Soil order knowledge as a driver in soil properties estimation from Vis-NIR spectral data – Case study from northern Karnataka (India)

Subramanian Dharumarajan, Cécile Gomez, Manickam Lalitha, Beeman Kalaiselvi, Ramakrishnappa Vasundhara, R. Hegde

Soil order knowledge as a driver in soil properties estimation from Vis-NIR spectral data – Case study from Northern Karnataka (India)

S. Dharumarajan*[1], C. Gomez[2,3], M. Lalitha[1], B. Kalaiselvi[1], R. Vasundhara[1], R. Hegde[1]

[1]ICAR-National Bureau of Soil Survey and Land Use Planning, Regional Centre, Hebbal, Bangalore-560024
[2]LISAH, Univ. Montpellier, IRD, INRAE, Institut Agro Montpellier, France
[3]Indo-French Cell for Water Sciences, IRD, Indian Institute of Science, Bangalore, India
*sdharmag@gmail.com

ABSTRACT

Visible and near-infrared (Vis-NIR, 350-2500 nm) laboratory spectroscopy has been proven to provide soil properties estimations, such as clay or organic carbon (OC). However, the performances of such estimations may be dependent on pedological and spectral similarities between calibration and validation datasets. The objective of this study was to analyse how the soil order knowledge can be used to increase regression models performance for soil properties estimation. For this purpose, Random Forest regression models were calibrated and validated from both regional database (called regional models) and subsets stratified by soil order from the regional database (called soil-order models). The regional database contained 482 soil samples belonging to four soil orders (Alfisols, Vertisols, Inceptisols and Entisols) and associated with Vis-NIR laboratory spectra and six soil properties: OC, sand, silt, clay, cation exchange capacity (CEC) and pH. First, regional models provided i) high accuracy of some soil properties estimations when considering the regional strategy in the validation step (e.g., R^2_{val} of 0.74, 0.76 and 0.74 for clay, CEC and sand, respectively) but ii) modest accuracy of these same soil properties when considering subsets stratified by soil order from the regional database in validation step (e.g., R^2_{val} of 0.48, 0.58 and 0.38 over Vertisol for clay, CEC and sand, respectively). So the estimation accuracy appreciation is highly depending on the validation database as there is a risk of over-appreciated prediction

accuracies at the soil-order scale when figures of merit are based on a regional validation dataset. Second, this work highlighted that the benefit of a soil-order model compared to a regional model for calibration depends on both soil property and soil order. So no recommendations for choosing between both models for calibration may be given. Finally, while Vis-NIR laboratory spectroscopy is becoming a popular way to estimate soil physicochemical properties worldwide, this work highlights that this technique may be used discreetly depending on the targeted scale and targeted soil type.

Key words: Visible Near-infrared, regional model, soil-order model, random forest, soil variability, prediction accuracy

1. Introduction

Visible and near-infrared (Vis-NIR, 350–2500 nm) laboratory spectroscopy provides a complementary method to wet chemistry methods for estimating soil properties (e.g., Viscarra Rossel et al., 2006; Demattê et al., 2004; Stenberg et al., 2010; McBride et al., 2022) and is non-destructive, rapid, low-cost, efficient, repeatable and reproducible with an acceptable degree of accuracy. Soil reflectance in the 350–2500 nm spectral region is the result of soil physical, chemical, and mineralogical properties and their compositions (Ben-Dor, 2002; Stenberg et al., 2010) as the soil spectrum is composed of absorption features of chemical constituents (e.g., absorption of OH of water molecules) and overall spectral shape of the physical properties (e.g., texture) (Ben-Dor and Banin, 1995a, 1995b). As explained by Chabrillat et al. (2019) a targeted soil property can be estimated accurately from Vis-NIR data if this targeted property follows the following rules: ‘Rule (1.1) the soil property S_i has a specific spectral signature due to a chemical or physical structure (e.g., OH- ion for clay) or

Rule (1.2) the soil property S_i is correlated with a soil property S_j having a specific spectral signature due to an associated chemical or physical structure (e.g., cation exchange capacity –CEC- correlated with clay content) (Ben-Dor et al., 2002); and additionally, Rule (1.3) the soil property S_i has to have a quite high amount of variability (Gomez et al., 2012a, b)'.

Soil properties are estimated from laboratory Vis-NIR spectroscopy using regression models, such as stepwise multilinear regression (Leone et al., 2012), multivariate adaptive regression splines (Bilgili et al., 2010), memory-based learning (Jaconi et al., 2019; Ng et al., 2022), Partial Least Square Regression (PLSR, Viscarra Rossel and Behrens, 2009; Gupta et al., 2018; Davari et al., 2021), cubist (Viscarra Rossel et al., 2016) and support vector machine (SVM, Stevens et al., 2010; Naibo et al., 2022) and random forest (RF, Hobley and Prater, 2019; Bao et al., 2020; Dharumarajan et al., 2022). Nawar and Mouazen (2019) used the RF model to compare the efficacy of in situ and field Vis-NIR spectroscopy on the estimation of soil properties and confirmed that the RF model could capture maximum variability (R^2=0.65-0.75) under both conditions. Morellos et al. (2016) reported that machine learning techniques, such as RF, are capable of making spectral variable selections more efficiently compared with PLSR. Ghasemi and Tavakoli (2013) studied the performance of the RF algorithm on Vis-NIR spectroscopy with PLSR and nonlinear SVM and concluded that RF performed well and has the potential for modelling linear and nonlinear multivariate calibrations.

For more than two decades, Vis-NIR laboratory spectroscopy has been extensively explored in various pedological contexts and based on these regression models to estimate various soil properties, such as pH (e.g., Shepherd and Walsh, 2002), soil organic carbon (SOC) (e.g., Bellon-Maurel et al., 2011; Hedley et al., 2015), texture or particle size fractions (e.g., Gomez et al., 2008), CEC (e.g., Shepherd and Walsh, 2002), exchangeable bases (e.g., Pinheiro et al., 2017), available nutrients (e.g., Cozzolino and Moron, 2003; Terra et al., 2015) and soil salinity (e.g., Farifteh et al., 2008).

Based on the high potential of this technique, Vis-NIR soil spectral libraries covering different extent (local, regional, country, continental, and global extents) have been developed these later years (Shepherd and Walsh, 2006; Vasques et al., 2008; Stevens et al., 2013; Viscarra Rossel et al., 2016). Large soil spectral libraries contain information from a wide variety of soils and benefit from a large range of contents for the targeted soil properties and correlations between soil properties, but they rarely reflect local specificities (Stevens et al., 2013; Gogé et al., 2014) unless they include a high density of spatial sampling (Viscarra Rossel et al., 2016). Numerous studies showed that estimations of soil properties over local areas using a large library can be improved by selecting an appropriate "local" subset from the large library to be used in the calibration step (Zeng et al., 2016). Several ways have been developed to build an "appropriate local subset" based on large libraries and calibrate regression models, such as considering calibration datasets constituted a subset of the large libraries based on i) the geographical locations which have to be close to the validation subset (e.g., Guerrero et al., 2010; Shi et al., 2015), ii) their spectral similarity with the local spectra (e.g., Wetterlind and Stenberg, 2010, Gogé et al., 2012; Nocita et al., 2014) or iii) environmental covariates similar to one of the local targeted samples, such as parent material (e.g., Peng et al., 2013; Xu et al., 2016) and land use type (e.g., Zeng et al., 2016). An additional procedure, called "spiking", considered calibration datasets constituted by both the large library and a subset of local samples (e.g., Brown, 2007; Sankey et al., 2008; Nawar and Mouazen, 2017).

While some studies have highlighted that local models (e.g., based on land use, parent material or soil groups) may outperform regional models (e.g., Vasques et al., 2010; Liu et al., 2018), the literature also contains studies showing that local models may not exhibit any advantages over regional models (e.g., Madari et al., 2005; McDowell et al., 2012). For example, Zeng et al. (2016) obtained better soil organic matter predictions for uplands based on local models (using calibration data restricted to land use types or spectral similarity) in comparison with regional models (using calibration data from a regional spectral library); inversely, they obtained better performances for paddy lands based on "regional" models compared to local models. Gomez and Coulouma (2018) showed that prediction models built at a regional database yielded good performances when they were validated at the same regional extent but poor to good performances when they were validated at a local extent (within-field in their case), depending on the model robustness.

In this context, the objective of this study was to analyze how the soil order knowledge can be used to increase regression models performance for soil properties estimation. Models were calibrated and validated from both regional database (regional model) and subsets stratified by soil order from the regional database (soil-order model). This work used a soil spectral library composed of 482 soil samples collected from the northern Karnataka Plateau in India, which is characterized by four soil orders.

2. Materials and methods

2.1. Study area

The study area extends across seven sub-watersheds belonging to five districts of Karnataka (Gulbarga, Koppal, Yadgir, Bidar and Gadag, Table 1) representing the northern Karnataka Plateau region (Fig. 1). These sub-watersheds cover an area from 1603 ha to 68131 ha. They experience semiarid climatic conditions with average annual rainfall and temperature of 633-866 mm and 22-33° C, respectively and is considered drought-prone. With the exception of August and September, the potential evapotranspiration exceeds the rainfall occurrence throughout the year. Predominantly, the seven sub-watersheds have the geology of the peninsular gneiss, basalt and schists. The length of the growing period across the studied area varied from <90 days for the Koppal district to 120-150 days for the Yadgir, Kalburgi, and Gadag districts. The major crops grown in the area are sorghum (*Sorghum bicolor*), maize (*Zea mays L*), cotton (*Gossypium sp*.), sunflower (*Helianthus annuus*), groundnut (*Arachis hypogaea*), red gram (*Cajanus cajan*), mango (*Mangifera indica*), pomegranate (*Punica granatum*), marigold (*Tagetes sp.)* and sapota (*Manilkara zapota*) under rainfed conditions. The sequence of dominant soil orders in the northern Karnataka Plateau is Alfisols, Inceptisols, Vertisols and Entisols (NBSS&LUP, 1998), based on the USDA classification system.

