https://doi.org/10.5194/soil-2019-48
Preprint. Discussion started: 17 September 2019

Estimation of effective calibration sample size using visible near infrared spectroscopy: deep learning vs machine learning

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, convolutional neural network (CNN) is regarded as a highly accurate model for predicting soil properties on a large database, however it has not been ascertained yet how large the sample size should be for CNN model to be effective. This paper 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 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 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 the performances of both PLSR and Cubist model were compared to the CNN model where the performance of CNN outweighed the PLSR and Cubist model at 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 to reach a plateau above sample size of 4200 and 5000 respectively. A sensitivity analysis was performed on the CNN model 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 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 pre-processing methods (Rinnan et al., 2009), as well as the size of calibration dataset and its representativeness (Kuang and Mouazen, 2012;Ng et al., 2018). Different orders and combinations of the pre-processing methods, which are developed to remove artefact in the spectral signal, will result in different model performances. Furthermore, the pre-processing techniques developed for a particular dataset might not work for 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).

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 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. 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 spectroscopy data (Cui and Fearn, 2018;Liu et al., 2018;Ng et al., 2019;Padarian et al., 2019a, b). In particular, recent studies (Ng et al., 2019;Padarian et al., 2019a) 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.

The question of how much samples are needed for the CNN model to perform better than the machine learning model using the spectroscopy data has yet to be determined. It is commonly depicted and hypothesized 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).

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. 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.

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 soil particle size was quantified by the pipette method as described in Donagema et al. (2011). The method consists on 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), 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 colorimetry 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 $\leq$ 2 mm. The sample was distributed homogeneously in 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, which is equivalent to 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).

3. 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 model, the spectra were only normalized with SNV prior to being fed into the model.

3.1. PLSR model

PLSR is one of the 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.

3.2. Cubist model

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) and "Cubist" package (Kuhn and Quinlan, 2018) for PLSR and Cubist modelling respectively.

3.3. CNN model

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 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 . 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. 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 of the CNN model is given in Ng et al. (2019) and will not be repeated here.

The CNN 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.
All the model performances are compared in terms of coefficient of determination (R^2), and the root mean square error (RMSE) values based on the validation dataset.

4. Results

4.1. 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.

4.2. Model performance comparison

The model performances for the validation dataset using the full calibration data (n_{site}= 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 Demattê et al. (2016), who observed good predictions for sand and clay content.

Demattê 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 Demattê 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.

4.3. 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.

4.4. Sensitivity analysis: evaluating important wavelengths

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:

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

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.

In 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:

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

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 $f(\overline{\boldsymbol{X}})$ is the prediction value using the mean values of the spectra 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 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 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 then fed into the pre-trained CNN model ($f()$). The variance between the average and modified average spectra were then compared to the actual variance of the target properties as a measure of the model sensitivity (Eq. 2).

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. The important wavelengths for the prediction of CEC are between the regions of 1600 – 2000 nm. This result is similar to the observations made by Lee et al. (2009) on the surface horizon dataset where 1772 and 1805 nm are important 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 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 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).
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 for 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 as predictors and conditions within the Cubist model, respectively. 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).
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. Nonetheless, some of the important regions overlapped.
Although all three methods used different ways to derive important wavelengths, PLSR model tends 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 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 around 1100 and 1700 nm.

5. Conclusion

In this paper, we assess the effective sample size and identify important wavelengths in predicting various soil properties using Cubist and CNN model. In general, the CNN method can perform 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. Here, we found that CNN is more accurate than a machine learning model when the number of samples is above 2000. The more complex and deeper network of a deep learning model, the most likely it will need a larger number of samples for training. PLSR and Cubist models performed less accurate than the CNN model as sample size increases, and it seems like they 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 model ever reached a plateau in accuracy.

Author contributions

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

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.

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.

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.

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.

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. F., 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, 18, Artn 3169 10.3390/S18093169, 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, 2019a.

Padarian, J., Minasny, B., and McBratney, A. B.: Transfer learning to localise a continental soil vis-NIR calibration model, Geoderma, 340, 279-288, 10.1016/j.geoderma.2019.01.009, 2019b.

