SUBMISSION

~~Establishing a~~Assessment of mid-infrared ~~spectral library for assessing~~ soil~~-~~ spectroscopy ~~fertility related properties~~ for estimating soil properties that are important for yam production ~~across~~ in ~~the~~ West Africa~~n yam belt~~

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

1 Department of Environmental Systems Science, ETH Zurich, 8092 Zurich, Switzerland
2 CSIRO Land and Water, Canberra, ACT 2601, Australia

¤Current Address: Sustainable Agroecosystems group, Institute for Agricultural Sciences, ETH Zu¨rich, Tannenstrasse 1, 8092 Zu¨rich, Switzerland
* philipp.baumann@usys.ethz.ch

Abstract

The yam belt of West Africa spans ~~across~~ the central zone of coastal countries in West Africa and provides 93 % of the ~~worldwide~~ world's yam production. Soil degradation and nutrient depletion are severe problems that occur in many regions between the humid forest zone and the northern Guinean savanna.~~;~~ ~~whereby s~~Therefore, soil ~~f~~infertility ~~constraints are~~ is a major factor~~s~~ limiting the yields of yam and other staple crops~~' yields~~. ~~Adapting cropping and fertilization strategies is necessary to foster maintenance or improvement of soil functions, but at same time effects on s~~Soil fertility needs to be assessed quantitatively in order to improve soil functions and cropping and fertilization strategies. ~~To be able to conduct soil diagnostics for future studies, w~~We developed a ~~cost-effective~~ mid-infrared (mid-~~ir~~IR) soil spectral library to enable timely and cost-efficient ~~time-efficiently evaluate~~ assessments of ~~intermediate and long-term responses of~~ soil physical and chemical ~~soil~~ properties that affect soil fertility. Our library included a total of 94 soil samples: 80 composite ~~soil~~ samples from 80 fields in four landscapes that are representative ~~for~~ of the climate~~ic~~ and soil ~~conditions within the~~ of the West African yam belt, and ~~additional samples from~~ 14 samples from ~~fields from~~ a sentinel site ~~that was sampled under~~ of the land degradation surveillance framework, collected by the World Agroforestry Center. All soil samples were measured with a mid-IR spectrometer and analyzed ~~using~~ with conventional soil ~~laboratory~~ analytical ~~analyses~~ methods to determine total carbon, nitrogen, sulfur, iron, aluminium, potassium, total calcium, exchangeable calcium, effective cation exchange capacity, DTPA extractable iron and clay content. We derived ~~and for all measured soil properties mid-ir~~ partial least square ~~pls~~ regressions (PLSR) to estimate the soil properties from the spectra. ~~models containing soils from all fields were trained and~~ The estimates were evaluated using ~~5~~ five times repeated 10-fold cross-validation.~~. We established well performing (rpd > 2; R^2 > 0.75) mid-ir spectroscopic prediction m~~ The models of total carbon, total nitrogen, total sulfur, total iron, total alumini um, total potassium, total calcium, exchangeable calcium, effective cation exchange capacity, DTPA extractable iron and clay content~~s for total carbon, total nitrogen, total sulfur, total iron, total aluminum, total potassium, total calcium, exchangeable calcium, effective cation exchange capacity, dtpa extractable iron and clay content~~ produced accurate estimates of the soil properties (R^2 > 0.75). We suggest that within the four landscapes across the climatic gradient of the West African yam belt, these soil properties can be accurately estimated by mid-IR soil spectroscopy. ~~Hence, we suggest that these soil properties can be quantitatively assessed by soil spectroscopy for new soils within the four landscapes across the climatic gradient of the West African yam~~

~~belt. Further,~~ The spectroscopic models for total ~~zink~~zinc, pH, exchangeable magnesium, ~~dtpa~~ DTPA extractable copper and manganese ~~had~~ produced ~~intermediate model performance~~ less accurate estimates ($R^2 > 0.5$), and are therefore suitable for semi-quantitative screening of ~~high~~ large ~~vs~~or~~.~~ ~~low~~ small values, which may be adequate for general assessments of fertility. ~~Based on the results of o~~Our study suggests that mid-IR spectroscopy could be used to ~~we therefore conclude that many~~ assess soil ~~fertility related~~ properties that affect the fertility of soil for yam production. ~~can be diagnosed~~ The technique is faster and cheaper ~~by soil infrared spectrosopy~~than conventional analytical techniques, ~~thereby~~ and can potentially reduce~~ing~~ the ~~amount~~ number of conventional soil analyses ~~for~~ in future ~~studies~~ assessments in the region.

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 between the humid forest zone and the northern Guinean savanna, and contributes to about 93 % of total world yam production, with a total tuber yield of 62 million tons in 2016 [1]. The West African yam belt has been experiencing increased land pressure due to accelerated population growth, which has in many places led to an expansion of cropping areas by deforestation, causing soil degradation and nutrient depletion. ~~For example,~~ ~~there~~There is a trend of shortened fallow periods in West Africa over the last decades and a~~.~~ ~~A~~s a consequence of intensive cropping and the lack of new soil organic matter (~~som~~SOM) restoration by short-term fallow periods, soil fertility has decreased across the yam belt (reference).

~~The som~~ Soil organic matter serves as the most important pool of plant-available macro- and micronutrients and plays a particularly important role in supplying cation exchange surfaces to plant nutrients in tropical soils [2, 3]. Traditionally, yam is grown without external input in the zone. Thus, maintaining or increasing C and other nutrient levels in ~~som~~ SOM is of utmost importance for crop productivity in West Africa's tropical soils [4]. Moreover, linking soil properties and yam yields [5] and accounting for soil macro- and micronutrient availability [6] will be key to improve crop and soil management strategies.

Due to interactions of climate, soil, land use and agronomic practices, context-specific management recommendations for optimized yam cultivation are required ~~in order~~ to close the gap between actual (8.8 t ha^{-1} fresh tuber; [1]) and potential yield (110 t ha^{-1}; [7]) of the crop [4, 7–9]. A range of soil management practices have potentially varying effects on soil fertility measures [5]. Here, we consider soil fertility as an integrative qualitative measure of long-term soil quality attributes and their interactions that support the agricultural production potential. Soil fertility can then be further attributed to three main components, the physical, chemical and biological aspects of soil fertility [10]. There is the need to evaluate innovative agronomic practices regarding intermediate and long-term responses of physical and chemical soil properties in the yam belt. To achieve this, we need more cost- and time-efficient methods to quantify soil properties than traditional wet chemistry laboratory analyses, which are often cost-intensive and slow. Infrared ~~(ir)~~ spectroscopy is a mature and reliable technology that can ~~address this challenge of providing~~provide soil measurements ~~affordable and fast soil measurements~~ rapidly and inexpensively[11].