Table. 1. Description of the seven sub-watersheds

District name	Sub-watershed name	Location	Area (ha)	Number of profiles
Gadag	Belhatti	75.63° E 15.31° N 75.58° E 15.24° N	1603	9
	Nilogal	75.69° E 15.13° N 75.58° E 15.02° N	10744	27
Koppal	Kavalur & Gudigere	76.34° E 15.49° N 75.87° E 15.16° N	68131	40
Yadgir	Kilankeri	77.48° E 16.80° N 77.15° E 16.48° N	60106	16
Bidar	Raipalli	77.27° E 17.69° N 77.20° E 17.62° N	3059	31
Gulburga	Sonath	77.10° E 17.67° N 77.02° E 17.59° N	3875	12
	Padsavali	76.49° E 17.62° N 76.42° E 17.57° N	2873	4

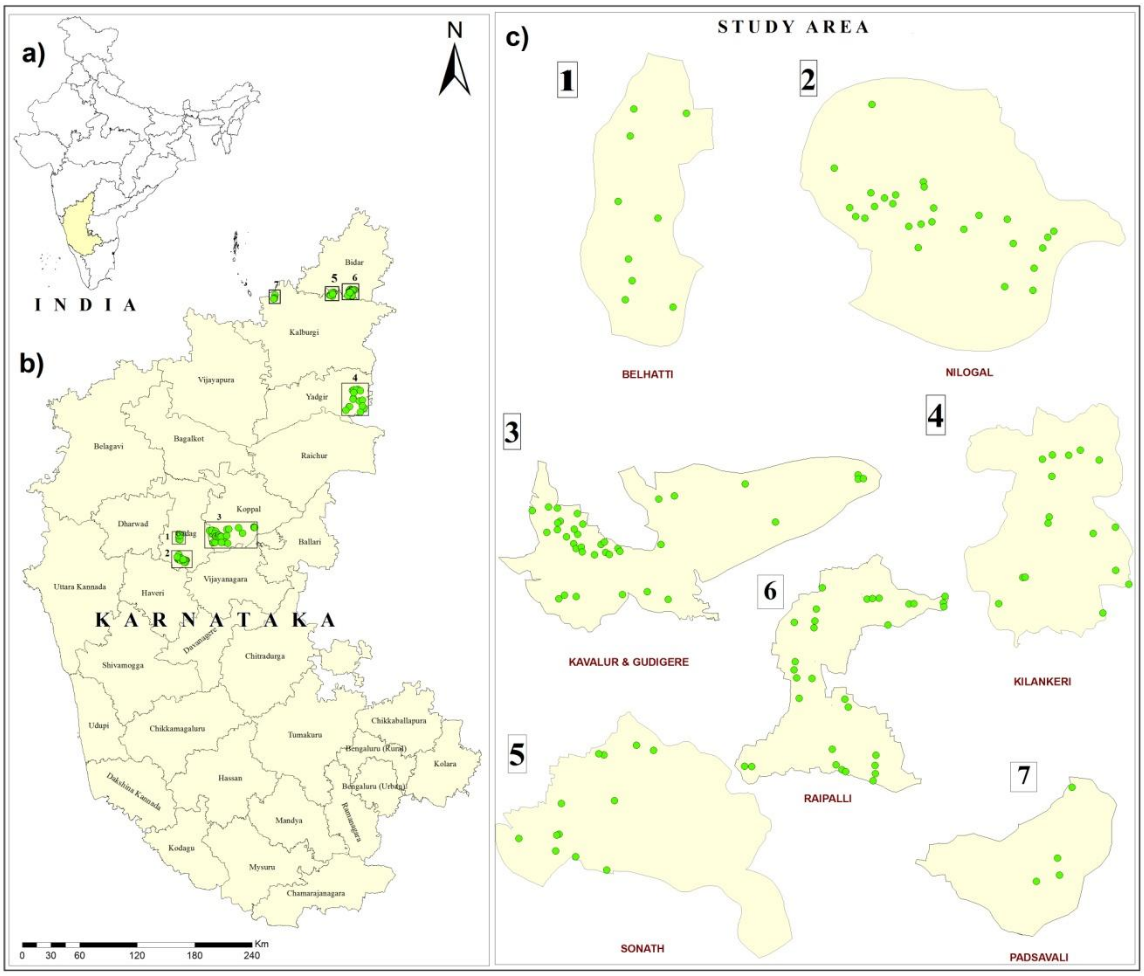

Fig. 1. Location of a) the Karnataka state in India, b) the seven sub-watersheds (black rectangles) over the state of Karnataka and c) the soil profile (green points) over each seven sub-watershed.

2.2 Datasets

Soil profiles collected under the Sujala III project (Hegde et al., 2018) were used for the present study. A total of 139 soil profiles were selected and dug until the hard rock was reached or up to 2 m, whichever occurred first based on the landform, slope and land use variability (Fig. 1b and c). The Belhatti, Nilogal, Kavalur & Gudigere, Kilankeri, Raipalli, Sonath, Padsavali sub-watersheds contain 9, 27, 40, 16, 31, 12, 4 soil profiles, respectively (Table 1) and the number of profiles depends on soil variability in the sub-watershed.

Horizon-wise soil samples (a total of 482 samples) were collected, air-dried, sieved through a 2 mm sieve and analyzed for soil properties. The studied soils were taxonomically grouped into soil orders, namely, Vertisols (20 profiles, 82 samples), Alfisols (59 profiles, 217 samples), Inceptisols (44 profiles, 152 samples) and Entisols (16 profiles, 31 samples), based on their morphological characteristics (Soil survey staff, 2014). Dominant soil characteristics of different soil orders are presented in supplementary information 1.

The samples were analyzed for particle-size distribution by the International Pipette method (Richards, 1954), and OC was estimated by the Walkley and Black (1934) method. Soil pH in 1:2.5 soil : water suspension and cation exchange capacity (CEC) were determined as described by Jackson (1973). The 482 samples constituted the regional dataset, while the samples stratified by soil order constituted four subsets (one subset per soil order). The correlation between soil properties were analysed using Pearson correlation coefficient.

2.3 Spectral data acquisition

An ASD pro-FR Portable Spectroradiometer (Analytical Spectral Devices Inc., Boulder, CO, USA) was used to measure the Vis-NIR spectral data of the soils under laboratory conditions. The processed soil samples (sieved and dried) were illuminated with four tungsten quartz halogen lamps that were fixed at an angle of 36°. The soil spectral reflectance was recorded with a field of view of 8° using a pistol grip. Between 350 and 1000 nm, the spectral sampling interval of the ASD spectrometer was originally 1.4 nm for a spectral resolution of 3 nm, while from 1000 to 2500 nm, the spectral sampling interval was originally 2 nm for a spectral resolution of 10 nm. The reflectance was oversampled by the ASD software to 1 nm in both spectral ranges, leading to a total number of spectral bands of 2151. White reference spectra were measured with a Spectralon® standard white panel after every 5 samples. A representative spectrum for each soil sample was obtained by the mean of measurements of the individual samples in triplicate.

2.4. Preprocessing of spectral data

Spectral data were pre-processed to correct for background effects and light scattering and to omit nonlinearities in the spectra (e.g., Nocita et al., 2013; Babaeian et al., 2015). The spectral absorbance obtained at ranges of 350-400 nm and 2450-2500 nm were removed to eliminate noises. All spectral data were first transformed into pseudo absorbance (log [1/reflectance]) values to achieve linearization between the spectra and soil properties by highlighting the edges of absorption (Stenberg et al., 2010). Then, the Savitzky–Golay filter was applied to eliminate high-frequency noise and pass low-frequency signals to achieve smooth soil spectra (Delwiche, 2010). This filter fits successive subsets (windows) of adjacent data points (7 nm) with a low-degree polynomial through the use of linear least squares.

2.5 Spectroscopic modelling

Random forest regression (RF) was used for soil property predictions from Vis-NIR spectra. The RF regression works on the principle of assemblages of a number of decision trees where random vectors are independently selected and equally distributed among all the trees (Breiman, 2001; Zeraatpisheh et al., 2021). The number of trees (n_{tree}), minimum number of samples at the terminal node n_{min} and the number of predictors used for fitting the tree (M_{try}) are the three parameters that decide the fitting of RF. A Random Forest 4.6 package in an R environment was used for the estimation of soil properties. The RF parameters were optimised using the *tune* function, and the parameters used for running the model are presented in Supplementary Information 2. The accuracy of the model is set by the mean square error (MSE_{OOB}) of the aggregated out-of-bag (OOB) predictions generated from the bootstrap subset and is calculated as follows:

$$\mathrm{MSE}_{OOB} = n^{-1} \sum_{i=1}^{n} (z_i - \hat{z}_i^{OOB})^2 \qquad (1)$$

where n is the number of observations, z_i is the average prediction of the i^{th} observation and $\hat{z}_i^{OOB}$ is the average prediction for the i^{th} observation from all trees for which the observation was OOB.

2.6 Bootstrap procedure

A bootstrap procedure was applied to each dataset (the entire dataset and the four subsets stratified by soil order) to define N sets of calibration and validation subsets, where N is equal to 50 (Efron and Tibshirani, 1993). Bootstrapping involved repeated random sampling for calibration and validation data. Each subset stratified by soil order was divided randomly into thirds; two third of the subset was used for calibration (providing four calibration subsets called *BD_cal_Ver*, *BD_cal_Alf*, *BD_cal_Inc* and *BD_cal_Ent*) and one third of the subset was used for validation (providing four validation subsets called *BD_val_Ver*, *BD_val_Alf*, *BD_val_Inc* and *BD_val_Ent*) (Fig. 2). Then, these four calibration subsets and four validation subsets were aggregated to constitute the *BD_Cal_Regional* dataset containing 328 samples and the *BD_Val_Regional* dataset containing 154 samples, respectively (Fig. 2).

