International Journal of Bioresource Science
Citation: IJBS: 5(1): 01-20, June 2018
DOI: 10.30954/2347-9655.01.2018.1

©2018 New Delhi Publishers. All rights reserved

Advanced Modeling of Soil Biological Properties Using Visible Near Infrared Diffuse Reflectance Spectroscopy

David C. Weindorf[1*], Somsubhra Chakraborty[2], Jennifer Moore-Kucera[1], Bin Li[3], Lisa Fultz[4], Veronica Acosta-Martinez[5] and Chenhui Li[1]

[1]Texas Tech University, Lubbock, TX, USA
[2]Indian Institute of Technology Kharagpur, Kharagpur, India
[3]Louisiana State University, Baton Rouge, LA, USA
[4]Louisiana State University Agricultural Center, Baton Rouge, LA, USA
[5]Agricultural Research Service, Lubbock, TX, USA

*Corresponding author: david.weindorf@ttu.edu

ABSTRACT

Although visible near infrared diffuse reflectance spectroscopy (VisNIR DRS) is an emerging, rapid, non-destructive, and cost effective technology to predict a host of soil biological parameters, the traditional chemometric partial least squares regression (PLS) model often poses challenges during sensor development. In an effort to identify alternatives to the PLS model, three multivariate machine learning algorithms [PLS, penalized spline regression (PSR), and random forest (RF) regression]in conjunction with two spectral preprocessing methods [Savitzky–Golay first derivative and absorbance (ABS)] were compared with respect to 12 soil biological parameters of 123 soil samples. The RF model with ABS spectra successfully predicted all biological parameters with residual prediction deviation (RPD) ranging from 2.60 to 3.60 and outperformed PSR and PLS models. The best PSR model was obtained for total bacteria with an RPD of 2.70 and an r^2 of 0.86 and among other variables, only Gram positive bacteria (RPD=2.63, r^2=0.85), Gram negative bacteria (RPD=2.58, r^2=0.85), and SOM (RPD=2.67, r^2=0.86) were satisfactorily predicted, exhibiting r^2>0.80 and RPD>2.5. Conversely, all variables except SOM (RPD=2.07) were poorly predicted by PLS models which had an RPD<2. Furthermore, linear discriminant analysis qualitatively clustered soils with different levels of microbial parameters. Summarily, the RF model with ABS spectra showed great promise in characterizing soil microbial communities with potential for such analysis in-situ.

Keywords: Microbial community, Linear discriminant analysis, Partial least squares, Penalized spline, Random forest, Visible near infrared diffuse reflectance spectroscopy

Soil biological processes have many profound implications on soil health. Soil is a complex and dynamic ecosystem that is home to abundant and diverse microbial communities, with billions of microorganisms inhabiting one gram of soil (Coleman and Whitman, 2005; Curtis and Sloan, 2005). In soil ecosystems, microorganisms play an important role in various functions such as: nutrient cycling, structural formation, regulation of soil organic matter (SOM) dynamics, C sequestration, and enhancement of plant growth (Millard and Singh, 2010). Given the complexity of these populations, a number of standard laboratory procedures are used to identify and quantify them. One approach, fatty acid profiling, involves extraction of signature lipids, present within the microbial cells, which are used to identify different taxonomical groups (Zelles *et al.* 1995). Two of the most common approaches to profile soil microbial fatty acids are the phospholipid fatty acid (PLFA) method and the ester-linked fatty acid methyl ester (EL-FAME) method (White *et al.* 1996; Zelles, 1999). Each method has advantages and disadvantages as have been reviewed and evaluated by Zelles (1999), Schutter and Dick (2000), Drenovsky *et al.* (2004), and Fernandes *et al.* (2013).

Visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) is a reliable, rapid, field-portable, non-destructive, and cost effective technique used for characterization of several soil properties (Ben-Dor and Banin, 1995; Reeves *et al.* 2000; Islam *et al.* 2003; Brown *et al.* 2006; Viscarra Rossel *et al.* 2006; Morgan *et al.* 2009; Vasques *et al.* 2009). Recently, soil microbiological parameters like microbial biomass carbon (MBC), basal respiration, mineralizable C and N have been rapidly estimated via infrared spectroscopy and frequently used as indicators of SOM quality (Pietkainen and Fritze, 1995; Chodak *et al.* 2002; Ludwig *et al.* 2002; Couteaux *et al.* 2003; Rinnan and Rinnan, 2007; Cecillon *et al.* 2008; Zornoza *et al.* 2008). When used in combination with reference estimation methods and multivariate algorithms, VisNIR DRS becomesa robust method for quantifying many soil properties (Viscarra Rossel *et al.* 2006). While overtones of OH^-, SO_4^{2-}, and CO_3^{2-} groups and combination bands of H_2O and CO_2 are responsible for unique spectral signatures of common clay minerals; O–H, C–N, N–H, and C=O groups are active bonds for soil organic matter in the NIR region (Hunt and Salisbury, 1970; Malley *et al.* 2002).

The operational theory of Vis NIR DRS utilizes a halogen light source to project light onto a soil sample. Simultaneously, some of that emitted light is reflected back and captured by the spectrometer probe, where the reflectance values are precisely recorded for analyses as a function of wavelength, especially in the visible and near infrared range (350-2500 nm). Different soil attributes such as moisture (Zhu *et al.* 2010), carbon, or microbial populations cause a spectral absorbance at different, unique wavelengths (Chang *et al.* 2001). The collected spectra are then subjected to a number of preprocessing methods and statistical calibration-validation approaches to associate parsed spectra with the soil parameter of interest. Spectral preprocessing typically involves transformation of raw reflectance spectra such as averaging of replicate scans followed by producing 1st and 2nd derivatives of reflectance spectra on desired intervals or discreet wavelet transformation (DWT). In the calibration stage, a subset of samples are used to develop regression models; while in the validation stage the remaining samples are used to evaluate the validity of regression models for estimating constituent concentrations. Calibration models traditionally involve partial least squares (PLS) regression, principal component regression, or stepwise multiple linear regression. Recently, researchers have used clustering techniques such as linear discriminant analysis (LDA), random forest (RF) classification, and support vector machines for qualitative discrimination of VisNIR spectra of diverse soil samples (Chakraborty *et al.* 2010; Chakraborty *et al.* 2012).

Of interest to this study is the calibration stage. Although principal component regression and PLS efficiently decrease the dimensionality of spectroscopic data and the transformed new variables (i.e., principal component and PLS latent factor) are de-correlated, these variables are hard to relate to the original spectral absorptions, and thus these techniques may not be appropriate for sensor development (Ge *et al.* 2007). More robust, state-of-the-art statistical techniques like random forest regression (RF)and penalized spline regression (PSR) remain essentially untested with regard to soil microbial populations and other biological properties. RF represents a highly advanced data mining technology which combines information from a number of decision trees (Breiman, 2001).

If advanced statistical approaches can render alternative predictive models of soil biological parameters, they can be applied to prescreen large sample sets to identify those samples deserving of additional laboratory analysis or to make rapid, on-site evaluations of microbial populations as a primer of optimized soil health. The objectives of this study were to i)test if two machine learning algorithms (RF and PSR) can improve PLS predictive models of broad taxonomic microbial groups based on FAME profiles and MBC measurements and ii) test if LDA can be used for qualitative VisNIR discrimination of the soils with different levels of microbial parameters.

