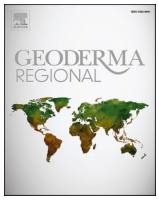
Geoderma Regional 37 (2024) e00805

Contents lists available at ScienceDirect

Geoderma Regional

journal homepage: www.elsevier.com/locate/geodrs



Multivariate regional deep learning prediction of soil properties from near-infrared, mid-infrared and their combined spectra

Rumbidzai W. Nyawasha [a,b,*], Alexandre M.J.-C. Wadoux [c], Pierre Todoroff [d,e], Regis Chikowo [a,f], Gatien N. Falconnier [a,b,e,f], Maeva Lagorsse [d,e], Marc Corbeels [e,g], Rémi Cardinael [a,b,e]

[a] Plant Production Sciences and Technology, University of Zimbabwe, Harare, Zimbabwe
[b] CIRAD, UPR AIDA, Harare, Zimbabwe
[c] LISAH, Univ Montpellier, AgroParisTech, INRAE, IRD, L'Institut Agro, Montpellier, France
[d] CIRAD, UPR AIDA, F-97410 Saint-Pierre, Réunion, France
[e] AIDA, Univ Montpellier, CIRAD, Montpellier, France
[f] International Maize and Wheat Improvement Centre (CIMMYT)-Zimbabwe, Harare, Zimbabwe
[g] IITA, International Institute of Tropical Agriculture, PO Box 30772, Nairobi 00100, Kenya

ARTICLE INFO

Keywords:
Infrared spectroscopy
Neural network
Sub-Saharan Africa
Soil organic carbon
Texture
Keras

ABSTRACT

Artificial neural network (ANN) models have been successfully used in infrared spectroscopy research for the prediction of soil properties. They often show better performance than conventional methods such as partial least squares regression (PLSR). In this paper we develop and evaluate a multivariate extension of ANN for predicting correlated soil properties: total carbon (C), total nitrogen (N), clay, silt, and sand contents, using visible near-infrared (vis-NIR), mid-infrared (MIR) or combined spectra (vis-NIR + MIR). We hypothesize that accounting for the correlation through joint modelling of soil properties with a single model can eliminate "pedological chimera": unrealistic values that may arise when properties are predicted independently such as when calculating ratio or soil texture values. We tested two types of ANN models, a univariate (ANN-UV) and a multivariate model (ANN-MV), using a dataset of 228 soil samples collected from Murehwa district in Zimbabwe at two soil depth intervals (0–20 and 20–40 cm). The models were compared with results from a univariate PLSR (PLSR-UV) model. We found that the multivariate ANN model was better at conserving the observed correlations between properties and consequently gave realistic soil C:N and C:Clay ratios, but that there was no improvement in prediction accuracy over using a univariate model (ANN or PLSR). The use of combined spectra (vis-NIR + MIR) did not make any significant improvements in prediction accuracy of the multivariate ANN model compared to using the vis-NIR or MIR only. We conclude that the multivariate ANN model is better suited for the prediction of multiple correlated soil properties and that it is flexible and can account for compositional constrains. The multivariate ANN model helps to keep realistic ratio values – with strong implications for assessment studies that make use of such predicted soil values.

1. Introduction

Soils play a vital role in nourishing life on earth, supporting food production and essential support services that are crucial for human well-being (Lal, 2016). A soil is characterized by its physical, biological, and chemical properties, such as pH, nutrients, and soil organic carbon (SOC) contents. Soils and soil properties vary over space in relation to the parent material, climate, topography, among others, and change over time in response to natural processes and human activities (Jenny, 1994; Beillouin et al., 2023). Sampling and monitoring of soils is costly and time consuming, as it usually requires a large number of measurements and laboratory analyses (Webster and Lark, 2013). To adequately capture the spatial and temporal variations of soils, effective and less costly methods of data collection and analysis have been developed, including the use of visible and near-infrared (vis-NIR) and mid-infrared (MIR) spectroscopy (Nocita et al., 2015).

* Corresponding author at: Plant Production Sciences and Technology, University of Zimbabwe, Harare, Zimbabwe.
E-mail address: rumbidzaiwnyawasha@yahoo.com (R.W. Nyawasha).

https://doi.org/10.1016/j.geodrs.2024.e00805
Received 13 November 2023; Received in revised form 2 May 2024; Accepted 6 May 2024
Available online 9 May 2024

Following the collection of vis-NIR or MIR infrared spectra from soil samples, statistical models can be employed to establish a predictive relationship between the spectral characteristics and values of soil properties for which corresponding laboratory measurements are available (Barra et al., 2021). The use of vis-NIR and MIR spectroscopy for soil analysis has been extensively documented in the literature (Janik et al., 2009; Knox et al., 2015; Wijewardane et al., 2018; Cambou et al., 2021) and findings have been summarized in reviews (Janik et al., 1998; Soriano-Disla et al., 2014; Nocita et al., 2015; Barra et al., 2021). A single spectrum contains numerous wavelengths, whose patterns are related to the chemical compounds contained in the soil sample (Wadoux et al., 2021). Different information is obtained depending on the spectral region; peaks found in the vis-NIR are less intense, usually made up of overtones or combinations of fundamental vibrations found in the MIR (Viscarra Rossel et al., 2006). As a result, it has been reported that more accurate predictive models can be built using MIR than vis-NIR (Soriano-Disla et al., 2017), but this strongly depends on the specific soil property in question; properties such as organic C, pH, clay silt and sand contents, or phosphorus are usually more accurately predicted using MIR whereas vis-NIR generally gives better results for exchangeable aluminium and exchangeable potassium (Viscarra Rossel et al., 2006; Johnson et al., 2019).

Partial least squares regression (PLSR) has become the most popular regression model in soil spectroscopy (Viscarra Rossel and Lark, 2009; Soriano-Disla et al., 2014) and has been shown to perform well in many situations (Janik et al., 1998; Viscarra Rossel et al., 2006; Cambou et al., 2016; Allo et al., 2020; Bachion de Santana and Daly, 2022). It is highly versatile, with the ability to handle multicollinearity, reducing data dimensionality and working effectively with small sample sizes (Wold et al., 2001). Usually, each soil property is modelled independently, ignoring the correlations that exist between properties. In cases where multiple dependent properties are predicted, this can result in inconsistent predictions and the occurrence of "pedological chimera" as defined by Lagacherie et al. (2022). For example, previous research has found that prediction accuracy of SOC increased by considering soil texture (Madari et al., 2006) and that ignoring the correlation between SOC and total nitrogen (TN), or cation exchange capacity (CEC) can lead to unrealistic ratio values of the soil properties in question (van der Westhuizen et al., 2023). As a solution, multivariate counterparts of PLSR have been developed, the most common being the PLS2 regression model, a modification of PLSR developed by Wold et al. (1983) and Martens and Naes (1987). PLS2 has several advantages as it enables the prediction of all the dependent variables simultaneously (Vandeginste et al., 1998), explicitly accounting for the correlation among the dependent variables. Inspection of the loadings of the dependent variables also gives useful interpretative information. However, in terms of predictive accuracy, PLS2 usually performs worse than a model fitted for an individual variable. Several studies, (Pedro and Ferreira, 2007; Blanco and Peguero, 2008; Mishra and Passos, 2022), acknowledged that the univariate model gave higher prediction accuracy than PLS2.

Recently, data-driven models and algorithmic tools from the field of machine learning have become popular for predicting soil properties from spectral data (Meza Ramirez et al., 2021). Machine learning algorithms can model complex, nonlinear relationships within the data (Jordan and Mitchell, 2015). Commonly used algorithms in soil spectroscopy are support vector machines (Demattê and da Silva Terra, 2014; Deiss et al., 2020), cubist (Minasny and McBratney, 2008; Clergue et al., 2023), random forest (Viscarra Rossel and Behrens, 2010; McDowell et al., 2012; Wadoux, 2023), and artificial neural networks (ANNs) (Daniel et al., 2003; Wijewardane et al., 2018). The use of ANNs has been successful for soil property prediction and showed better performance than conventional methods such as PLSR in several studies (Daniel et al., 2003; Viscarra Rossel and Behrens, 2010; Ng et al., 2019; Padarian et al., 2019). The main advantages of ANNs over conventional regression models are the ability to extract relevant information in high-dimensional datasets, the modelling of non-linear relationships between spectra and soil properties, and a flexibility in the definition of the algorithm and objective function (Ludwig et al., 2019; Margenot et al., 2020). Despite its flexibility, to date very few studies have attempted to understand whether a multivariate ANN model accounts for the correlations that exist among soil properties, although promising results were found in Mishra and Passos (2022), Ng et al. (2019), and Ramsundar et al. (2015). Ng et al. (2019) tested various implementations of convolutional neural networks, a variant of ANNs, that uses images as inputs, to predict several soil properties simultaneously.

