Estimation of soil properties with mid-infrared soil spectroscopy across yam production landscapes in West Africa

Philipp Baumann[1¤], Juhwan Lee[2], Laurie Paule Schönholzer[1], Emmanuel Frossard[1], Johan Six[1]

1 Department of Environmental Systems Science, ETH Zurich, 8092 Zurich, Switzerland
2 School of Molecular and Life Sciences, Curtin University, GPO Box U1987, Perth WA 6845, Australia

¤Current Address: Universitätsstrasse 2, 8092 Zürich, Switzerland
* philipp.baumann@usys.ethz.ch

Abstract

Soil infertility is a major problem limiting the sustainable production of yam and other staple crops in the yam belt of West Africa. Its quantitative assessment is needed to improve soil functions and crop fertilization management, but is still lacking for the region. We developed and tested a mid-infrared (mid–IR) soil spectral library to enable timely and cost-efficient assessments of soil properties. Our library included 80 field composite soil samples in four landscapes that are representative of the West African yam belt, and additional 14 samples from a sentinel site of the Land Degradation Surveillance Framework. We derived partial least square regression models to estimate the soil properties from the spectra. Five times repeated 10-fold cross-validation was used to evaluate the models. The models produced accurate estimates of total carbon, total nitrogen, total sulfur, total iron, total aluminium, total potassium, total calcium, exchangeable calcium, effective cation exchange capacity, DTPA extractable iron and clay content ($R^2 > 0.75$). The models for total zinc, pH, exchangeable magnesium, DTPA extractable copper and manganese produced less accurate estimates ($R^2 > 0.5$). Our results suggest that mid-IR spectroscopy can be used to reliably assess the landscape-level variation in soil properties related to the fertility and yam production potential of soils across the environment of the West African yam belt.

Introduction

Yam (*Dioscorea* spp.) is an important food and cash crop in West Africa. The yam belt of West Africa spans across the central zone of coastal countries in West Africa, located across the humid forest zone and northern Guinean savanna. It contributes to about 92 % of total world yam production, e.g. a total yield of 73 million tons in 2017 [1]. The cropping area in the West African yam belt has been expanded with accelerated population growth, which has in many places caused soil degradation and nutrient depletion. There is a trend of shortened fallow periods in the cropping areas of West Africa over the last decades, which has further exacerbated a decline in soil fertility across the yam belt. Traditionally, yam is grown without external input in the areas. Therefore, the production of yam and other crops grown in the region depends on soil organic matter (SOM) [2], which serves as a main pool of plant-available nutrients and

provides cation exchange surfaces to soil nutrients [3, 4]. Particularly, a strong positive relationship between high organic matter stocks and and yam productivity is reported after fallow and no fertilizer input [5, 6]. Currently, maintaining or increasing C and other nutrient levels in SOM is of utmost importance for sustainable production of West Africa's soils [7]. Linking soil properties and yam yields [8] and accounting for soil macro- and micronutrient status [9] is fundamental to improving crop and soil management strategies.

Soil fertility is considered an integrative measure of soil quality attributes and their interactions that support the long-term agricultural production potential. Soil fertility is commonly decomposed into three main components, the physical, chemical and biological aspects of soil fertility [10]. Here, it is important to interpret soil fertility in the form of soil conditions and functions at an adequate resolution over time and space, and in relation to the crop of interest. For yam, low tuber yields are attributed to the soil's status that often have an unbalanced ratio of essential nutrients (i.e. N, P, K) and receive mineral fertilizer in inadequate amounts, decreasing tuber yields or priming organic matter mineralization [7]. Yet, the relation between soil properties and tuber yield is not well studied [8]. The reason is that yam response to mineral fertilization is highly variable because of confounding environmental and management variables, such as climate, micronutrient deficiencies, seed tuber quality and planting density or disease pressure across the yam belt [9, 11]. Further, there are no soil fertility recommendations specific for yam under West African conditions. For this reason, establishing yam field trials designed with different organic and mineral fertilization strategies within different yam growing regions is required to optimize yam fertilization targeting regional soil and environmental conditions [8]. Despite the importance of soil fertility, there are challenges of quantifying it with respect to yam performance. For example, there are many influential soil measures required in high temporal resolution, which is costly. Soil fertility is usually reported in the form of fertility classes, or a set of properties with qualitative categorization and indices derived from minimum data sets are used, all of which require extensive soil analysis.

Soil fertility in the yam belt is under threat, and there is an increasing demand to evaluate the intermediate and long-term effects of innovative agronomic practices promoting soil conditions [8]. In order to quickly assess key properties such as soil organic carbon (SOC) and the cation exchange capacity (CEC), we need more cost- and time-efficient methods in alternative to traditional wet chemistry laboratory analyses, which are often cost-intensive and slow. Proximal soil sensing is a method that can provide reliable soil measurements rapidly and inexpensively [12]. Soil visible and near infrared (vis–NIR), and mid-infrared (mid-IR) diffuse reflectance spectroscopy has gained popularity over the past 30 years to assess soil properties to complement conventional laboratory analytical methods [13]. Previous studies have shown successful spectroscopic predictions of soil properties that affect soil fertility, such as organic C, texture, cation exchange capacity (CEC), and exchangeable K [13–16], however we are not aware of a published spectral library that is targeted to yam production landscapes in West Africa. Many soil chemical composition and physical properties, such as soil mineralogy, the concentration, forms and distribution of SOM, are closely associated with IR spectral diversity. However, its mineral and organic composition tends to vary considerably at landscape or larger scales. This landscape-level soil heterogeneity requires the need for calibrating relationships between soil composition and spectra across the landscape.

Soil assessment with IR spectroscopy often needs laboratory reference analysis data for model development and calibration. Further, a library that includes a broad range of soil types and variability found in the region in which it is used needs to be established. Depending on the study scale — field (e.g. [17]), region, country (e.g. [18]), continent (e.g. [16]), world (e.g. [19]), various statistical predictive modeling strategies are

typically employed to account for regional variability in soil properties and determine empirical relationships between spectra and soil attributes.

For soil spectroscopy, particular regions in spectra are characteristic for functional groups of soil components. Thus, elucidating spectral features that are important for the prediction of a particular soil attribute helps to understand and validate the mechanisms based on which these models predict.

Our main objectives of this study are to (1) develop and evaluate mid-IR spectroscopic models to estimate soil properties for selected landscapes representing major soil and climatic conditions in the West African yam belt, (2) to determine important spectral features and respective soil properties, and (3) to build a new soil spectral library in the landscapes for soil prediction and assessment.

