Kuhn and Quinlan, 2018) for PLSR and Cubist modelling respectively.~~

Cubist has been used successfully in soil spectroscopy studies and in many cases found to perform better than PLSR and other machine learning models (Dangal et al., 2019). Cubist creates one or more rules, in which if the rules are met, a certain linear model can be utilized to predict the target task. The model was evaluated using the "Cubist" package (Kuhn and Quinlan, 2018) in R.

2.4.3. CNN model

Formatted: Heading 4, Space After: 12 pt, Line spacing: 1.5 lines

The CNN model is composed of three types of layers: convolutional, pooling and fully-connected layer. The convolutional layer extracts feature from the inputs, the pooling layer reduces the dimensionality of the input feature, and the fully connected layer connects the outputs from previous layers to the desired target outputs.

The CNN model utilized in this study ~~is~~was derived from our previous study (Ng et al., 2019), where the spectra data were fed into the model as a one-dimensional data. The architecture of the CNN model is included in Table 2 and ~~Figure 2 .~~Figure 2. Some of the layers within the network are shared to enable simultaneous output predictions.

The CNN model was trained with an initial learning rate of 0.001 and Adam optimizer~~.~~ (Kingma and Ba, 2014). The network was trained a batch size of 50, and a maximum epoch of 200. For model optimization purposes, the calibration data is further divided into 75% train and 25% test set. Dropout, early stopping and reduced learning rates are used as a regularization technique to prevent network overfitting. ~~Details~~For further details of the CNN model, the reader is ~~given in~~referred to Ng et al. (2019) ~~and will not be repeated here.~~

. The CNN model was implemented in Python (v3.5.1; Python Software Foundation, 2017) using Keras library (v2.1.2; Chollet, 2015) and Tensorflow (v1.4.1; Abadi et al., 2015) backend.

Formatted: Space After: 12 pt

All the model performances are compared in terms of coefficient of determination (R^2), and the root mean square error (RMSE), bias and ratio of performance to inter-quartile distance (RPIQ) values based on the validation dataset. Generally, larger values of R^2 and RPIQ and smaller bias and RMSE indicate better model performance.

1. Results

Formatted: Space After: 12 pt, Line spacing: 1.5 lines

Visualization of the CNN

An attempt to take a look at what the CNN model actually learns is conducted. The reflectance spectrum data was fed into the first convolutional layer. The filter in the first layer encodes various pre-processing of the input spectra data. Some of the filters shown in the first convolution layer looks like the input spectra pattern (filter #3, 4 and 10), and some of them looks like transformation pattern: absorbance (filter #1, 5, 6, 7, 9, 13 and 16) and derivatives (filter # 2, 8, 11, 12, 14 and 15). The spectrum becomes smoother when they passed through the second convolutional layer, where some filters only accentuate certain peaks (Figure 3). Thus, the ability of the convolutional layers to represent various transformation of the spectra make CNN a robust model that does not require any spectra pre-processing.

Model performance comparison

The model performances for the validation dataset using the full calibration data (n_{cal}= 3188, N=9027) with all the models are first presented in Table 3. Among all the properties predicted, the sand and clay content showed the best performance with R^2 values greater than 0.75 regardless of the types of model used. This finding is in agreement with the ones from Dematt$\hat{e}$ et al. (2016), who observed good predictions for sand and clay content.

Dematt$\hat{e}$ et al. (2016) reported R^2 values ranging from 0.51 and 0.86 for sand (0.86), silt (0.51, clay (0.85), organic matter (0.63) and CEC (0.66) using PLSR model with 4790 out of 7185 samples as calibration samples. The performances of our PLSR and Cubist model are lower than those reported by Dematt$\hat{e}$ et al. (2016) could probably due to the larger variation of the dataset used here. Furthermore, representative sampling using conditioned Latin hypercube sampling was used in selecting the calibration samples prior to the model development. Nonetheless, the overall CNN model used here still performs better.

Effect of sample training size: sub-setting the calibration data

A total of eight subset models based on the unique sample sizes were generated. The performance comparison of CNN and Cubist model based on average R^2 values is illustrated in Figure 4. The reported R^2 values are the average performance prediction for all five properties of all ten replicates. The value for sample size 9027 is from a single data random split for validation of the data.

In general, the PLSR and Cubist model tend to perform better when the sample size is relatively small (<2000). When the sample size is approximately 1800, there is not much difference in the performances for all models. However, when the sample size is further increased (>2000), the CNN model starts to show better performance in comparison to both PLSR and Cubist model. The performance of PLSR and Cubist model reaches plateau at approximately 4000 and 5500 samples respectively,

~~while the performance of CNN is still increasing, as depicted in the theoretical curve (Figure 1). The slight drop in Cubist's performance for sample size 9027 is because there is only one realization of data split (75% of the data).~~

~~We further compared the average model performance based on the RMSE ratios of machine learning models against the CNN model (Figure 5).~~ ~~This comparison was developed using the model performance for each unique property, and the variances presented was based on ten simulations.~~ ~~If the machine learning model performs better than the CNN model, the RMSE ratios of a particular machine learning model to CNN model should be less than one.~~

~~Based on the RMSE ratios of PLSR against the CNN model, we can observe that PLSR perform better than CNN when the sample is less than 1400 (Figure 5). Similar performance is achieved when the sample size is approximately 1415. In terms of RMSE ratios, overall CNN model seems to perform better in comparison to the Cubist model regardless of sample size. Nonetheless, the model performance for a smaller sample size seems to vary a lot (longer whisker). When the sample size is approximately 850, both models seem to perform similarly. A portion of the model performs better, while the remaining perform worse. As the calibration sample size increases, the CNN model performs better in comparison to the Cubist model. Thus, It can be recommended that deep learning is most efficient for spectral modelling for sample size above 2000.~~

2.5. Sensitivity analysis: evaluating important wavelengths

Formatted: Space After: 12 pt, Line spacing: 1.5 lines

To uncover how CNN predicts different soil properties, a sensitivity analysis was conducted to assess the importance of each wavelength in contributing to predictions. Evaluating the sensitivity of the model can be done in several ways, for example, Cui and Fearn (2018) calculated the sensitivity of a CNN model for NIR by taking a numerical partial derivative of the output with respect to each wavelength. For wavelength i, the sensitivity S was calculated as:

Formatted: Space After: 12 pt

Formatted: Space After: 12 pt

$$S_i = \frac{f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_i + \varepsilon, \dots, \boldsymbol{X}_n) - f(\boldsymbol{X})}{\varepsilon} \quad \text{(Eq. 1)}$$

where $\boldsymbol{X}$ is the reflectance spectra, and $f(\boldsymbol{X})$ is the CNN prediction using the spectra, ε is a small number. The idea is that if wavelength i has an important contribution to the prediction, a small perturbation to the reflectance value will create a large change in the prediction.

Formatted: Space After: 12 pt

In our previous study (Ng et al., 2019), we calculated the sensitivity as a function of the variance of the model for each window of spectra. Here, we calculate the sensitivity based on the variance principle as an alternative approach:

$$S_i = \frac{Var\,(f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_i, \dots, \boldsymbol{X}_n) - f(\bar{\boldsymbol{X}}))}{Var\,(\boldsymbol{Y})} \quad \text{(Eq. 2)}$$

Where Var is the variation calculation, $f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_i, \dots, \boldsymbol{X}_n)$ is the prediction of spectra due to variation in wavelength i with other wavelengths held constant at their mean values, and $f(\bar{\boldsymbol{X}})$ is the prediction value using the mean values of the spectra

Formatted: Space After: 12 pt

and $\boldsymbol{Y}$ is the observed values of the target variable. In essence, we calculated how the model varied in comparison to the observations as a function of wavelength.