For each bootstrap iteration, a regional RF model was fitted for predicting each soil property, based on the *BD_Cal_Regional* and validated using the *BD_Val_Regional* dataset and the four validation subsets stratified by soil order (*BD_val_Ver*, *BD_val_Alf*, *BD_val_Inc* and *BD_val_Ent*). As well, for each bootstrap iteration, a soil-order RF model for each soil property was built based on each calibration subset stratified by soil order (*BD_cal_Ver*, *BD_cal_Alf*, *BD_cal_Inc* and *BD_cal_Ent*) and validated on the validation data of the same order.

2.7 Model evaluation

The performance of the RF models was evaluated based on the 50 iterations for each validation dataset using four accuracy estimates (Bellon-Maurel et al., 2010), the coefficient of determination (R^2_{val}), root mean square error ($RMSE_{val}$), mean error (ME_{val}), and ratio of performance to interquartile distance ($RPIQ_{val}$), based on the following equations:

$$R^2_{val} = 1 - \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(o_i - \overline{o_i})^2} \quad (2)$$

$$ME_{val} = \frac{1}{n}\sum_{i=1}^{n}(o_i - p_i) \quad (3)$$

$$RMSE_{val} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \quad (4)$$

where p_i and o_i are the predicted and observed values, respectively and $\overline{o_i}$ is the means of the observed values.

$$RPIQ_{val} = \frac{IQ}{RMSE_{val}} \quad (5)$$

where IQ is the difference between the third quartile Q3 and the first quartile Q1. A larger RPIQ value indicates improved model performance. The reliability of the prediction was evaluated based on the RPIQ, for which a RPIQ lower than 1.5 may be consider as a poor performance, RPIQ from 1.5 to 3.0 may be consider as a acceptable performance, and RPIQ up to 3.0 may be consider as a good performance (Veum et al., 2015).

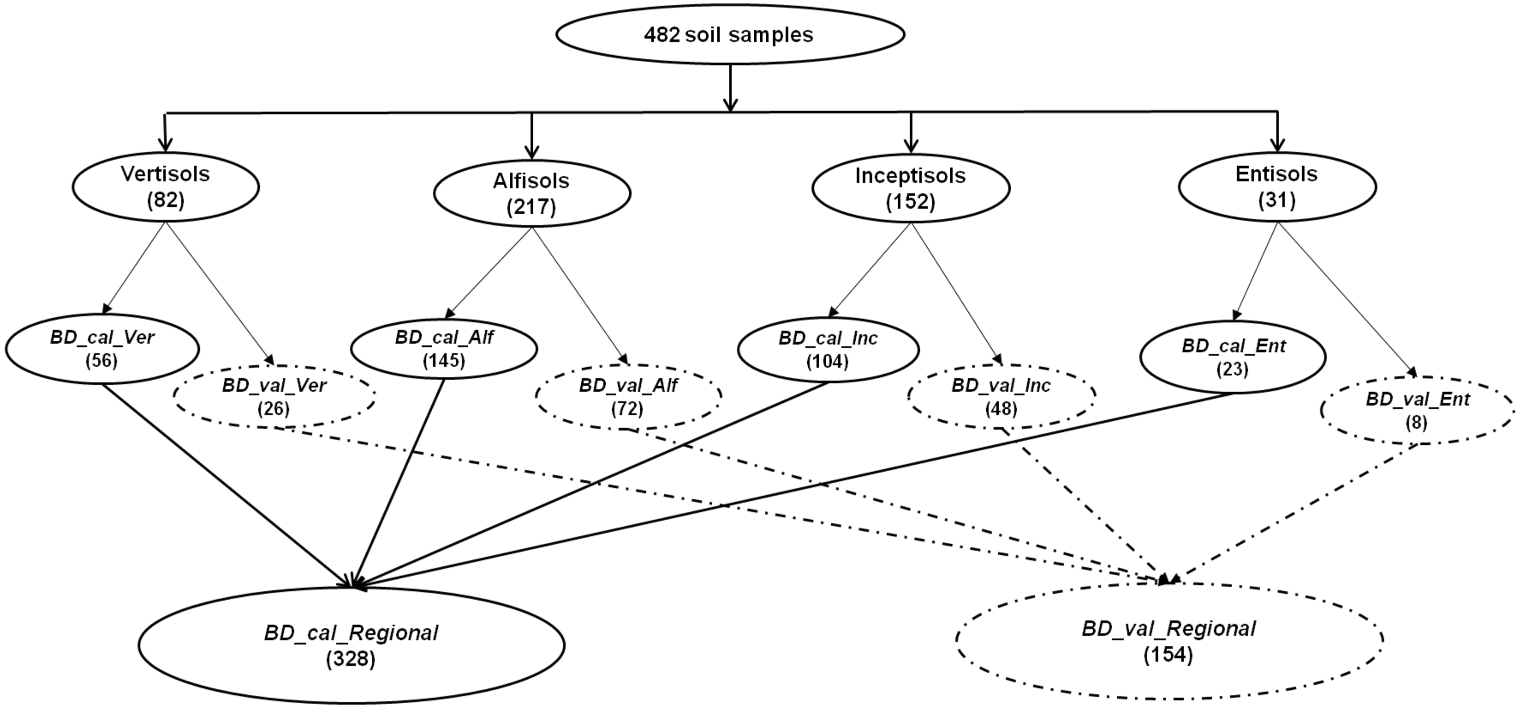

Fig. 2. Construction of calibration and validation datasets for regional and soil-order models (number of samples in parentheses)

3. Results

3.1 Preliminary analysis of soil properties and spectra

3.1.1 Based on the entire dataset

The clay, sand and silt of the entire soil dataset (482 samples) ranged from 1.2 to 77.2%, 2.7 to 93.4% and 2.4 to 39.4%, respectively, with means of 42.8, 40.8 and 16.3%, respectively (Table 2). The soil pH ranged from 4.7 to 11.2 with mean of 8.0. The SOC content ranged from 0.03 to 1.6% with a mean of 0.6%. The mean CEC of the northern Karnataka Plateau soils was 29.5 cmol (+) $kg^{-1,}$ with a 66.4% coefficient of variation.

Based on the entire soil dataset, clay had a high negative correlation with sand (*r* = -0.95), a high positive correlation with CEC (*r* = 0.71) and a modest correlation with silt (*r* = 0.42) (Supplementary Information 3). Sand had a high negative correlation with silt (*r* = -0.68). CEC had a positive correlation with silt (*r* = 0.64) and a negative correlation with sand (*r* = -0.79). Finally, no correlations existed between the other properties of the overall soil dataset. The sand content was positively correlated with the average reflectance along the Vis-NIR spectral range, while the clay content was negatively correlated with the average reflectance along the Vis-NIR spectral range (Fig. 3). The CEC and silt content also followed correlation patterns similar to clay along the Vis-NIR spectral range. Finally, there was no significant correlation between pH and OC with the average reflectance.

Table 2. Statistical summary of soil properties for the entire dataset and each subset stratified per soil order.

		sand (%)	silt (%)	clay (%)	pH	SOC (%)	CEC (cmol (+) kg^{-1})
Entire samples (N=482)	Min	2.7	2.4	1.2	4.7	0.03	1.7
	Max	93.4	39.4	77.2	11.2	1.60	80.9
	Mean	40.8	16.3	42.8	8.0	0.58	29.5
	SD	23.4	7.9	19.0	1.1	0.27	19.6
	CV (%)	57.4	48.5	44.4	13.8	47.5	66.4
Vertisols (N=82)	Min	2.7	10.5	37.7	6.7	0.16	10.2
	Max	51.8	36.5	77.2	9.5	1.29	80.9
	Mean	15.5	22.1	62.4	8.5	0.59	51.5
	SD	10.6	5.2	8.75	0.6	0.25	17.9
	CV (%)	68.4	23.4	14.0	6.6	43.1	34.8
Alfisols (N=217)	Min	6.3	2.4	2.3	4.7	0.12	1.7
	Max	93.4	34.5	76.1	9.9	1.55	54.0
	Mean	49.3	12.0	38.6	7.5	0.56	18.1
	SD	19.3	6.2	17.7	1.1	0.26	9.8
	CV (%)	39.2	51.6	45.9	14.7	46.2	54.1
Inceptisols (N=152)	Min	3.2	3.8	4.6	5.4	0.08	3.4
	Max	88.4	37.9	73.3	11.2	1.26	80.4
	Mean	39.8	19.2	41.1	8.6	0.55	35.4
	SD	22.5	7.2	18.0	1.0	0.28	18.8
	CV (%)	56.5	37.5	43.8	11.6	51.0	53.1
Entisols (N=31)	Min	9.97	2.6	1.2	6.0	0.03	2.03
	Max	94.0	39.4	58.3	8.7	1.60	51.9
	Mean	53.2	17.6	29.1	7.6	0.61	21.5
	SD	28.0	10.6	18.3	0.8	0.33	16.7
	CV (%)	52.8	60.2	62.9	10.7	52.4	77.7