In this paper we develop, further expand, and test the multivariate extension of ANNs for predicting soil properties from their vis-NIR, MIR and combined spectra (vis-NIR + MIR). After model development, we investigate the ability of the multivariate model to predict correlated soil properties, as compared to a model that predicts each property individually. The methodology is tested for total carbon, total nitrogen, sand, silt, and clay contents in soils from Murehwa district located in the sub-humid region of Zimbabwe. We hypothesize that combined modelling of several soil properties can eliminate "pedological chimera" by accounting for the correlations between the properties. The comparison between observed and predicted soil properties from a univariate and a multivariate model is made using vis-NIR, MIR or combined vis-NIR + MIR spectra.

2. Materials and methods

2.1. Study area

The study site is in Murehwa district (17°39′S, 31°47′E), a smallholder farming area situated about 80 km northeast of Harare, the capital city of Zimbabwe. The study site is located about 1300 m above sea level and is situated in Zimbabwe's Agroecological Region II – a zone of high potential for agricultural production (Mugandani et al., 2012). The area receives annual rainfall of between 750 and 1000 mm. The dominant soil type in the district are granitic derived sands (Lixisols) which have inherently low fertility. There are small sporadic areas with more fertile clay soils (Luvisols) resulting from dolerite intrusions (Zingore et al., 2007).

Soil samples were collected in three villages that were randomly selected from Ward 28 of the district. Fifty percent of the households in the three villages were randomly selected to give a total of 183 farming households. Soil samples were collected from all agricultural fields including gardens and fields under fallow for each of the selected households. Soil samples were collected between June and July 2021 at two depths i) 0–20 cm ii) 20–40 cm. Sampling was carried out following a zig-zig transect covering each field, with a sub-sample being collected at 10 m distance using an auger and all the sub-samples were mixed to obtain a composite per field and depth. Through a participatory process involving focus group discussions with key informants from each village as well as transect walks, the common lands that can be used for grazing, or also to collect firewood, litter, and wild fruits, were identified as miombo woodlands, vleis/grasslands (these are seasonal wetlands), gumtree plantations and fallow or abandoned fields (now part of the common grazing area). Common land areas exceeding 1 km in length were sampled by taking a composite sample at every 100 m distance. At this point 10 sub-samples were collected using an auger within a 10 m radius to make a composite sample. A total of 677 georeferenced locations were sampled to give 1354 soil samples, 1046 samples from croplands and 308 samples from common lands. All soils were air dried and sieved through a 2 mm sieve.

2.2. Spectral acquisition

Spectra were acquired at the laboratory of the French Agricultural Research Centre for International Development (CIRAD) in Saint Denis, La Réunion, on all soil samples ground to 200 μm. The MIR spectra were measured using an Agilent 4300 handheld FTIR spectrometer (Agilent

Technologies, Santa Clara, CA). The device head (DRIFT) was placed directly on the ground soil and one spectrum was collected per sample. Each spectrum was the average of 80 internal spectra acquired over a wavenumber range between 650 and 4000 cm^{-1} with a measurement interval of 4 cm^{-1}. A reference spectrum ("Silver" reference) was obtained at the beginning of every spectral acquisition series and subsequently at hourly intervals. Spectral data was recorded as absorbance. The vis-NIR spectra were measured using a LabSpec 5000 (Analytical Spectral Devices, Inc. Boulder, CO, USA) with an optical fibre connected to the internal light (adapted to small sample sizes) over a wavelength of 350–2500 nm and spectral resolution of 3 nm at 700 nm and 10 nm at 1400/2100 nm. Three spectra were collected per sample and later averaged to obtain one spectrum to use for analysis. Spectra data was recorded as reflectance and then log-transformed [log(1/R)] to convert them to absorbance.

2.3. Spectral pre-processing and analysis

Spectral pre-processing was done to ensure the removal of any variations caused by light scattering and to enhance some features within the spectra (Wadoux et al., 2021). The MIR spectra were trimmed to remove the noise at the edges leaving the range between 800 and 4000 cm^{-1}. Firstly, smoothing was done using the Savitzky Golay filter with a window size of 11 and a polynomial order of 2 in the *signal* and *plyer* packages (Signal Developers, 2023) of R (R Core Team, 2023), then to correct for light scattering Standard Normal Variate (SNV) was used and finally resampling was carried out at wavelength of 10 nm.

The vis-NIR spectra were trimmed to 20,000–4080 cm^{-1}. Smoothing was also done using the Savitzky Golay filter with a window size of 11 and a polynomial order of 2 in the *signal* and *plyer* packages of R as well as the SNV to correct for light scattering. Resampling was done at 2 nm.

The MIR and vis-NIR datasets were combined using spectra concatenation to create one dataset ranging between 10,000–800 cm^{-1} and this new dataset was resampled at 8 nm wavelength before analysis (Fig. 1). Smoothing and scatter correction were also done using the same process as for the MIR spectra.

2.4. Laboratory analysis

A subset of 230 soil samples, corresponding to 17% of the total number of samples, was selected for laboratory analysis. This number of samples was determined by striking a balance between the cost of the soil analysis and the quantity required to obtain accurate estimates with spectroscopy. The selection was based on spectra similarity and the most representative spectra were chosen using the Kennard Stone algorithm as implemented in the Unscrambler X 10.5 Software (CAMO Software Inc., Oslo, Norway). Total carbon and total nitrogen were determined by the Dumas elemental dry combustion method using an Elementar VarioMax Cube. Soil texture analysis was done using the hydrometer method following Gee and Bauder (1986). Following the laboratory analyses, two soil samples were identified as outliers and were excluded from further analysis (i.e. they had unrealistic high carbon and nitrogen values). Consequently, a total of 228 samples were used for the model building in the next step. Descriptive statistics of the laboratory analyses are summarized in Table 1.

2.5. Modelling

2.5.1. Artificial neural networks

An ANN model is an interconnected network of numerous processing units called neurons (Hastie et al., 2009). Neurons are grouped together to form a layer; they are further connected to neurons in adjacent layers but not to neurons in the same layer. The number of neurons is a user-defined hyperparameter, and an ANN structure has an input layer, one or more hidden layers, and an output layer. The number of units in the input and output layers is determined by the data, but the hidden layer can be adjusted by the user. The connection strength between two neurons is determined by a parameter called weight complemented by a bias component. Mathematical functions known as activation methods are applied to the weighted sum of inputs in a neuron to introduce non-linearity in the output. These activation functions allow the ANN to model the complex relationship between the inputs and outputs. Several activation functions are available, and, in this study, we used the Rectified Linear Units (ReLU) function, which is known to output zero when a value is negative and keeps the input itself when it is positive. Model training involves finding the set of weights and bias that give the optimal predictions using an objective function as criterion. Optimization algorithms are used to optimize the weights and bias so they minimize the error of prediction. In this study we used a feed forward ANN model, which involves successive feed forward pathways flowing unidirectionally from the input to output layers through several hidden layers (Hastie et al., 2009). The Adaptative Moment Estimation (Adam) logarithm was used to train the model.

Table 1
Summary statistics of the measured properties of the 228 soil samples.

Property	Min	Q_1	Median	Q_3	Mean	Max
Total C (g kg^{-1})	1.85	3.94	6.90	14.38	10.47	47.20
Total N (g kg^{-1})	0.10	0.23	0.39	0.95	0.69	3.52
Sand (%)	28.44	72.29	83.84	88.22	77.76	94.92
Clay (%)	3.08	6.08	9.16	16.10	13.78	62.16
Silt (%)	2.00	5.00	6.70	11.40	8.46	39.40
C:N ratio	9.97	13.53	15.40	18.18	16.87	51.98
C:Clay ratio	0.01	0.04	0.07	0.12	0.09	0.54

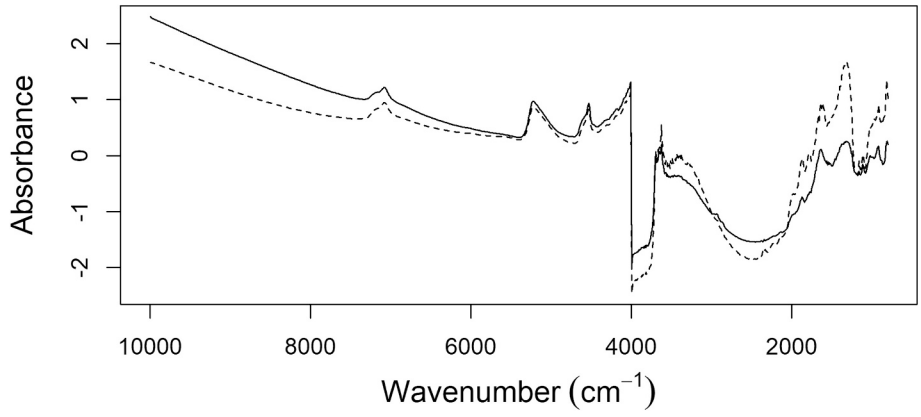


Fig. 1. Example of two combined vis-NIR and MIR spectra used in this study.