Soil visible and near infra red(~~Vis~~vis~~-nir~~–NIR), and mid-infrared (mid-~~ir~~IR) diffuse reflectance spectroscopy has been increasingly applied over the past 30 years to assess soil properties ~~cost-effectively and thereby~~to complement or in some cases replace conventional ~~many wet~~laboratory analytical ~~chemistry~~ methods [12]. ~~A large number of~~Many studies have shown successful spectroscopic ~~ir-based~~ predictions ~~for~~of soil ~~fertility~~ properties that affect soil fertility, such as organic C, texture, cation exchange capacity (cec), exchangeable K, and electrical conductivity [13] [14] [12]. Because of soil being a complex mixture of biochemical compounds and spectroscopy being characterized by complex interactions between matter and radiation, spectral processing and statistical modeling are essential for spectroscopy-based soil diagnostics. Signal preprocessing is widely used in soil spectroscopy to reduce spectral noise and light scattering effects, thereby potentially decreasing model complexity and improving its accuracy [15]. In addition, for a specific soil type or study area, laboratory reference analysis data as well

MAIN COMMENTS ABOUT INTRODUCTION :
- backgournd paragraph ok however the bits about SOM are
Raphael Viscarra Ros …
07/02/2018 23:02

I don't understand what you are trying to say here and the relevance because you are not measuring soil fertility. You are measuring some soil properties that affect fertility – I
Raphael Viscarra Ros …
07/02/2018 22:23

This is too general
Raphael Viscarra Ros
07/02/2018 22:23

Which ones are the important ones?
Raphael Viscarra Ros
07/02/2018 22:24

I don't really like these preidctions of EC as i think they are mostly due to texture and not EC as such
Raphael Viscarra Ros
07/02/2018 22:33

as calibration libraries need to be available to predict the properties of new soil samples. Depending on the study scale — field (e.g. [16]), region, country (e.g. [17]), continent (e.g. [18]), world (e.g. [19]), various statistical predictive modeling strategies have been employed to account for scale-dependent variability in soil properties. For example, when sample number is low and when targeting a ’general’ or ’global’ calibration for a soil spectral library with rather low soil variability, dimensionality reduction techniques such as partial least squares ~~(pls)~~ regression (PLSR), which can handle multicollinearity in spectra well, are commonly used to predict soil properties.

Soil ir spectral models can be used to determine empirical relationships between spectra and soil properties from a predictive point of view. However, besides solely targeting prediction, these models allow ~~to~~ interpretation and understanding of mechanisms that deliver soil predictions based on spectroscopy. In our study, we specify mechanistic relations as spectra resulting in response of a complex function involving interaction of mid-~~ir~~ IR radiation with the chemical components and physical properties (e.g. particle size distribution) of soil. Hence, when modeling soil properties by ~~ir~~ mid-IR spectroscopy, we infer relations between soil properties and physicochemical by finding relevant spectral features. Soil chemical composition and physical properties are major factors that drive ir spectral diversity, in relation with soil mineralogy, the concentration, forms and distribution of soil organic matter and soil fertility. Mechanisms that allow for an estimation of soil properties by spectroscopy and subsequent statistical modeling can be elucidated by a statistically sound model importance assessment. Model type specific variable importance measures can deliver the relative importance of a spectral regions (predictor or explanatory variable space) in a model trained based on chemical reference data and spectra.

This should go in the methods
Raphael Viscarra Ros
07/02/2018 22:37

Our ~~main~~ objectives of this study ~~are~~ were to (1) develop and evaluate mid-~~ir~~IR spectroscopic models to measure soil properties for selected landscapes representing major soil and climatic conditions in the West African yam belt, (2) to ~~infer spectral features and~~interpret the spectroscopic models and provide understanding ~~respective soil constituents that allow to predict~~in terms of the specific frequencies used in the ~~soil properties~~models to estimate the different soil properties, and (3) to propose recommendations for the application of the presented models in the studied landscapes for predicting new soils.

Note that there is nothing here about construction of a spectral library so I suggest to remove from title.
Raphael Viscarra Ros
07/02/2018 22:41

Materials and methods

Landscapes and soil sampling

Our ~~choice of study landscapes and sampling strategies~~study area ~~targeted to~~ covered the major climatic and soil biophysical conditions representative of the West African yam belt. ~~For this study, a total of~~ We selected four pilot landscapes ~~were selected in the West African yam belt, of which~~ two ~~are located~~ in the Ivory Coast and ~~the other~~ two ~~are~~ in Burkina Faso (S1 Fig). Each landscape represents a diverse geographic ecoregion in a 10 km x 10 km area. 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~~Ferralsols (WRB? reference). The landscape Ti´eningbou´e in Ivory Coast is at 8.14 °N and belongs to the forest savannah transitional zone. Dominant soil types are ~~nitrisols~~ Nitrisols and ~~lixisols~~Lixisols (WRB reference). The landscape Midebdo is at 9.97 °N and in the sub-humid savannah. Dominant soil types are ~~lixisols~~Lixisols, G~~g~~leysols, and L~~l~~eptosols. The landscape L´eo is at 11.07 °N and in the northern Guinean savannah. Dominant soil types are ~~lixisols~~ Lixisols and ~~vertisols~~Vertisols. The mean annual rainfall were approximately 1300 mm in Liliyo, and 900 mm in Ti´eningbou´e, Midebdo, and L´eo, respectively.