Materials and methods

Landscapes and soil sampling

Our study area covered the climatic and soil biophysical conditions representative of the West African yam belt. We selected four landscapes, two in the in Ivory Coast and two in Burkina Faso (S1 Fig). Each landscape (10 km x 10 km) represents a diverse geographic ecoregion. The landscapes cover a gradient between humid forest and the northern Guinean savannah. Specifically, the landscape Liliyo in Ivory Coast is at 5.88 °N and in the humid forest zone. The predominant soil types are Ferralsols [20]. The landscape Tiéningboué in Ivory Coast is at 8.14 °N and belongs to the forest savannah transitional zone. The soils are dominated by Nitrisols and Lixisols [20]. The landscape Midebdo is at 9.97 °N and in the sub-humid savannah. Its dominant soil types include Lixisols, Gleysols, and Leptosols [20]. The landscape Léo is at 11.07 °N and in the northern Guinean savannah and has Lixisols and Vertisols [20]. The mean annual rainfall were approximately 1300 mm in Liliyo, and 900 mm in Tiéingboué, Midebdo, and Léo, respectively.

During July and August 2016, we sampled the soil from a total of 80 fields used for growing yams across the four landscapes. In each landscape, we sampled soils from 20 yam fields (S2 Fig). The fields were selected in advance by taking into account visual variation in soil color and texture in yam fields across the landscape. The yam fields selected contained the maximum soil variability based on the soil colour and cropping history, taking into account both local farmers' knowledge on soil fertility and agronomic extension expertise. Yam is typically planted on soil mounds, ranging from 5000 to 10000 mounds per hectare with a single yam plant per mound. Within each field, we sampled the soil at four adjacent mounds in square arrangement, which were spaced between 0.5 m and 2 m. At each mound 6 to 8 small auger cores (2.5 cm in diameter) at the 30 cm depth were taken at a radius between 15 and 30 cm away from the center of a mound, depending on the size of the mounds. Then the soils from the four mounds were combined into one composite sample per field (around 500 to 1000 g of soil).

An additional set of 14 composite soil samples was collected by the International Center for Research in Agroforestry (ICRAF) at Liliyo from one sentinel site called "Petit-Bouaké" [12]. Sampling took place between 25 and 29 August, 2015 at positions that were previously selected for the Land Degradation Surveillance Framework (LDSF) in a spatially stratified manner [21]. The soil samples received from ICRAF were within the same landscape as the sampled soils in Liliyo within YAMSYS, but sampled from different positions. All soil samples were air-dried and stored in plastic bags until further analysis.

Soil reference analyses

The air-dried soil samples were crushed and sieved at 2 mm for all. About 60 to 70 g of the sieved soil was oven-dried at 60 °C for 24 hours, of which 20 g were ball-milled. All chemical analyses except soil pH were conducted both on the soils sampled in yam fields ($n = 80$) and the LDSF soils obtained from ICRAF ($n = 14$).

The milled soils were analyzed for total C and macronutrient (N and S) concentrations using an elemental analyzer (vario PYRO cube, Elementar Analysensysteme GmbH, Germany), coupled to a mass spectrometer (IsoPrime100, Isoprime Ltd., UK). For each of the four landscapes, two soils were selected and analyzed based on three analytical replicates for quantifying within-sample variance of the elemental analysis. For the remaining samples, the analysis was not repeated. Sulfanilamide was used as a calibration standard for the dry combustion. For pH determination 10 g of air-dried soil per sample was placed in a 50 mL Falcon tube and 20 mL of de-ionized water was added. The samples were shaken in a horizontal shaker for 1.5 hours and measured for pH using a pH electrode (Benchtop pH/ISE meter model 720A, Orion Research Inc., USA).

Resin-extractable P was used as an indicator of plant-available P, as it correlates with P uptake by plants [22]. Inorganic P was extracted using an anion exchange resin membrane [23]. The extraction method was slightly modified for each sample using only one instead of two resin strips of 6 cm × 2 cm (55164 2S, BDH Laboratory Supplies, Poole, England) saturated with $CO_3{}^{-2}$, and 2 g instead of 4 g of dried soil was weighted. No fumigation step to determine microbial P was performed, as the soils from Burkina Faso and Ivory Coast had been dried and had storage periods longer than one month between sampling and analysis. In the resin eluates (a mixture of 0.1 M NaCl and 0.1 M HCl), the concentrations of inorganic P were measured colorimetrically using the malachite green method [24]. Available micronutrient (Fe, Mn, Zn, and Cu) concentrations were determined with the diethylenetriaminepentaacetic acid (DTPA) extraction, as described in [25]. The extracting solution consisted of 0.0005 M DTPA, 0.01 M $CaCl_2$, and 0.1 M triethanolamine. Briefly, 10 g of the sieved (<2 mm) soils were extracted with 20 mL of DTPA solution. Micronutrient concentrations in the filtrates were measured by inductively coupled plasma optical emission spectroscopy (ICP-OES, a Shimandzu Plasma Atomic Emission Spectrometer ICPE-9820). Final DTPA extractable concentrations of Fe, Mn, Zn, and Cu were calculated back to kg per dry soil. For each landscape, two soils were selected and analyzed in triplicates to assess analytical errors. For the remaining soils the analysis was not repeated.

The concentrations of total element (Fe, Si, Al, K, Ca, P, Zn, Cu, and Mn) in the soil was assessed by energy dispersive X-ray fluorescence spectrometry (ED-XRF) measurements on a SPECTRO XEPHOS instrument (SPECTRO Analytical Instruments GmbH, Germany). For each sample 4 g of the milled soil was used. Exchangeable cations (Ca^{2+}, Mg^{2+}, K^+, Na^+, and Al^{3+}) were analyzed by the $BaCl_2$ method [26]. About 2 g of the air-dried soil (<2 mm) were extracted by shaking for 2 hours with 30 mL of 0.1 M BaCl2 on a horizontal shaker (120 cycles min^{-1}). The suspension was filtered through no. 40 filter paper (Whatman, Brentford, UK). For each landscape, two soils were analyzed in analytical triplicates. The concentrations of exchangeable cations in the $BaCl_2$ extract were determined by inductively coupled plasma optical emission spectroscopy (ICP-OES, Shimandzu Plasma Atomic Emission Spectrometer ICPE-9820). Different $BaCl_2$ extract dilutions were used in order to obtain an optimal signal intensity for the quantification of specific elements across all samples. Concentration of H^+ per kg dry soil was calculated based on the pH measured in the $BaCl_2$ extractant. The $BaCl_2$ extraction does only slightly modify pH and is therefore an appropriate method to calculate effective CEC (CEC_{eff}) at native soil pH. Using the concentrations of the $BaCl_2$-extractable cations (i.e. Ca^{2+}, Mg^{2+}, K^+, Na^+, Al^{3+} and H^+), CEC_{eff} was

calculated as sum of exchangeable cations in cmol of cation charge per kg dry soil. Exchangeable acidity was defined by the sum of exchangeable Al^{3+} and H^{+}. Base saturation in % was calculated as ratio of the sum of basic cations (Ca^{2+}, Mg^{2+}, K^{+}, Na^{+}) in cmol(+) per kg soil to the CEC_{eff} multiplied by 100.