2.5.2. Partial least squares regression

PLSR is a commonly used modelling technique to predict a set of dependent variables (y) from a set of predictor variables (X) usually when there are many highly collinear predictor variables (Wold et al., 2001). The PLSR algorithm searches for a set of factors that perform simultaneous decomposition of the predictor and dependent variables whilst maximizing the covariance between them with the aim of finding a few factors that explain most of the variation. Here, X and y are decomposed into factors (T) and factor loadings (P and q) in such a way that the first few factors explain the most variation. The factors that remain can be ignored, that is why residuals E and f are added.

$$X = TP + E$$

$$Y = Tq + f$$

This way, predictions of the dependent variable (Y) can be made using the combination of factor scores and factor loadings of the new spectrum (X) (Viscarra Rossel et al., 2006). In this study we used the orthogonal score algorithm for a single response PLSR model using ten components (Wadoux et al., 2021).

2.6. Model building and fitting

Two types of ANN models were built, a univariate model which predicts one soil property at a time, and a multivariate model which predicts more than one property at the same time (Fig. 2). The univariate model was made up of one input layer, three hidden layers and one output layer. The first two hidden layers had 128 and 64 neurons, respectively, and they were followed by a dropout layer with a dropout rate of 0.2 to help in minimizing issues of overfitting. The next hidden layer had 32 neurons and it was followed by an output layer with a single neuron. The adaptive moment estimation (Adam) optimizer was used to update the model weights. The loss function used was the mean squared error, with a batch size of 64 and a maximum of 150 epochs.

The multivariate model was made up of one input layer, four hidden layers and an output layer predicting five outputs simultaneously. The first two hidden layers had 320 and 128 neurons, respectively, and they were separated by a dropout layer with a dropout rate of 0.2. The subsequent layers had 64 and 32 neurons. For each of the hidden layers the ReLU activation was used. The five output layers had one neuron each and the linear activation was used. The three soil texture fractions sand, clay and silt are reported in percentage values, and they need to add to 100. This was achieved in the model by passing the input vector of the previous layer through a softmax layer which returns an output of similar length with each value ranging between 0 and 1 and the vector adding up to 1 (Wadoux, 2019). Subsequently, a lambda function was then defined to multiply the three outputs by 100. The weights were estimated with the Adam optimizer using the mean absolute percentage error loss function. This function is independent of the units of the soil properties, so that all properties have similar importance during model fitting. The model was trained with a batch size of 64 and 150 epochs. The choice of the number and type of layers and neurons was based on trial and error.

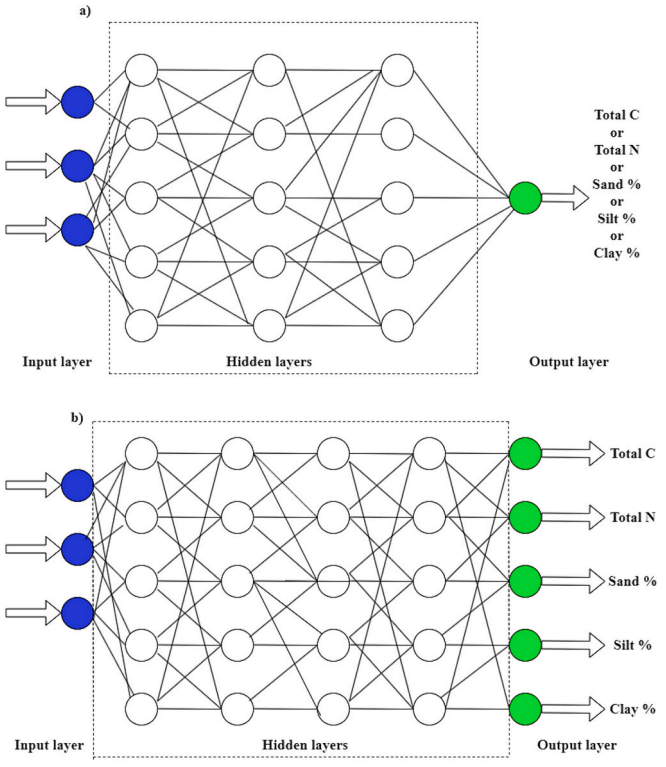


Fig. 2. Schematic representation of the (a) univariate and (b) multivariate models built in this study.

The models were trained using vis-NIR, MIR and the combined vis-NIR + MIR data. The two ANN models were compared to a univariate PLSR model to gauge their performance against a conventional model. The ANN models in this study were built using the *keras* package (Allaire and Chollet, 2023) in R with *tensorflow* as backend (Allaire and Tang, 2023) and the PLSR was built using the *pls* package (Liland et al., 2023) also in R.

2.7. Evaluating the quality of predictions

The measured values of the soil properties from the laboratory analyses used to fit the models were split into training and validation sets using k-fold cross-validation to assess prediction accuracy of the model predictions on unseen data. Ten approximately equal-sized folds were created. Nine folds were used as a calibration set with the remaining fold being used for validation. The procedure was repeated until each of the ten folds had been used once as a validation set. Each validation fold had the predictions computed. The set.seed () function in R was used to ensure the production of consistent and reproducible sets of random numbers. We set the value once at the start of the loop sequence as such each loop sequence was expected to yield the same sequence of random numbers. Doing this, ensured the same validation/calibration sets were used across models and iterations of the cross-validation strategies. The validation statistics hereafter were calculated from the pairwise comparison of measured and predicted values obtained from all folds.

We calculated the mean error (ME), the root mean square error (RMSE) and the coefficient of determination R^2. Each represent a specific aspect of prediction quality. The indices were calculated as follows:

The ME:

$$ME = \frac{1}{n}\sum_{i=1}^{n} obs_i - pred_i \quad (1)$$

Where obs_i and $pred_i$ represent the measured and predicted values, respectively, and n the total number of measured values. The ideal ME value is 0, with positive or negative values indicating systematic over or under prediction, respectively.

The RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{(obs_i - pred_i)^2}{n}} \quad (2)$$

where obs_i are the measured values and $pred_i$ are the predicted values. The RMSE indicates the magnitude of error in the unit of the soil property, it has an optimal value of 0. The R^2, was calculated as:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (3)$$

where RSS is Residual Sum of Squares and TSS is Total Sum of Squares. It has an optimal value of 1 but can be negative if the model is a worse predictor than the mean of the measured values taken as prediction. The RSS and TSS are calculated as follows:

$$RSS = \sum (obs_i - pred_i)^2$$

$$TSS = \sum (obs_i - \overline{obs_i})^2$$

Where:

$\overline{obs_i}$ represents the mean value of the observed data.

3. Results

In section 3.1 we present the prediction results obtained by the three models: univariate PLSR (PLSR-UV), univariate ANN (ANN-UV), multivariate ANN (ANN-MV) using either the vis-NIR, MIR or the combined vis-NIR + MIR spectra.

3.1. Model comparison using vis-NIR, MIR and combined vis-NIR + MIR spectra

The best prediction models were obtained using MIR spectra, followed by vis-NIR + MIR spectra and lastly by vis-NIR spectra (Table 2). Model predictions based on MIR spectra had consistently higher R^2 values and lower RMSE values, and this difference was significant when compared to predictions based on vis-NIR spectra. For example, for prediction of total C with the PLSR-UV model, the R^2 and RMSE values were 0.91 and 2.87 g kg^{-1} vs 0.74 and 4.78 g kg^{-1} for MIR and vis-NIR respectively. However, the difference was smaller, when comparing predictions based on MIR spectra vs combined vis-NIR + MIR spectra, particularly for the predictions of soil sand and clay contents where the differences in R^2 and RMSE were small. For example, the prediction of soil clay content with the PLSR-UV model had R^2 and RMSE values of 0.74 and 5.68% vs 0.73 and 5.82% for MIR spectra and combined vis-NIR + MIR spectra, whilst those for vis-NIR spectra were 0.59 and 7.15%.

The PLSR-UV model performed marginally better than the two ANN models for the prediction of total C, total N, and soil sand and clay contents (Table 2), although in some instances the differences were not significant. To give an example, using MIR spectra, for total carbon predictions, the PLSR-UV model had an R^2 of 0.91 and RMSE of 2.87 g kg^{-1}, whereas for the ANN-MV model the values were 0.89 and 3.09 g kg^{-1}, and for the ANN-UV 0.89 and 3.13 g kg^{-1}. A similar trend was also observed for total nitrogen; the PLSR-UV model had an R^2 of 0.87 and 0.24 g kg^{-1}, followed by the ANN-MV model with 0.85 and 0.26 g kg^{-1}, and the ANN-UV model with 0.83 and 0.28 g kg^{-1}. The same trend was observed for the prediction of soil clay content. For predictions of the soil sand content, the PLSR-UV model performed best with R^2 of 0.8 and RMSE of 6.68%, however the two ANN models had similar R^2 (0.77), the difference between them being that the ANN-MV had a lower RMSE of 7.15% compared to the 7.28% of the ANN-UV model.