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.

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.

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., 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.

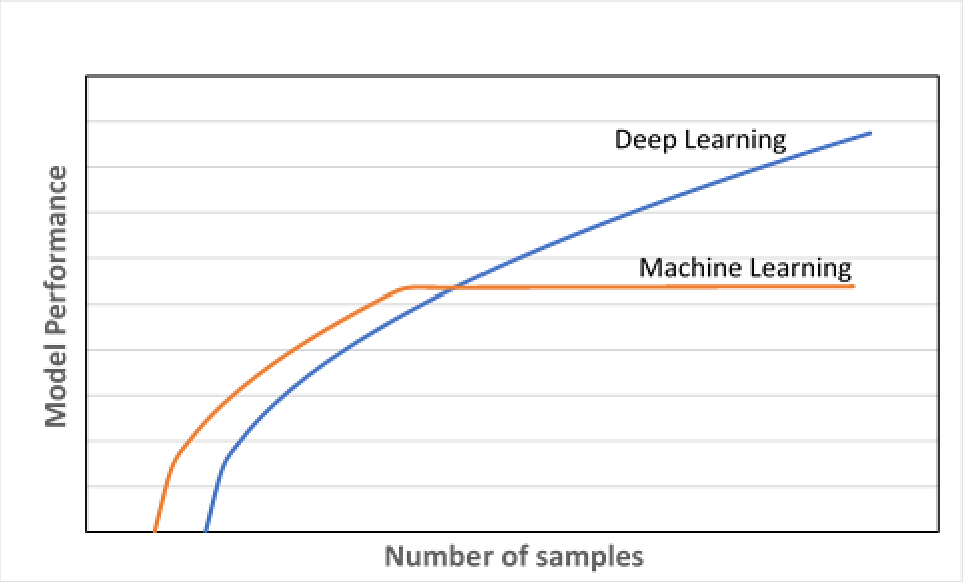


Figure 1. Model performance of deep learning vs other machine learning algorithms as a function of number of samples.