During July and August 2016, ~~soil sampling was done in a total of~~we sampled the soil from 80 ~~yam~~ fields used for growing yams across the four landscapes. Within each of the four landscapes, soils from 20 yam fields were sampled (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 area of about 10 km x 10 km. The yam fields had been previously selected to be equally distributed across the landscape. Yam is planted on soil mounds, typically there are 5000 to 10000 mounds per hectare with a single yam plant per mound. Within each field, four adjacent mounds in square arrangement were sampled. At each of the four selected mounds 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, whereby the sampling

So what was the design of the sample? I find it difficult to understand. So you stratified by colour and texture and then selected fields purposively or on even distances (ie a sort of grid?)??? This needs a bit better explanation
Microsoft Office User
07/03/2018 19:33

How far apart were these mounds? So what area does each composite represent? I understand the reasons for compositing, of course, but one of the advantages of spectroscopy is that you can collect many more samples with spatial registration so that then you can look at the data spatially. With compositing, you potentially lose that - depending on the area that the composite represetns.
Microsoft Office User
07/03/2018 19:35

radius depended on the size of the mounds. Soils from the four mounds were then combined into one composite sample per field (around 500 to 1000 g of soil).

An additional set of 14 soil samples was collected by the International Center for Research in Agroforestry (icraf) at Liliyo from one sentinel site called ”Petit-Bouak´e”; [11]. 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 [20]. The soil samples received from icraf were within the same landscape as the sampled soils in Liliyo within yamsys, but sampled from different positions.

Sampled soils were air-dried and stored in plastic bags until further analysis.

Soil reference analyses

The ~~air-dried~~air-dried soil samples were ~~gently ~~crushed and sieved at 2 mm for all conducted analyses. For each soil sample, about 60 to 70 g of the sieved soil was oven-dried at 60 °C for 24 hours. ~~About 20~~Twenty grams of the oven-dried soils were then 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 analyzed based on three analytical replicates for quantifying within-sample ~~uncertainty ~~variance of the elemental analysis. For the remaining samples, the analysis was not repeated. Sulfanilamide was used as a calibration standard.

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 ([21]). Inorganic P was extracted using an anion exchange resin membrane [22]. The extraction method was ~~slightly ~~modified slightly 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 dry 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), concentrations of inorganic P were measured colorimetrically using the malachite green method [23].

Available micronutrient (Fe, Mn, Zn, and Cu) concentrations were determined with the diethylenetriaminepentaacetic acid (dtpa) extraction, as described in [24]. 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. 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

These were composite samples too?
Microsoft Office User
07/03/2018 19:39

MAIN COMMENT: this section could be much reduced. I presume that most of the methods you used (except
Microsoft Office User
07/03/2018 22:14 ...

This is confusing. Please write clearly what you did and which samples you used. I guess you first dried, then
Microsoft Office User
07/03/2018 21:55 ...

For the mass spec?
Microsoft Office User
07/03/2018 22:03

???
Microsoft Office User
07/03/2018 22:03

No settling time?
Microsoft Office User
07/03/2018 22:04

Was this also to assess analytical error?
Microsoft Office User
07/03/2018 22:14

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 [25]. 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. Concentrations of exchangeable cations in the BaCl2 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 BaCl2 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 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 [26]. Briefly, 51 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.

Soil 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~~ 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 an ~~ir~~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.

Wow this is really excessive – we use 8 cm-1 – we did some experiments to compare and we found that any finer is simply adding redundant data that increases computation time. If you use 16 cm-1 it is ok even but some peaks start disappearing – the effect is minimal on the calibrations, but we still settled on 8
Microsoft Office User
07/03/2018 22:18

Often better to do this after certain number of measurements, unless of course you are measuring consistently throughout the day J
Microsoft Office User
07/03/2018 22:22

~~Soil spectral~~Spectroscopic modeling

The entire analysis was performed using the `R` statistical computing language and environment version 3.4.2 [27]. Custom `R` functions tailored to spectroscopy modeling from the `simplerspec` package and other R packages were used for spectral data management, processing and model building steps (see S1 Appendix).

Lets put this at the end. Lets describe the science and then the technical stuff
Microsoft Office User
07/03/2018 22:23

Processing of soil spectra

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

This can go in the end with the R packages. Here the reference should be to the original Savitzky-Golay paper.

Also is this on the the smoothing? Or did you use the derivative with it?
Microsoft Office User
07/03/2018 22:27

Often I do not do the scaling because all of the spectra were recorded by the same instrument. So mean centering is enough.
Microsoft Office User
07/03/2018 22:26

Model building and evaluation

~~Soil~~ The soil ~~reference data~~ properties were modeled by applying partial least squares ~~(pls)~~ regression (PLSR) ~~based on preprocessed~~with the spectra as predictors. More details on pls regression and general remarks on predictive modeling can be found in S2 Appendix.

No need for this. Simply cite the original PLS paper by Wold or the Martens and Naes, etc. This method is now well known and so a citation will suffice. However, what we
Microsoft Office User
07/03/2018 22:32 ...

The models were fitted using the orthogonal scores pls regression algorithm (pls1, single soil property response vector to predict), which is implemented in the pls [29] R package, as described by [30]). Based on the caret (classification and regression training [31]) and the pls R packages, a set of pls regression candidate models were built using cross-validation resampling. In particular, 5-times repeated 10-fold cross-validation was performed to determine the number of components (ncomp) in the final models and to estimate model performance. The repeated K-fold cross-validation procedure provides unbiased and precise internal estimates of predictive performance, targeting both most efficient use of the available data set and realistic model evaluation measures [32], [33]. For each soil property, ~~ncomp~~ the number of factors for final ~~pls~~ the most accurate ~~regression (pls1) model~~PLSR 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 ~~ncomp~~ with the number of PLSR factors (ncomp) $= 1, 2, ..., 10$. Within each of the $r \times$ 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.

Again, I would suggest that you describe what you did and cite the original references for the methods and then at the end say that the implementation of those methods was
Microsoft Office User
07/03/2018 22:43 ...

Lets write in English and not in R-inglish ;-)
Microsoft Office User
07/03/2018 22:57

$$\mathrm{rmse} =$$

$$_{i=1}(\hat{y}_i - y_i)^2 \qquad (1)$$

$$n$$

Based on this performance profile, minimal ncomp among the models whose performance was within a single standard error (“One standard error” rule, [34]) of the lowest numerical value of rmse was selected as final ncomp for the respective soil properties.

Model evaluation at the finally chosen ncomp was conducted based on estimates of the measures rmse, R^2 (coefficient of determination) obtained via linear least-squares regression on predicted vs. observed data and ratio of performance to deviation (rpd). The rpd index is the ratio of the chemical reference data standard deviation to the rmse of prediction and is a scaled index to compare goodness-of-fit across data sets with difference variances in observed values.

$$\mathrm{rpd} = \frac{s_y}{\mathrm{rmse}} \qquad (2)$$

This index is rather redundant but I know it is used in spectroscopic papers so use it. However, important to note that if your data were not normally
Microsoft Office User
07/03/2018 22:59 ...