The best model for prediction of soil silt content was the ANN-UV model based on vis-NIR + MIR spectra with an R^2 of 0.71 and RMSE of 4.99. The ANN-UV model also performed better using NIR spectra where it had significantly higher R^2 of 0.68. Using MIR spectra, the ANN-MV model had the best results followed by the PLSR-UV model and lastly the ANN-UV model (Table 2).

The results presented hereafter focus on the ANN models developed using MIR spectra as these were providing the best results in nearly all cases (Table 2).

3.2. Prediction residuals plots

Fig. 3 shows a comparison of the residuals plots for both the ANN-UV and ANN-MV models. For the prediction of total C both models show a narrow pattern around the zero line, although the ANN-MV model has an equal distribution of residual values above and below this line unlike the ANN-UV model for which there is a high number of residual values below the zero line. Between 0 and 20 g C kg^{-1} both models show values close to the zero line. As the soil C values increase, however, the scatter of the residuals also increases. A similar pattern is evident for the predictions of total N, where at values between 0 and 1.2 g N kg^{-1} the residuals are scattered close to the zero line, but the scatter increases for larger values. The ANN-UV model also has a significant number of values below the zero line.

For the prediction of soil texture both models show comparable patterns, with most values scattered around the zero line. For the soil clay and silt predictions the scatter increases after 25% and 10% respectively whilst for sand content there is a large scatter between 0 and 70% which decreases as the sand content increases.

3.3. Correlations between soil properties

The ANN-UV and ANN-MV models showed similar patterns of correlation to the measured data for all soil properties using the Pearson's linear correlation coefficient (r) (Fig. 4). There are strong correlations between measured properties: total C vs total N, sand vs clay content as well as silt vs clay content with both the ANN-UV and ANN-MV models capturing these patterns well. On the other hand, there were weaker correlations in the measured data for total C vs clay content ($r = 0.32$) and total N vs clay content ($r = 0.32$), with high linear correlations for

Table 2
Comparison of univariate PLSR (PLSR-UV), univariate (ANN-UV) and multivariate (ANN-MV) neural network models for three spectral data set (i.e. vis-NIR, MIR or combined vis-NIR + MIR spectra using the mean error (ME), root mean square error (RMSE) and the coefficient of determination R^2.

	vis-NIR				MIR				vis-NIR + MIR			
	Model	ME	RMSE	R^2	Model	ME	RMSE	R^2	Model	ME	RMSE	R^2
Total C	PLSR-UV	0.07	4.78	0.74	PLSR-UV	0.09	2.87	0.91	PLSR-UV	0.05	3.99	0.82
	ANN-UV	−0.24	5.41	0.66	ANN-UV	0.52	3.13	0.89	ANN-UV	0.06	4.99	0.71
	ANN-MV	−2.69	6.66	0.49	ANN-MV	−0.79	3.09	0.89	ANN-MV	−1.13	4.23	0.79
Total N	PLSR-UV	0.00	0.35	0.72	PLSR-UV	0.01	0.24	0.87	PLSR-UV	0.01	0.31	0.78
	ANN-UV	0.05	0.43	0.59	ANN-UV	0.01	0.28	0.83	ANN-UV	0.02	0.41	0.64
	ANN-MV	−0.18	0.48	0.48	ANN-MV	−0.09	0.26	0.85	ANN-MV	−0.07	0.31	0.78
Sand	PLSR-UV	0.12	9.19	0.63	PLSR-UV	0.19	6.68	0.80	PLSR-UV	0.05	7.08	0.78
	ANN-UV	−2.86	10.38	0.52	ANN-UV	−1.04	7.28	0.77	ANN-UV	−0.48	9.38	0.61
	ANN-MV	2.63	10.2	0.54	ANN-MV	0.17	7.15	0.77	ANN-MV	1.39	7.72	0.74
Clay	PLSR-UV	−0.07	7.15	0.59	PLSR-UV	−0.07	5.68	0.74	PLSR-UV	−0.03	5.82	0.73
	ANN-UV	−0.13	7.53	0.55	ANN-UV	−0.65	6.18	0.69	ANN-UV	−0.98	6.62	0.65
	ANN-MV	−1.75	7.83	0.51	ANN-MV	−0.08	5.81	0.73	ANN-MV	0.98	5.92	0.72
Silt	PLSR-UV	−0.06	4.15	0.36	PLSR-UV	−0.02	3.49	0.55	PLSR-UV	0.05	3.67	0.50
	ANN-UV	−0.27	5.29	0.68	ANN-UV	−0.56	3.56	0.53	ANN-UV	0.06	4.99	0.71
	ANN-MV	−0.83	4.31	0.31	ANN-MV	−0.04	3.42	0.57	ANN-MV	−0.35	3.71	0.49

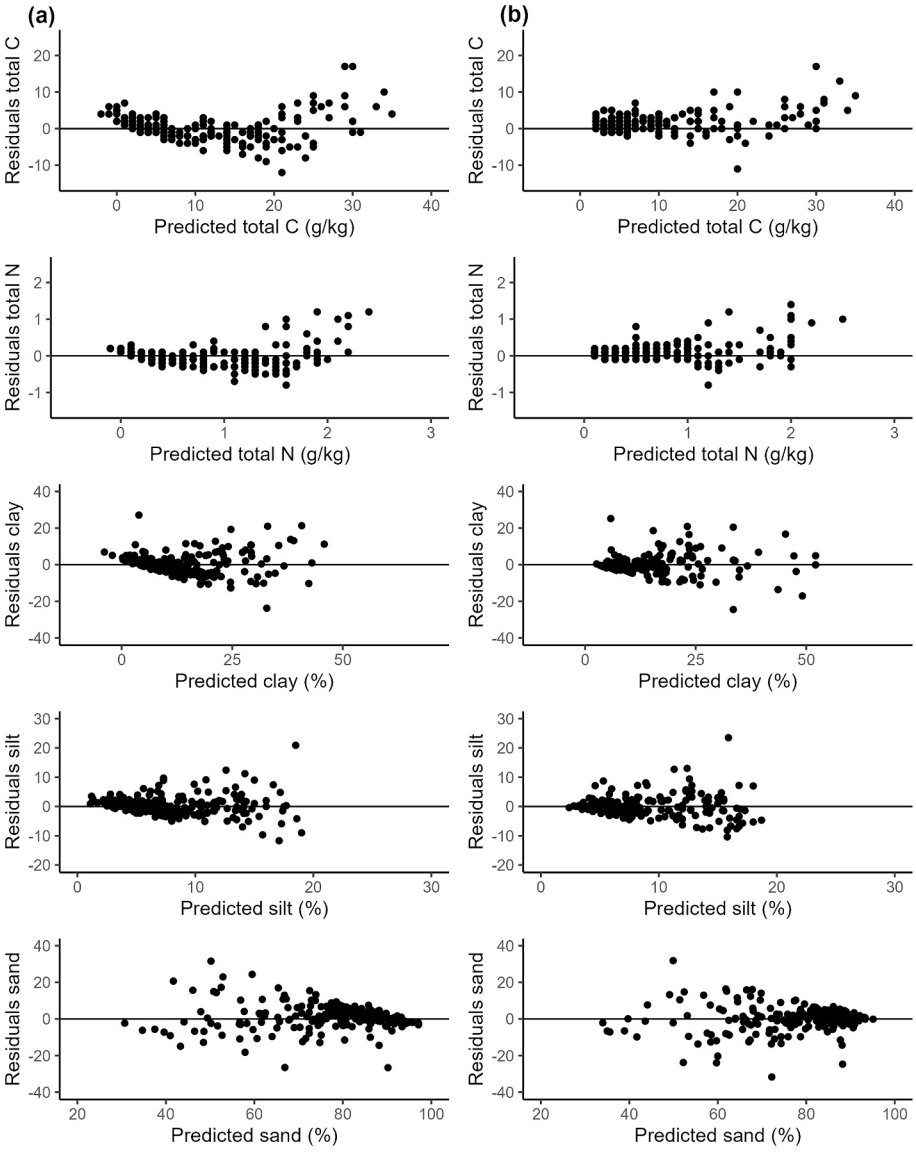


Fig. 3. Residual plots for total C, total N and clay, sand and silt contents as predicted by a) univariate (ANN-UV), and b) multivariate (ANN-MV) artificial neural network models.

both properties at clay values between 0 and 15%. This pattern is similar for both the ANN-UV and ANN-MV models.

Fig. 5 shows the soil C:N and C:Clay ratios for the two ANN models compared to the ratios obtained from the measured values. The soil C:N ratio of the measured values ranges between 10 and 25 with a pattern similar to that of the ratio obtained from the predicted values using the ANN-MV and ANN-UV models. However, there are several large values predicted by the ANN-UV model (i.e. values larger than 75) as well as several negative values. The ANN-MV model better captured the soil C:N ratio with all the values within 10 and 25, showing less variability and fewer large and no unrealistic (i.e. negative) values.