Particle size analysis was conducted as described in [27]. Briefly, 50 g of dried 2 mm sieved soil was stirred with 50 mL sodium hexametaphosphate and 100 mL of deionized water. Readings with a hydrometer (ASTM 152 H) were taken after letting it stand in the suspension for 30 minutes.

Spectroscopic measurements

The milled soils ($n = 94$) were measured on a Bruker ALPHA DRIFT spectrometer (Bruker Optics GmbH, Ettingen, Germany), which was equipped with a ZnSe optics device, a KBr beamsplitter, and a DGTS (deuterated tri-glycine sulfate) detector. Mid-IR Spectra were recorded between 4000 cm^{-1} and 500 cm^{-1} with a spectral resolution of 4 cm^{-1} and a sampling resolution of 2 cm^{-1}. Reflectance (R) spectra were transformed to apparent absorbance (A) using $A = \log_{10}(1/R)$ and corrected for atmospheric CO_2 using macros within the OPUS spectrometer software (Bruker Corporation, US). The spectra were referenced to a IR-grade fine ground potassium bromide (KBr) powder spectrum, which was measured prior the first soil sample and repeatedly measured every hour. All spectra were recorded by averaging 128 measurements for each of the three sample repetitions per soil.

Spectroscopic modeling

Processing of soil spectra

Three replicates of spectra were averaged for each sample. The spectra were transformed by using a Savitzky-Golay smoothed first derivative using a third-order polynomial and a window size of 21 points (42 cm^{-1} at spectrum interval of 2 cm^{-1}) [28]. Prior to spectral modeling, Savitzky-Golay preprocessed spectra were further mean centered and scaled (divided by standard deviation) at each wavenumber.

Model development and validation

The measured soil properties were modeled by applying partial least squares regression (PLSR) [29] with the preprocessed spectra as predictors. PLSR is a dimensionality reduction technique that works well for small data sets with correlated predictor variables. The models were fitted using the orthogonal scores PLSR algorithm. The PLSR development was done using cross-validation. In particular, 5-times repeated 10-fold cross-validation was performed to provide unbiased and precise assessment of predictive model performance [30], [31]. For each individual soil property, the number of factors for the most accurate PLSR model was tuned separately. For each soil property model, the sample set was repeatedly randomly split into $k = 10$ (approximately) equally-sized subsets without replacement for all repeats $r = 1, 2, .., 5$ and all candidate values in the tuning grid with the number of PLSR factors (ncomp) $= 1, 2, ..., 10$. Within each of the $r \times \text{ncomp} = 5 \times 10 = 50$ resampling data set splits, each of the 10 possible held-out and model fitting set combinations (folds) was subjected to candidate model building at the respective ncomp, using $k - 1 = 9$ out of 10 subsets and remaining held-out samples were predicted based on the fitted models. The root mean square error (RMSE, eq. (1)). of the held-out samples was calculated by aggregating all repeated K-fold cross-validation predictions ($\hat{y}_i$) and corresponding observed values (y_i) grouped by ncomp, which resulted in a cross-validated performance profile RMSE vs. ncomp.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

Based on this performance profile, the minimal ncomp among the models whose performance was within a single standard error ("One standard error" rule, [32]) of the lowest numerical value of RMSE was selected.

Model assessment was done with the best factors for each properties using cross-validation hold outs. We reported the cross-validated measures RMSE, R^2 (coefficient of determination) obtained via linear least-squares regression, and ratio of performance to deviation (RPD), after averaging predictions across repeats. The RPD index is the ratio of the chemical reference data standard deviation to the RMSE of prediction.

$$\text{RPD} = \frac{s_y}{\text{RMSE}} \quad (2)$$

Besides calculating the above listed performance indexes, accuracy (bias) and precision (variance) of resampling-based held-out predictions was expressed and depicted on an individual soil sample basis. Particularly, prediction means and 95 % confidence intervals by cross-validation (Eq. 3 and 4; $n = r = 5$) were compared against observed values in order to give prediction uncertainties from the cross-validation and show that the chosen resampling was appropriate.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{\hat{y}_i}\right)^2 \quad (3)$$

$$\overline{\hat{y}_i} \pm t(n-1, 1-\alpha/2)\frac{S_n}{\sqrt{n}}; \alpha = 0.05 \quad (4)$$

In order to cover the full training data space in the models for future sample predictions, the final PLSR models were rebuilt using the entire training set and the respective values of optimal final ncomp determined by the procedure described above.

Besides the above mentioned model evaluation metrics, mean squared error (MSE) and its partitioned additive components squared bias (SB), non-unity slope (NU), and lack of correlation (LC) were computed as described by [33].

$$\text{MSE} = \text{SB} + \text{NU} + \text{LC} \quad (5)$$

$$= \text{RMSE}^2 \quad (6)$$

$$\text{SB} = \text{Bias}^2 = \left[\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i\right]^2 \quad (7)$$

$$\text{NU} = (1-b)^2 \times \overline{\hat{y}_i} \quad (8)$$

$$\text{LC} = (1-r^2) \times \overline{y^2} \quad (9)$$

where b denote the slope and r^2 the coefficient of determination of the least-squares regression of observed (y_i) on predicted ($\hat{y}_i$) data. In short, the three additive components allow to find prominent basic types of model errors. These are translation (SB), rotation (NU) and scatter (LC).

Model interpretation

Vis–NIR spectra contain complex information about the relationship to soil composition and its properties. To establish a predictive relationship, statistical models need to find relevant spectral features for each soil property. Model interpretation requires a variable importance assessment to decide on the contribution of spectral variables to prediction and to explain spectral mechanisms. Therefore, we conducted model interpretation based on the variable importance in projection (VIP) method [34, 35], using the model at respective best number of factors (ncomp). The VIP measure v_j was calculated for each wavenumber variable j as

$$v_j = \sqrt{p \sum_{a=1}^{A} \left[\mathrm{SS}_a \left(w_{aj} / \|w_{aj}\| \right)^2 \right] / \sum_{a=1}^{A} (\mathrm{SS}_a)} \tag{10}$$

where w_{aj} are the PLSR weights for the a^{th} component for each of the wavenumber variables and SS_a is the sum of squares explained by the a^{th} component:

$$\mathrm{SS}_a = q_a^2 \, t_a^{\mathsf{T}} \, t_a \tag{11}$$

where q_a are the scores of the predicted variable y and t_a are the scores of the predictors X. These VIP scores account for multicollinearity found in spectra and are considered as robust measure to identify relevant predictors. Important wavenumbers were classified with a VIP score above 1. A variable with VIP above 1 contributes more than average to the model prediction. For model interpretation, we only computed VIP at the respective finally chosen number of PLS components a_{final} for each considered model. We focused on a selection of four well performing models with $R^2 \geqslant 0.8$ (RPD $\geqslant 2.3$) to illustrate model interpretation. These were total C, total N and clay content.