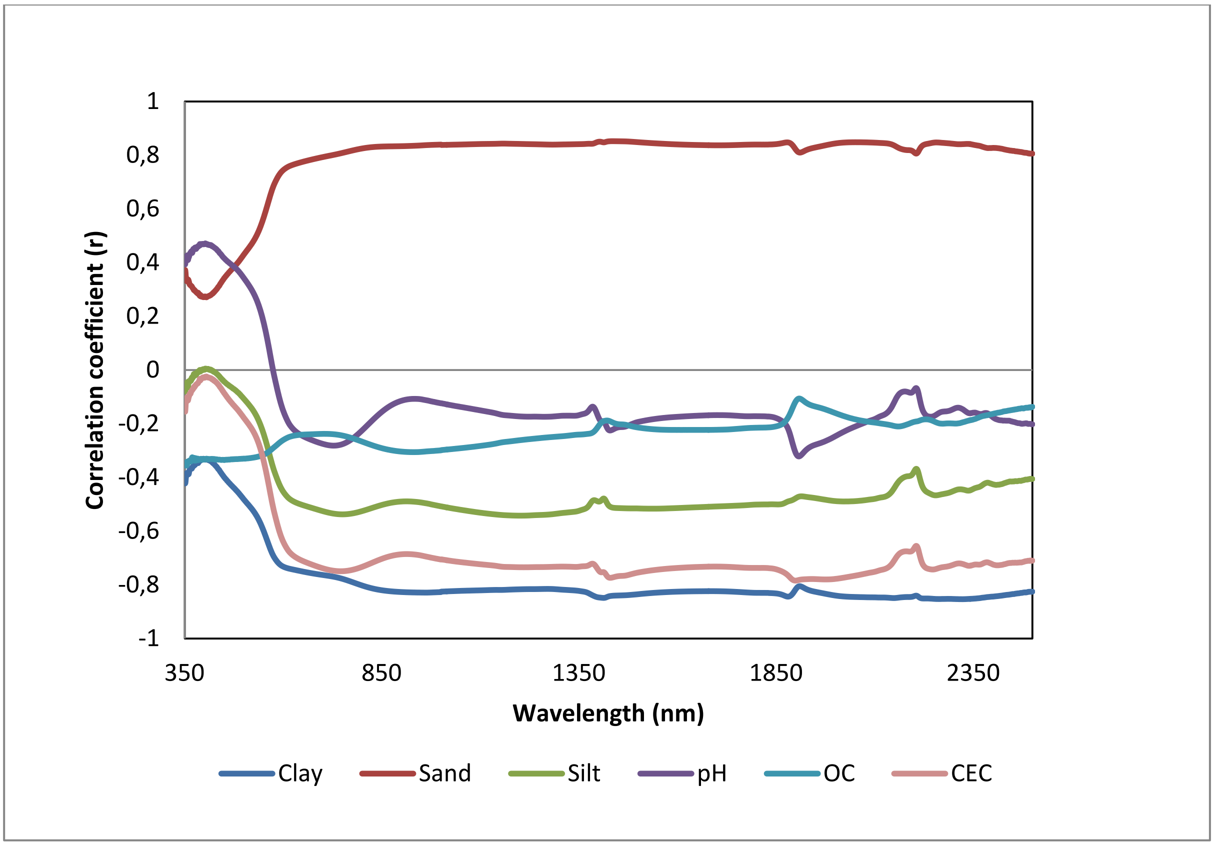

Fig. 3. Correlation coefficient (*r*) between soil properties and mean reflectance at each wavelength based on the entire dataset.

3.1.2 Based on subsets stratified per soil order

The Vertisols and Inceptisols were characterized by a higher content of clay (mean > 40%), CEC (mean > 35 cmol (+) kg^{-1}) and pH (mean > 8.5) than Alfisols and Entisols (Table 2). The high CEC in Vertisols and Inceptisols may be due to the presence of highly weatherable minerals derived from basaltic parent materials and these soils have abundant 2:1 type clay minerals. The Alfisols and Entisols were characterized by high contents of sand (mean > 49%) and CEC (mean of 18.1 and 21.5 cmol (+) kg^{-1}, respectively). The SOC range and distribution were similar from one soil order to another (Table 2).

Regardless of the soil order, clay had a high negative correlation ($r < -0.87$) with sand (Supplementary Information 4 to 7). Clay and CEC had a high positive correlation in Inceptisols and Entisols ($r > 0.89$) and a modest correlation in Alfisols and Vertisols (r from 0.43 to 0.46). Clay and silt were highly correlated in Entisols ($r = 0.85$), slightly correlated in Inceptisols ($r = 0.50$), and had no correlation in either of the other soil orders. OC and pH had a modest negative correlation in Vertisols and Inceptisols (r of -0.57 and -0.56, respectively) and poor correlations in the other soil orders.

3.1.3 Vis-NIR spectra per soil order

The mean spectra measured for Entisols and Alfisols presented the highest absorption band centred at 2207 nm (Fig. 4), which corresponds to the combination of OH stretching and OH-Al bending modes observed in clay (Chabrillat et al., 2002). Vertisols recorded relatively poor reflectance irrespective of the bandwidth, which might be due to the presence of smectite clay minerals in Vertisols and high moisture-holding capacities (Baumgardner et al., 1985; Demattê et al., 2017). The higher reflectance of Entisols and Alfisols might be attributed to the predominance of highly weatherable minerals (Poppiel et al., 2018) and sand contents (Viscarra Rossel et al., 2006) which may have increased their albedo. Alfisols and Entisols had broad absorption features between 850-1100 nm related to the specific absorption shoulder of goethite and haematite (Srivastava et al., 2004). These particular iron oxide absorption bands were not observed in the reflectance spectra of other soil orders because iron oxides are underdeveloped in Inceptisols and Vertisols (Poppiel et al., 2018).

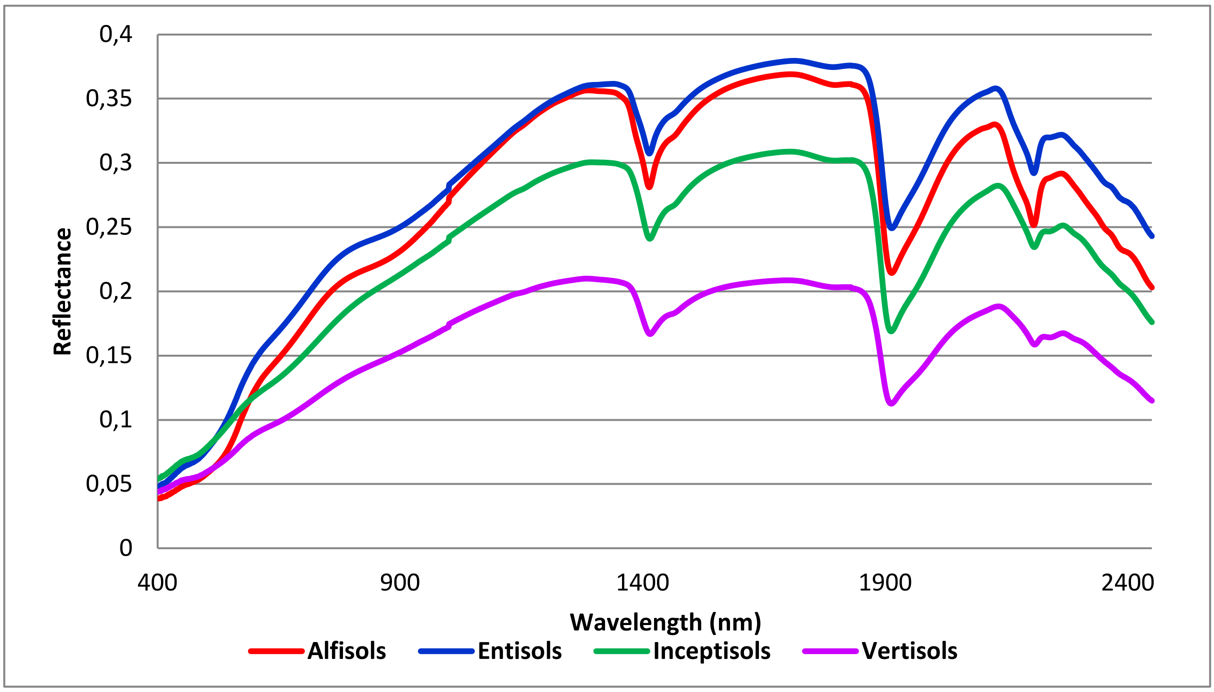

Fig. 4. Mean spectral reflectance of soil samples stratified per soil order.

3.2 Prediction performance of regional models

3.2.1. Analysis based on the entire database

Fifty regional models were built from a *BD_Cal_Regional* dataset for each soil property and validated using a *BD_Val_Regional* dataset. The RF regional models for CEC estimates provided good performances, with R^2_{val} and $RPIQ_{val}$ values of 0.76 and 3.00, respectively (Fig. 5c), as the RF regional models for clay and sand which provided good performances with R^2_{val} values of 0.74 and $RPIQ_{val}$ values of 3.17 and 3.14, respectively (Fig. 5a and b). The RF regional models for silt and pH estimates provided modest performances, with R^2_{val} and $RPIQ_{val}$ values above 0.5 and 1.5, respectively (Fig. 5d and e). Finally, the regional models for SOC estimates yielded poor performances, with R^2_{val} value lower than 0.5 (Fig. 5f). The variations in performances based on 50 iterations (standard deviation) were modest, regardless of the studied soil property (Supplementary Information 8).