A similar pattern was observed for the C:Clay ratios, where the ANN-UV model had several large values; some higher than 0.5 as well as some values below zero. The ANN-MV model showed less variability, and like the measured data, it had few numbers of large values. Generally, the C:Clay ratios ranged between 0.01 and 0.5 across all the models except for the ANN-UV model.

4. Discussion

4.1. Comparison of vis-NIR, MIR, and vis-NIR + MIR

The best results for the models were observed using the MIR spectra followed by the combined vis-NIR + MIR spectra and lastly the vis-NIR spectra. This result could be explained by the fact that in the MIR region there are fundamental vibrations whereas only overtones and combinations bands are present in the vis-NIR regions. Several other studies report similar results, particularly for soil carbon predictions where MIR outperforms vis-NIR (Viscarra Rossel et al., 2006; Vohland et al., 2014; Wijewardane et al., 2018). The soils used in this study have on average a high sand content (78%) (Table 1). This is a common characteristic in the district, primarily due to the prevalence of granitic-derived soils (Zingore et al., 2011). In their study, Viscarra Rossel et al. (2006) suggest that MIR spectroscopy has a good ability to discriminate quartz and clay minerals thereby allowing for good characterization of soil properties. Indeed, functional groups related to quartz minerals have distinct peaks in the MIR spectra (Janik et al., 1998). This observation could explain the superior results obtained using MIR spectroscopy in our study.

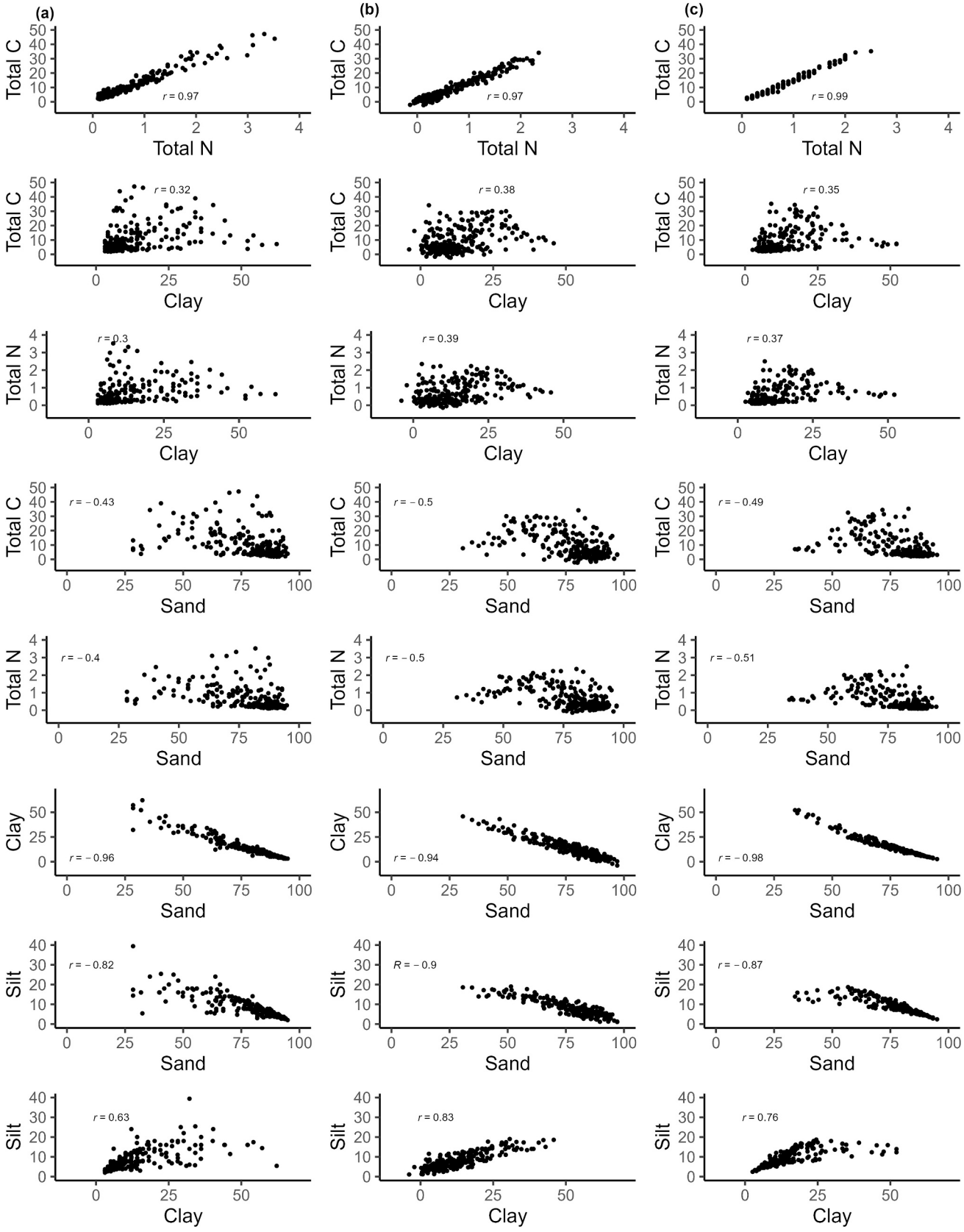


Fig. 4. Scatterplot showing the relationship between total carbon, total nitrogen and clay, sand, and silt contents for a) measured values, and b) the univariate (ANN-UV), c) the multivariate (ANN-MV) artificial neural network models.

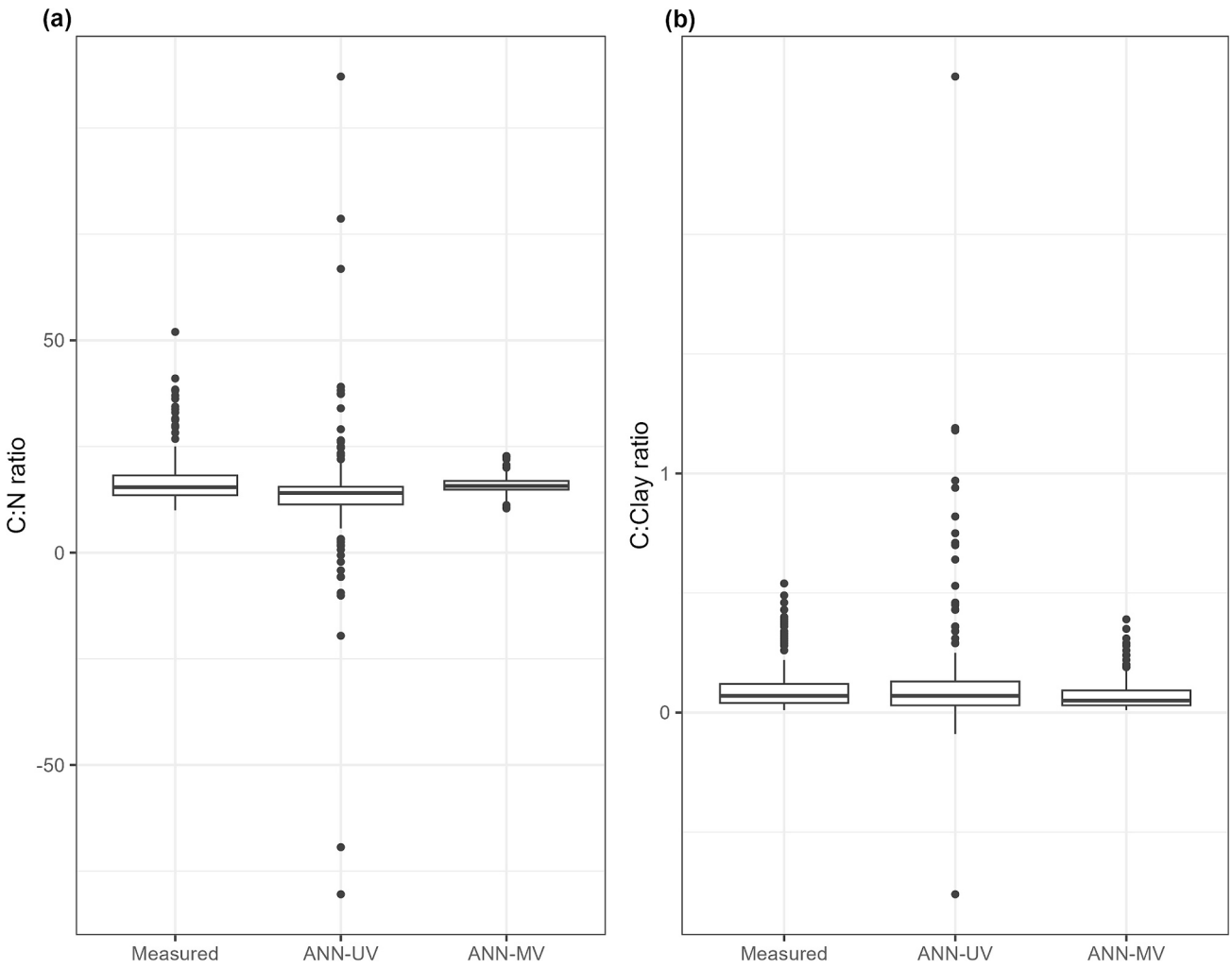


Fig. 5. Boxplots of the (a) soil C:N ratios and (b) C:Clay ratios as calculated with measured values and predicted by the ANN-UV and ANN-MV models.