The current sensitivity analysis (Eq. 2) considers the actual variance of the data for a better approximation of wavelengths sensitivity. To calculate the variance sensitivity, two new data frames ~~were~~are created. The first data frame contains data which is the average of all the validation spectra data ($\overline{\boldsymbol{X}}$) and the second contains modified average spectra data ($\overline{\boldsymbol{X}}_i$), in which some of the average measurements ~~were~~are replaced with the actual spectral reflectance at a wavelength width of 5 nm.

The illustrations of the process of deriving new data frames are included in Figure 6. Both data frames ~~were~~are then fed into the pre-trained CNN model ($f()$). The variance between the average and modified average spectra ~~were~~are then compared to the actual variance of the target properties as a measure of the model sensitivity (Eq. 2).

~~3. The sensitivity analysis of the CNN model in predicting each property is illustrated in Figure 7. Only certain parts of the spectra are used by the CNN model for prediction, which corresponds to the soil properties and composition.~~Results

Formatted: Space After: 12 pt, Line spacing: 1.5 lines

3.1. VIS-NIR-SWIR spectral characteristics

Large variability within the soil properties and texture could potentially influence the soil spectral characteristics (shown in Figure 3). In general, there was an increase in reflectance between 400 - 1000 nm, with several prominent absorption features at 1400, 1900 and 2200nm. Absorption features in the VIS-NIR (400 - 1000 nm) which is related to iron oxides, such as haematite (Fe_2O_3) and goethite (FeOOH) (Clark, 1999). Absorption near 1400 nm is associated with the first overtone of an O-H stretch vibration of water or metal-O-H vibration, while absorption is 1900 nm is combination vibrations of water related to H-O-H bend and O-H stretch (Viscarra Rossel et al., 2009). Absorption in the 2100-2400 nm region is related to the combination vibrations of minerals. Generally, spectra that have a higher clay content would show smaller reflectance (greater absorption) values in comparison to those with lower clay content. The representative samples of the VIS-NIR-SWIR spectra before and after pre-processing were included in Figure 3.

3.2. Visualization of the spectra within CNN model

An attempt to take a look at what the CNN model actually learns was conducted. As the raw reflectance spectrum was fed into the CNN model, it passed through a convolutional layer which extracted information from the spectra. Filters from the first two convolutional layers were included in Figure 4. Though only raw spectra were fed into the CNN model, we could see that the spectra underwent some spectra pre-processing within each filters of the layers. Some of the filters shown in the first convolution layer looked like the input spectra pattern (filter #3, 4 and 10), and some of them mimicked like transformation pattern: absorbance (filter #1, 5, 6, 7, 9, 13 and 16) and derivatives (filter # 2, 8, 11, 12, 14 and 15). The spectrum became

smoother when they passed through the second convolutional layer, where some filters only accentuated certain peaks (Figure 4).

3.3. Prediction of soil properties and model comparison

The model performances for the validation dataset using the full calibration data (n_{site}= 3188, N=9027) for various soil properties and chemometrics model were presented in Table 3. CNN model outperformed both Cubist and PLSR model (in terms of higher R^2 and RPIQ and lower RMSE).

The performance achieved using the CNN model with the prediction of sand (R^2= 0.85; RPIQ =1.52), silt (R^2=0.58; RPIQ =0.75), clay (R^2=0.86; RPIQ =1.05), organic matter (R^2=0.69; RPIQ =0.91) and CEC (R^2=0.68; RPIQ =0.69). Both the PLSR and Cubist had similar performance for the prediction of the various properties. PLSR model achieved R^2 of 0.79, 0.47, 0.80, 0.48 and 0.52, and RPIQ of 1.29, 0.67, 0.87, 0.70, and 0.57 for the prediction of sand, silt, clay, organic matter, and CEC respectively. Meanwhile, Cubist model achieved R^2 of 0.78, 0.45, 0.81, 0.54 and 0.52 and RPIQ of 1.19, 0.67, 0.92, 0.70 and 0.59 for the prediction of sand, silt, clay, organic matter, and CEC respectively. Nonetheless, on some cases, the CNN model prediction yielded higher bias on the prediction of some soil properties, such as OM and CEC (bias = -0.11 and -0.76 respectively), than PLSR model (bias = 0.04 and -0.17) for the same properties. The Cubist model yielded bias of -0.22 and -0.17 for the prediction of OM and CEC respectively.

Among all the properties predicted, the sand and clay content showed the best performance with R^2 values greater than 0.75 regardless of the types of model used ranging from (0.78 – 0.85 and 0.8 – 0.86) respectively. This finding is in agreement with the ones from Demattê et al. (2016), who observed good predictions for sand and clay content with R^2 of 0.86 and 0.85. Pinheiro et al. (2017) reported the prediction accuracy of 0.62 and 0.78 for the sand and clay content, respectively. The low performance of the silt predicted can be linked to error associated with the laboratory analysis method, where the silt content is derived from the difference of the soil mass after the sand and clay content are determined. The prediction for OM content in our study ranges from R^2 of 0.48 – 0.69. Shibusawa et al. (2001) reported R^2 of 0.65 for the prediction of OM using slightly different wavelength region (400-2400nm). Our prediction of CEC ranges from R^2 of 0.52 – 0.68. Chang et al. (2001) and Islam et al. (2003) reported R^2 of 0.81 and 0.88, respectively for the prediction of CEC. Although some prediction accuracies are slightly lower than other studies, they are still within an acceptable range.

3.4. Effect of sample training size: learning curve

A total of nine subset models based on the unique sample sizes were generated to investigate the effect of training sample size. The performance comparison of all the models expressed as average R^2 values is illustrated as a learning curve in Figure 5. The depicted R^2 values are the average performance prediction for all five properties of all ten replicates, except for the largest sample size (N = 9027) where a single data random split for validation of the data is used. The learning curve generally follows

the common pattern found in machine learning studies (Figueroa et al., 2012), the performance increased rapidly with an increase in the size of the training set from around 350 to 1400. For PLSR and Cubist, the growth in performance became slower after it reached 2800 samples. PLSR performance reached a plateau after 4000 samples while the increase in performance in Cubist was marginal after 5500 samples.

In general, the PLSR and Cubist model tend to perform better when the sample size was relatively small (<1500). When the sample size was approximately 1800, there was only a small difference in the performances for all models. However, when the sample size was further increased (>2000), the CNN model started to show better performance in comparison to both PLSR and Cubist model. The effectiveness of PLSR and Cubist model reached a plateau at approximately 4000 and 5500 samples, respectively, while the performance of CNN was still increasing, as depicted in the theoretical curve (Figure 1). The slight drop in Cubist's performance at sample size 9027 was because there was only one realization of data split (75% of the data).

We further compared the average model performance based on the RMSE ratios of machine learning models against the CNN model (Figure 6). This comparison was developed using the model performance for each unique property, and the variances presented was based on ten simulations. If a particular model X performs better than the Y model it is compared against, the RMSE ratios of X/Y should be less than one.

Upon comparing the RMSE ratios of PLSR/Cubist model, we found that PLSR performed better than the Cubist model when the sample size is less than 1400. Cubist model performed better than the PLSR model as the sample size was increased. Using the RMSE ratios of PLSR/CNN model, PLSR was found to perform better than CNN when the sample is less than 1400 (Figure 5). Similar performance of both PLSR and CNN model was achieved when the sample size is approximately 1400. In terms of RMSE ratios of Cubist/CNN, overall CNN model performed better in comparison to the Cubist model regardless of sample size. This was slightly different than the one that was observed when only R^2 parameter was utilized. The RMSE ratios of Cubist/CNN seemed to vary more for a smaller sample size (longer whisker). When the sample size is approximately 850, both models seemed to perform similarly. A portion of the model performed better, while the remaining performed worse. As the calibration sample size increased, the CNN model performed better in comparison to the Cubist model. Thus, it can be recommended that the current CNN model structure is most efficient for VIS-NIR-SWIR spectral modelling with sample size above 2000. CNN still can be used for small number of samples, but its performance is not better than PLSR or Cubist.