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 detect eventual model instabilities and trends in uncertainty patterns among sample predictions.

What do you mean?
Microsoft Office User
07/03/2018 23:01

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

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

In order to cover the full training data space in the models for future sample predictions, the final pls regression 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 [35].

I have not read this paper but I am still unsure how you did this.
Microsoft Office User
07/04/2018 09:48

$$\mathrm{mse} = \mathrm{sb} + \mathrm{nu} + \mathrm{lc} \qquad (5)$$

$$= \mathrm{rmse}^2 \qquad (6)$$

$$\mathrm{sb} = \mathrm{Bias}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i\right)^2 \qquad (7)$$

$$\mathrm{nu} = (1-b)^2 \times \hat{y}_i \qquad (8)$$

$$\mathrm{lc} = (1-r^2) \times y^2 \qquad (9)$$

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).

OK so you built models with all the 94 data. But I don't get how you evaluated them – this is unclear.

On another point – I like quantifying the error by using this:

RMSE^2 = ME^2 + SDE^2 where the ME is the mean error (bias), the SDE is the standard deviation of the error (imprecision) – both of which are encompassed in the root mean square error (inaccuracy)
Microsoft Office User
07/03/2018 23:04

Model interpretation

After model tuning, model interpretation was conducted based on the variable importance in projection (vip). The vip is also known as variable influence on projection scores, which is a measure of variable importance tailored for ~~pls regression~~PLSR.

The vip implementation as first described [36] and further investigated by [37] was applied. Thereby, vip scores were calculated from the ~~the pls regression~~PLSR parameters taking multicollinearity into account, which is likely to occur because of the nature of spectroscopic data. vip scores are considered as robust measure to identify relevant predictors, in this case wavenumber variables. Important wavenumbers were classified with a vip score above 1. A variable with vip above 1 contributes more than average to the model prediction. The vip measure v_j was calculated for each wavenumber variable j as

$$v_j = \sqrt{p\sum_{a=1}^{A} SS_a\left(w_{aj}/\lVert w_{aj}\rVert\right)^2 \Big/ \sum_{a=1}^{A}(SS_a)} \qquad (10)$$

where w_{aj} are the pls regression 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.

The sums of squares SS_a for the a^{th} component were calculated from the scores q_a of the predicted variable y and the t_a scores of the spectral matrix X, given by the expression

$$\mathrm{SS}_i = q_a^2\, t_a^{\mathrm{T}}\, t_a \qquad (11)$$

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 � 0.8 (rpd � 2.3) to illustrate model interpretation. These were total C, total N and clay content.

OK but I would also suggest to summarise this a bit and a citation or two as this is a well known index
Microsoft Office User
07/04/2018 09:37

Data and code availability

ô

Ô Add a paragraph on reproducible data and modeling workflow (Open science using reproducible software tools); I will add and tag (digital identifier) raw data and scripts on a common open science repository such as Dataverse; scripts are also on github.

In this section say you used R with package x for this and package y for that, etc etc.
Microsoft Office User
07/03/2018 23:06

Results

Soil chemical analyses and mid-IR spectra

I would summarise all of this as it is useful but not the main point of this paper – show a table with some descriptive statistics and that should help to significantly reduce this section.

It would be nice to see a correlation matrix of the soil data, which might help you with
Microsoft Office User
07/03/2018 23:09
...

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 g C kg^{-1} soil to 24.7 g C kg^{-1} soil. Total C values at the landscape scale were the lowest (median) in L´eo and the highest in Ti´eningbou´e. Soils from yam fields in the two landscapes from Ivory Coast had relatively high total C compared to the fields in the landscapes in Burkina Faso ((6.1 ± 3.6) g C kg^{-1}soil (mean ± standard deviation) vs. (13.0 ± 5.4) g C kg^{-1}soil). The range of total soil C concentrations within individual landscapes was similar for L´eo, Midebdo, and Ti´enigbou´e. Total C across the fields in L´eo had the smallest range with 2.9 g C kg^{-1} soil to 9 g C kg^{-1} soil. The median value and variation of total C exhibited similar patterns across the landscapes to cec$_{\mathrm{eff}}$. Total N concentrations across all fields ranged from 0.18 g N kg^{-1} soil to 2.48 g N kg^{-1} soil. Total N within and across the four landscapes exhibited a similar pattern as total C. Generally, the two 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 Ti´eningbou´e (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 two landscapes of Bukina Faso had on average more than two times higher total S than the other two landscapes. Total P concentrations were in a similar range for the landscapes L´eo, Midebdo, and Liliyo. In Ti´eningbou´e, 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 L´eo.

Soil resin P concentrations varied between 0.8 mg P kg^{-1} soil to 33.1 mg P kg^{-1} soil. In Ti´eningbou´e 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 Ti´eningbou´e and the lowest in L´eo. 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 Ti´eningbou´e (= 6.4), Liliyo (= 6.5), and Midebdo (= 6.5). Median pH(H_2O) of yam fields in L´eo (= 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 Ti´eningbou´e 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 Ti´eningbou´e. The cec$_{\mathrm{eff}}$ ranged from 0.9 cmol(+) kg^{-1} soil to 14.6 cmol(+) kg^{-1} soil across all fields and landscapes. Median cec$_{\mathrm{eff}}$ tended to increase in the following order across landscapes: L´eo > Midebdo > Liliyo > Ti´eningbou´e. The interquartile range of cec$_{\mathrm{eff}}$ was also the highest in Ti´eningbou´e and the lowest in L´eo.

Soil mid-IR spectroscopic models

~~Spectroscopic pls~~ Partial least squares regressions ~~regression models~~ were developed and evaluated for all soil attributes together with the chemical reference data (S1 Table).

I guess you said this in the methods already
Microsoft Office User
07/03/2018 23:13

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} � 0.8 criterion and 11 when applying a threshold of rpd � 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$_{\mathrm{eff}}$). Figure S4 Fig depicts the model evaluation summary and mean cross-validated predictions (5 × 10-fold cross-validation) including resampling confidence intervals of best performing models with rpd � 2. The resampling prediction intervals were very narrow, showing all pls regression models were stable. The chosen repeated K -fold cross-validation procedure was therefore appropriate because low bias and low variance properties.