MATERIALS AND METHODS

Soil Sampling

Soil samples for this study were collected as part of an on-going study to examine soil microbial dynamics in relation to biogeochemical cycling and C sequestration in fields under Conservation Reserve Program (CRP) contracts and dryland

annual cropping systems (typical crop is *Gossypium hirsutum* L., cotton). Fields were located across seven counties in the semi-arid Southern High Plains of Texas (Fig. 1), with average annual temperature of 16.0°C and annual precipitation of 480 mm. A total of 19 CRP grasslands, seven long-term dryland cropping systems, three dryland cropping systems that had previously been under CRP contract, and three native rangelands were sampled in July 2012, June 2013, or November 2013 based upon prior study objectives. Soil samples were of the Amarillo series (Fine-loamy, mixed, superactive, thermic Aridic Paleustalfs) or Patricia series (Fine-loamy, mixed, superactive, thermic Aridic Paleustalfs) (Soil Survey Staff, 2012). Amarillo is a benchmark soil of this region commonly characterized by fine sandy loam in the upper 30 cm.

A total of 123 soil samples were utilized in this study. In July 2012, 46 samples were collected from 16CRPfields (32 samples) and seven croplands (14 samples). Half of the samples from each management were collected from 0-10cm and the other half from 10-30cm depth. In June 2013, six samples (three from 0-10cm and three from 10-30cm) were collected from native rangelands. In November 2013, 36 samples were collected from three long-term CRP fields and 35 were collected from recently converted CRP fields. At each long-term or converted CRP field samples were collected using a hydraulic probe (Giddings Mfg., Colorado, USA) to a depth of 100 cm. Samples were separated into depths of 0-10, 10-30, 30-50, and 50-100 cm and thoroughly homogenized. A total of three homogenized samples per depth per field were collected. All samples were stored on ice in the field and transported the same day to the Texas Tech University Soil and Environmental Microbiology laboratory in sealed plastic bags. Soils were disaggregated to pass through a 4.75 mm sieve and field-moist samples were further analyzed for microbial communities (via FAME profiling) and MBC and microbial biomass N (MBN) within three days of sample collection.

Biological analysis

In this study, AR grade (Sigma) chemicals were used without further purification. All solutions were prepared with MilliQ™ (18.2 MΩ) water. Soil microbial communities were characterized using direct extraction of microbial fatty acids according to the ester-linked fatty acid methyl ester (EL-FAMEs) method described by Schutter and Dick (2000) using 3g of field-moist equivalent soil. Briefly, lipids from the microbial cells were extracted by saponification at 37°C under alkaline conditions. The fatty acids were then methylated to form FAMEs which were further separated in an organic solvent. Upon addition of Methyl tert-butyl ether and Hexane (1:1) containing Methylnonadecanoate (19:0) as an internal standard, samples were quantified using a 6890 GC series II (Hewlett Packard, Wilmington, DE). The temperature program in the GC was ramped from 170°C to 250°C at 50°C min^{-1}, then to 300°C for 2 min to clear the column between samples(Acosta Martinez *et al.* 2004). The fatty acids were identified by comparison of retention times and peak areas with automated MIDI peak identification software (Microbial ID, Inc., Newark, DE). Total and individual FAME concentrations (nmolFAME-C g^{-1} soil) were calculated by comparing peak areas to an analytical standard (19: 0, Sigma Chemical Co., St. Louis, MO) calibration curve. The microbial communities were characterized into: Gram positive bacteria (GM+), Gram negative bacteria (GM-), actinomycetes, arbuscularmycorrhizal fungi (AMF), and saprophytic fungi using indicator fatty acids (Table 1). FAMEs produced were described using standard nomenclature: total number of carbon atoms, total number of double bonds followed by colon and position of double bonds from the methyl end of the group. The total bacterial biomass was calculated by adding GM+, GM-, and actinomycetes biomarkers while fungal biomass was calculated by adding saprophytic fungi biomarkers. The fungi to bacteria ratio was calculated by relating the 18:2ω6,9c biomarker with total bacteria.