3.5. Sensitivity Analysis

The critique of CNN is that it is a complex model and a black box. To uncover how the CNN model works, a sensitivity analysis was conducted to show how CNN is predicting each of the soil properties, illustrated in Figure 7. Only certain parts of the spectra were used by the CNN model for prediction, which corresponded to the soil properties and composition. The important wavelengths for the prediction of CEC are between the regions of 1600 – 2000 nm. This result is similar to the

Formatted: Space After: 12 pt

observations made by Lee et al. (2009) on the surface horizon dataset where 1772 and 1805 nm are ~~important~~essential in predicting the CEC. The presence of high CEC is often linked to the presence of organic matter (OM) and clay content. It is interesting that the same region is important in predicting organic matter but not clay content. Aside from the same region used by CEC, wavelengths' region between 1100 – 1200 nm are also deemed ~~important~~relevant by the CNN model for the prediction of OM content. This finding is slightly different to those reported by Lee et al. (2009) in which the important wavelengths reported are at 1772, 1871, 2069, 2246, 2351 and 2483 nm for the profile dataset and 1871, 2072 and 2177 nm for the surface horizon dataset. ~~It is also worth noting that the model does not use the visible part of the spectra for prediction. In comparison to the sensitivity of MIR spectra data on previous study (Ng et al., 2019), the NIR model's sensitivity index is much broader, which reflects NIR's characteristic broad peak.~~

Similar wavelength regions are deemed to be important in predicting the soil texture although the importance slightly varied among the type of texture of interest (sand, silt and clay) at wavelengths between 500 and 1800 nm. The important wavelengths for the prediction of sand and clay content share a higher similarity in comparison to that of silt content prediction. The most ~~important~~crucial wavelength identified is around 850 nm for the prediction of sand and clay content, and around 1100 nm for the prediction of silt content. These observations are also different from those reported by Demattê (2002) and Lee et al. (2009) where the important wavelengths for the prediction of soil texture are at 1800 – 2400 nm. In particular, the soil texture prediction found in the CNN model is strongly related to hematite and/or goethite, -OH and Al-OH groups from kaolinite (Viscarra Rossel and Behrens, 2010;Pinheiro et al., 2017;Fang et al., 2018).

We also compare important wavelengths from the machine learning models against the one from the deep learning model for the prediction of OM as an example. Common wavelengths found to be related to the organic matter predictions are 1100, 1600, 1700 – 1800, 2000, 2200 – 2400 nm (Dalal and Henry, 1986;Stenberg et al., 2010).

Formatted: Space After: 12 pt

As a comparison, we calculated important wavelengths used in the PLSR and Cubist models. The important wavelengths utilized in the PLSR model was derived based on the absolute value of the regression coefficients. The height of the line indicates the importance of ~~a~~ particular ~~wavelength~~wavelengths for the determination of organic matter content in the soil. Important wavelengths identified for the prediction of organic matter were 500 – 700, 1400 and 1715 nm.

The wavelengths used in the Cubist were derived based on model usage ~~(Figure 8). The blue and pink lines represent which wavelengths are used~~either as predictors ~~and~~(blue lines) or conditions ~~within the Cubist model, respectively.~~(pink lines) (Figure 9). Some of the wavelengths used in the Cubist model are similar to those observed in the PLSR model, in particular the visible (500 – 700 nm), and shortwave infrared regions (1400 and 1900 nm).

4. Discussion

4.1. Understanding the CNN models

While conventional PLSR and machine learning models require spectra pre-processing for the spectra data input, CNN model takes raw spectra as inputs. CNN has been shown to be a successful end-to-end learning model which learn feature automatically while minimizing hand-crafted pre-processing process. Upon taking a closer look at the various filters within the convolutional layers, we found that the filters behaved like spectra pre-processing method. It is interesting to note that using the raw spectral input, various spectral pre-processing that was commonly used within spectroscopy could be observed within the layer itself. Given the various complexity within the CNN model, the use of spectra pre-processing prior to being fed is unnecessary. This advantage opens up possibilities of developing highly accurate chemometrics model, which also plays a role in automatic spectral pre-processing.

CNN have been proven to be extremely successful, however how they work remains largely a mystery as they are buried in layers of computations. Sensitivity analysis enabled us to see better the inner workings of the CNN model. We could understand better which wavelengths features are essential from the spectra when used in developing the regression prediction. Important wavelengths derived from the sensitivity analysis based on the CNN model looked slightly different from those of PLSR and Cubist models. Wavelengths around the 1700 nm region were deemed to be the most important, followed by those between the 1150 nm region. Nonetheless, some of the important regions overlapped. It was also worth noting that the model did not use the visible part of the spectra for prediction. In comparison to the sensitivity of MIR spectra data on previous study (Ng et al., 2019), the NIR model's sensitivity index ~~Important wavelengths derived from the sensitivity analysis based on the CNN model look slightly different to those of PLSR and Cubist models. Wavelengths around the 1700 nm region is deemed to be the most important, followed by those between the 1150 nm region.~~ was much broader, which reflected NIR's characteristic broad peak.

~~Nonetheless, some of the important regions overlapped.~~

Although all three methods used different ways to derive important wavelengths, PLSR model ~~tends~~tended to use most parts of the spectra. When irrelevant wavelengths are included in model development, it may reduce the model performance. The Cubist model ~~seems~~seemed more selective in terms of wavelengths used, however this example showed that it also used most parts of the VIS-NIR-SWIR spectra. CNN model used wavelengths between 800-~~ ~~~~2000 nm, with ~~particular emphasize~~emphasis around 1100 and 1700 nm.

Formatted: Space After: 12 pt

~~Conclusion~~

4.2. The effect of calibration sample size to model performance

PLSR, Cubist and CNN represent models with increased complexity. By combining results from 5 soil properties, we can show better a generalisation of the performance of the models as a function of training sample size. Simpler models (PLSR) performed better at a smaller sample size (< 1400). Cubist outperformed PLSR at sample size > 2000, while CNN outweighed other models when sample size > 2500. The increase in the accuracy of machine learning models (PLSR and Cubist) became insignificant when the number of samples was greater than 5000. This trend of plateauing of performance (maximized up to a certain point) with an increase in sample size as had been observed by several authors (Shepherd and Walsh, 2002;Kuang and Mouazen, 2012;Ramirez-Lopez et al., 2014;Ng et al., 2018). This trend is related to the complexity of the model, as a simpler model (such as PLSR) cannot capture all variation in the data. Thus, a more complex model is suitable when the number of samples is large.

Previous studies by Ng et al. (2019) and Padarian et al. (2019) had shown that CNN performed better than PLSR and Cubist when the model was trained with more than 10,000 samples. However, there were also studies using CNN with a small number of training samples. This study showed that CNN model only outperformed PLSR and Cubist models when the sample size is greater than 2000. As sample size increases, the efficiency of CNN model is increased. We observed a larger reduction in RMSE (CNN compared to the other 2 models) with increasing calibration sample size. Thus, we recommend using a minimum of 2000 samples to train CNN model for the VIS-NIR-SWIR spectra. To further improve the performance of the CNN model, simultaneous prediction of soil properties could also be implemented within the model.

The advantage of using deep learning on a small number of samples is minimal as CNN is a data-hungry model; it is also more computationally expensive than the typical machine learning models. While our results pertain to the spectral dataset from Brazil and a particular structure of the CNN, we believe our results can serve as a guide on the number of samples needed to create a better deep learning model. Future research could test this idea on larger and more variable datasets (e.g. a global spectral library with more than 100,000 samples) and to see if a more complex and deeper network of CNN can handle such dataset.