Total C was accurately predicted, with an rpd of 3.7 and a rmse of 1.6 g C kg^{-1} soil. The developed models were 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.

Exchangeable K (rpd = 1.1), resin P (rpd = 1.3) and bs$_{\mathrm{eff}}$ (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 pls components.

Among the mid-ir pls regression models for the measured soil attributes, 1c 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 expressed in vip (bottom panel), which is superimposed by the preprocessed spectra (Savitzky-Golay first derivative prior to scaling and centering) and the raw absorbance spectra (top panel; average of 3 spectra replicates per sample). Prominent peaks were selected by local maxima within a span of 10 spectral points (20 cm^{-1}). Some of these peaks were removed because they were identified as spectral noise. Where present peaks matched fundamental mid-ir absorptions described in the literature [38–40], peak wavenumbers and corresponding functional groups/bonds involved in vibrations were annotated in S5 Fig.

A ~~high~~ large proportion of ~~the spectral range exceeded~~ absorptions? Had vip > 1 ~~across all wavenumbers~~ for each the total C, total N and percent clay models (S5 Fig). Important wavenumbers (vip > 1) for total C were predominantly between 3140 cm^{-1} and 1230 cm^{-1}. Our vip results showed that not only prominent peaks from fundamental bands of in previous studies assigned vibrations of soil constituents played a important role for the prediction of the selected soil properties, but also regions in between prominent spectral features. 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 pls regression models (Figure S5). 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

Performance of spectroscopic models

Despite covering different soil types and climatic zones, mid-ir accurately predicts C ~~with high accuracy~~ (rmse = 1.6 g kg^{-1} soil) , in a range that are usually only reported for field-scale models. The chosen strategy to develop a mid-ir model library specific for the selected landscapes resulted in more accurate predictions than continental-scale global spectral models found in the literature [18, 41, 42]. This spatial scaling effect on model performance has been widely reported in the literature, with a trend that models covering a smaller geographical range and/or lower variation in som resulted in lower prediction errors [43]. However, despite mixing different soil types and sampling across a wide geographic scale, the strategy of developing global pls regression models was successful in our study.

Error decomposition analysis by measures proposed by [35] revealed that errors in pls regression models for the studied soil properties were predominantly attributed to lack of correlation (lc), therefore showing no systematic bias. On the other hand, no or little contribution of the other two error components, squared bias (sb) and non-unity slope, directly may arise from inherent properties of the pls regression algorithm. Particularly, decomposition of preprocessed spectra as predictors is simultaneously done by maximizing covariance between component scores and the response (soil property) as well as maximally summarizing the variation in the predictors at each of the decomposition iterations.

We conclude that signals of poorly predicted soil properties may be either not represented in the spectra, or may be masked by other effects such as light scattering. We found some minor improvements for resin P and Fe(dpta) models when applying log transformation prior to modeling.

OK so as I mention in the methods please say that you used
Microsoft Office User
07/03/2018 23:14

I think that there was too much reduncandy in your spectra
Microsoft Office User
07/04/2018 09:40

Because they are likely to be very correlated – check this and it should
Microsoft Office User
07/04/2018 09:42

MAIN COMMENTS:
-ok to discuss performance of
Microsoft Office User
07/04/2018 10:09

You do not really show the development of a spectral library so
Microsoft Office User
07/04/2018 09:44

Do not use global as this is confusing – use ‘large’ or
Microsoft Office User
07/04/2018 09:45

Pls explain – if there is no correlation then bias does not matter
Microsoft Office User
07/04/2018 09:48

I don’t understand how you assessed this from the models with all the
Microsoft Office User
07/04/2018 09:49

Is this discussion? Or methods?
Microsoft Office User
07/04/2018 09:50

This is not a conclusion section – also, the log transform was to
Microsoft Office User
07/04/2018 09:50

Interpretation of spectroscopic models

All mid-ir spectra measured for soils in the four landscapes exhibited a similar characteristic pattern of absorbance (S5 Fig). When comparing our spectra to different pure mineral sub-spaces matched by [18], the spectra measured in the soils across the landscapes mostly resemble the subspace of quartz dominated soils and were spectrally relative homogeneous. The quartz dominated spectral features might result from 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) [38]. 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.

Some of this could go in the first section of the results – where I proposed to show some representative spectra from the different soil types or landscapes…
Microsoft Office User
07/04/2018 09:53

Although extraction-based and surface chemistry determined soil nutrients generally have variable predictive performance because of complex relationships between soil composition and soil matrix exchange processes [12], we found reasonable prediction accuracies for Zn(dtpa) and Cu(dtpa) using mid-mir spectroscopy and pls regression models.

[44] remark that good prediction of cec by mid-mir spectroscopy depends on clay type, proportion of clay and organic matter contents, whose spectral features can be well detected in the mid-ir. By considering vip as measure of pls variable importance, relevant spectral features of clay and organic matter were also found in our study. In accordance to [44], our findings confirm that the prediction of clay content is affected by the spectral features that are related to soil organic C and also clay lattice primary mid-ir vibrations.

The contribution of wavenumber variables was almost identical for total C and total N pls regression models. This suggests that almost the same spectral features are used to predict total N and total C. The close relation between total C and N was also evident in the correlation of chemical reference values ($R^2 = 0.85$). The positive relationship between total C and N can be explained by a large proportion of N that is bound to som. In contrast, total C and % clay models exploited in some different spectral features across the mid-ir ranges. Further, total C had a relatively weak linear relationship with % clay ($R^2 = 0.51$), compared to that with total N. Model interpretation by vip revealed that the regions around well-defined kaolinite peaks were important for the prediction of % clay. Therefore, this confirms that mid-ir reflectance spectroscopy is able to sufficiently discriminate major soil physical constituents such as clay. More importantly, the models can be straight-forward to interpret when using vip, and they incorporate mechanistic relations between soil composition, related spectral patterns and properties.