Soil MBC and MBN were quantified by the chloroform fumigation extraction technique (Vance *et al.* 1987) using 15g oven dry equivalent field moist soil. Fumigated (24 h) and non-fumigated (control) samples were extracted with 0.5 *M* K_2SO_4 and filtered through Whatman No. 42 filter paper. The total organic C and N in the filtrate were analyzed with a TOC/TN analyzer (Shimadzu Model TOCV/CPH-TN, Japan). The MBCand MBN were calculated by subtracting non-fumigated sample values from fumigated samples considering k_{EC}=0.45 for C (Wu *et al.* 1990) and k_{EN}=0.54 for N (Jenkinson, 1988) as

constants. SOM was estimated via loss-on-ignition method (Nelsonand Sommers, 1996). Samples were dried overnight (~15 hours) at 105°C before being placed in the muffle furnace at 400°C for 8 hours.

VisNIR scanning and spectral pretreatments

Field-moist samples were scanned using a PSR-3500® portable VisNIR spectroradiometer (Spectral Evolution, Lawrence, MA, USA) with a spectral range of 350 to 2500 nm. The spectroradiometer had a 2nm sampling interval and a spectral resolution of 3.5, 10, and 7nm from 350 to 1000 nm, 1500 nm, and 2100 nm, respectively. Scanning was facilitated with a contact probe featuring a 5W built-in light source. Samples were brought to room temperature, evenly distributed in an opaque polypropylene sample holder and scanned from the top with the contact probe connected to the PSR-3500® with a metal-clad fiber optic cable. Full contact with the sample was ensured to avoid outside interference. Quadruplet scans were taken per sample with a 90° rotation between scans to obtain an average spectral curve. Each individual scan was an average of 10 internal scans over a time of 1.5 seconds. The detector was white referenced (after each sample) using a 12.7cm × 12.7cm NIST traceable radiance calibration panel, ensuring that fluctuating down-welling irradiance would not saturate the detector.

Raw reflectance spectra were processed via a statistical analysis software package, R version 2.11.0 (R Development Core Team, 2008) using custom "R" routines (Chakraborty *et al.* 2013). These routines involved (i) a parabolic splice to correct for "gaps" between detectors, (ii) averaging replicate spectra, and (iii) fitting a weighted (inverse measurement variance) smoothing spline to each spectra with direct extraction of smoothed reflectance at 10nm intervals.

This study used two spectral pretreatments to prepare the smoothed soil spectra for analysis, and three multivariate algorithms to develop the VisNIR predictive models. Spectral pretreatments reduce the influence of the side information contained in the spectra. The pretreatment transformations applied were Savitzky–Golay (SG) first derivative using a first-order polynomial across a ten band window, and optical density or absorbance (ABS) [log (1/reflectance)]. Both pretreatment transformations were implemented in the Unscrambler®X 10.3 software (CAMO Software Inc., Woodbridge, NJ). Subsequently, both SG and ABS spectra were included as candidate explanatory variables for biological parameters in following VisNIR models.

Machine learning

For each spectral pretreatment, three multivariate methods were tested including PLS, penalized spline regression (PSR), and random forest regression (RF) (Halaand and Thomas, 1988; Breiman, 2001). Samples with missing values were removed *a priori* before modeling. The whole dataset was used with leave-one-out-cross validation (LOOCV) to prevent over-fitting and evaluation of model generalizing capability. When using small data sets (40–120 samples) in quantitative multivariate modeling, LOOCV provides the best estimate of the predictive performance of an obtained model (Martens and Dardenne, 1998). For PSR, the cubic B-spline was used via R version 2.14.1 (R Development Core Team, 2008) as the base function with100 equally spaced knots. The order of the penalty was set to the default value of three. The optimal value for the penalty-tuning parameter was selected by minimizing the LOOCV error on the training set. Moreover, the 'Random Forest' package was used in R to build the RF model. The number of trees in RF was set to the default value of 500. The coefficient of determination (r^2), cross-validation root mean squared error (RMSEcv), residual prediction deviation(RPD), and bias were used as rubrics for judging model predictability. Subsequently, both aforementioned models were compared with PLS to test whether PSR or RF can improve DRS predictability. The optimum number of PLS latent factors (rotations of principal components for a slightly different optimization criterion) was selected on the basis of the number of factors with the smallest total residual validation Y-variance or highest total explained validation Y-variance (CAMO Software Inc., Woodbridge, NJ). Since RPD is the ratio of standard deviation (SD) and RMSE, model generalization capacity increases when validation set SD highly surpasses the RMSE.