The use of combined vis-NIR + MIR spectra did not improve the predictive accuracy of soil properties in this study. It is important to state that several studies report different results about this. On the one hand, a study conducted by Johnson et al. (2019) reported an improved accuracy with combined spectra for several soil properties. On the other hand, it has been shown that because the predictions with MIR spectra alone are already highly accurate, combining spectra either results in slightly worse results (Viscarra Rossel et al., 2006; Shao and He, 2011; Ng et al., 2019) or produces results that are equally comparable to MIR alone (Knox et al., 2015).

4.2. Model performance of PLSR vs ANN

The two ANN models developed in this study, namely the univariate and the multivariate models, provided accurate prediction of soil properties and their performance was comparable to the commonly used PLSR model. The PLSR model had slightly higher R^2 values and lower RMSE for the five predicted soil properties. This is similar to the findings by Kuang et al. (2015) and Margenot et al. (2020) who showed that the PLSR and ANN models tend to give comparable predictions in their studies that had a rather small and homogeneous dataset. On the other hand, several other studies have concluded that ANN models outperform conventional models like PLSR (Wijewardane et al., 2018; Ludwig et al., 2019; Margenot et al., 2020). The fact that this was not the case in our study may be attributed to the small dataset (i.e. 228 samples) used to fit the models. Artificial neural networks are popular in soil spectroscopy research because they can deal with complex non-linear patterns found in data, can handle correlated predictors, and perform well in many situations. However, they have many parameters, so that large datasets (n > 20,000) are usually necessary to estimate them (Jordan and Mitchell, 2015). The process of building and training an ANN model is also complex because it involves finding an appropriate structure (i.e. number and types of layers and neurons per layer) which is often a subjective process (Hastie et al., 2009), although some optimization techniques exist such as Bayesian optimization (Wadoux, 2019; Shen and Viscarra Rossel, 2021). Overall, from our study it is not clear whether the time spent, and complexity involved in building an ANN model outweighs the added value of using a non-linear machine learning model.

4.3. Comparison of multivariate and univariate ANN models

The possibility for a model to predict several variables simultaneously is interesting since the soil properties are usually correlated. The ANN-MV model performed better than its univariate version in predictions of total carbon and total nitrogen as it had consistently higher R^2 and lower RMSE values (Table 2). The model performance is also confirmed by the residual plots (Fig. 3) where no atypical behaviour is seen unlike for the univariate model where there seems to be a pattern. The correlations between total C and total N were captured by the multivariate model (Fig. 4) as evidenced by the high linear correlation coefficient (r = 0.99) and the similar patterns of the predicted and

measured values. We can thus deduce that when making simultaneous predictions a multivariate model learns the correlation between the data and uses this information to enhance predictions. A similar conclusion was made in Ng et al. (2019) and Ramsundar et al. (2015). In Ng et al. (2019), the multitask convolutional neural network was able to make simultaneous predictions of multiple outputs whilst maintaining their correlation. In our study, for the prediction of soil texture, particularly for sand and clay, the univariate and multivariate models had comparable results, with similar correlation coefficient between predicted and measured properties, albeit with the multivariate showing lower RMSE values (Table 2). In their study using a multi-task convolutional network model, Ng et al. (2019) reported that although they did not add an explicit accounting of the correlations between properties during their model fitting process, multi-task models were better than single task models at maintaining correlations. This is a particularly important trait especially when considering soil properties like texture, where maintaining accurate percentages of the texture variables is crucial since soil texture plays an important role in the stabilization of organic carbon in a soil (Hassink, 1997; Laub et al., 2023).

Though there exists a multivariate version of PLSR, known as PLS2, we opted not to test it because several studies have shown that the univariate version of PLSR had better predictive capabilities (Vandeginste et al., 1998; Blanco and Peguero, 2008). Brereton (2000) argued that the PLS2 model's ability of providing all predictions in a single model can be advantageous, simplifying computations. However, in many instances the predictions tend to be less accurate than when each variable is independently predicted with PLSR. Here we used the univariate version of PLSR as a benchmark and focused instead on testing the multivariate and univariate versions of the ANN model.

4.4. Estimating ratios and compositions

We studied the predictions of two key ratios: the soil C:N ratio, which is calculated using total carbon and total nitrogen values and is a sensitive indicator of soil quality and for assessing the carbon and nitrogen nutrition balance of soils. The second ratio is the C:Clay ratio, calculated using soil carbon and clay content, which has been proposed as an indicator for soil organic carbon status and soil structure quality (Poeplau and Don, 2023). We observed that the predictions made by the ANN-MV model gave significantly better results for both ratios (Fig. 5). The range of values for the soil C:N ratio were all within the range between 10 and 25, comparable to the measured values, whereas the ANN-UV model gave more unrealistic values including some negative ones. Previous studies in the study area have shown that soil carbon concentrations in the most fertile soils rarely exceed 10 g C kg^{-1} (Masvaya et al., 2010; Zingore et al., 2011).

A similar trend was also observed when calculating the C:Clay ratio, with unrealistic values being predicted by the ANN-UV model including negative values. Soil clay content plays an important role in the formation of soil organic carbon since clay minerals have a high specific surface area and carry a charge, enabling them to bind, and thereby chemically stabilize, organic matter. Clay aggregates also provide micropores for the physical protection of soil organic carbon (Wattel-Koekkoek et al., 2001). The C:Clay ratios obtained in this study range between 1:10–1:13 and sometimes even lower, which suggests that these soils are degraded (Poeplau and Don, 2023). However, it is worth noting that the soils in the study area are generally low in clay and high in sand content (Table 1), as they are granitic derived. When the soil clay plus silt fraction is low, usually little physical protection of organic matter occurs to influence soil physical properties (Feller and Beare, 1997; Nyamangara et al., 2014). Additionally, soil clay content does not consistently serve as an accurate predictor of SOC, particularly in tropical soils that have high concentrations of aluminium and iron oxides (Khomo et al., 2017; Kirsten et al., 2021).

We observed that the ANN-MV model was able to eliminate the problem of unrealistic values of soil C:N and C:Clay ratios making it a better model for predictions of the two ratios. Whilst we predicted the soil properties separately and then calculated the ratios, an alternative method would be to predict the ratios directly from the models, a common approach in digital soil mapping studies (van der Westhuizen et al., 2023). This would require fitting specific parameters during model training to ensure that the results are generated as ratios. Further studies should aim to explore this method to ascertain its feasibility.

Compositional data such as soil texture need to add up to a 100 (Jaconi et al., 2019), and ideally this is done at the prediction phase (i.e. not with *ex-post* correction on the particle size fractions). When sand, silt and clay fractions are predicted independently, as is the case with univariate models, it tends to give results that do not sum to 100. In this study, we observed that the predicted values of sand, silt and clay contents using the univariate model did not sum up to 100 even though the laboratory data did. A common strategy in this case is to use the additive log-ratio transformation (Odeh et al., 2003), where the model is fitted on transformed variables and a back-transformation is made. In our case, we considered the ANN-MV model flexible since we were able to create a constraint for texture prediction as described by Wadoux (2019) (i.e., by passing the input vector through a softmax layer that returns values between 0 and 1, and using a lambda function to multiply these values by 100). This ensures that all soil texture predictions summed to 100, eliminating the need for a subsequent step of back-transformation.

5. Conclusion

We tested two versions of an artificial neural network (ANN) model: a multivariate model which predicts all soil properties at once, and a univariate model, which predicts one soil property at a time. We applied both versions in the prediction of five soil properties, namely, total C, total nitrogen, sand, clay, and silt contents using their vis-NIR, MIR and combined vis-NIR and MIR spectra. The multivariate model had constraints to allow the prediction of compositional variables. We tested the two models in terms of reproduction of correlations between properties and quality of predictions, and through comparison with a univariate partial least-squares model (PLSR). We also tested different sets of input data, using either vis-NIR, MIR and a combination thereof. From the results and discussion, we draw the following conclusions:

- There was no improvement in prediction accuracy when using ANN compared to PLSR. For predicted soil properties, total C, total N, sand and clay content, PLSR-UV had similar prediction accuracy to the two ANN models. Although we used a small dataset and the differences between the two models were not significant, it remains unclear whether time spent, and complexity involved in building an ANN model outweighs the added value of using a non-linear machine learning model.
- The multivariate ANN model produced slightly better results for predictions of nearly all properties compared to its univariate counterpart. It consistently had higher R^2 and low RMSE values and was better at maintaining the correlation patterns observed between the soil properties.
- The multivariate model was also better at giving realistic values for soil C:N and C:Clay ratios, and we found it more flexible as we could add a constraint to have the soil texture predictions add up to 100 without requiring back-transformation.
- Best models were obtained using MIR spectra for all the soil properties and there was no added advantage on the use of combined vis-NIR + MIR spectra. We speculate that the ability of MIR to discriminate quartz and clay minerals allows for good characterization of soil properties, and since these soils are high in quartz there was no advantage of using the combined vis-NIR + MIR spectra.