This will be evident in the results if you add a correlation matrix there
Microsoft Office User
07/04/2018 09:55

The infrared spectroscopy literature has reported many band assignments of functional groups found in a broad range of constituents of som. Peaks associated with SOM are predominantly between 1725 cm^{-1} (C _ O of carboxylic acids) and 1050 cm^{-1} (C–O of carbohydrates) (for a compilation see e.g. [39]). Moreover, humic and fulvic acids have also many signals within the mid-ir range, with pronounced absorption peaks at 1720 cm^{-1} and 1620 cm^{-1} (fulvic acids), and 1270 cm^{-1} (humic acids) [45]. However, our vip analysis did not show these characteristic spectral bands below 1250 cm^{-1} that had high importance for the prediction of total C, which would be correlated to the sum of various pools of som. Furthermore, the total C model might rely on pyrogenic C spectral features such as the aromatic C _ C stretch found at 1430 cm^{-1} to 1380 cm^{-1} [46], as high vip values were found in this region. The detection of pyrogenic C spectral signal was likely attributed to small pieces of charred organic matter found in many of the fields, which resulted from the slash-and-burn practices in the past.

Is this relevant for your study?
Microsoft Office User
07/04/2018 09:56

Our model interpretation results suggest that not only spectral absorbance features of known compounds are relevant for the prediction of soil properties by mid-ir spectroscopy, but also regions between prominent peaks. One possible explanation of non-peak features lies in the general knowledge that signals of various organic and mineral compounds can overlap and mask each other. In tropical soils such as in the yam belt of West Africa, it is likely that mineral signals overlap organic because total C and hence som concentration was quite low, ranging between 2.4 g C kg^{-1} soil and 24.7 g C kg^{-1} soil. Furthermore, the developed pls regression models could also have picked up spectral features of various soil compounds that are correlated to the modeled respective soil property. For example, it is known that cec is affected by som, iron oxides and clays in tropical soils, each of the latter compound have unique spectral features. Lastly, spectral preprocessing methodologies can enhance features in spectra and emphasize different aspects of spectra. As an example, the applied first derivative Savitzky-Golay filter accentuates changes in spectra or peak shapes across wavenumbers. As a consequence, the relationship between spectra and the soil property to be modeled is modified. As far as we know, the effects of preprocessing in respect to comparisons of important features vs. known pure compound spectral peak assignments are rarely critically discussed in the literature.

Important direct and indirect relations between soil chemical composition and other properties were expected to be represented in the spectral models since signatures of major mineral and organic components had been detected in the mid-ir range in various studies. Therefore, combining an importance ranking of wavenumber predictors (preprocessed absorbance) with corresponding model contribution and a literature comparison of pure soil compound assignments is practical to infer major mechanisms enabling prediction of soil properties in the presented data set.

Application of spectroscopic models

The results of this study can enable us to derive recommendations for an implementation of mid-ir soil spectroscopy as a diagnostic tool in future studies in the selected landscapes of the yam belt. For example, the presented spectroscopy models can be applied in experimental and on-farm studies within the selected landscapes. First, one should take into account the geographical scale and soil types of the reference samples used for the construction of the model calibration library [12]. Secondly, the application range to quantify a specific soil property is limited to the range of analytical values used in the calibration. In some cases, predictions might be biased when soils in the study area are not in the range of the variation considered for calibration [43]. Many studies showed that extrapolation may not be possible with spectral models for soil property prediction [43]. Thus, when applying our 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, in cases with biased extrapolations.

Merging soils from different origins into soil propertey-specific ’global’ models instead of developing site-targeted calibrations will simplify model handling and prediction for future studies. For soils originating from within the area of the study landscapes sampled here or soils with similar ranges in soil properties and soil types mid-ir spectra can be measured, processed and plugged into the trained models presented here, whereby several soil properties can be estimated at once in a quantitative manner. Further, we would suggest to implement the ReSampling-Local (rs-local) data-driven method developed by [47], when the soil spectral library is extended with many new samples.

The developed final pls regression models had between 1 and 9 components, which is a relatively small number. A low model complexity and parsimonious models guarantee robust models that do not over-fit. For example, less complex models predict

I think this is one of the main points for discussion – obviously minerals and organics are the main consitutents with spectral response – everything else is due to some correlation to these. So I would reorganize the
Microsoft Office User
07/04/2018 09:57 ...

I think that this is a pilot study and potentially the start of work to build a spectral library for the afican yam belt. Take a look at work we did to develop a mid-IR spectral library for regions in Australia for
Microsoft Office User
07/04/2018 09:59 ...

We have little evidence of this – the issue is that we have only got cross validation assessments and a relatively small dataset, which may not have been sampled optimally…. So difficult to make such statements
Microsoft Office User
07/04/2018 10:00

As before – use ‘general’ or regional
Microsoft Office User
07/04/2018 10:02

more accurately when spectral noise, e.g. from scattering effects, occur in spectra of new samples to be predicted.

This study has laid the methodological foundation that will allow to evaluate soil effects of innovative nutrient management strategies for yam fields in West Africa. Specifically, we propose to implement a spectroscopy-driven approach to both test whether soil management innovations maintain or increase soil fertility in a high number of site environment trials and to quantify soil properties from new fields with unknown soil properties. Our approach is novel in terms of that we provide the complete raw data and metadata, processing and modeling framework and all finally derived models in a fully reproducible manner. Data, code, and model sharing will bring further benefits besides reproducibility and traceability. Data analysis and modeling was conducted in the open source statistical computing environment R, and a customized R package, ”simplerspec” has been released to simplify spectral data import, including data processing, modeling and prediction workflows. For example, the use of an integrative and standardized soil diagnostic method can promote collaboration across researchers in different institutions when embedded in a tidy soil spectroscopy modeling framework that brings consistency in the data management. Also, the burden in conducting different routine methods on many laboratory devices with cumbersome data cleaning and analysis workflows can be to a large extent removed and novel scientists can be trained in soil spectroscopy and the use of multivariate statistics. Finally, this allows a scientist to directly apply the calibrated models for new soils and hence employ spectroscopy as a productive tool.