5. Conclusions

In this paper, we ~~assess~~assessed the ~~effective~~effect of training sample size and ~~identify~~identified important wavelengths in predicting various soil properties using Cubist and CNN model. In general, the CNN ~~method can perform~~ model performed better than the Cubist when the sample size is relatively large~~. The number of calibration samples is also affected by the structure of the CNN model. The number of samples reported in this study might not apply to other CNN models but can serve as a guide on the number of samples needed to create a better deep learning model.~~ (>2000). Here, we found that with its current model structure, CNN is more accurate than a machine learning model when the number of calibration samples is above 2000. The more complex and deeper network of a deep learning model, the ~~most~~more likely it will need a larger number of samples for training. PLSR and Cubist models ~~performed~~perform less accurate than the CNN model as sample size

Formatted: Space After: 12 pt

increases, and ~~it seems like they~~ both models reached a plateau after a sample size of 4000~~ –~~ 5000. Meanwhile, the performance of CNN still increases until the maximum number of data used in this study (N = 9000). Future studies should explore larger dataset to see the generalization of the accuracy vs sample size and to explore if the deep learning CNN model ever ~~reached~~reaches a plateau in accuracy.

Author contributions

Wartini Ng was responsible for the data analysis, and prepared the manuscript; Budiman Minasny contributed ~~in~~to the idea, data analysis and editing the manuscript; Wanderson de Sousa Mendes and José A.M.Demattê contributed to the idea, provided the data and ~~contributed in~~ editing the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This study was financed in part by the ARC Linkage Project LP150100566 - Optimised field delineation of contaminated soils. The authors would also like to thank members of the Geotechnologies in Soil Science Group (https://esalqgeocis.wixsite.com/geocis) and Sao Paulo Research Foundation (FAPESP, grant numbers 2014/22262-0 and 2016/26124-6). BM is a member of a consortium supported by LE STUDIUM Loire Valley Institute for Advanced Studies through its LE STUDIUM Research Consortium Programme.

References

Acquarelli, J., van Laarhoven, T., Gerretzen, J., Tran, T. N., Buydens, L. M. C., and Marchiori, E.: Convolutional neural networks for vibrational spectroscopic data analysis, Anal Chim Acta, 954, 22-31, 10.1016/j.aca.2016.12.010, 2017.
Barnes, R. J., Dhanoa, M. S., and Lister, S. J.: Standard Normal Variate Transformation and De-Trending of near-Infrared Diffuse Reflectance Spectra, Appl Spectrosc, 43, 772-777, Doi 10.1366/0003702894202201, 1989.
Bellinaso, H., Demattê, J. A. M., and Romeiro, S. A.: Soil Spectral Library and Its Use in Soil Classification, Rev Bras Cienc Solo, 34, 861-870, Doi 10.1590/S0100-06832010000300027, 2010.
Bendor, E., and Banin, A.: Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties, Soil Sci Soc Am J, 59, 364-372, DOI 10.2136/sssaj1995.03615995005900020014x, 1995.
Bouyoucos, G. J.: The hydrometer as a new method for the mechanical analysis of soils, Soil Sci, 23, 343-353, Doi 10.1097/00010694-192705000-00002, 1927.
Chang, C. W., Laird, D. A., Mausbach, M. J., and Hurburgh, C. R.: Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties, Soil Sci Soc Am J, 65, 480-490, 2001.

Formatted: Line spacing: 1.5 lines
Formatted: Space After: 12 pt
Formatted: Line spacing: 1.5 lines
Formatted: Space After: 12 pt
Formatted: Line spacing: 1.5 lines
Formatted: Space After: 12 pt
Formatted: Line spacing: 1.5 lines

Chen, H. Z., Liu, Z. Y., Gu, J., Ai, W., Wen, J. B., and Cai, K.: Quantitative analysis of soil nutrition based on FT-NIR spectroscopy integrated with BP neural deep learning, Anal Methods-Uk, 10, 5004-5013, 10.1039/c8ay01076e, 2018.
Clark, R. N.: Chapter 1: Spectroscopy of Rocks and Minerals, and Principles of Spectroscopy, in: Manual of Remote Sensing, edited by: Rencz, A. N., Remote Sensing for the Earth Sciences, John Wiley and Sons, New York, 3- 58, 1999.
Cui, C. H., and Fearn, T.: Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration, Chemometr Intell Lab, 182, 9-20, 10.1016/j.chemolab.2018.07.008, 2018.
Dalal, R. C., and Henry, R. J.: Simultaneous Determination of Moisture, Organic-Carbon, and Total Nitrogen by near-Infrared Reflectance Spectrophotometry, Soil Sci Soc Am J, 50, 120-123, DOI 10.2136/sssaj1986.03615995005000010023x, 1986.
Dangal, S., Sanderman, J., Wills, S., and Ramirez-Lopez, L.: Accurate and Precise Prediction of Soil Properties from a Large Mid-Infrared Spectral Library, Soil Systems, 3, 11, https://doi.org/10.3390/soilsystems3010011, 2019.
Demattê, J. A. M.: Characterization and discrimination of soils by their reflected electromagnetic energy, Pesqui Agropecu Bras, 37, 1445-1458, Doi 10.1590/S0100-204x2002001000013, 2002.
Demattê, J. A. M., Bellinaso, H., Araujo, S. R., Rizzo, R., and Souza, A. B.: Spectral regionalization of tropical soils in the estimation of soil attributes, Rev Cienc Agron, 47, 589-598, 2016.
Donagema, G. K., de Campos, D. B., Calderano, S. B., Teixeira, W., and Viana, J. M.: Manual de métodos de análise de solo, Embrapa Solos-Documentos (INFOTECA-E), 2011.
Fang, Q., Hanlie, H., Zhao, L., Kukolich, S., Yin, K., and Wang, C.: Visible and near-infrared reflectance spectroscopy for investigating soil mineralogy, 2018.
Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., and Ngo, L. H.: Predicting sample size required for classification performance, BMC Medical Informatics and Decision Making, 12, 8, 10.1186/1472-6947-12-8, 2012.
He, K. M., Zhang, X. Y., Ren, S. Q., and Sun, J.: Deep Residual Learning for Image Recognition, 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr), 770-778, 10.1109/Cvpr.2016.90, 2016.
Islam, K., Singh, B., and McBratney, A.: Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy, Aust J Soil Res, 41, 1101-1114, https://doi.org/10.1071/SR02137, 2003.
Kingma, D. P., and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Lake Tahoe, Nevada, 2012.
Kuang, B., and Mouazen, A. M.: Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale, Eur J Soil Sci, 63, 421-429, 10.1111/j.1365-2389.2012.01456.x, 2012.
LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436-444, 10.1038/nature14539, 2015.
Lee, K. S., Lee, D. H., Sudduth, K. A., Chung, S. O., Kitchen, N. R., and Drummond, S. T.: Wavelength Identification and Diffuse Reflectance Estimation for Surface and Profile Soil Properties, T Asabe, 52, 683-695, 2009.
Liu, L., Ji, M., and Buchroithner, M.: Transfer Learning for Soil Spectroscopy Based on Convolutional Neural Networks and Its Application in Soil Clay Content Mapping Using Hyperspectral Imagery, Sensors-Basel, 2018.
Why Deep Learning over Traditional Machine Learning?: https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063, 2018.
Ng, W., Minasny, B., Malone, B., and Filippi, P.: In search of an optimum sampling algorithm for prediction of soil properties from infrared spectra, Peerj, 6, 10.7717/peerj.5722, 2018.
Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., and McBratney, A. B.: Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra, Geoderma, 352, 251-267, https://doi.org/10.1016/j.geoderma.2019.06.016, 2019.
Padarian, J., Minasny, B., and McBratney, A. B.: Using deep learning to predict soil properties from regional spectral data, Geoderma Regional, 16, e00198, https://doi.org/10.1016/j.geodrs.2018.e00198, 2019.
Pinheiro, E. F. M., Ceddia, M. B., Clingensmith, C. M., Grunwald, S., and Vasques, G. M.: Prediction of Soil Physical and Chemical Properties by Visible and Near-Infrared Diffuse Reflectance Spectroscopy in the Central Amazon, Remote Sens-Basel, 9, 10.3390/rs9040293, 2017.
Quinlan, J. R.: C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Mateo, California, 1993.