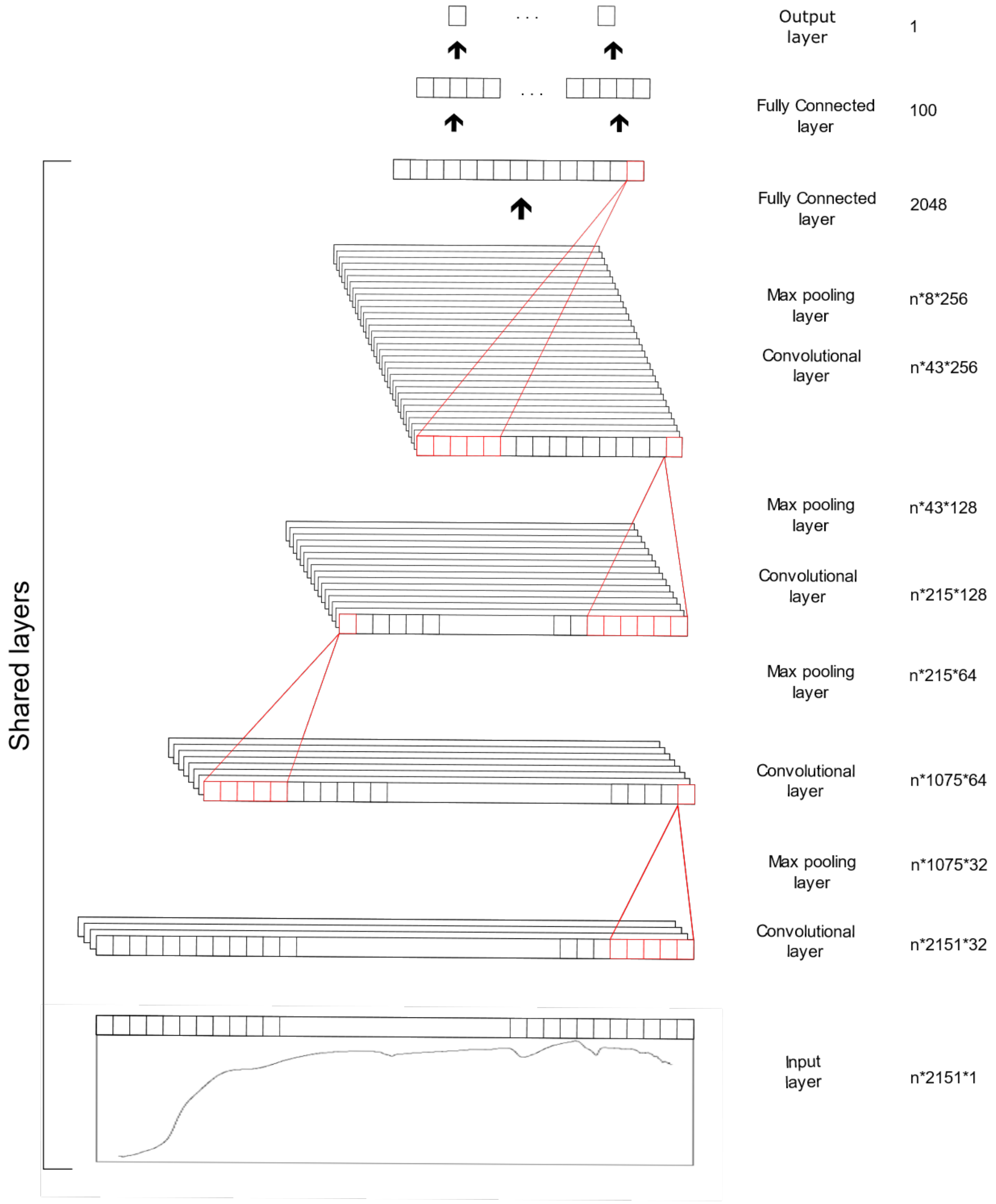


Figure 2. Architecture of the one-dimensional Convolutional Neural Network (CNN) model.