Tone this down a bit – this is a pilot study that you hope to develop further. Also, I have seen little discussion on the nutrient requirements of yams and how the soil properties that you modelled help with that.
Microsoft Office User
07/04/2018 10:03

To me, novelty needs to be in the science, not the more technical aspects. So I would remove this if you want to submit to a scientific journal. You have already said that the data and code are available – that is enough
Microsoft Office User
07/04/2018 10:04

If you want to publish your package, I would leave this out of here. It is not needed. We can discuss this in the paper about the package! This is not relevant to this study, which is about modelling to predict fertility properties for yams
Microsoft Office User
07/04/2018 10:06

When predicting new soils based on the models presented in this study one should follow our protocol of soil spectroscopy sample preparation, measurement settings, data handling and processing steps. The importance of methodological consistency is well reported in the literature (e.g. [48]). We suggest to use a Bruker alpha device, as these models were based on spectra from this device and because it is known that different spectrometer types and diffuse reflection attachments create differing spectra on identical samples (e.g. scattering effects, path length differences; [49]). If a different spectrometer type is used, we encourage to perform a re-calibration because calibration transfer only pays off for huge spectral libraries when remeasuring all spectra from a huge soil sample library is more laborious. When applying the prediction models presented here on new locations within the study landscape or different locations in the yam belt, we strongly recommend to include a validation set covering about 10 to 15 % of new samples. For estimating soil properties in new campaigns we recommend to select validation samples using a a sampling strategy that ensures validation samples are evenly distributed across the component space, such as the Kennard-Stone algorithm. Applying appropriate calibration sampling algorithms on spectral data ensure that validation samples comprise the major variability in soil spectra [50].

Lastly, uncertainties derived at model evaluation should always be included when reporting new predictions based on our spectroscopy models. As our models were unbiased and model error was quite uniform across the range of observed values, rmse_{rcv} of the respective calibration models can be used for this purpose.

Conclusion

I can read this once you have revised
Microsoft Office User
07/04/2018 10:09

This study involved measuring many of the agronomically relevant soil chemical and physical properties across the four landscapes in the yam belt of West Africa. These are routinely quantified by wet chemistry and applied widely for agricultural diagnostics. As an alternative to the traditional approach, mid-ir spectroscopy models were developed and tested to cost-effectively and rapidly determine the distribution and variability of important soil properties. Based on the results of this study, many of these models can potentially be applied in future studies with sufficient accuracy. Model evaluation ultimately showed 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 that can be used to predict these soil fertility properties. We therefore conclude that this study laid the foundation of applying soil spectroscopy within the studied landscapes of the West African yam belt.

Why are all figures supporting info???
Microsoft Office User
07/04/2018 10:08

Microsoft Office User
07/04/2018 10:09

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. Evaluation of mid-IR PLS regression 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 cm^{-1}) for the selected wavenumbers. Fundamental mid-ir vibrations that are well described in the literature [38, 39, 42] 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 cm^{-1}); 3rd order polynomial fit). The bottom panel shows variable importance in the projection (vip) for three selected well performing pls regression 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 $y = 0$) pls regression β coefficients.

S1 Appendix. Spectra processing and modeling framework in R.

I also think this is unnecessary – we can include this in the paper for your package – but up to you
Microsoft Office User
07/04/2018 10:08

The spectral data processing pipeline was mainly built on top of the `tidyverse` set of packages [51] and the `data.table` package [52]. Graphical output was created based on the `ggplot2` package [53].

S2 Appendix. Some remarks on PLS regression.

Unnecessary!

Microsoft Office User
07/04/2018 10:07

In pls regression, there is a single tuning parameter called the number of components (ncomp) or latent variables. This model parameter cannot be directly estimated from the data because there is no analytical formula to calculate it. The choice of ncomp in the model determines the model complexity and how adaptive the final pls regression model for a soil property is to the training spectra. Along with many other modern predictive modeling techniques, pls regression potentially bears the risk to over-fit the structure in predictor data (e.g., spectra) in relation to the response (e.g., total organic C) to predict, particularly in cases with low number of observations and high number of predictions. Hence in over-fitting situations, due to each sample's unique noise learned besides general patterns during training, the model does not generalize well to new samples to predict. Model tuning aims to find model parameter values that yield best and realistic prediction accuracy. The below described model building approach embraces both model tuning and evaluation and thereby avoids over-fitting.

S1 Table. Descriptive summary of soil reference data and evaluation results of cross-validated pls regression 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 pls regression components (ncomp).

Acknowledgments

References

1. Food and Agriculture Organization of the United Nations. FAOSTAT statistics database; 2016. Available from: `www.fao.org/faostat/`.

2. 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.

3. 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.

4. 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.

5. Frossard E, Aighewi BA, Ak´e 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.

6. 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.

7. Lebot V. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. No. 17 in Crop production science in horticulture series. Wallingford, UK ; Cambridge, MA: CABI; 2009.

8. Hgaza VK, Diby LN, Oberson A, Tschannen A, Ti´e BT, Sangakkara UR, et al. Nitrogen Use by Yam as Affected by Mineral Fertilizer Application. Agronomy Journal. 2012;104(6):1558. doi:10.2134/agronj2011.0387.

9. Diby LN, Hgaza VK, Ti´e TB, Assa A, Carsky R, Girardin O, et al. How does soil fertility affect yam growth? Acta Agriculturae Scandinavica, Section B - Plant Soil Science. 2011;61(5):448–457. doi:10.1080/09064710.2010.505578.

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. Shepherd KD, Vagen TG, Gumbricht T, Walsh MG, Shepherd G, United Nations Environment Programme, et al. Land health surveillance: an evidence-based approach to land ecosystem management : illustrated with a case study in the West Africa Sahel; 2012.

12. Nocita M, Stevens A, van Wesemael B, Aitkenhead M, Bachmann M, Barth`es 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`.

13. C´ecillon L, Barth`es 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.

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. Rinnan `[U+FFFD]` Berg Fvd, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry. 2009;28(10):1201–1222. doi:10.1016/j.trac.2009.07.007.

16. Cambou A, Cardinael R, Kouakoua E, Villeneuve M, Durand C, Barth`es 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.

17. Clairotte M, Grinand C, Kouakoua E, Th´ebault 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.

18. 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.

19. Viscarra Rossel RA, Behrens T, Ben-Dor E, Brown DJ, Dematt^e 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. 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`.

21. Nuernberg NJ, Leal JE, Sumner ME. Evaluation of an anion[U+2010]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.

22. 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.

23. 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.

24. 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.

25. 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.

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

27. 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/.

28. Stevens A, Ramirez–Lopez L. An introduction to the prospectr package. 2014;.

29. Mevik BH, Wehrens R, Liland KH. pls: Partial Least Squares and Principal Component Regression; 2016. Available from: https://CRAN.R-project.org/package=pls.