Data and code availability

The entire analysis was performed using the `R` statistical computing language and environment (version 3.4.2) [36]. We used the pls [37] package for PLSR, as described by [38]. Cross-validation resampling, model tuning, and assessment was done using the caret package [39]. Custom functions from the `simplerspec` package were used for spectroscopic modeling [40]. The data and the code to reproduce the results of this study is available online via Zenodo [41].

Results

Chemical and physical properties of yam soils

The distribution of soil properties at the yam fields shows a wide variation across the landscapes (S3 Fig). Total C concentrations across all fields ranged from $2.4\,\mathrm{g\,C\,kg^{-1}}$ soil to $24.7\,\mathrm{g\,C\,kg^{-1}}$ soil. Total C values at the landscape scale were the lowest (median) in Léo and the highest in Tiéningboué. Soils from yam fields in the two landscapes from Ivory Coast (13.0 ± 5.4 g C $\mathrm{kg^{-1}}$ soil; mean $\pm$ standard deviation) had relatively high total C compared to the fields in the landscapes in Burkina Faso (6.1 ± 3.6 g C $\mathrm{kg^{-1}}$ soil). The range of total soil C concentrations within individual landscapes was similar for Léo, Midebdo, and Tiénigboué. Total C across the fields in Léo had the smallest range with $2.9\,\mathrm{g\,C\,kg^{-1}}$ soil to $9.0\,\mathrm{g\,C\,kg^{-1}}$ soil. The median value and variation of $\mathrm{CEC_{eff}}$ exhibited similar patterns across the landscapes to total C. Total N concentrations across all fields ranged from $0.18\,\mathrm{g\,N\,kg^{-1}}$ soil to $2.48\,\mathrm{g\,N\,kg^{-1}}$ soil. Total N within and across the four landscapes exhibited a similar pattern as total C.

Generally, the landscapes in Burkina Faso were low in total N compared to those from Ivory Coast (0.44 ± 0.24 g N kg^{-1} soil vs. 1.09 ± 0.46 g N kg^{-1} soil). Median total N concentrations were almost identical for Liliyo and Tiéningboué ($1.1\,g\,N\,kg^{-1}$ soil). Total S concentrations varied between $41\,mg\,S\,kg^{-1}$ soil to $242\,mg\,S\,kg^{-1}$ soil across all fields, and showed a similar pattern as total C and N. The yam fields in the landscapes of Bukina Faso had on average more than two times higher total S than the other landscapes. Total P concentrations were in a similar range for the landscapes Léo, Midebdo, and Liliyo. In Tiéningboué, total P values were on average almost two times higher than the remaining fields ($817\,mg\,S\,kg^{-1}$ soil vs. $453\,mg\,S\,kg^{-1}$ soil), with more within-landscape variation.

Total Fe, total Al, total Ca, total Zn, and total Cu concentrations in the soil tended to be high for the landscapes in Ivory Coast, compared to the soils in Burkina Faso. In general, their ranges and interquartile ranges represented more variation in the micronutrients for the landscapes in Ivory Coast. Total K concentration was highly variable within and across the landscapes. The largest range of total K was found in Liliyo. The median and variation of total K concentration were the lowest in Midebdo, while the highest total K median was measured for yam fields in Léo.

Soil resin P concentrations varied between $0.8\,mg\,P\,kg^{-1}$ soil to $33.1\,mg\,P\,kg^{-1}$ soil. In Tiéningboué resin-extractable P was on average higher than in soils within the other landscapes. Median extractable Fe and its interquartile ranges were comparable across the landscapes. However, there were some fields where extractable Fe reached values higher than $100\,mg\,Fe\,kg^{-1}$ soil. Median extractable Zn values showed a similar pattern as total C, with the highest median values and interquartile range in Tiéningboué and the lowest in Léo. In comparison, the highest median values and interquartile range of extractable Cu and Mn were found in Liliyo. For extractable Zn, Cu, and Mn median values and interquartile range were higher in the two landscapes in Ivory Coast than the two landscapes in Burkina Faso.

Across all samples and landscapes, soil pH(H_2O) varied between 4.7 to 8.4. Median pH(H_2O) was comparable in Tiéningboué (= 6.4), Liliyo (= 6.5), and Midebdo (= 6.5). Median pH(H_2O) of yam fields in Léo (= 6) was lower than in the other landscapes. Exchangeable K, Ca, and Mg concentrations showed similar geographic patterns across the four landscapes. In Burkina Faso, each of the exchangeable cations showed relatively low median concentrations across the fields and less landscape-level variation than in Ivory Coast. In general, the highest median and variation of exchangeable cations were measured for the yam field soils in Tiéningboué among the landscapes. Median exchangeable Al values were comparable among the landscapes, although there were some outliers with exchangeable Al $> 20\,mg\,kg^{-1}$ soil for Midebdo, Liliyo, and Tiéningboué. The CEC_{eff} ranged from $0.9\,cmol(+)\,kg^{-1}$ soil to $14.6\,cmol(+)\,kg^{-1}$ soil across all fields and landscapes. Median CEC_{eff} tended to increase in the following order across landscapes: Léo > Midebdo > Liliyo > Tiéningboué. The interquartile range of CEC_{eff} was also the highest in Tiéningboué and the lowest in Léo.

Soil mid-IR spectroscopic models

Cross-validated spectroscopic predictions of all soil attributes together with the chemical reference data are shown in S1 Table.

Among the measured soil properties, models for total K (RPD = 6.4) and total Al (RPD = 6.2) were best performing. Out of a total of 27 soil attributes, 9 were well quantified by the models when considering categorization judged upon on an $R^2_{rcv} \geqslant 0.8$ criterion and 11 when applying a threshold of RPD $\geqslant 2$. Within the latter group, 4 soil attributes are directly related to the mineralogy (total Fe, Al, K and Ca), 3 are related to soil organic matter (total C, N and S), 1 texture (clay fraction), 1 to plant nutrition (exchangeable Fe), and 2 related to mineralogy and plant nutrition (exchangeable Ca

and CEC_{eff}). Fig S4 Fig shows the model evaluation summary and mean cross-validated predictions including resampling confidence intervals of best performing models with RPD $\geqslant 2$. The resampling prediction intervals were very narrow, showing all PLSR models were stable.