Finally, the Fisher's LDA approach was applied with best performing spectral transformation for dimensionality reduction and qualitative VisNIR discrimination of the soils with different levels of microbial parameters. Each biological property

was transformed *a priori* into discrete classes [1st quartile (Q_1), 2nd quartile (Q_2), 3rd Quartile (Q_3), and 4th quartile (Q_4)] for classification purposes. Furthermore, to evaluate classification results, kappa (κ) coefficients were computed (Thompson and Walter, 1988).

RESULTS AND DISCUSSION

The summary statistics of all measured soil biological properties are provided in Table 2. Soil samples varied widely in their biological properties, likely a result of differences in land use, vegetation cover, and sampling depth. Compared to other properties, SOM (6-fold), saprophytic fungi (9-fold), and fungi: bacteria (18-fold) exhibited lower ranges of variation (Table 2). The ranges of variation of other properties were markedly larger [58 (MBC) to 408-fold (AMF)]. Not shown, the MBC: SOC ratio varied from 0.001 to 0.067 with a mean of 0.016. Except for saprophytic fungi, fungi: bacteria, and MBC, all other variables were significantly ($p<0.05$) correlated with each other (Table 3). The ratio of fungi: bacteria varied greatly (0.22-3.99) and was not correlated with SOM.

Machine learning

Since, ABS spectra exhibited higher RPD than SG spectra for most of the response variables, and because in the ABS spectra the intensities are linearly linked to the concentrations of interest (Bellon-Maurel and McBrateney, 2011), the modeling results reported here all usethe ABS spectra between 350 and 2500nm (Table 4). Among the three multivariate algorithms tested (PSR, PLS, and RF), all biological parameters were estimated with greatest accuracy by RF. Lab-measured versus RF predicted models for all parameters showed close out-of-bag prediction r^2 (similar to LOOCV prediction on each of the training points), ranging from 0.85 to 0.92. In general, PSR models for all response variables showed underestimation at higher values and overestimation at lower values (Fig. 2). Exhibiting a similar trend and deviation from the 1:1 line, the prediction decreased further for PLS models (Fig. 3). Conversely, RF models closely approximated the 1:1 line and improved prediction accuracy (Fig. 4).

Notably, Chang *et al.* (2001) categorized the accuracy and stability of their spectroscopy models based on the RPD values of the validation set. An RPD >2.0 was considered a stable and accurate predictive model; an RPD value between 1.4 and 2.0 indicated a fair model that could be improved by more accurate predictive techniques; an RPD value <1.4 indicated poor predictive capacity. In this study, RF models successfully predicted all biological parameters with RPDs ranging from 2.60 (saprophytic fungi) to 3.60 (SOM), exceeding the respective RPDs produced by both PSR and PLS counterparts (Table 4). Applying an even more stringent model evaluation rubric of Sayes *et al.* (2005), excellent predictions were obtained for SOM (RPD=3.60, r^2=0.92), GM+ (RPD=3.27, r^2=0.91), total FAMEs (RPD=3.21, r^2=0.90), total bacteria (RPD=3.22, r^2=0.90), actinomycetes (RPD=3.18, r^2=0.90), and GM-(RPD=3.10, r^2=0.90). Moreover, RF models for total fungi and fungi: bacteria exhibited very high prediction scores with RPD values of 3.03 and 3.06, respectively, and an r^2 of 0.89, which was slightly less than the prescribed cutoff r^2 of 0.9 (Sayes *et al.* 2005; Zornoza *et al.* 2008). RF models for AMF, saprophytic fungi, MBC, and MBN were accurate with r^2>0.80 and RPD>2.5. In contrast, PSR models produced intermediate generalization capabilities. The best PSR model was obtained for total bacteria with an RPD of 2.70 and an r^2 of 0.86 and among other variables, only GM+(RPD=2.63, r^2=0.85), GM-(RPD=2.58, r^2=0.85), and SOM (RPD=2.67, r^2=0.86)