Overall, we found a clear advantage of using a multivariate neural network for the prediction of correlated soil properties. The multivariate

model helps to keep realistic ratio values and this has strong implications for assessment studies that make use of such predicted soil values.

CRediT authorship contribution statement

Rumbidzai W. Nyawasha: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alexandre M.J.-C. Wadoux:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Pierre Todoroff:** Writing – review & editing, Supervision, Methodology, Investigation. **Regis Chikowo:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Gatien N. Falconnier:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis. **Maeva Lagorsse:** Validation, Methodology, Formal analysis. **Marc Corbeels:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Rémi Cardinael:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Rumbidzai Nyawasha reports financial support was provided by Agropolis Foundation. Rumbidzai Nyawasha reports financial support was provided by Fondation TotalEnergies. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This study was funded by the DSCATT project “Agricultural Intensification and Dynamics of Soil Carbon Sequestration in Tropical and Temperate Farming Systems” (N° AF 1802-001, N° FT C002181), supported by the Agropolis Foundation (“Programme d’Investissement d’Avenir” Labex Agro, ANR-10-LABX- 0001-01) and by the TOTAL Foundation within a patronage agreement. We thank Admire Muwati for his help in soil sampling.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geodrs.2024.e00805.

References

Allaire, J., Chollet, F., 2023. Keras: R Interface to ‘Keras’. R package version 2.13.0.9000. https://tensorflow.rstudio.com/.

Allaire, J., Tang, Y., 2023. tensorflow: R Interface to ‘TensorFlow’. R package version 2.14.0.9000. https://github.com/rstudio/tensorflow.

Allo, M., Todoroff, P., Jameux, M., Stern, M., Paulin, L., Albrecht, A., 2020. Prediction of tropical volcanic soil organic carbon stocks by visible-near- and mid-infrared spectroscopy. Catena 189, 104452. https://doi.org/10.1016/j.catena.2020.104452.

Bachion de Santana, F., Daly, K., 2022. A comparative study of MIR and NIR spectral models using ball-milled and sieved soil for the prediction of a range soil physical and chemical parameters. Spectrochim. Acta Part A 279, 121441. https://doi.org/10.1016/j.saa.2022.121441.

Barra, I., Haefele, S.M., Sakrabani, R., Kebede, F., 2021. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: recent advances–a review. TrAC Trends Anal. Chem. 135, 116166 https://doi.org/10.1016/j.trac.2020.116166.

Beillouin, D., Corbeels, M., Demenois, J., Berre, D., Boyer, A., Fallot, A., Feder, F., Cardinael, R., 2023. A global meta-analysis of soil organic carbon in the Anthropocene. Nat. Commun. 14, 1–10. https://doi.org/10.1038/s41467-023-39338-z.

Blanco, M., Peguero, A., 2008. An expeditious method for determining particle size distribution by near infrared spectroscopy: comparison of PLS2 and ANN models. Talanta 77, 647–651. https://doi.org/10.1016/j.talanta.2008.07.015.

Brereton, R.G., 2000. Introduction to multivariate calibration in analytical chemistry. Analyst 125, 2125–2154. https://doi.org/10.1039/b003805i.

Cambou, A., Cardinael, R., Kouakoua, E., Villeneuve, M., Durand, C., Barthès, B.G., 2016. Prediction of soil organic carbon stock using visible and near infrared reflectance spectroscopy (VNIRS) in the field. Geoderma 261, 151–159. https://doi.org/10.1016/j.geoderma.2015.07.007.

Cambou, A., Allory, V., Cardinael, R., Vieira, L.C., Barthès, B.G., 2021. Comparison of soil organic carbon stocks predicted using visible and near infrared reflectance (VNIR) spectra acquired in situ vs. on sieved dried samples: synthesis of different studies. Soil Secur. 5, 100024 https://doi.org/10.1016/j.soisec.2021.100024.

Clergue, T.C., Saby, N.P.A., Wadoux, A.M.J.-C., Barthès, B.G., Lacoste, M., 2023. Estimating soil aggregate stability with infrared spectroscopy and pedotransfer functions. Soil Secur. 11, 100088 https://doi.org/10.1016/j.soisec.2023.100088.

Daniel, K.W., Tripathi, N.K., Honda, K., 2003. Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). Aust. J. Soil Res. 41, 47–59. https://doi.org/10.1071/SR02027.

Deiss, L., Margenot, A.J., Culman, S.W., Demyan, M.S., 2020. Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. Geoderma 365, 114227. https://doi.org/10.1016/j.geoderma.2020.114227.

Demattê, J.A.M., da Silva Terra, F., 2014. Spectral pedology: a new perspective on evaluation of soils along pedogenetic alterations. Geoderma 217–218, 190–200. https://doi.org/10.1016/j.geoderma.2013.11.012.

Feller, C., Beare, M.H., 1997. Physical control of soil organic matter dynamics in the tropics. Geoderma 79, 69–116. https://doi.org/10.1016/S0016-7061(97)00039-6.

Gee, G.W., Bauder, J.W., 1986. Particle-size analysis. In: Klute, A. (Ed.), Methods of Soil Analysis, Part 1. Physical and Mineralogical Methods-Agronomy. Agronomy Society of America/Soil Science Society of America, Madison, Wisconsin, pp. 384–411.

Hassink, J., 1997. The capacity of soils to preserve organic C and N by their association with clay and silt particles. Plant Soil 191, 77–87. https://doi.org/10.1023/A:1004213929699.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer. https://doi.org/10.1007/978-0-387-84858-7.

Jaconi, A., Vos, C., Don, A., 2019. Near infrared spectroscopy as an easy and precise method to estimate soil texture. Geoderma 337, 906–913. https://doi.org/10.1016/j.geoderma.2018.10.038.

Janik, L.J., Merry, R.H., Skjemstad, J.O., 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? Aust. J. Exp. Agric. 38, 681–696. https://doi.org/10.1071/EA97144.

Janik, L.J., Forrester, S.T., Rawson, A., 2009. The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. Chemom. Intell. Lab. Syst. 97, 179–188. https://doi.org/10.1016/j.chemolab.2009.04.005.

Jenny, H., 1994. Factors of Soil Formation a System of Quantitative Pedology. Courier Corporation.

Johnson, J.M., Vandamme, E., Senthilkumar, K., Sila, A., Shepherd, K.D., Saito, K., 2019. Near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice fields in sub-Saharan Africa. Geoderma 354, 113840. https://doi.org/10.1016/j.geoderma.2019.06.043.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349, 255–260. https://doi.org/10.1126/science.aaa8415.

Khomo, L., Trumbore, S., Bern, C.R., Chadwick, O.A., 2017. Timescales of carbon turnover in soils with mixed crystalline mineralogies. Soil 3, 17–30. https://doi.org/10.5194/soil-3-17-2017.

Kirsten, M., Mikutta, R., Vogel, C., Thompson, A., Mueller, C.W., Kimaro, D.N., Bergsma, H.L.T., Feger, K.H., Kalbitz, K., 2021. Iron oxides and aluminous clays selectively control soil carbon storage and stability in the humid tropics. Sci. Rep. 11, 1–12. https://doi.org/10.1038/s41598-021-84777-7.

Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B., Harris, W.G., 2015. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. Geoderma 239–240, 229–239. https://doi.org/10.1016/j.geoderma.2014.10.019.

Kuang, B., Tekin, Y., Mouazen, A.M., 2015. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. Soil Tillage Res. 146, 243–252. https://doi.org/10.1016/j.still.2014.11.002.

Lagacherie, P., Buis, S., Constantin, J., Dharumarajan, S., Ruiz, L., Sekhar, M., 2022. Evaluating the impact of using digital soil mapping products as input for spatializing a crop model: the case of drainage and maize yield simulated by STICS in the Berambadi catchment (India). Geoderma 406, 115503. https://doi.org/10.1016/j.geoderma.2021.115503.

Lal, R., 2016. Soil health and carbon management. Food Energy Secur. 5, 212–222. https://doi.org/10.1002/fes3.96.

Laub, M., Corbeels, M., Couëdel, A., Ndungu, S.M., Mucheru-Muna, M.W., Mugendi, D., Necpalova, M., Waswa, W., Van De Broek, M., Vanlauwe, B., Six, J., 2023. Managing soil organic carbon in tropical agroecosystems: evidence from four long-term experiments in Kenya. Soil 9, 301–323. https://doi.org/10.5194/soil-9-301-2023.