Total C was accurately predicted, with an RPD of 3.7 and a RMSE of $1.6\,g\,C\,kg^{-1}$ soil. The models were also able to predict total N well (RPD = 3; RMSE = $0.2\,g\,C\,kg^{-1}$ soil. Prediction accuracy of total S was slightly lower than for total C, but its RPD and RMSE suggest that the model was reliable for prediction. However, exchangeable K (RPD = 1.1), resin P (RPD = 1.3) and BS_{eff} (RPD = 1) were poorly predicted (S1 Table). Predictions for percent clay were reliable ($R^2 = 0.8$; RMSE = 2 %), whereas predictions for percent sand ($R^2 = 0.45$; RMSE = 8 %) and percent silt ($R^2 = 0.43$; RMSE = 6 %) were not accurate.

Finally chosen models of all soil attributes had between 1 and 9 PLSR components. Among the mid-IR PLSR models for the measured soil attributes, LC contributed between 97 % and 100 % to the MSE (mean squared error). There was no contribution of SB to MSE and NU made only marginal contribution (0 % to 3 % to MSE).

Model interpretation

S5 Fig shows variable importance for spectral predictors, which is superimposed by the preprocessed spectra and the raw absorbance spectra. A large proportion of absorptions had VIP > 1 for each the total C, total N and percent clay models (S5 Fig). Important wavenumbers (VIP > 1) for total C were mostly between $3140\,cm^{-1}$ and $1230\,cm^{-1}$. Besides clear absorption peaks, there were relatively continous spectral features that were important to the models. For example, the relatively continuous and smooth spectral region between the alkyl C–H vibrations at $2855\,cm^{-1}$ and $2362\,cm^{-1}$ had comparable contribution to the model as peak regions associated with total C prediction. Variable importance patterns across wavenumbers were almost identical for total C and N PLSR models, and its reference measurements were strongly correlated ($r = 0.94$; Fig S1 Appendix). In contrast, the clay content model deviated from the total C model in particular regions, for example around the kaolinite OH^- feature at $3620\,cm^{-1}$ or at kaolinite Al–O–H vibrations at $934\,cm^{-1}$ and $914\,cm^{-1}$.

Discussion

Timely and accurate estimates of multiple soil properties are required to better understand and predict soil constraints across the yam belt in West Africa. The soil spectral library from our study, which includes four landscapes of the yam belt, can be practical to monitor and manage soil infertility that is considered a major constraint for yam production in West Africa. Specifically, our results show that the total amount of C, nutrients, and exchangeable cations can be accurately estimated with mid-IR spectra in the selected yam growing landscapes. To estimate the availability of specific nutrients, however, more efforts need to be made to measure them in fine temporal and spatial resolution. Nevertheless, the spectroscopic models can be used to predict the potential of new soil samples for C and nutrient storage in the region. As soil sensing/spectral modeling complement conventional soil measurements and monitoring, its cheap high-throughput screening of the yam soils contribute effectively to practically describing their fertility status and other broader conditions. Our library includes new spectral models to build a capacity to assess broader soil conditions in experimental and on-farm studies within the selected landscapes. Furthermore, it can support to improve yam cropping practices and the management of soils' functions.

The mid-IR model accurately estimated C (RMSE = 1.6 g kg^{-1} soil). Typically, only field-scale spectroscopic models achieve such accuracy in that range [13, 42] compared to the predictive accuracy reported for larger-scale application of spectroscopic models [16, 43, 44]. Models covering a wide geographical range of soils often result in high prediction errors [45]. Despite different soil types and climate regimes across a wide geographic spacing between the calibration fields, we achieved an accurate spectroscopic estimation of total C. The model was also able to reliably estimate a range of important soil properties in addition to total C. Specifically, candidate soil variables eligible for a mid-IR quantification include total C, total N, total S, total Ca, total K, total Al, exchangeable Ca, Fe DTPA, $CEC_{eff.}$, and clay content ($R^2 > 0.75$). Reference measurements for total N, C, S, exchangeable Ca and $CEC_{eff.}$ were highly correlated to total C (Fig S1 Appendix). As a result, these properties can as well be estimated accurately with mid-IR modeling. Total Ca, Al, and clay content correlated a bit less to total C ($r > 0.70$). Total K and Fe DTPA were poorly correlated to total C. Nevertheless, their spectroscopic estimates were relatively accurate. This suggests that the mid-IR prediction of yam soils is closely based on correlation to organic carbon as well as other absorption features of many organic and mineral soil components. We also found reasonable prediction accuracy for Zn(DTPA) and Cu(DTPA), despite that soil nutrients that are extraction-based or dependent on surface chemistry usually have variable predictive performance [46] . Because relationships between soil composition and soil matrix exchange processes are typically complex, it may not be represented in the models in a straight-forward manner [13, 46]. There was good prediction of CEC by mid-MIR spectroscopy as it depends on clay type, proportion of clay and organic matter contents, whose spectral features can be well detected in the mid-IR [46]. Overall, the PLSR models were unbiased and errors were mostly attributed to a lack of correlation.

The VIP assessment of the total C, N, and clay models identified a broad set of important spectral features across the mid-IR range for their prediction (S5 Fig). This is was is typically reported literature as the mid-IR contains many relevant predictive spectral variables [47]. For example, clay content was characterized with the absorbance bands related to soil organic C as well clay lattice primary mid-IR vibrations [46]. The regions of Kaolinite O – H lattice absorptions near 3996 cm^{-}1 and 3620 cm^{-}1 as well as K – OH absorptions at 12 345 cm^{-}1 and 914 cm^{-}1 were particularly important for clay content estimation. For total C and total N, the model contribution of spectal variables was almost identical. This is probably due to high stoichiometric correlation between C and N ($R^2 = 0.85$, Fig S1 Appendix) and effects from a high proportion of N that is bound to SOM.The models for total C and clay content exploited similar spectral predictors between 3850 cm^{-}1 and 1750 cm^{-}1, although there were different influential spectral features in the remaining regions. Total C had a less pronounced linear relationship with % clay ($R^2 = 0.51$, Fig S1 Appendix), which explains only partial resemblance of the corresponding important spectral variables. In the tropical soils in the yam belt of West Africa, it is common that mineral signals overlap organic because total C and hence SOM content is generally low, i.e. ranging between 2.4 g C kg^{-1} soil and 24.7 g C kg^{-1} soil . Furthermore, various soil compounds correlated with key properties of interest, which are also themselves correlated. For example, CEC is affected by SOM, iron oxides and clay types in the soils, where these compounds have unique absorptions. Hence, we expected that these spectral features would be useful for predicting the soil properties.