Raij, B. V., Andrade, J. C., and H. Cantarella, J. A. Q.: Análise química para avaliação de solos tropicais, IAC, Campinas, 2001.
Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Dematte, J. A. M., and Scholten, T.: Sampling optimal calibration sets in soil infrared spectroscopy, Geoderma, 226, 140-150, 10.1016/j.geoderma.2014.02.002, 2014.
Rinnan, A., van den Berg, F., and Engelsen, S. B.: Review of the most common pre-processing techniques for near-infrared spectra, Trac-Trend Anal Chem, 28, 1201-1222, https://doi.org/10.1016/j.trac.2009.07.007, 2009.
Savitzky, A., and Golay, M. J. E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal Chem, 36, 1627-1639, 10.1021/ac60214a047, 1964.
Shepherd, K. D., and Walsh, M. G.: Development of Reflectance Spectral Libraries for Characterization of Soil Properties, Soil Sci Soc Am J, 66, 988-998, https://doi.org/10.2136/sssaj2002.9880, 2002.
Shibusawa, S., Imade Anom, S. W., Sato, S., Sasao, A., and Hirako, S.: Soil mapping using the real-time soil spectrophotometer. In: Grenier, G., Blackmore, S. (Eds.), ECPA, Third European Conference on Precision Agriculture. : Agro Montpellier, 1. Montpellier, France, pp. 497–508., 2001,
Simonyan, K., and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, CoRR, abs/1409.1556, 2014.
Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., and Wetterlind, J.: Chapter Five - Visible and Near Infrared Spectroscopy in Soil Science, in: Adv Agron, edited by: Sparks, D. L., Academic Press, 163-215, 2010.
Szegedy, C., Liu, W., Jia, Y. Q., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going Deeper with Convolutions, Proc Cvpr Ieee, 1-9, 2015.
Terra, F. S., Dematte, J. A. M., and Rossel, R. A. V.: Proximal spectral sensing in pedological assessments: vis-NIR spectra for soil classification based on weathering and pedogenesis, Geoderma, 318, 123-136, 10.1016/j.geoderma.2017.10.053, 2018.
Viscarra Rossel, R. A., Cattle, S. R., Ortega, A., and Fouad, Y.: In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy, Geoderma, 150, 253-266, https://doi.org/10.1016/j.geoderma.2009.01.025, 2009.
Viscarra Rossel, R. A., and Behrens, T.: Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma, 158, 46-54, 10.1016/j.geoderma.2009.12.025, 2010.
Walkley, A., and Black, I. A.: An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method, Soil Sci, 37, 29-38, Doi 10.1097/00010694-193401000-00003, 1934.
Wold, S., Martens, H., and Wold, H.: The Multivariate Calibration-Problem in Chemistry Solved by the Pls Method, Lect Notes Math, 973, 286-293, 1983.
Yang, J., Xu, J. F., Zhang, X. L., Wu, C. Y., Lin, T., and Ying, Y. B.: Deep learning for vibrational spectral analysis: Recent progress and a practical guide, Anal Chim Acta, 1081, 6-17, 10.1016/j.aca.2019.06.012, 2019.
Zhu, X., Vondrick, C., Fowlkes, C. C., and Ramanan, D.: Do We Need More Training Data?, International Journal of Computer Vision, 119, 76-92, 10.1007/s11263-015-0812-2, 2016.

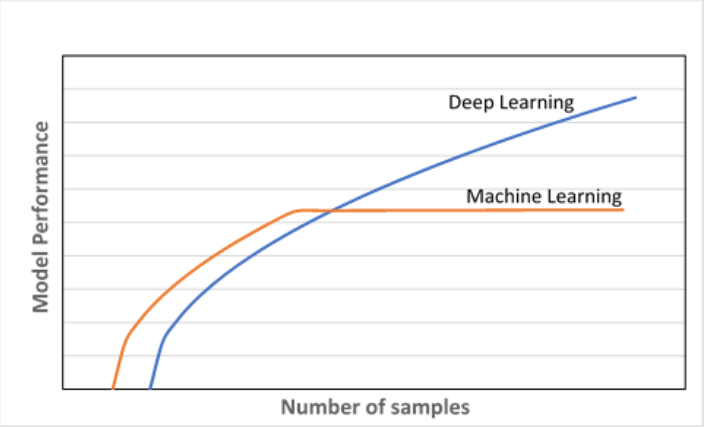


Figure 1. Model performance of deep learning vs other machine learning algorithms as a function of number of samples.

Formatted: Left, Indent: Left: 0 cm, Hanging: 1 cm

Formatted: Line spacing: 1.5 lines

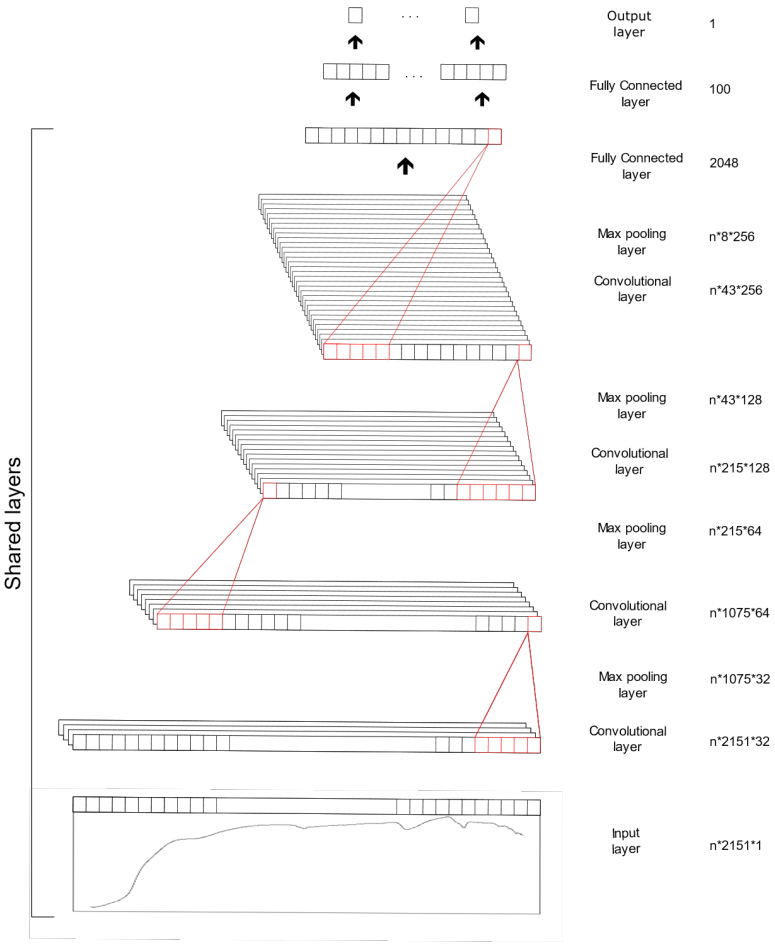
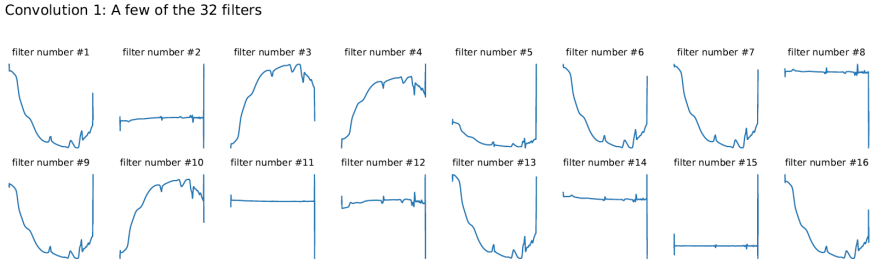
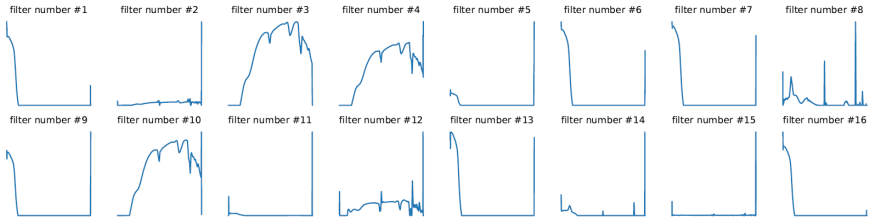


Figure 2. Architecture of the one-dimensional Convolutional Neural Network (CNN) model.