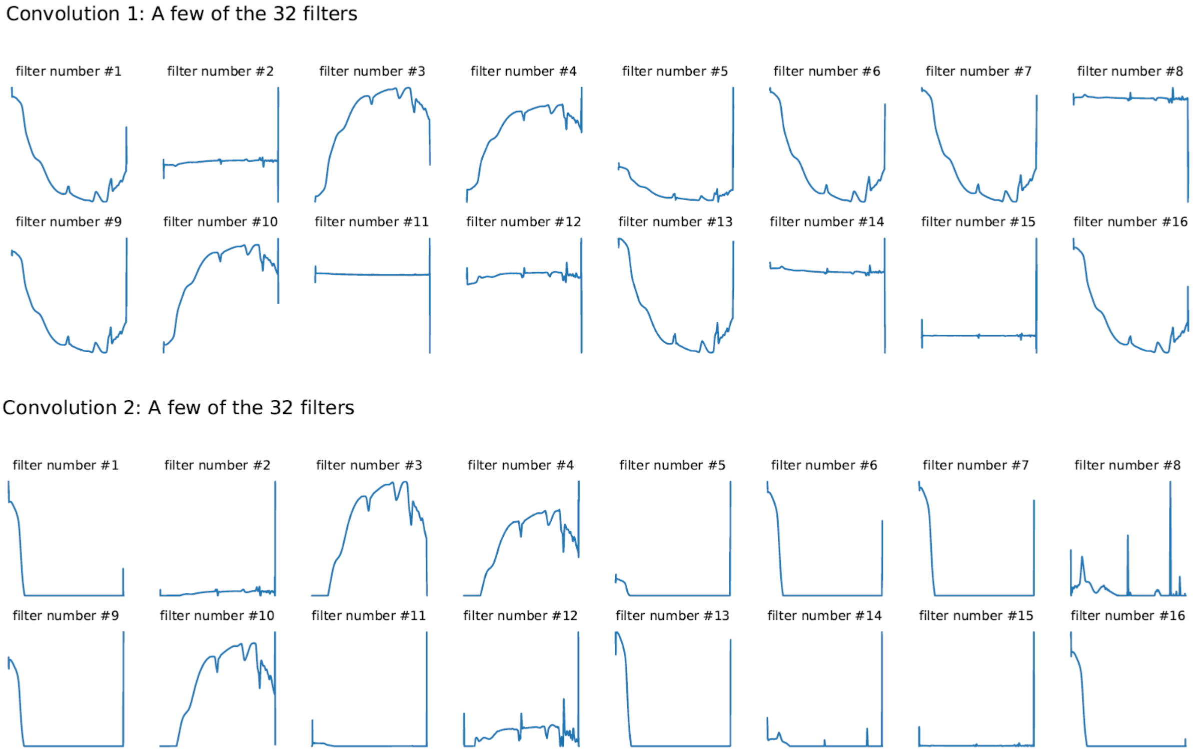
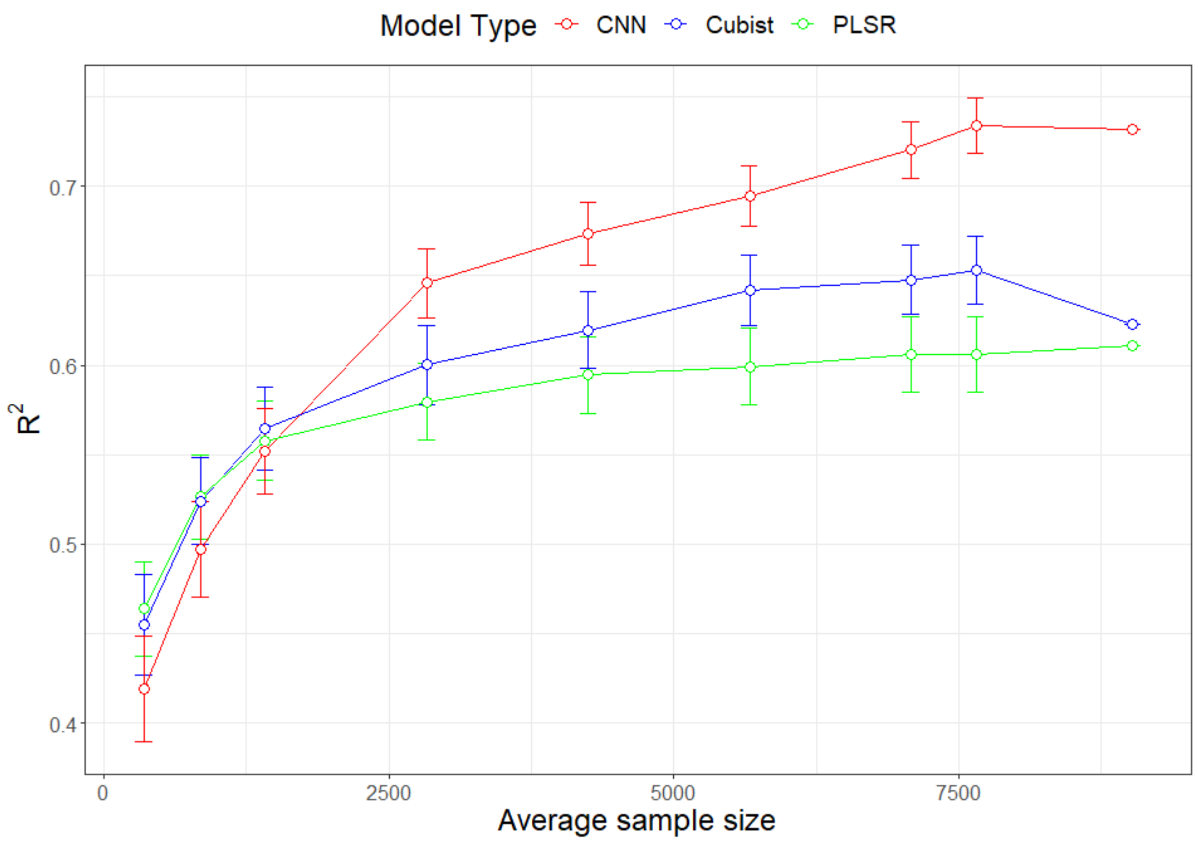


Figure 3. Visualization of the filters within the convolutional layers within Convolutional Neural Network (CNN) with the visible, near, and shortwave infrared (VIS-NIR-SWIR) spectra data.