All mid-IR spectra measured for soils in the four landscapes exhibited a similar characteristic pattern of absorbance (S5 Fig). The spectra measured in the soils across the landscapes mostly resembled quartz dominated soils in the minearal sub-spaces matched by Sila et al. [16] and were spectrally relative homogeneous. The quartz dominated spectral features can be explained by high sand contents across the

landscapes (range 30 % to 92 %, median 76 %). Besides quartz features, the spectra also showed prominent kaolinite peaks at 3695 cm^{-1} (surface OH^- groups), 3620 cm^{-1} (inner OH^- groups), 914 cm^{-1} (inner OH^- groups), and 936 cm^{-1} (outer OH^- groups) [48]. Separated kaolinite Si-O bands at 1034 cm^{-1} and 1011 cm^{-1} were not detected in the spectra most likely because broad quartz Si-O features superimposed these peaks.

Our study potentially serves as the framework to build a more representative soil spectroscopic library for the West African yam belt. At the current stage, we propose to implement a spectroscopy-driven approach to diagnose soils from the other locations within soil management and yam innovation trials. However, the data set presented here is relatively small and no randomized spatial sampling strategy was used for selecting locations. Hence, this effort to enhance the library to achieve better spatial coverage is required as our study clearly showed both opportunities and limitations of mid-IR soil spectroscopy for future diagnostics in the selected landscapes of the yam belt. To continue developing a reliable soil spectroscopic library, three requirements should be be fulfilled as follows according to [49]. First, the soil spectroscopic library should cover more soil variation of the region. Second, the workflow from sampling to the measurements should be consistently and thoroughly done. Finally, laboratory reference analyses required for empirical calibration should allow to be certified and standardized.

When applying the spectroscopic models in the presented landscapes, predictions exceeding the range of observed laboratory reference values should be analyzed chemically in order to verify models. After this, the calibration library can be updated with the newly characterized samples. Once such a spectroscopic library targeted to yam production fields is growing, a general model as the one presented here can be replaced by a more location-targeted approach. For example, a subset or subspace of the spectral library can be selected for each new location to be predicted. This can lead to more targeted and accurate models that are adapted to regional soils. We suggest to implement the ReSampling-Local (RS-LOCAL) data-driven method developed by [50], once the soil spectral library is enlarged with many new samples from yam growing regions.

Conclusion

We developed and tested models with mid-MIR spectra to estimate soil chemical and physical properties relevant to yam production at the four landscapes in the yam belt of West Africa. The selected properties are applied widely for agronomic performance evaluation but routinely quantified by wet chemistry. As an alternative to the traditional approach, we showed that mid-IR spectroscopy models have the potential to cost-effectively and rapidly determine the distribution and variability of important soil properties across yam production landscapes. Model evaluation supports that a list of the soil properties can be accurately predicted in a soil diagnostic mid-IR monitoring system. Specifically, total C, total N, total S, total Fe, total Al, total K, total Ca, exchangeable Ca, CEC_{eff}, Fe(DTPA), % clay, can be potentially suitable for quantification (RPD > 2; $R^2 > 0.75$) when aiming to predict in the range of soil property values found in the environmental conditions covered by this study. This study delivered parsimonious, unbiased and accurate mid-IR spectroscopy-based models to monitor and predict soil fertility. Therefore, this study laid the foundation of starting a soil spectral library within the studied landscapes of the West African yam belt.

Supporting information

S1 Fig. Studied landscapes. The location of four sampled landscapes in the yam belt of West Africa. Liliyo and Tieningboue are in Ivory Coast, and Midebdo and Leo are in Burkina Faso.

S2 Fig. Spatial distribution of yam fields by studied landscape. A total of 20 yam fields were selected in each landscape. The size of yam fields ranges from approximately to 0 is 3 ha.

S3 Fig. Soil chemical properties per landscape. The number of soils analyzed for each individual property is indicated above the top whiskers.

S4 Fig. Assessment of mid-IR PLSR models Predicted (from 5 × repeated 10-fold cross-validation) vs. observed soil properties (determined by laboratory reference analyses). Only soil properties modeled with RPD > 2 ($R^2 > 0.75$) are shown.

S5 Fig. Variable Importance in the Projection (vip) scores of PLS regression models for total soil C, total N and % clay , including overlaid raw and preprocessed spectra. Top panel shows resampled mean sample absorbance spectra ($n = 94$). Prominent peaks were identified by a picking local maxima with a span of 10 points ($20\,\mathrm{cm}^{-1}$) for the selected wavenumbers. Fundamental mid-IR vibrations that are well described in the literature [44, 48, 51] were added as labels when identified peaks matched literature assignments. (Q) stands for quartz and (K) for kaolinite. The middle panel depicts preprocessed spectra (Savitzky-Golay first derivative with a window size of 21 points ($42\,\mathrm{cm}^{-1}$); 3rd order polynomial fit). The bottom panel shows variable importance in the projection (VIP) for three selected well performing PLSR models (total C, total N and % clay; RPD > 2). The black horizontal line at VIP $= 1$ indicates the threshold above where absorbance at the wavenumbers explain more than average to the prediction of a certain soil property. Dashed points closely below the $y = 0$ line of the VIP graph visualize positive (above $y = 0$) and negative (below $y = 0$) PLSR β coefficients.

S1 Appendix. Correlation matrix of measured soil properties. Soil properties were measured with conventional laboratory analyses ($n = 94$). Pearson correlation coefficients (r) were rounded to 1 digit.

S1 Table. Descriptive summary of soil reference data and evaluation results of cross-validated plsr models. All samples across the four landscapes were aggregated into a single model per respective soil property. Model evaluation was done on held-out predictions of 5 times repeated 10-fold cross-validation (abbreviated by rcv) at the finally selected number of PLSR components (ncomp).

Acknowledgments

References

1. Food and Agriculture Organization of the United Nations. FAOSTAT statistics database; 2019. Available from: `www.fao.org/faostat/`.

2. Padwick GW. Fifty Years of *Experimental Agriculture* II. The Maintenance of Soil Fertility in Tropical Africa: A Review. Experimental Agriculture. 1983;19(4):293–310. doi:10.1017/S001447970001276X.