Table 1: Indicator fatty acids used to evaluate microbial groups

Microbial group	Indicator fatty acids
Saprophytic fungi	18:3ω6c (6,9,12), 18:1ω9c, 18:2ω6,9c
AMF[a]	16:1ω5c
GM+[b]	14:0 iso, 15:0 iso, 15:0 anteiso, 16:0 iso, 17:0 iso, 17:0 anteiso
GM-[c]	17:0 cyclo, 19:0 cyclo ω8c, 18:1ω7c / 18:1ω6c
Actinomycetes	18:0 10-methyl, 17:0 10-methyl, 16:0 10-methyl / 17:1 iso ω9c

[a]*AMF, arbuscular mycorrhizal fungi;* [b]*GM+, Gram positive bacteria;* [c]*GM-, Gram negative bacteria.*

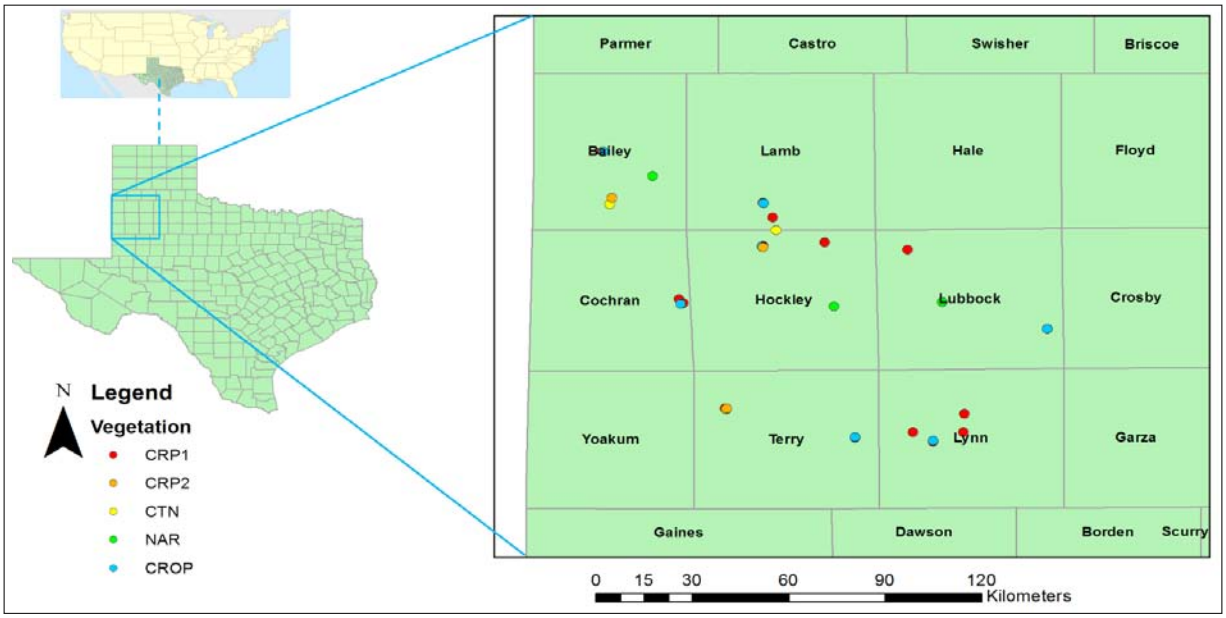

Fig. 1: Soil sampling location in Texas, USA. A total of 123 soil samples were collected and used in regression analysis. Field types are Conservation Reserve Program (CRP1, CRP2), cotton (CTN), natural rangeland (NAR), and other croplands (CROP)