342

343 **Figure 4. Model performances (in terms of average R^2 for five soil properties) as a function of sample size using Partial Least Squares**
344 **Regression (PLSR), Cubist and Convolutional Neural Network (CNN) model based on ten simulations. The values for the largest**
345 **sample size (n=9027) is a single realization 75% of the data.**

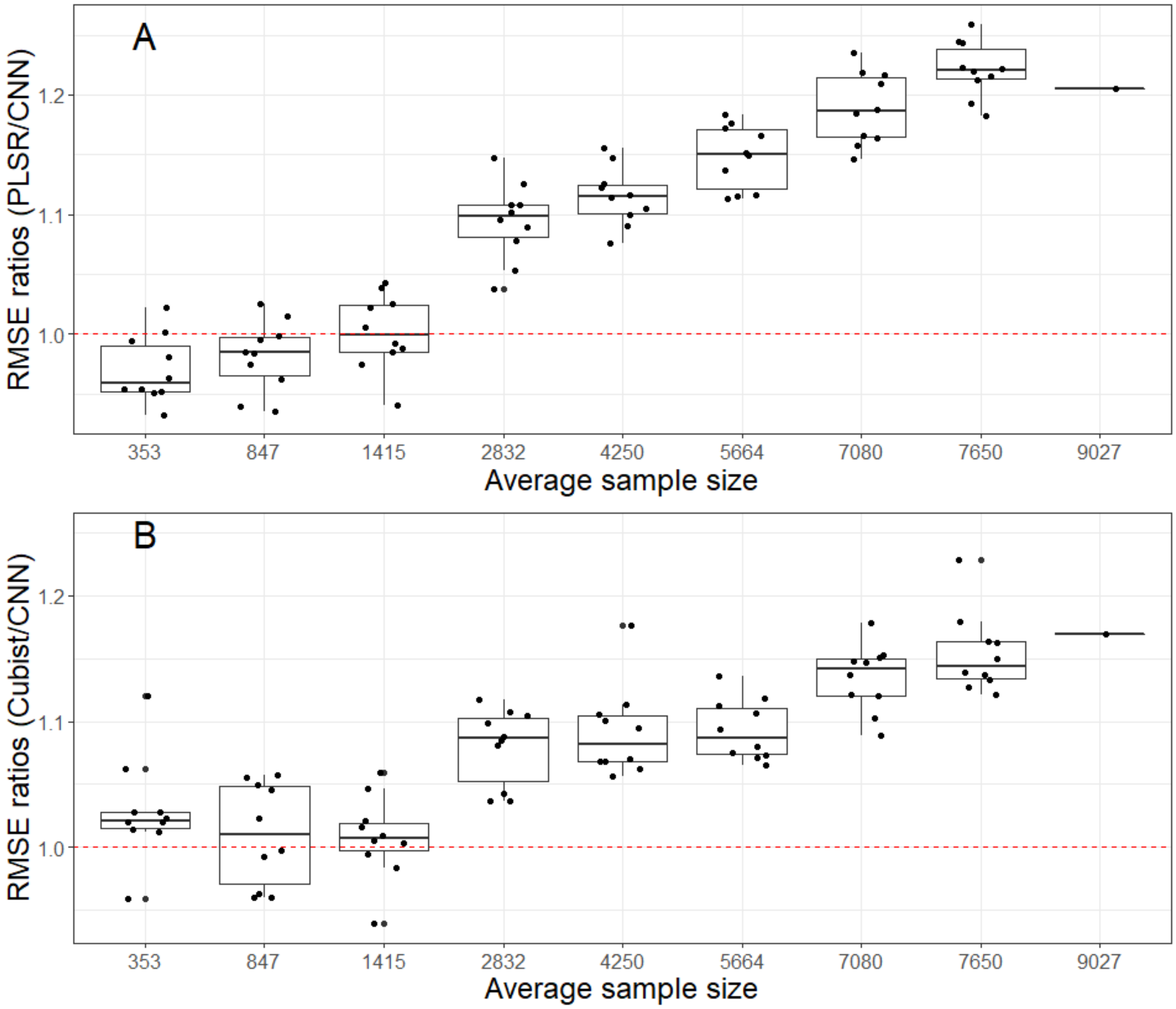
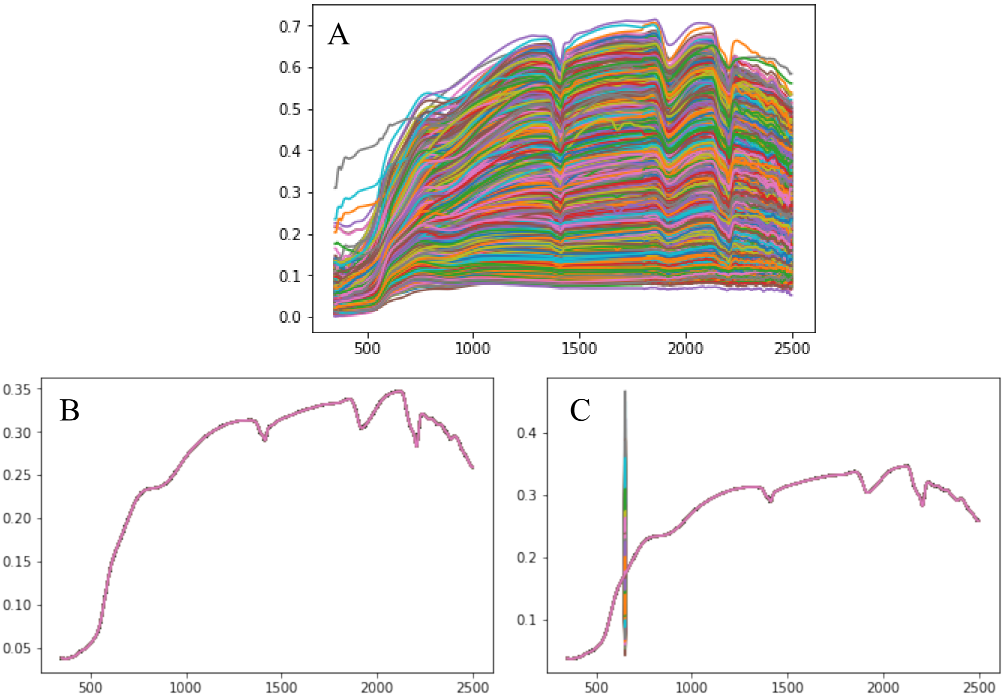