3. Syers JK, Campbell AS, Walker TW. Contribution of organic carbon and clay to cation exchange capacity in a chronosequence of sandy soils. Plant and Soil. 1970;33(1-3):104–112. doi:10.1007/BF01378202.

4. Soares MR, Alleoni LRF. Contribution of Soil Organic Carbon to the Ion Exchange Capacity of Tropical Soils. Journal of Sustainable Agriculture. 2008;32(3):439–462. doi:10.1080/10440040802257348.

5. Diby LN, Hgaza VK, Tie TB, ASSA A, Carsky R, Girardin O, et al. Productivity of Yams (Dioscorea Spp.) as Affected by Soil Fertility. Journal of Animal & Plant Sciences. 2009;5(2):494–506.

6. Kassi SPAY, Koné AW, Tondoh JE, Koffi BY. Chromoleana Odorata Fallow-Cropping Cycles Maintain Soil Carbon Stocks and Yam Yields 40 Years after Conversion of Native- to Farmland, Implications for Forest Conservation. Agriculture, Ecosystems & Environment. 2017;247:298–307. doi:10.1016/j.agee.2017.06.044.

7. Carsky RJ, Asiedu R, Cornet D. Review of soil fertility management for yam-based systems in west africa. African Journal of Root and Tuber Crops. 2010;8(2):1.

8. Frossard E, Aighewi BA, Aké S, Barjolle D, Baumann P, Bernet T, et al. The Challenge of Improving Soil Fertility in Yam Cropping Systems of West Africa. Frontiers in Plant Science. 2017;8. doi:10.3389/fpls.2017.01953.

9. O'Sullivan JN, Jenner R. Nutrient Deficiencies in Greater Yam and Their Effects on Leaf Nutrient Concentrations. Journal of Plant Nutrition. 2006;29(9):1663–1674. doi:10.1080/01904160600851569.

10. Abbott LK, Murphy DV, editors. Soil Biological Fertility: A Key to Sustainable Land Use in Agriculture. Springer Netherlands; 2007. Available from: //www.springer.com/de/book/9781402017568.

11. Cornet D, Sierra J, Tournebize R, Gabrielle B, Lewis FI. Bayesian Network Modeling of Early Growth Stages Explains Yam Interplant Yield Variability and Allows for Agronomic Improvements in West Africa. European Journal of Agronomy. 2016;75:80–88. doi:10.1016/j.eja.2016.01.009.

12. UNEP. Land Health Surveillance: An Evidence-Based Approach to Land Ecosystem Management. Illustrated with a Case Study in the West Africa Sahel; 2012.

13. Nocita M, Stevens A, van Wesemael B, Aitkenhead M, Bachmann M, Barthès B, et al. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. In: Advances in Agronomy. vol. 132. Elsevier; 2015. p. 139–159. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0065211315000425.

14. Viscarra Rossel RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma. 2006;131(1–2):59–75. doi:10.1016/j.geoderma.2005.03.007.

15. Cécillon L, Barthès BG, Gomez C, Ertlen D, Genot V, Hedde M, et al. Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). European Journal of Soil Science. 2009;60(5):770–784. doi:10.1111/j.1365-2389.2009.01178.x.

16. Sila AM, Shepherd KD, Pokhariyal GP. Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. Chemometrics and Intelligent Laboratory Systems. 2016;153:92–105. doi:10.1016/j.chemolab.2016.02.013.

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

18. Clairotte M, Grinand C, Kouakoua E, Thébault A, Saby NPA, Bernoux M, et al. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. Geoderma. 2016;276:41–52. doi:10.1016/j.geoderma.2016.04.021.

19. Viscarra Rossel RA, Behrens T, Ben-Dor E, Brown DJ, Demattê JAM, Shepherd KD, et al. A global spectral library to characterize the world's soil. Earth-Science Reviews. 2016;155:198–230. doi:10.1016/j.earscirev.2016.01.012.

20. FAO. World Reference Base for Soil Resources 2014: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. FAO; 2014.

21. Vagen TG, Shepherd KD, Walsh MG, Winowiecki L, Desta LT, Tondoh JE. AfSIS technical specifications: Soil Health Surveillance.; 2010. Available from: http://www.worldagroforestry.org/sites/default/files/afsisSoilHealthTechSpecs_v1_smaller.pdf.

22. Nuernberg NJ, Leal JE, Sumner ME. Evaluation of an anion-exchange membrane for extracting plant available phosphorus in soils. Communications in Soil Science and Plant Analysis. 1998;29(3-4):467–479. doi:10.1080/00103629809369959.

23. Kouno K, Tuchiya Y, Ando T. Measurement of soil microbial biomass phosphorus by an anion exchange membrane method. Soil Biology and Biochemistry. 1995;27(10):1353 – 1357. doi:http://dx.doi.org/10.1016/0038-0717(95)00057-L.

24. Ohno T, Zibilske LM. Determination of Low Concentrations of Phosphorus in Soil Extracts Using Malachite Green. Soil Science Society of America Journal. 1991;55(3):892. doi:10.2136/sssaj1991.03615995005500030046x.

25. Lindsay WL, Norvell WA. Development of a DTPA soil test for zinc, iron, manganese, and copper. Soil science society of America journal. 1978;42(3):421–428.

26. Hendershot WH, Duquette M. A simple barium chloride method for determining cation exchange capacity and exchangeable cations. Soil Science Society of America Journal. 1986;50(3):605–608.

27. Bouyoucos GJ. A recalibration of the hydrometer method for making mechanical analysis of soils. Agronomy journal. 1951;43(9):434–438.

28. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry. 1964;36(8):1627–1639. doi:10.1021/ac60214a047.

29. Wold S, Martens H, Wold H. The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. In: Kågström B, Ruhe A, editors. Matrix Pencils. vol. 973. Springer Berlin Heidelberg; 1983. p. 286–293. Available from: http://link.springer.com/10.1007/BFb0062108.

30. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21(15):3301–3307. doi:10.1093/bioinformatics/bti499.

31. Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics & Data Analysis. 2009;53(11):3735–3745. doi:10.1016/j.csda.2009.04.009.

32. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis; 1984. Available from: https://books.google.ch/books?id=JwQx-WOmSyQC.

33. Gauch HG, Hwang JT, Fick GW. Model evaluation by comparison of model-based predictions and measured values. Agronomy Journal. 2003;95(6):1442–1446.

34. Wold S, Johansson E, Cocchi M. PLS-partial least squares projections to latent structures. 3D QSAR in drug design. 1993;1:523–550.

35. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems. 2005;78(1-2):103–112. doi:10.1016/j.chemolab.2004.12.011.

36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.R-project.org/.