Formatted: Line spacing: 1.5 lines





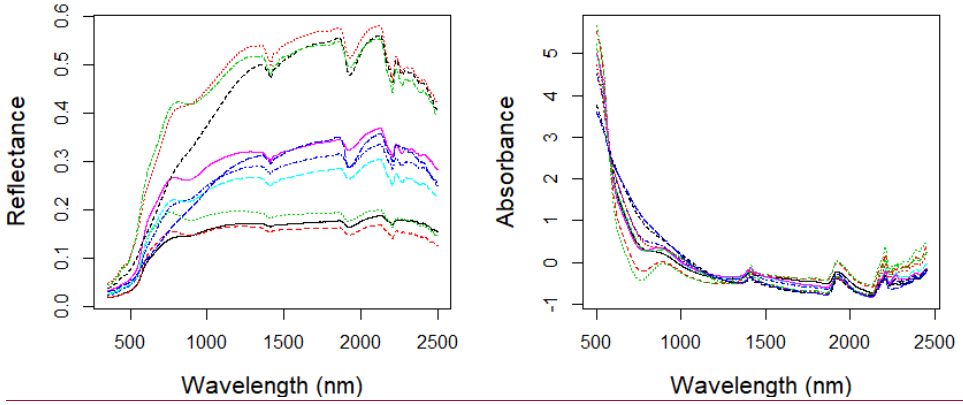
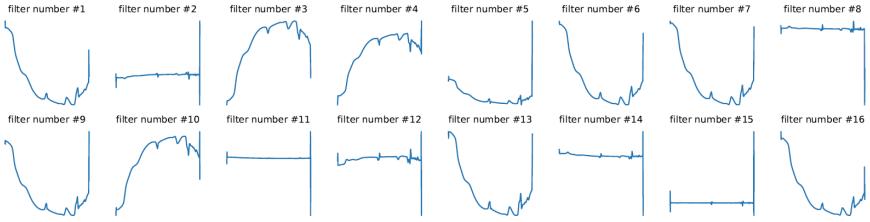


Figure 3. Visible, near and shortwave infrared (VIS-NIR-SWIR) spectra of 10 soil samples without spectra pre-processing (left) and with spectra pre-processing (right).

543

Convolution #1: A few of the 32 filters



Convolution #2: A few of the 64 filters

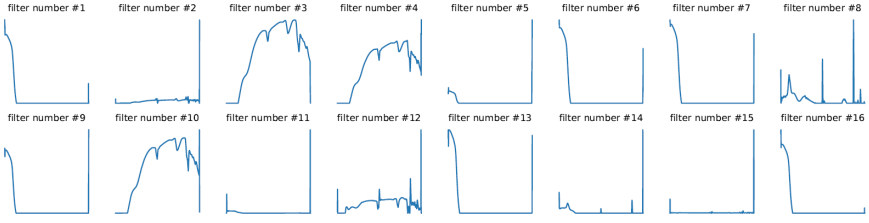


Figure 4. Visualization of the filters ~~within~~in the first two convolutional layers within the one-dimensional Convolutional Neural Network (CNN) ~~with~~model of the visible, near, and shortwave infrared (VIS-NIR-SWIR) spectra data.

Formatted: Line spacing: 1.5 lines

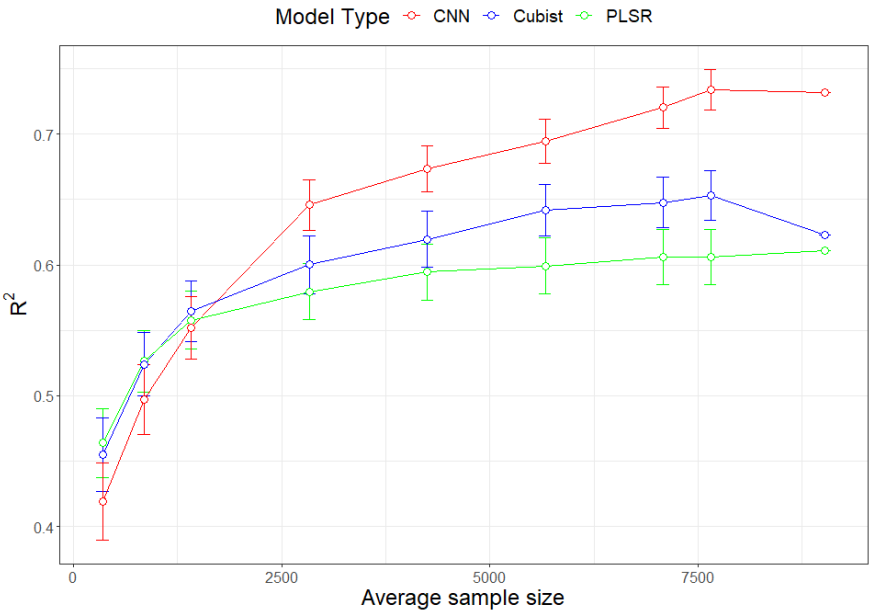
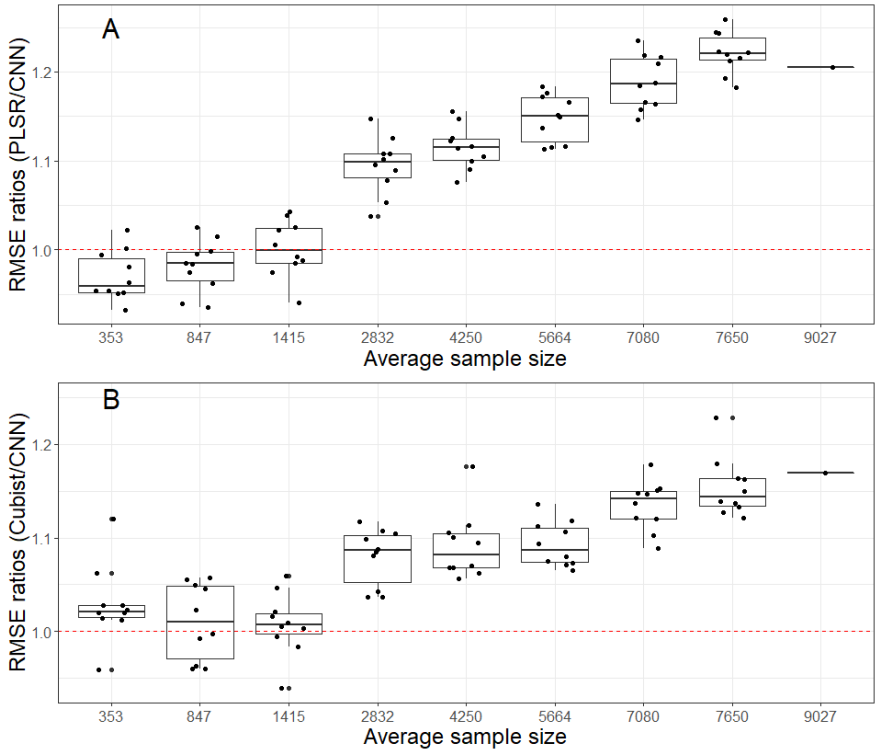


Figure 5. Model performances (in terms of average R^2 for five soil properties) as a function of sample size using Partial Least Squares Regression (PLSR), Cubist and Convolutional Neural Network (CNN) model based on ten simulations. The ~~values~~value for the largest sample size (~~n~~=N = 9027) is a single realization 75% of the data.