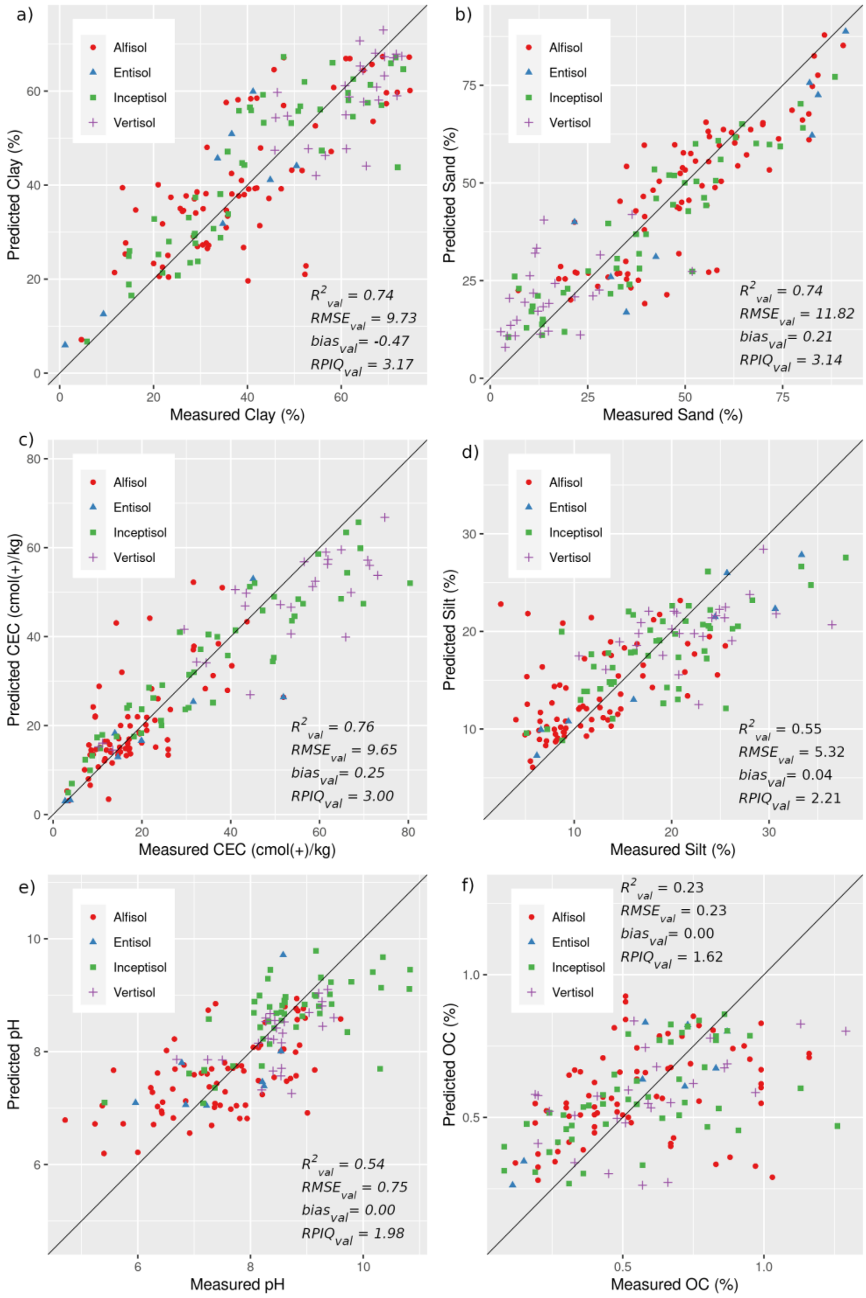

Fig. 5. Scatter plots of predicted versus observed soil properties obtained for the *BD_Val_Regional datasets.*

3.2.2. Analysis based on the soil order subsets

The 50 regional models built from the samples of *BD_Cal_Regional* for each soil property were then tested on samples of specific soil orders: *BD_val_Ver*, *BD_val_Alf*, *BD_val_Inc*, *BD_val_Ent*. While the regional models for clay and sand prediction provided good performances over the entire dataset (Fig. 5a and b), both models yielded acceptable (R^2_{val} >

0.50, $RPIQ_{val}$ > 1.50) to good (R^2_{val} > 0.70, $RPIQ_{val}$ > 3.00) performances for soil samples belonging to Alfisols, Inceptisols and Entisols (Table 3, Fig. 5a and b) and poor performances for Vertisols (R^2_{val} < 0.50, Table 3, purple points on Fig. 5a and b), which were characterized by the smallest clay and sand ranges among the four soil orders (SD of 8.75% and 10.6%, respectively, Table 2). Additionally, while the regional models for CEC prediction provided good performances over the entire dataset (Fig. 5c), it yielded acceptable (R^2_{val} > 0.50, $RPIQ_{val}$ > 1.50) performances for Vertisols (Table 3, purple points on Fig. 5c), good performances (R^2_{val} > 0.70, $RPIQ_{val}$ > 3.00) for Inceptisols and Entisols (Table 3, green and blue points in Fig. 5c) and poor performances for Alfisols (Table 3, red points on Fig. 5c).

The regional models for silt prediction yielded acceptable (R^2_{val} > 0.50, $RPIQ_{val}$>1.50) to good (R^2_{val} > 0.70, $RPIQ_{val}$ > 3.00) performances for soil samples belonging to Inceptisols and Entisols (Table 3, Fig. 5d), but performed poorly over Vertisols and Alfisols (Table 3, Fig. 5d) where the silt range was small (SD of 5.2 and 6.2%, respectively, Table 2). Finally, the regional models for the prediction of pH and SOC yielded poor performances regardless of the soil order (Table 3, Fig. 5e and f). Therefore, although the regional models for pH prediction provided acceptable performances over the entire dataset (Fig. 5e), it did not provide accurate predictions at the soil-order level (Table 3).

3.3 Prediction performance of soil-order model

Fifty soil-order models were built from calibration samples of each soil order (*BD_Cal_Ver, BD_Cal_Alf, BD_Cal_Inc and BD_Cal_Ent*, Fig. 2) for each soil property and validated using validation samples for each soil order (*BD_Val_Ver, BD_Val_Alf, BD_Val_Inc and BD_Val_Ent*, Fig. 2). The soil-order models for clay and CEC estimates built from Vertisols and Alfisols and tested on the same soil order yielded acceptable predictions (R^2_{val} > 0.50,

$RPIQ_{val} > 1.50$), while the soil-order models built from Inceptisols and Entisols for clay and CEC resulted in good predictions ($R^2_{val} > 0.70$, $RPIQ_{val} > 3.00$) (Table 3).

The soil-order models for sand estimation built from Alfisols and tested on the same soil order yielded acceptable predictions ($R^2_{val} > 0.50$, $RPIQ_{val} > 1.50$), while those built from Inceptisols and Entisols and tested on these same two soil orders yielded good predictions ($R^2_{val} > 0.70$, $RPIQ_{val} > 3.00$) (Table 3). For Vertisol, the soil-order models for sand estimation and tested on this same soil order yielded poor predictions ($R^2_{val} < 0.50$, $RPIQ_{val} < 1.50$) (Table 3). The soil-order models built from Entisols predicted silt content with acceptable accuracy ($R^2_{val} > 0.50$, $RPIQ_{val} > 1.50$), and the three other soil-order models built for silt estimation provided poor performances (Table 3). Regardless of the soil order, the soil-order models for SOC yielded poor predictions ($R^2_{val} < 0.50$, $RPIQ_{val} < 1.50$) (Table 3).

In accordance with the R^2_{val} and $RMSE_{val}$ values, these models calibrated from subsets stratified by soil orders for clay prediction outperformed the regional model when applied to each validation dataset of the corresponding soil order (Table 3). Similarly, the soil-order models for Vertisols, Alfisols and Inceptisols performed better than the regional models for the prediction of CEC. Although both regional and soil-order models performed well for the prediction of the sand contents of Alfisols, Inceptisols and Entisols, with respect to RPIQ, the soil-order model ($RPIQ_{val}$ of 2.12) slightly outperformed the regional model ($RPIQ_{val}$ of 2.04) for Alfisols (Table 3). In addition, the regional models outperformed the soil-order models in all other situations.

Table 3. Performance of regional and soil-order models (50 iterations) for the prediction of soil properties of different orders (standard deviation in parenthesis). (Models that yielded R^2_{val} values from 0.50 to 0.70 are highlighted in bold. Models that yielded R^2_{val} values above 0.70 are highlighted in bold and underlined).