37. Mevik BH, Wehrens R, Liland KH. pls: Partial Least Squares and Principal Component Regression; 2019. Available from: https://CRAN.R-project.org/package=pls.

38. Martens H, Naes T. Multivariate Calibration. Wiley Chichester; 1989.

39. from Jed Wing MKC, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al.. caret: Classification and Regression Training; 2019. Available from: https://CRAN.R-project.org/package=caret.

40. Baumann P. philipp-baumann/simplerspec: Beta release simplerspec 0.1.0 for zenodo; 2019. Available from: https://doi.org/10.5281/zenodo.3303637.

41. Baumann P. philipp-baumann/yamsys-soilspec-publication: Pre- release of spectral models for the publication of the soil spectral library of the West African yam belt; 2018. Available from: https://doi.org/10.5281/zenodo.1174869.

42. Guerrero C, Wetterlind J, Stenberg B, Mouazen AM, Gabarrón-Galeote MA, Ruiz-Sinoga JD, et al. Do We Really Need Large Spectral Libraries for Local Scale SOC Assessment with NIR Spectroscopy? Soil and Tillage Research. 2016;155:501–509. doi:10.1016/j.still.2015.07.008.

43. Rossel RAV, Webster R. Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. European Journal of Soil Science. 2012;63(6):848–860. doi:10.1111/j.1365-2389.2012.01495.x.

44. Stevens A, Nocita M, Tóth G, Montanarella L, van Wesemael B. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. PLoS ONE. 2013;8(6):e66409. doi:10.1371/journal.pone.0066409.

45. Stenberg B, Rossel RAV. Diffuse Reflectance Spectroscopy for High-Resolution Soil Sensing. In: Rossel RAV, McBratney AB, Minasny B, editors. Proximal Soil Sensing. Progress in Soil Science. Springer Netherlands; 2010. p. 29–47. Available from: http://link.springer.com/chapter/10.1007/978-90-481-8859-8_3.

46. Janik LJ, Skjemstad JO, Merry RH. Can mid infrared diffuse reflectance analysis replace soil extractions? Australian Journal of Experimental Agriculture. 1998;38(7):681. doi:10.1071/EA97144.

47. Madari BE, Reeves JB, Machado PLOA, Guimarães CM, Torres E, McCarty GW. Mid- and near-Infrared Spectroscopic Assessment of Soil Compositional Parameters and Structural Indices in Two Ferralsols. Geoderma. 2006;136(1):245–259. doi:10.1016/j.geoderma.2006.03.026.

48. Madejová J, Kečkés J, Pálková H, Komadel P. Identification of components in smectite/kaolinite mixtures. Clay Minerals. 2002;37(2):377–388. doi:10.1180/0009855023720042.

49. Rossel RAV, Jeon YS, Odeh IOA, McBratney AB. Using a Legacy Soil Sample to Develop a Mid-IR Spectral Library. Soil Research. 2008;46(1):1. doi:10.1071/SR07099.

50. Lobsey CR, Viscarra Rossel RA, Roudier P, Hedley CB. rs-local data-mines information from spectral libraries to improve local calibrations: rs-local improves local spectroscopic calibrations. European Journal of Soil Science. 2017;doi:10.1111/ejss.12490.

51. Rossel RAV, Behrens T. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma. 2010;158(1-2):46–54. doi:10.1016/j.geoderma.2009.12.025.

Soil attribute	n	Soil reference analyses					mid-IR PLS regression (5 × rep. 10-fold CV)			
		$Min_{obs.}$	$Max_{obs.}$	$Med_{obs.}$	$Mean_{obs.}$	$CV_{obs.}$	ncomp	$RMSE_{rcv}$	$R^2{}_{rcv}$	RPD_{rcv}
Total Fe [g kg^{-1}]	94	4	35	10	12	54	5	3	0.81	2.3
Total Si [g kg^{-1}]	94	200	363	262	262	12	3	20	0.61	1.6
Total Al [g kg^{-1}]	94	10	102	48	53	42	5	4	0.97	6.0
Total K [g kg^{-1}]	94	1	34	6	10	91	7	2	0.96	5.1
Total Ca [g kg^{-1}]	94	0.3	7.6	1.4	1.9	70	5	0.6	0.78	2.1
Total Zn [mg kg^{-1}]	94	10	72	19	23	49	1	7	0.64	1.7
Total Cu [mg kg^{-1}]	94	0	29	5	7	87	3	3	0.70	1.8
Total Mn [mg kg^{-1}]	94	59	1146	222	308	74	4	118	0.73	1.9
Sand [%]	80	29.8	91.6	75.6	74.2	14	2	8.0	0.44	1.3
Silt [%]	80	3.9	54.1	12.0	14.1	60	2	6.7	0.38	1.3
Clay [%]	80	4.5	26.1	10.1	11.6	42	2	2.2	0.79	2.2
pH_{H_20}	80	4.7	8.4	6.4	6.4	11	8	0.5	0.60	1.6
K (exch.) [mg kg^{-1}]	94	0	868	104	145	95	1	118	0.30	1.2
Ca (exch.) [mg kg^{-1}]	92	98	2170	604	774	70	5	242	0.80	2.2
Mg (exch.) [mg kg^{-1}]	93	18	432	76	113	84	3	58	0.62	1.6
Al (exch.) [mg kg^{-1}]	94	0	47	0	4	258	2	9	0.21	1.1
CEC_{eff} [cmol(+) kg^{-1}]	91	0.9	14.6	4.2	5.3	67	6	1.4	0.84	2.5
BS_{eff} [%]	91	79	100	100	98	4	1	3	0.25	1.1
Total C [g kg^{-1}]	94	2.4	24.7	8.5	9.9	58	6	1.6	0.92	3.5
Total N [g kg^{-1}]	94	0.2	2.5	0.7	0.8	61	5	0.2	0.89	3.0
Total S [mg kg^{-1}]	94	41	242	99	111	46	3	20	0.85	2.6
Total P [mg kg^{-1}]	94	240	1631	467	530	40	3	132	0.61	1.6
log(P resin) [mg kg^{-1}]	92	-0.2	3.5	1.4	1.4	57	2	0.6	0.43	1.3
log(Fe(DTPA)) [mg kg^{-1}]	92	1.0	6.7	2.7	2.9	38	9	0.5	0.76	2.0
Zn (DTPA) [mg kg^{-1}]	87	0.2	11.5	1.9	2.8	89	3	2.1	0.26	1.1
Cu (DTPA) [mg kg^{-1}]	92	0.1	1.5	0.2	0.4	89	6	0.2	0.73	1.9
Mn (DTPA) [mg kg^{-1}]	92	2.5	31.4	6.5	8.6	69	3	4.0	0.55	1.5