Liland, K., Mevik, B.-H., Wehrens, R., Hiemstra, P., 2023. pls: Partial Least Squares and Principal Component Regression [WWW Document], 2.8-2. URL. https://github.com/khliland/pls (accessed 8.23.23).

Ludwig, B., Murugan, R., Parama, V.R.R., Vohland, M., 2019. Accuracy of estimating soil properties with mid-infrared spectroscopy: implications of different chemometric approaches and software packages related to calibration sample size. Soil Sci. Soc. Am. J. 83, 1542–1552. https://doi.org/10.2136/sssaj2018.11.0413.

Madari, B.E., Reeves, J.B., Machado, P.L.O.A., Guimarães, C.M., Torres, E., McCarty, G. W., 2006. Mid- and near-infrared spectroscopic assessment of soil compositional parameters and structural indices in two Ferralsols. Geoderma 136, 245–259. https://doi.org/10.1016/j.geoderma.2006.03.026.

Margenot, A., O'Neill, T., Sommer, R., Akella, V., 2020. Predicting soil permanganate oxidizable carbon (POXC) by coupling DRIFT spectroscopy and artificial neural networks (ANN). Comput. Electron. Agric. 168, 105098 https://doi.org/10.1016/j.compag.2019.105098.

Martens, H., Naes, T., 1987. Multivariate calibration by data compression. In: Williams, P., Norris, K. (Eds.), Near-Infrared Technology in the Agricultural and Food Industries. American Association of Cereal Chemists, St. Paul, MN, pp. 57–87.

Masvaya, E.N., Nyamangara, J., Nyawasha, R.W., Zingore, S., Delve, R.J., Giller, K.E., 2010. Effect of farmer management strategies on spatial variability of soil fertility and crop nutrient uptake in contrasting agro-ecological zones in Zimbabwe. Nutr. Cycl. Agroecosyst. 88, 111–120. https://doi.org/10.1007/s10705-009-9262-y.

McDowell, M.L., Bruland, G.L., Deenik, J.L., Grunwald, S., Knox, N.M., 2012. Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. Geoderma 189–190, 312–320. https://doi.org/10.1016/j.geoderma.2012.06.009.

Meza Ramirez, C.A., Greenop, M., Ashton, L., Rehman, I., 2021. Applications of machine learning in spectroscopy. Appl. Spectrosc. Rev. 56, 733–763. https://doi.org/10.1080/05704928.2020.1859525.

Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemom. Intell. Lab. Syst. 94, 72–79. https://doi.org/10.1016/j.chemolab.2008.06.003.

Mishra, P., Passos, D., 2022. Multi-output 1-dimensional convolutional neural networks for simultaneous prediction of different traits of fruit based on near-infrared spectroscopy. Postharvest Biol. Technol. 183, 111741 https://doi.org/10.1016/j.postharvbio.2021.111741.

Mugandani, R., Wuta, M., Makarau, A., Chipindu, B., 2012. Re-classification of agro-ecological regions of Zimbabwe in conformity with climate variability and change. Afr. Crop. Sci. J. 20, 361–369.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. 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.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben, Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E. K., Vargas, R., Wetterlind, J., 2015. Soil spectroscopy: an alternative to wet chemistry for soil monitoring. Adv. Agron. 132, 139–159. https://doi.org/10.1016/bs.agron.2015.02.002.

Nyamangara, J., Marondedze, A., Masvaya, E.N., Mawodza, T., Nyawasha, R., Nyengerai, K., Tirivavi, R., Nyamugafata, P., Wuta, M., 2014. Influence of basin-based conservation agriculture on selected soil quality parameters under smallholder farming in Zimbabwe. Soil Use Manag. 30, 550–559. https://doi.org/10.1111/sum.12149.

Odeh, I.O.A., Todd, A.J., Triantafilis, J., 2003. Spatial prediction of soil particle-size fractions as compositional data. Soil Sci. 168, 501–515. https://doi.org/10.1097/01.ss.0000080335.10341.23.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. Geoderma Reg 16, e00198. https://doi.org/10.1016/j.geodrs.2018.e00198.

Pedro, A.M.K., Ferreira, M.M.C., 2007. Simultaneously calibrating solids, sugars and acidity of tomato products using PLS2 and NIR spectroscopy. Anal. Chim. Acta 595, 221–227. https://doi.org/10.1016/j.aca.2007.03.036.

Poeplau, C., Don, A., 2023. A simple soil organic carbon level metric beyond the organic carbon-to-clay ratio. Soil Use Manag. 39, 1057–1067. https://doi.org/10.1111/sum.12921.

R Core Team, 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing [WWW Document]. URL. https://www.r-project.org.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., Pande, V., 2015. Massively multitask networks for drug discovery. In: International Conference on Machine Learning. https://doi.org/10.48550/arXiv.1502.02072.

Shao, Y., He, Y., 2011. Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy. Soil Res. 49, 166–172. https://doi.org/10.1071/SR10098.

Shen, Z., Viscarra Rossel, R.A., 2021. Automated spectroscopic modelling with optimised convolutional neural networks. Sci. Rep. 11, 208. https://doi.org/10.1038/s41598-020-80486-9.

Signal Developers, 2023. signal: Signal processing [WWW Document]. URL. http://r-forge.r-project.org/projects/signal/ (accessed 8.23.23).

Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., MacDonald, L.M., McLaughlin, M. J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49, 139–186. https://doi.org/10.1080/05704928.2013.811081.

Soriano-Disla, J.M., Janik, L.J., Allen, D.J., McLaughlin, M.J., 2017. Evaluation of the performance of portable visible-infrared instruments for the prediction of soil properties. Biosyst. Eng. 161, 24–36. https://doi.org/10.1016/j.biosystemseng.2017.06.017.

van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P., 2023. Multivariate random forest for digital soil mapping. Geoderma 431, 116365. https://doi.org/10.1016/j.geoderma.2023.116365.

Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., 1998. Multivariate Calibration, Data Handling in Science and Technology. Elsevier Ltd. https://doi.org/10.1016/S0922-3487(98)80046-4

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025.

Viscarra Rossel, R.A., Lark, R.M., 2009. Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. Eur. J. Soil Sci. 60, 453–464. https://doi.org/10.1111/j.1365-2389.2009.01121.x.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007.

Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B., 2014. Determination of soil properties with visible to near- and mid-infrared spectroscopy: effects of spectral variable selection. Geoderma 223–225, 88–96. https://doi.org/10.1016/j.geoderma.2014.01.013.

Wadoux, A.M.J.-C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70. https://doi.org/10.1016/j.geoderma.2019.05.012.

Wadoux, A.M.J.-C., 2023. Interpretable spectroscopic modelling of soil with machine learning. Eur. J. Soil Sci. 74, 1–14. https://doi.org/10.1111/ejss.13370.

Wadoux, A.M.J.-C., Malone, B., Minasny, B., Fajardo, M., McBratney, A.B., 2021. Soil Spectral Inference with R: Analysing Digital Soil Spectra Using the R Programming Environment. Springer International Publishing AG, Cham.

Wattel-Koekkoek, E.J.W., Van Genuchten, P.P.L., Buurman, P., Van Lagen, B., 2001. Amount and composition of clay-associated soil organic matter in a range of kaolinitic and smectitic soils. Geoderma 99, 27–49. https://doi.org/10.1016/S0016-7061(00)00062-8.

Webster, R., Lark, R.M., 2013. Field Sampling for Environmental Science and Management. Routledge, New York. https://doi.org/10.4324/9780203128640.

Wijewardane, N.K., Ge, Y., Wills, S., Libohova, Z., 2018. Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library. Soil Sci. Soc. Am. J. 82, 722–731. https://doi.org/10.2136/sssaj2017.10.0361.

Wold, S., Albano, C., Dunn III, W.J., Esbensen, K., Hellberg, S., Johansson, E., Sjorstrom, S., 1983. Pattern recognition: Finding and using regularities in multivariate data. In: Martens, H., Russwurm Jr., H. (Eds.), Food Research and Data Analysis. Applied Science Publishers, London, pp. 147–188.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Zingore, S., Murwira, H.K., Delve, R.J., Giller, K.E., 2007. Influence of nutrient management strategies on variability of soil fertility, crop yields and nutrient balances on smallholder farms in Zimbabwe. Agric. Ecosyst. Environ. 119, 112–126. https://doi.org/10.1016/j.agee.2006.06.019.

Zingore, S., Tittonell, P., Corbeels, M., van Wijk, M.T., Giller, K.E., 2011. Managing soil fertility diversity to enhance resource use efficiencies in smallholder farming systems: a case from Murewa District. Zimbabwe Nutr. Cycl. Agroecosyst. 90, 87–103. https://doi.org/10.1007/s10705-010-9414-0.