		Validation Dataset															
Properties	Model	*BD_val_Ver*(26)				*BD_val_Alf*(72)				*BD_val_Inc*(48)				*BD_val_Ent*(8)			
		R^2_{val}	$RMSE_{val}$	$bias_{val}$	$RPIQ_{val}$	R^2_{val}	$RMSE_{val}$	$bias_{val}$	$RPIQ_{val}$	R^2_{val}	$RMSE_{val}$	$bias_{val}$	$RPIQ_{val}$	R^2_{val}	$RMSE_{val}$	$bias_{val}$	$RPIQ_{val}$
clay (%)	regional models	0.48 (0.10)	9.09 (1.48)	-5.21 (1.41)	1.25 (0.26)	**0.63 (0.06)**	**10.65 (0.91)**	**-0.49 (1.26)**	**1.95 (0.17)**	**0.78 (0.05)**	**8.52 (0.92)**	**1.25 (0.86)**	**3.24 (0.37)**	**0.84 (0.06)**	**8.85 (2.12)**	**4.77 (2.18)**	**3.43 (1.29)**
	soil-order models	**0.54 (0.11)**	**6.17 (0.93)**	**0.40 (1.15)**	**1.83 (0.31)**	**0.64 (0.06)**	**10.47 (0.71)**	**-0.54 (1.39)**	**1.98 (0.14)**	**0.79 (0.04)**	**8.42 (0.88)**	**-0.43 (1.18)**	**3.28 (0.37)**	**0.80 (0.08)**	**8.43 (1.98)**	**-0.13 (2.80)**	**3.63 (1.47)**
CEC (cmol (+) kg^{-1})	regional models	**0.58 (0.14)**	**12.69 (1.66)**	**-4.46 (1.64)**	**1.90 (0.29)**	0.46 (0.09)	8.96 (1.42)	2.78 (1.08)	1.17 (0.18)	**0.82 (0.04)**	**8.31 (0.77)**	**-1.45 (0.89)**	**3.88 (0.37)**	**0.72 (0.21)**	**9.88 (4.20)**	**3.12 (2.64)**	**3.04 (1.61)**
	soil-order models	**0.51 (0.14)**	**12.90 (1.58)**	**0.24 (2.08)**	**1.86 (0.21)**	**0.61 (0.05)**	**6.21 (0.69)**	**-0.01 (0.73)**	**1.67 (0.19)**	**0.83 (0.04)**	**7.94 (0.83)**	**0.11 (1.23)**	**4.07 (0.47)**	0.68 (0.14)	9.45 (2.02)	-0.10 (2.69)	2.85 (0.94)
sand (%)	regional models	0.38 (0.13)	11.61 (2.25)	6.46 (1.98)	0.99 (0.30)	**0.59 (0.06)**	**12.78 (1.00)**	**-1.21 (1.33)**	**2.04 (0.15)**	**0.79 (0.05)**	**10.32 (1.08)**	**-0.42 (1.28)**	**3.48 (0.38)**	**0.87 (0.06)**	**10.78 (2.01)**	**-3.42 (2.44)**	**4.31 (1.44)**
	soil-order models	0.45 (0.13)	8.25 (1.52)	0.13 (1.97)	1.39 (0.32)	**0.60 (0.06)**	**12.29 (0.76)**	**0.28 (1.32)**	**2.12 (0.14)**	**0.76 (0.05)**	**11.18 (1.07)**	**0.07 (0.14)**	**3.21 (0.34)**	**0.75 (0.08)**	**14.05 (2.60)**	**0.37 (4.45)**	**3.26 (0.97)**
pH	regional models	0.43 (0.11)	0.53 (0.07)	-0.22 (0.08)	1.44 (0.23)	0.41 (0.08)	0.85 (0.06)	0.13 (0.07)	2.00 (0.13)	0.50 (0.08)	0.69 (0.07)	-0.09 (0.08)	1.47 (0.17)	0.30 (0.21)	0.76 (0.14)	0.15 (0.19)	1.93 (0.49)
	soil-order models	0.45 (0.11)	0.45 (0.06)	0.00 (0.08)	1.69 (0.27)	0.39 (0.08)	0.86 (0.07)	-0.01 (0.07)	1.99 (0.15)	0.41 (0.08)	0.76 (0.07)	0.02 (0.10)	1.34 (0.15)	0.12 (0.14)	0.79 (0.09)	-0.03 (0.13)	1.74 (0.26)
SOC (%)	regional models	0.32 (0.11)	0.21 (0.02)	0.01 (0.03)	1.34 (0.17)	0.16 (0.06)	0.24 (0.02)	0.01 (0.02)	1.52 (0.11)	0.30 (0.07)	0.24 (0.02)	-0.02 (0.02)	1.83 (0.17)	0.44 (0.30)	0.26 (0.14)	-0.05 (0.07)	1.56 (0.77)
	soil-order models	0.34 (0.10)	0.21 (0.02)	0.00 (0.03)	1.38 (0.17)	0.14 (0.06)	0.24 (0.02)	0.00 (0.02)	1.50 (0.12)	0.28 (0.08)	0.24 (0.02)	0.00 (0.03)	1.81 (0.17)	0.40 (0.28)	0.28 (0.10)	0.00 (0.10)	1.27 (0.48)
silt (%)	regional models	0.19 (0.12)	4.97 (0.66)	-1.18 (0.53)	1.50 (0.21)	0.31 (0.10)	5.44 (0.53)	1.30 (0.50)	1.20 (0.10)	**0.50 (0.08)**	**5.29 (0.46)**	**-0.89 (0.56)**	**1.77 (0.18)**	**0.86 (0.07)**	**5.11 (1.02)**	**-1.78 (1.02)**	**3.35 (0.90)**
	soil-order models	0.27 (0.13)	4.53 (0.63)	0.15 (0.69)	1.65 (0.24)	0.30 (0.13)	5.25 (0.55)	-0.24 (0.41)	1.25 (0.14)	0.41 (0.08)	5.62 (0.46)	0.00 (0.78)	1.67 (0.16)	**0.53 (0.17)**	**7.84 (1.79)**	**0.48 (2.31)**	**2.24 (0.93)**

4. Discussion

4.1. Predictions at the regional scale based on regional models

The soil properties which were successfully predicted based on regional models, were characterized by i) a high variability (e.g., clay contents from 1.2 to 77.2% with a SD of 19%; Table 2) and ii) either a spectral response due to physicochemical responses (e.g., clay which is characterized by a absorption band at 2208 nm corresponding to the combination of OH stretch and OH-Al bending modes, Chabrillat et al., 2019) or a correlation to one property which was successfully predicted (e.g., sand which was correlated to clay, Supplementary Information 3). These results are in accordance with the three rules defined by Ben-Dor et al. (2002) and then Gomez et al. (2012a, b), presented in Chabrillat et al. (2019) and recalled in our Introduction section. Conversely, soil properties characterized by a short variability of values (e.g., SOC with a mean of 0.6% and SD of 1.1%, Table 2) were poorly predicted at the regional scale by the regional models (Fig. 5f).

The accurate clay estimations might be due to the use of wavelengths in RF models related to clay including the bands around 2208 nm corresponding to the combination of OH stretch and OH-Al bending modes (Chabrillat et al., 2002). The accurate predictions of CEC might be attributed to the correlation between CEC and clay and the large range of CEC values at the regional scale (Table 2), as CEC does not have a primary response to spectral reflectance (Leone et al., 2012; Xu et al., 2018). Similar levels of performance were observed for the various models for the prediction of clay, sand and CEC in the literature. Ahmadi et al. (2021) stated that the mean coefficients of determination (R^2) for various Vis-NIR prediction studies for sand and clay were 0.76 and 0.70, respectively. Terra et al. (2015) emphasised that the promising results of models for the prediction of sand (R^2_{cal} from 0.85 to 0.90) and clay contents (R^2_{cal} from 0.85 to 0.88) may effectively replace the analysis of soil particle size by conventional methods.

Silt content was predicted with reliable accuracy (R^2_{val}=0.55, $RPIQ_{val}$=2.21 and $RMSE_{val}$= 5.32%), which was in agreement with Viscarra Rossel et al. (2006). Additionally, pH was predicted with reliable accuracy (R^2_{val}=0.54, $RPIQ_{val}$=1.98 and $RMSE_{val}$= 0.75), which is difficult to explain because pH does not have any spectral response or correlation to a property having a spectral response due to physical or chemical structures (Supplementary Information 3). The low range for SOC content might be the cause of the poor prediction of SOC (Dalal and Henry, 1986), which was confirmed with Fig. 4, where no significant absorption was observed near 500 and 800 nm (Latz et al., 1984).

4.2. Predictions at the soil order scale based on regional models

Based on regional models, the prediction performances obtained over each subset stratified per soil order differed from those obtained at the regional scale (Table 3 and Fig. 5). While clay and sand contents may be considered correctly predicted at the regional scale (Fig. 5a, b), both soil properties were poorly predicted over Vertisols samples (Table 3), for which these properties were characterized by a small range (SD of 8.75% and 10.6%, respectively, Table 2) and thus do not follow the rule (1.3) stated by Chabrillat et al. (2019). Additionally, while CEC may be considered correctly predicted at the regional scale (Fig. 5c), CEC was poorly predicted for Alfisols samples (Table 3), which was characterized by a small CEC range (SD of 9.8 cmol (+) kg^{-1}, Table 2) and thus does not follow the rule (1.3) stated by Chabrillat et al. (2019).

So models based on the regional database for calibration can be considered as providing high accuracy of some soil properties estimations when considering the regional strategy in the validation step but modest accuracy of these same soil properties when considering subsets stratified by soil order from the regional database in validation step. These results are in accordance with Gomez and Coulouma (2018), who showed that while their prediction models were accurate at a regional scale, the prediction model performances at within-field scales depended on the specific soil property. As the estimation accuracy appreciation is depending on the validation database, the appreciation of prediction accuracies can be done both at regional and soil-order scale to reinforce the performance analysis.

4.3. Predictions at the soil order scale based on soil order models

The soil-order models dedicated to Entisols and Inceptisols predict clay contents (R^2_{val} of 0.80 and 0.79, respectively, Table 3) with more accuracy than the soil-order models dedicated to Vertisols (R^2_{val} of 0.54, Table 3), as the presence of smectite clay minerals and the high moisture-holding capacity of Vertisols may reduce the relative spectral reflectance at 1300–1400, 1800–1900, and 2200–2500 nm bands (Baumgardner et al., 1985; Babaeian et al., 2015; Demattê et al., 2017). The prediction of CEC was on par with clay for different soil orders, which might be due to a positive correlation between clay and CEC. The trends in CEC prediction for the soil orders were similar to the trends in the correlation coefficients between clay and CEC (Supplementary Information 4-7). The higher performances for sand prediction ($R^2_{val} \geq 0.75$, Table 3) in Inceptisols and Entisols might be explained by the higher sand content in these soils which are at the inception of soil development (Santos et al., 2013). A relatively better prediction of silt content was achieved through a soil-order model for Entisols, which might be attributed to the predominance of highly weatherable minerals in these soils that alter their albedo (Poppiel et al., 2018).

4.4. Regional model *versus* soil-order model

For Vertisols, the soil-order models for clay and sand estimates significantly outperformed the regional models (Table 3), while both the soil-order and regional models for other soil property predictions provided a similar range of performances. For Alfisols, the soil-order model for CEC estimates significantly outperformed the regional model (Table 3), while both the soil-order and regional models for the other soil property predictions provided a similar range of performances. Over Inceptisols, the regional models for pH and silt estimates significantly outperformed the soil-order models (Table 3), while both the soil-order and regional models for other soil property predictions provided a similar range of performances. For Entisols, the regional models for CEC, sand and silt estimates significantly outperformed the soil-order models (Table 3), while both the soil-order and regional models for the other soil property predictions provided a similar range of performances.

Therefore, these results did not allow us to conclude whether a regional model or a soil-order model is the best strategy for predicting different properties across different soils. The literature is also not unanimous on this point, as some works have shown that regional models outperform soil-order models (e.g., Vasques et al., 2010; Liu et al., 2018), while other works have shown the opposite (e.g., Madari et al., 2005; McDowell et al., 2012). Therefore, while our results did not enable any recommendations for choosing between a regional or soil-order model, they highlight the risk of overestimating prediction accuracy at the soil-order scale when figures of merit are based on a validation dataset built at the regional scale.