Figure 5. Model performances (in terms of root mean square error (RMSE) ratios of (A) Cubist over Convolutional Neural Network (CNN) model and (B) Partial Least Squares Regression (PLSR) over CNN as an average of five soil properties) based on various sample size using ten simulations. The red – dotted line represents a 1:1 RMSE ratio.

https://doi.org/10.5194/soil-2019-48
Preprint. Discussion started: 17 September 2019



351 **Figure 6. Illustration of sensitivity analysis process: (A) represents the validation spectra data, (B) represents the overall average of**
352 **the validation spectra data and (C) represents the modified average of the validation spectra data.**

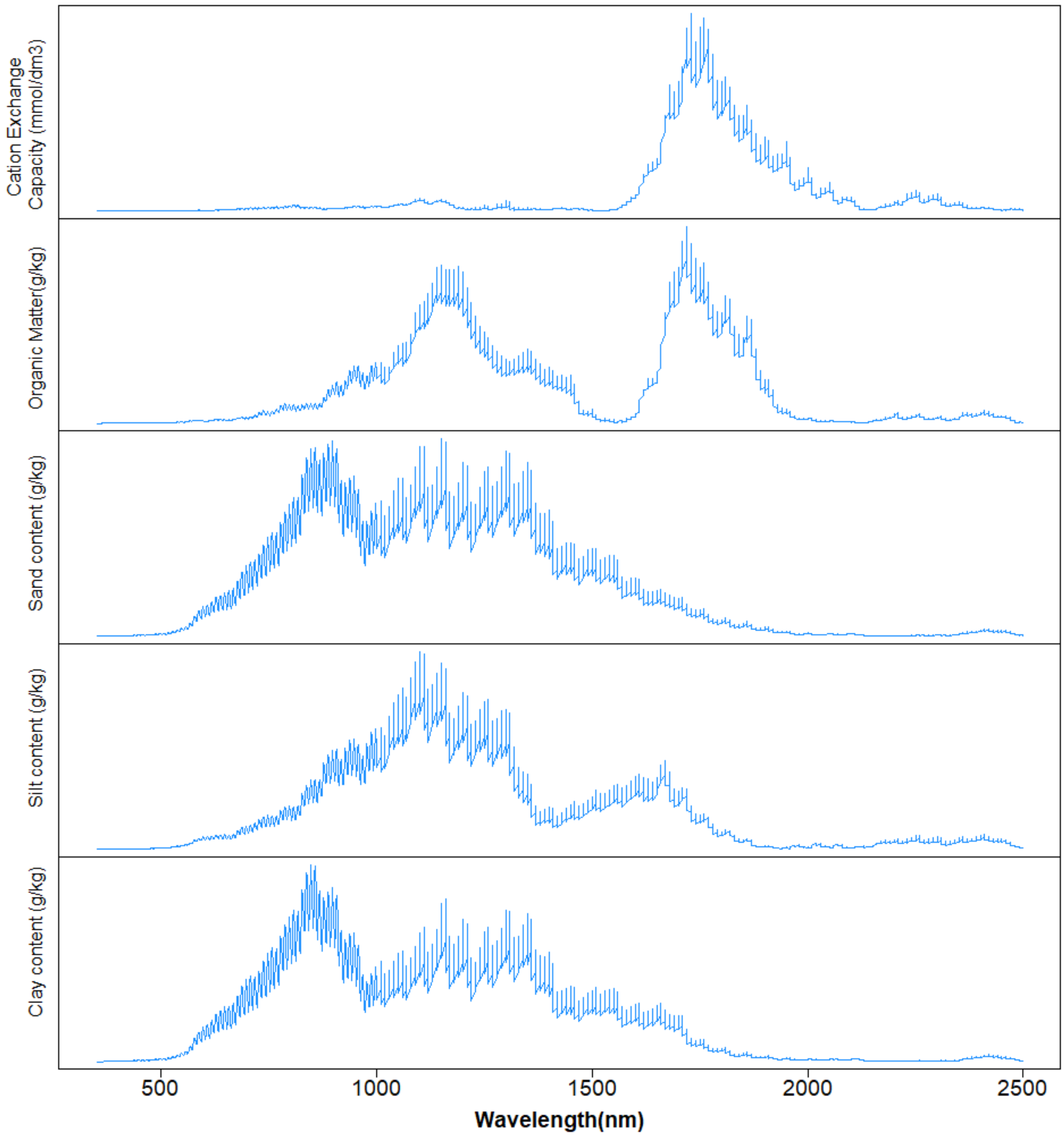
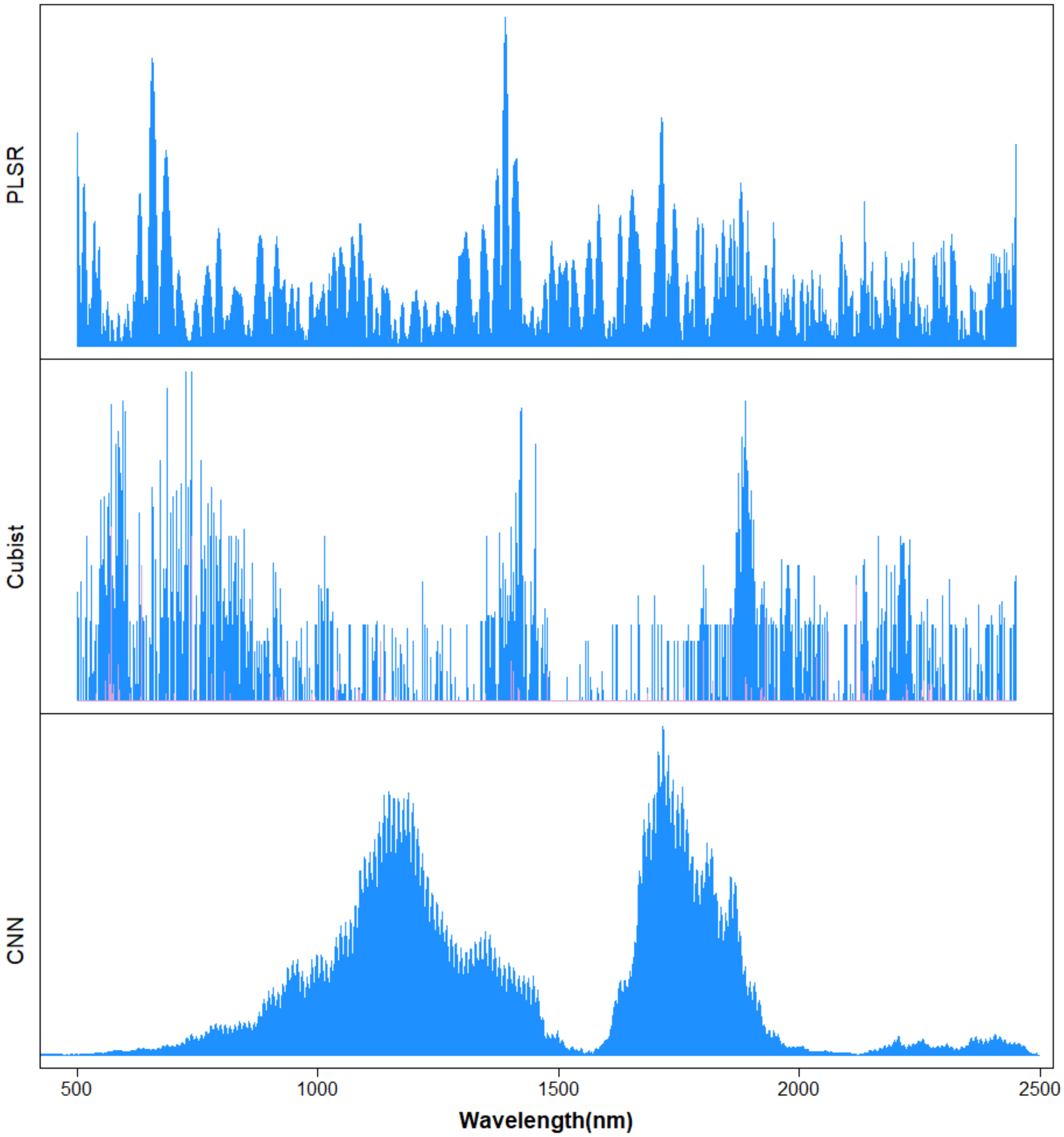


Figure 7: Sensitivity analysis of the visible, near and shortwave infrared (VIS-NIR-SWIR) spectra in predicting various soil properties using the Convolutional Neural Network (CNN) model. The graph depicts sensitivity index (calculated from(Eq. 2)) for different soil properties as a function of wavelength.



358

359 **Figure 8: Important wavelengths for the prediction of organic matter (OM) content using Partial Least Squares Regression (PLSR),**
360 **Cubist and Convolutional Neural Network (CNN) model.**

361

Table 1: Descriptive statistics of the soil properties measurements.

	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

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

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