Formatted: Line spacing: 1.5 lines


A
RMSE ratios (PLSR/CNN)
1.2
1.1
1.0
353
847
1415
2832
4250
5664
7080
7650
9027
Average sample size
B
RMSE ratios (Cubist/CNN)
1.2
1.1
1.0
353
847
1415
2832
4250
5664
7080
7650
9027
Average sample size

555

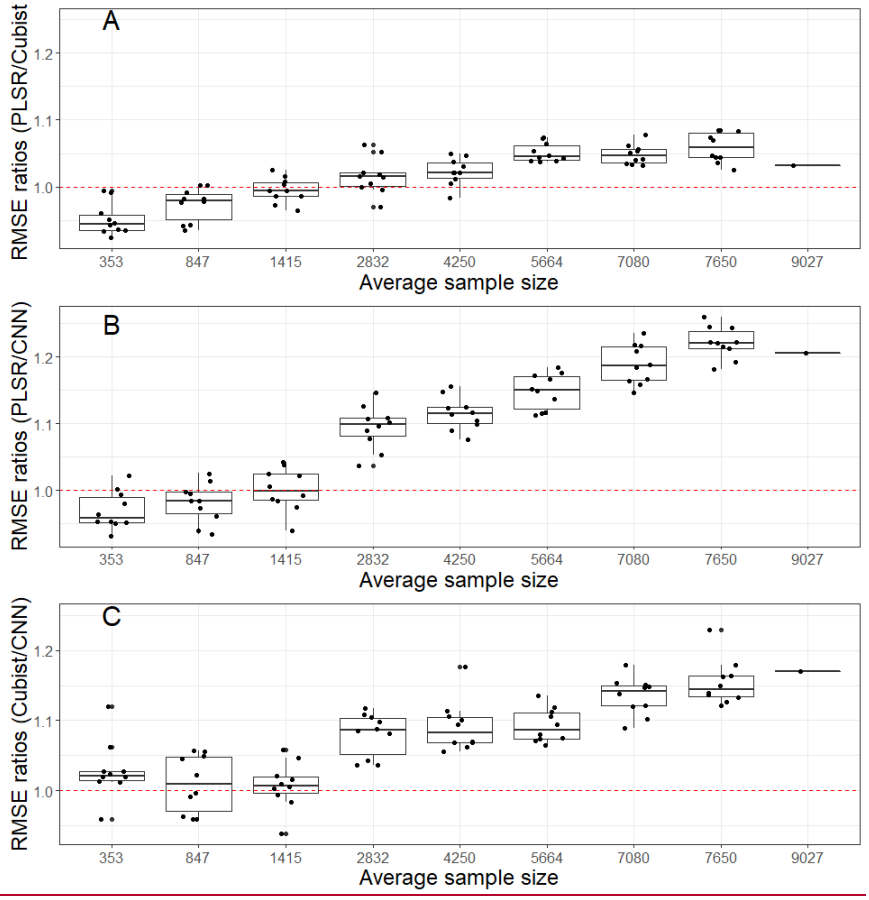


Figure 6. Model performances (in terms of root mean square error (RMSE) ratios of (A) Partial Least Squares Regression (PLSR) over Cubist model (B) PLSR over Convolutional Neural Network (CNN) model and (~~B) Partial Least Squares Regression (PLSR)~~C) Cubist over CNN as an average of five soil properties) based on various sample size using ten simulations. The red – dotted line represents a 1:1 RMSE ratio.

Formatted: Line spacing: 1.5 lines

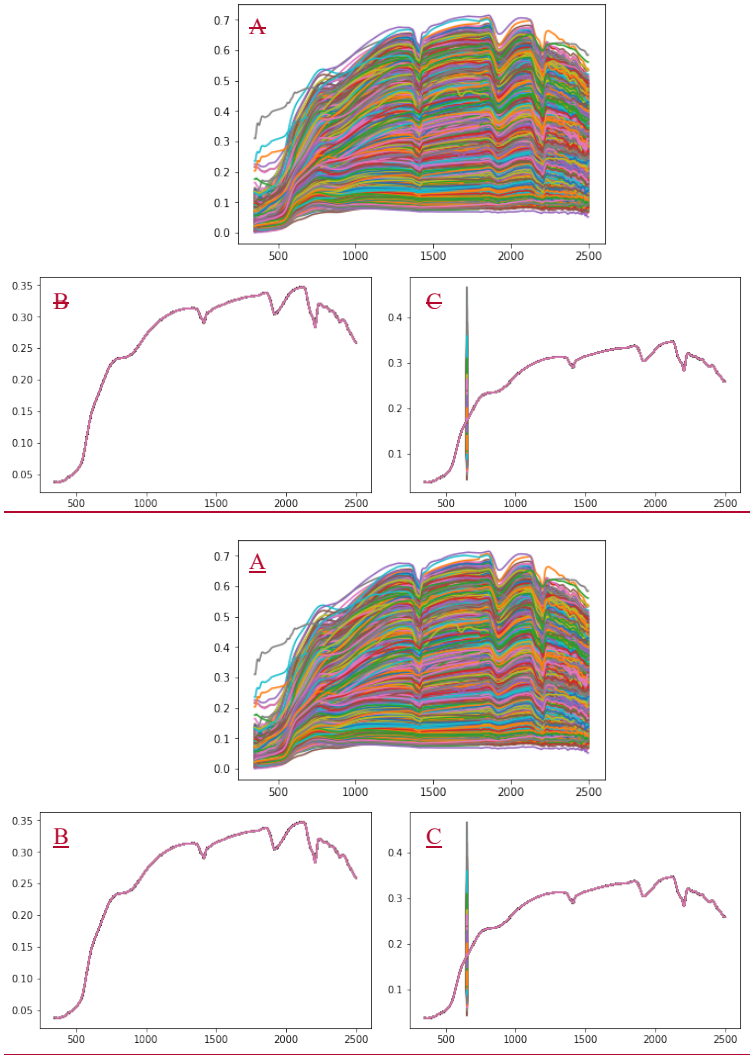
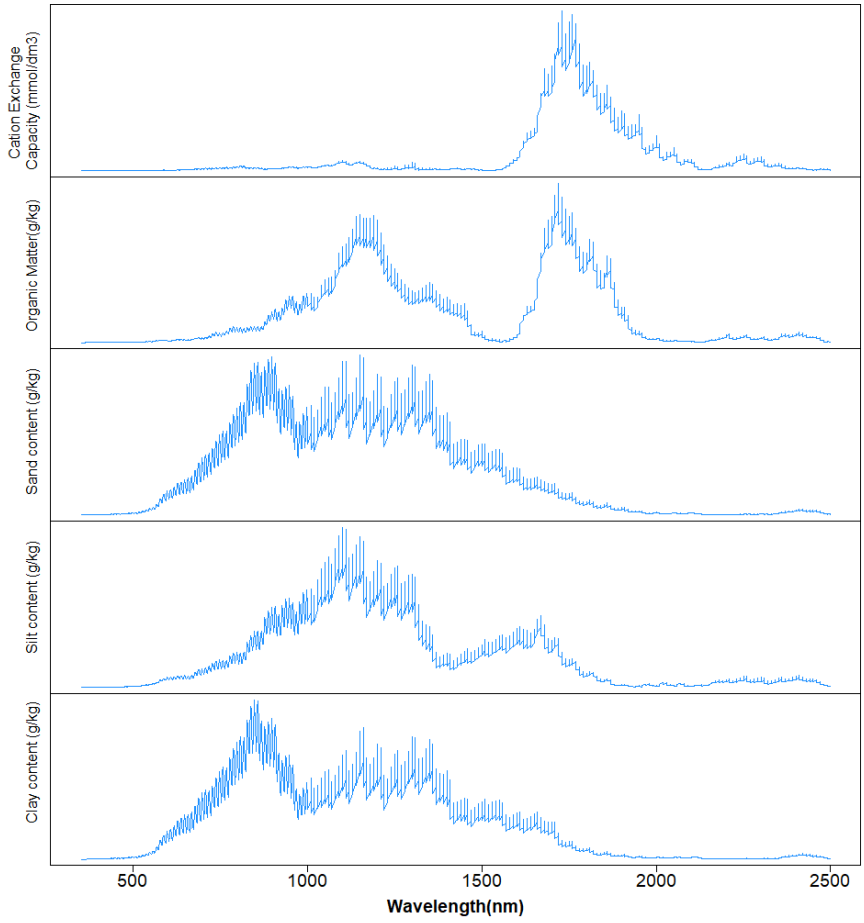