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

31. Wing MKCfJ, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. caret: Classification and Regression Training; 2017. Available from: https://CRAN.R-project.org/package=caret.

32. 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.

33. 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.

34. 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.

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

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

37. 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.

38. Madejov´a J, Keˇck´es J, P´alkov´a H, Komadel P. Identification of components in smectite/kaolinite mixtures. Clay Minerals. 2002;37(2):377–388. doi:10.1180/0009855023720042.

39. Viscarra Rossel RA~~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.

40. Cornell RM, Schwertmann U. The iron oxides: structure, properties, reactions, occurrences, and uses. 2nd ed. Weinheim: Wiley-VCH; 2003.

41. Viscarra Rossel RA~~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.

42. Stevens A, Nocita M, T´oth 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.

43. Stenberg B, Viscarra Rossel RA~~Rossel RAV~~. Diffuse Reflectance Spectroscopy for High-Resolution Soil Sensing. In: Viscarra Rossel RA~~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`.

44. 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.

45. Baes AU, Bloom PR. Diffuse reflectance and transmission Fourier transform infrared (DRIFT) spectroscopy of humic and fulvic acids. Soil Science Society of America Journal. 1989;53(3):695–700.

46. Chatterjee S, Santos F, Abiven S, Itin B, Stark RE, Bird JA. Elucidating the chemical structure of pyrogenic organic matter by combining magnetic resonance, mid-infrared spectroscopy and mass spectrometry. Organic Geochemistry. 2012;51:35–44. doi:10.1016/j.orggeochem.2012.07.006.

47. Lobsey CR, Viscarra Rossel RA, Roudier P, Hedley CB. <span style=”font-variant:small-caps;”>rs-local</span> data-mines information from spectral libraries to improve local calibrations: <span style=”font-variant:small-caps;”>rs-local</span> improves local spectroscopic calibrations. European Journal of Soil Science. 2017;doi:10.1111/ejss.12490.

48. Wetterlind J, Stenberg B, Viscarra Rossel RA~~Rossel RAV~~. Soil Analysis Using Visible and Near Infrared Spectroscopy. In: Maathuis FJM, editor. Plant Mineral Nutrients. Totowa, NJ: Humana Press; 2013. p. 95–107. Available from: `http://link.springer.com/10.1007/978-1-62703-152-3_6`.

49. Pimstein A, Notesco G, Ben-Dor E. Performance of Three Identical Spectrometers in Retrieving Soil Reflectance under Laboratory Conditions. Soil Science Society of America Journal. 2011;75(2):746. doi:10.2136/sssaj2010.0174.

50. Ramirez-Lopez L, Schmidt K, Behrens T, van Wesemael B, Dematt^e JAM, Scholten T. Sampling optimal calibration sets in soil infrared spectroscopy. Geoderma. 2014;226–227:140–150. doi:10.1016/j.geoderma.2014.02.002.

51. Wickham H. tidyverse: Easily Install and Load 'Tidyverse' Packages; 2017. Available from: https://CRAN.R-project.org/package=tidyverse.

52. Dowle M, Srinivasan A. data.table: Extension of ‘data.frame‘; 2017. Available from: https://CRAN.R-project.org/package=data.table.

53. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2009. Available from: http://ggplot2.org.

Soil attribute	n	Soil reference analyses					mid-ir pls regression (5 × rep. 10-fold cv)			
		Min$_{ob}$	Max$_{ob}$	Med$_{ob}$	Mean$_{ob}$	cv$_{ob}$	ncom	rmse$_r$	R^2_{rcv}	rpd$_{rcv}$
Total Fe [g kg^{-1}]	94	4	35	10	12	54	5	3	0.80	2.3
Total Si [g kg^{-1}]	94	200	363	262	262	12	3	19	0.65	1.7
Total Al [g kg^{-1}]	94	10	102	48	53	42	6	4	0.97	6.2
Total K [g kg^{-1}]	94	1	34	6	10	91	7	1	0.98	6.4
Total Ca [g kg^{-1}]	94	0.3	7.6	1.4	1.9	70	5	0.6	0.79	2.2
Total Zn [mg kg^{-1}]	94	10	72	19	23	49	6	6	0.66	1.7
Total Cu [mg kg^{-1}]	94	0	29	5	7	87	8	3	0.72	1.9
Total Mn [mg kg^{-1}]	94	59	1146	222	308	74	4	125	0.70	1.8
Sand [%]	80	29.8	91.6	75.6	74.2	14	2	7.9	0.45	1.3
Silt [%]	80	3.9	54.1	12.0	14.1	60	2	6.4	0.43	1.3
Clay [%]	80	4.5	26.1	10.1	11.6	42	2	2.2	0.80	2.2
pH$_{H_20}$	80	4.7	8.4	6.4	6.4	11	8	0.4	0.68	1.8
K (exch.) [mg kg^{-1}]	94	0	868	104	145	95	1	121	0.24	1.1
Ca (exch.) [mg kg^{-1}]	92	98	2170	604	774	70	6	228	0.82	2.4
Mg (exch.) [mg kg^{-1}]	93	18	432	76	113	84	2	61	0.58	1.5
Al (exch.) [mg kg^{-1}]	94	0	47	0	4	258	2	9	0.17	1.1
cec$_{eff}$ [cmol(+) kg^{-1}]	91	0.9	14.6	4.2	5.3	67	6	1.4	0.85	2.6
BS$_{eff}$ [%]	91	79	100	100	98	4	1	4	0.05	1.0
Total C [g kg^{-1}]	94	2.4	24.7	8.5	9.9	58	6	1.6	0.93	3.7
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	143	0.55	1.5
log(P resin) [mg kg^{-1}]	92	-0.2	3.5	1.4	1.4	57	2	0.6	0.40	1.3
log(Fe(DTPA)) [mg kg^{-1}]	92	1.0	6.7	2.7	2.9	38	9	0.5	0.83	2.4
Zn (DTPA) [mg kg^{-1}]	87	0.2	11.5	1.9	2.8	89	3	2.1	0.27	1.2
Cu (DTPA) [mg kg^{-1}]	92	0.1	1.5	0.2	0.4	89	6	0.2	0.74	1.9
Mn (DTPA) [mg kg^{-1}]	92	2.5	31.4	6.5	8.6	69	3	4.0	0.55	1.5