5. Conclusion

In the present study, the effectiveness of using Vis-NIR spectroscopy for the prediction of soil properties was analyzed based on soil order knowledge in both calibration and validation steps. While these results did not enable any recommendations for choosing between a regional or soil-order model when validating on soil-order datasets, they highlighted the risk of overestimating prediction accuracy at the soil-order scale when figures of merit are based on a validation dataset built at a regional scale. As large soil spectral libraries are currently highly developed, this work showed that soil-order knowledge may be useful to avoid misestimating soil properties. In future, this work could be completed by an analysis of how land use or other environmental covariates may be used to improve soil properties prediction models.

Acknowledgement

The authors thank the Karnataka Watershed Development Department and the World Bank for funding the Sujala III project. The authors thank the ATCHA, ANR-16-CE03-0006 project for supporting the work. The authors also thank Sebastien Troiano from INRAE, UMR LISAH, for his help in setting up the spectral laboratory. The authors also acknowledge Dr. Laurent Ruiz, Indo-French Cell for Water Sciences, Bangalore for his guidance in developing the spectral library of Karnataka. The authors also thank Dr. Arti Koyal, CTO, NBSS&LUP for helping with recording spectral data.

Declaration of Competing Interest

The authors declare that there are no known competing interests.

References

Ahmadi, A., Emami, M., Daccache, A., He, L., 2021. Soil Properties Prediction for Precision Agriculture Using Visible and Near-Infrared Spectroscopy: A Systematic Review and Meta-Analysis. Agron. 11, 433. https://doi.org/10.3390/agronomy11030433.

Asgari, N., Ayoubi, S., Demattê, J.A.M., Dotto. A.C., 2020. Carbonates and organic matter in soils characterized by reflected energy from 350–25000 nm wavelength. J. Mt. Sci. 17, 1636–1651. https://doi.org/10.1007/s11629-019-5789-9

Babaeian, E., Homaee, M., Vereecken, H., Montzka, C., Norouzi, A.A., van Genuchten, M.T., 2015. A Comparative Study of Multiple Approaches for Predicting the Soil-Water Retention Curve: Hyperspectral Information vs. Basic Soil Properties. Soil Sci. Soc. Am. J. 79(4), 1043-1058. https://doi.org/10.2136/sssaj2014.09.0355.

Bao, Y., Meng, X., Ustin, S.L., Wang, X., Zhang, X., Liu, H., Tang, H., 2020. Vis-SWIR spectral prediction model for soil organic matter with different grouping strategies. Catena 195, 104703. https://doi.org/10.1016/j.catena.2020.104703.

Baumgardner, M.F., Silva, L.F., Biehl, L.L., Stoner, E.R., 1985. Reflectance properties of soils. Adv. Agron. 38, 1–44. doi:10.1016/S0065-2113(08)60672-0.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A., 2010. Prediction of soil attributes by NIR spectroscopy. A critical review of chemometric indicators commonly used for assessing the quality of the prediction. Trends Anal. Chem. (TRAC) 29 (9), 1073–1081. https://doi.org/10.1016/j.trac.2010.05.006.

Bellon-Maurel, V., McBratney, A., 2011. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils e Critical review and research perspectives. Soil Biol. Biochem. 43(7), 1398-1410. DOI: 10.1016/j.soilbio.2011.02.019.

Ben-Dor, E., 2002. Quantitative remote sensing of soil properties. Adv. Agron. 75, 173–243. doi:10.1016/S0065-2113(02)75005-0.

Ben-Dor, E., Patkin, K., Banin, A., Karnieli, A. 2002. Mapping of several soil properties using DAIS-7915 hyperspectral scanner data: a case study over clayey soils in Israel. Int. J. Remote Sens. 23, 1043–1062

Ben-Dor, E., Banin, A., 1995a. Near infrared analysis (NIRA) as a method to simultaneously evaluate spectral featureless constituents in soils. Soil Sci. 159(4), 259–270. https://doi.org/10.1097/00010694-199504000-00005.

Ben-Dor, E, Banin, A., 1995b. Near infrared analysis (NIRA) as a rapid method to simultaneously evaluate several soil properties. Soil Sci. Soc. Am. J. 59, 364–372. 10.2136/sssaj1995.03615995005900020014x

Bilgili, A.V., van Es, H.M., Akbas, F., Durka, A., Hively, W.D., 2010. Visible near-infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. Arid Environ. 74, 229–238. doi:10.1016/j.jaridenv.2009.08.011

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5-32. https://doi.org/10.1023/A:1010933404324.

Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140, 444–453. DOI:10.1016/j.geoderma.2007.04.021.

Chabrillat, S., Goetz, A.F.H., Krosley, L., Olsen, H.W., 2002. Use of hyperspectral images in the identification and mapping of expansive clay soils and the role of spatial resolution. Remote Sens. Environ. 82, 431–445. https://doi.org/10.1016/S0034-4257(02)00060-3.

Chabrillat, S., Gholizadeh, A., Neumann, C, Berger, D., Milewski, R., Ogen, Y., Ben-Dor, E., 2019. Preparing a soil spectral library using the Internal Soil Standard (ISS) method: Influence of extreme different humidity laboratory conditions. Geoderma 355, 113855. https://doi.org/10.1016/j.geoderma.2019.07.013.

Cozzolino, D., Morón, A., 2003. The potential of near-infrared reflectance spectroscopy to analyse soil chemical and physical characteristics. J. Agric. Sci. 140(1), 65-71. https://doi.org/10.1017/S0021859602002836.

Dalal, R.C., Henry, R.J., 1986. Simultaneous determination of moisture, organic carbon, and total nitrogen by near infra-red reflectance spectrophotometry. Crop Sci. Soc. Am. 50, 120–123. https://doi.org/10.2136/sssaj1986.03615995005000010023x.

Davari, M., Karimi, S.A., Bahrami, H.A., Taher Hossaini, S.M., Fahmideh, S., 2021. Simultaneous prediction of several soil properties related to engineering uses based on laboratory Vis-NIR reflectance spectroscopy. Catena 197, 104987. https://doi.org/10.1016/j.catena.2020.104987.

Delwiche, S.R., 2010. A graphical method to evaluate spectral preprocessing in multivariate regression calibrations: Example with Savitzky-Golay filters and partial least squares regression. Appl. Spectrosc. 64, 73–82. https://doi.org/10.1366/000370210790572007.

Demattê, J.A., Campos, R.C., Alves, M.C., Fiorio, P.R., Nanni, M.R., 2004. Visible–NIR reflectance: a new approach on soil evaluation. Geoderma 121, 95-112. https://doi.org/10.1016/j.geoderma.2003.09.012

Demattê, J.A.M., Horák-Terra, I., Beirigo, R.M., Terra, F. da S., Marques, K.P.P., Fongaro, C.T., Silva, A.C., Vidal-Torrado, P., 2017. Genesis and properties of wetland soils by VIS-NIR-SWIR as a technique for environmental monitoring. J. Environ. Manage. 197, 50–62. https://doi.org/10.1016/j.jenvman.2017.03.014.

Dharumarajan, S., Lalitha, M., Gomez, C., Vasundhara, R., Kalaiselvi, B., Hegde, R. 2022. Prediction of soil hydraulic properties using VIS-NIR spectral data in semi- arid region of Northern Karnataka Plateau. Geoderma Reg. https://doi.org/10.1016/j.geodrs.2021.e00475.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman and Hall, London, UK.

Farifteh, J., Meer, F.D., Meijde, M.V., Atzberger, C., 2008. Spectral characteristics of salt-affected soils: A laboratory experiment. Geoderma 145, 196-206. https://doi.org/10.1016/j.geoderma.2008.03.011.

Ghasemi, J. B., Tavakoli, H., 2013. Application of random forest regression to spectral multivariate calibration. Anal. Methods 5, 1863-1871. https://doi.org/10.1039/C3AY26338J.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L., 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemom. Intell. Lab. Syst. 110 (1), 168–176.

Gogé, F., Gomez, C., Jolivet, C., Joffre, R., 2014. Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? Geoderma 213, 1–9. http://dx.doi.org/10.1016/j.geoderma.2013.07.016.

Gomez, C., Coulouma, G., 2018. Importance of the spatial extent for using soil properties estimated by laboratory VNIR/SWIR spectroscopy: Examples of the clay and calcium carbonate content, Geoderma 330, 244–253. https://doi.org/10.1016/j.geoderma.2018.06.006.

Gomez, C., Lagacherie, P., Bacha, S., 2012b. Using Vis–NIR hyperspectral data to map topsoil properties overbare soils in the Cap Bon region, Tunisia. In: Digital soil assessments and beyond—proceedings of the fifth global workshop on digital soil mapping, pp 387–392.

Gomez, C., Lagacherie, P., Coulouma, G., 2008. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. Geoderma 148, 141–148. https://doi.org/10.1016/j.geoderma.2008.09.016.

Gomez, C., Lagacherie, P., Coulouma, G., 2012a. Regional predictions of eight common soil properties and their spatial structures from hyperspectral Vis–NIR data. Geoderma 189–190, 176–185. http://dx.doi.org/10.1016/j.geoderma.2012.05.023.

Guerrero, C., Zornoza, R., Gómez, I., Mataix-Beneyto, J., 2010. Spiking of NIR regional models using samples from target sites: effect of model size on prediction accuracy. Geoderma 158, 66–77.

Gupta, A., Hitesh B. V., Das, B. S., Choubey, A. K., 2018. Local modeling approaches for estimating soil properties in selected Indian soils using diffuse reflectance data over visible to near-infrared region. Geoderma 325, 59–71. https://doi.org/10.1016/j.geoderma.2018.03.025

Hedley, C., Roudier, P., Maddi, L., 2015. VNIR Soil Spectroscopy for Field Soil Analysis. Commun. Soil Sci. Plant Anal. 46, 104–121. DOI: 10.1080/00103624.2014.988582.