Figure 7. Illustration of sensitivity analysis process: (A) represents the validation spectra data, (B) represents the overall average of the validation spectra data and (C) represents the modified average of the validation spectra data.

Formatted: Line spacing: 1.5 lines



564

565 **Figure 8: Sensitivity analysis of the visible, near and shortwave infrared (VIS-NIR-SWIR) spectra in predicting various soil**
566 **properties using the Convolutional Neural Network (CNN) model. The graph depicts sensitivity index (calculated from(Eq. 2)) for**
567 **different soil properties as a function of wavelength.**

568

Formatted: Line spacing: 1.5 lines

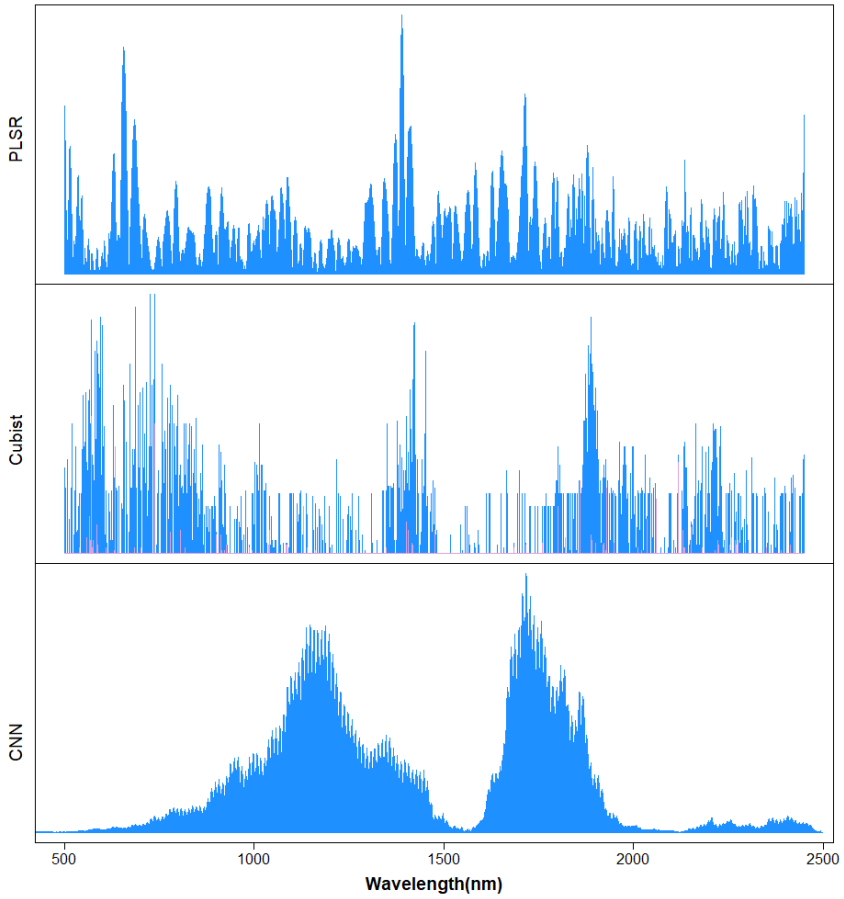


Figure 9: Important wavelengths for the prediction of organic matter (OM) content using Partial Least Squares Regression (PLSR), Cubist and Convolutional Neural Network (CNN) model.

Formatted: Indent: Left: 0 cm, First line: 0 cm, Line spacing: 1.5 lines

573
574
575 **Table 1: Descriptive statistics of the soil properties measurements.**

Formatted: Line spacing: 1.5 lines

	Sand	Silt	Clay	OM	CEC
		g kg^{-1}			$mmol_c$ kg^{-1}
Minimum	50.0	0.0	5.0	2.0	3.4
1st Quartile	644.0	31.0	112.0	6.0	22.9
Median	757.0	57.0	174.7	9.4	32.7
Mean	703.8	69.7	226.5	11.2	37.7
3rd Quartile	839.0	93.5	283.3	14.3	46.3
Maximum	969.0	562.0	840.0	69.0	375.7

576
577

Table 2: Architecture of the convolutional neural network.

Type	Shared	Filter size	# Filters	Activation
Convolutional	Yes	20	32	ReLU
Max-pooling	Yes	2	-	-
Convolutional	Yes	20	64	ReLU
Max-pooling	Yes	5	-	-
Convolutional	Yes	20	128	ReLU
Max-pooling	Yes	5	-	-
Convolutional	Yes	20	256	ReLU
Max-pooling	Yes	5	-	-
Dropout (0.4)	Yes	-	-	-
Flatten	Yes	-	-	-
Fully-connected	No	-	100	ReLU
Dropout (0.2)	No	-	-	-
Fully-connected	No	-	1	Linear

*ReLU: rectified linear units

Formatted: Line spacing: 1.5 lines

Formatted: Line spacing: single

Table 3: Results of model validation for the prediction of various soil attributes using the full calibration dataset.

Model	Properties	Unit	R^2	RMSE	bias	RPIQ
PLSR	Sand	g kg^{-1}	0.79	91.47	2.74	1.29
	Silt		0.47	41.58	-1.78	0.67
	Clay		0.80	73.01	-0.65	0.87
	OM		0.48	4.98	0.04	0.70
	CEC	$mmol_c$ kg^{-1}	0.52	16.77	-0.17	0.57
Cubist	Sand	g kg^{-1}	0.78	89.66	1.28	1.19
	Silt		0.45	38.68	-2.06	0.67
	Clay		0.81	69.65	-0.23	0.92
	OM		0.54	4.83	-0.22	0.70
	CEC	$mmol_c$ kg^{-1}	0.52	17.03	-0.93	0.59
CNN	Sand	g kg^{-1}	0.85	77.28	-0.16	1.52
	Silt		0.58	37.09	-1.74	0.75
	Clay		0.86	60.78	-0.53	1.05
	OM		0.69	3.83	-0.11	0.91
	CEC	$mmol_c$ kg^{-1}	0.68	13.73	-0.76	0.69

OM = organic matter; CEC = cation exchange capacity

Formatted: Line spacing: 1.5 lines

Formatted: Line spacing: single

Formatted: Line spacing: single

Formatted: Line spacing: single

Formatted: Line spacing: single

Formatted: Line spacing: single

Formatted: Line spacing: single

Formatted: Line spacing: single