Hobley, E.U., Prater, I., 2019. Estimating soil texture from vis–NIR spectra. Eur. J. Soil Sci. 70, 83-95, 10.1111/ejss.12733

Hegde, R., Niranjana, K. V., Srinivas, S., Danorkar, B. A., Singh. S. K., 2018. Site-specific land resource inventory for scientific planning of Sujala watersheds in Karnataka. Current Sci. 115(4), 645–652. http://dx.doi.org/10.18520/cs/v115/i4/644-652.

Jackson, M.L., 1973. Soil Chemical Analysis. Prentice Hall of India Pvt. Ltd. New Delhi.

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

Latz, K., Wesimiller, R.A., Van Scoyoc, G.E., Baumgarnder, M.F., 1984. Characteristic variation in spectral reflectance of selected eroded Alfisols. Soil Sci. Soc. Am. J. 48, 1130–1134. https://doi.org/10.2136/sssaj1984.03615995004800050035x.

Leone, A.P., Viscarra-Rossel, R.A., Amenta, P., Buondonno, A., 2012. Prediction of soil properties with PLSR and Vis-NIR spectroscopy: application to Mediterranean soils from Southern Italy. Curr. Analy. Chem. 8, 283–299. https://doi.org/10.2174/157341112800392571.

Liu, Y., Shi, Z., Zhang, G., Chen, Y., Li, S., Hong, Y., Shi, T., Wang, J., Liu, Y. 2018. Application of spectrally derived soil type as ancillary data to improve the estimation of soil organic carbon by using the Chinese soil Vis-NIR Spectral Library. Remote Sen. 10(11), 1747. https://doi.org/10.3390/rs10111747.

Lobsey, C. R., Viscarra Rossel, R. A., Roudier, P., Hedley, B., 2017. RS-local data-mines information from spectral libraries to improve local calibrations. Eur. J. Soil Sci. 68, 840–852. doi: 10.1111/ejss.12490

Madari, B.E., Reeves, J.B., Coelho, M.R., Machado, P.L., De-Polli, H., Coelho, R.M., Benites, V.M., Souza, L.F., McCarty, G.W., 2005. Mid and near-infrared spectroscopic determination of carbon in a diverse set of soils from the Brazilian national soil collection. Spectrosc. Lett. 38, 721–740. https://doi.org/10.1080/00387010500315876.

McBride, M. B. 2022. Estimating soil chemical properties by diffuse reflectance spectroscopy: Promise versus reality. European J. Soil Sci. 73, e13192. https://doi.org/10.1111/ejss.13192.

McDowell, M. L., Bruland, G.L., Deenik, J.L., Grunwald, S., 2012. Effects of subsetting by carbon content, soil order, and spectral classification on prediction of soil total carbon with diffuse reflectance spectroscopy. Appl. Environ. Soil Sci. https://doi.org/10.1155/2012/294121.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., Mouazen, A. M., 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using Vis-NIR spectroscopy. Biosyst. Eng. 152, 104–116. http://dx.doi.org/10.1016/j.biosystemseng.2016.04.018.

Naibo, G., Ramon, R., Pesini, G., Moura-Bueno, J.M., Barros, C.A., Caner, L., Silva, Y.J., Minella, J.P., dos Santos, D.R., Tiecher, T., 2022. Near-infrared spectroscopy to estimate the chemical element concentration in soils and sediments in a rural catchment. Catena 213, 106145. https://doi.org/10.1016/j.catena.2022.106145.

Nawar, S., Mouazen, A., 2019. On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. Soil Till. Res. 190, 120–127. https://doi.org/10.1016/j.still.2019.03.006.

Nawar, S., Mouazen, A., 2017. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. Catena 151, 118–129. https://doi.org/10.1016/j.catena.2016.12.014.

NBSS&LUP, 1998. Soils of Karnataka for Optimising Land Use. NBSS Publ., 47b. ISBN:81-85460-45-0.

Naimi, S., Ayoubi, S., Di Raimo, L. A. D. L., Dematte, J. A. M., 2022. Quantification of some intrinsic soil properties using proximal sensing in arid lands: Application of Vis-NIR, MIR, and pXRF spectroscopy. Geoderma Reg. 28, e00484. https://doi.org/10.1016/j.geodrs.2022.e00484

Nocita, M., Stevens, A., Toth, G., Panagos, P., vanWesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347.

Ng, W., Minasny, B., Jeon H., McBratney, A., 2022. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. Soil Security 100043, https://doi.org/10.1016/j.soisec.2022.100043.

Peng, Y., Knadel, M., Gislum, R., Deng, F., Norgaard, T., de Jonge, L.W., Moldrup, P., Greve, M.H., 2013. Predicting soil organic carbon at field scale using a national soil spectral library. J. Near Infrared Spectrosc. 21, 213–222.

Pinheiro, E. F. M., Ceddia, M. B., Clingensmith, C. M., Grunwald, S., Vasques, G. M., 2017. Prediction of Soil Physical and Chemical Properties by Visible and Near-Infrared Diffuse Reflectance Spectroscopy in the Central Amazon. Remote Sens. 9 (4). DOI:10.3390/rs9040293.

Poppiel, R.R., Lacerda, M.P.C., Oliveira Junior, M.P., Demattê, J.A.M., Romero D.J., Sato, M.V., Almeida Júnior, L.R., Cassol, L.F.M., 2018. Surface spectroscopy of Oxisols, Entisols and Inceptisol and relationships with selected soil properties. Rev. Bras. Ciênc. Solo 42, e0160519.

Richards, L. A., 1954. Diagnosis and improvement of saline and alkali soils. USDA Handbook, 60. USDA, Washington. D.C., USA

Sankey, J. B., Brown, D. J., Bernard, M. L., Lawrence, R. L., 2008. Comparing local vs. global visible and near-infrared (visnir) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma 148, 149-158. http://dx.doi.org/10.1016/j.geoderma.2008.09.019.

Santos, H.G., Jacomine, P.K.T., Anjos, L.H.C., Oliveira, V.A., Lumbreras, J.F., Coelho, M.R., Almeida, J.A., Cunha, T.J.F., Oliveira, J.B., 2013. Sistema brasileiro de classificação de solos. 3a ed. Brasília, DF: Embrapa Solos.

Shepherd, K.D., Walsh, M. G., 2002. 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.

Shi, Z., Ji, W., Viscarra Rossel, R.A., Chen, S., Zhou, Y., 2015. Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis–NIR spectral library. Eur. J. Soil Sci. 66, 679–687.

Soil Survey Staff, 2014. Keys to soil taxonomy. 12th ed. Washington, DC: United States Department of Agriculture, Natural Resources Conservation Service.

Srivastava, R., Prasad, J., Saxena, R.K., 2004. Spectral reflectance properties of some shrink-swell soils of Central India as influenced by soil properties. Agropedology 14, 45-54.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. Adv. Agron. 107, 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS One 8, e66409

Terra, F.S., Demattê, J.A.M., Rossel, R.A.V., 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing Vis–NIR and mid-IR reflectance data. Geoderma 255–256, 81-93. http://doi.org/10.1016/j.geoderma.2015.04.017.

Vasques, G.M., Grunwald, S.J.O.S., Sickman, J.O., 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146 (1), 14–25.

Vasques, G.M., Grunwald, S., Harris, W.G., 2010. Spectroscopic models of soil organic carbon in Florida, USA. J. Environ. Qual. 39, 923–934. https://doi.org/10.2134/jeq2009.0314.

Veum, K.S., Sudduth, K.A., Kremer, R.J., Kitchen, N.R., 2015. Estimating a Soil Quality Index with VNIR Reflectance Spectroscopy. Soil Sci. Soc. Am. J. 79, 637-649. https://doi.org/10.2136/sssaj2014.09.0390.

Viscarra Rossel, R.A., Behrens, T., 2009. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1), 46–54.

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.

Viscarra Rossel, R.A., Webster, R., 2011. Discrimination of Australian soil horizons and classes from their visible–near infrared spectra. Eur. J. Soil Sci. 62, 637–647. https://doi.org/10.1111/j.1365-2389.2011.01356.x.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D. et al. 2016. A global spectral library to characterize the world's soil. Earth Sci. Rev. 155, 198–230.

Walkley, A., Black, I.A. 1934. An estimation of the method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Sci. 37, 29-38.

Wetterlind, J., Stenberg, B., 2010. Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples. Eur. J. Soil Sci. 61, 823–843. https://doi.org/10.1111/j.1365-2389.2010.01283.x.

Xu, S., Shi, X., Wang, M., Zhao, Y., 2016. Effects of subsetting by parent materials on prediction of soil organic matter content in a hilly area using Vis–NIR spectroscopy. PLoS ONE 11(3): e0151536Zeng, R., Zhao, Y.-G., Li, D.-C., Wu, D.-W., Wei, C.-L., Zhang, G.-L., 2016. Selection of “Local” models for prediction of soil organic matter using a regional soil Vis-NIR Spectral Library. Soil Sci. 181, 13–19. http://dx.doi.org/10.1097/SS.0000000000000132 .

Zeng, R., Zhao, Y.-G., Li, D.-C., Wu, D.-W., Wei, C.-L., Zhang, G.-L. 2016. Selection of “Local” models for prediction of soil organic matter using a regional soil Vis-NIR Spectral Library. Soil Sci. 181, 13–19. http://dx.doi.org/10.1097/SS.0000000000000132.

Zeraatpisheh, M., Ayoubi, S., Mirbagheri, Z., Mosaddeghi, M. R., Xu, M., 2021. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. Geoderma Reg. 27, e00440. https://doi.org/10.1016/j.geodrs.2021.e00440