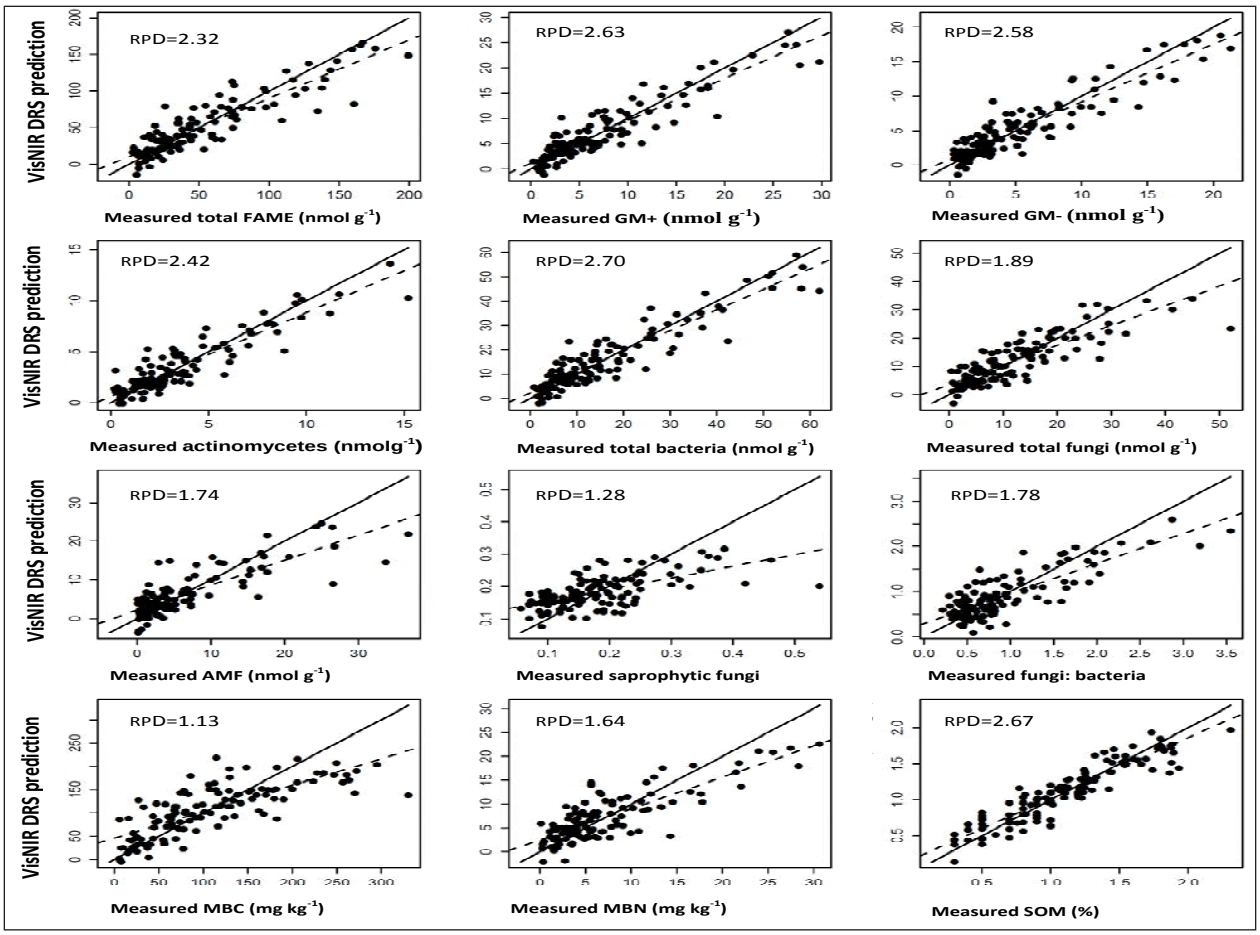

Fig. 2: Lab-measured vs. visible near infrared diffuse reflectance spectroscopy (VisNIR-DRS) penalized spline regression (PSR) predicted soil biological parameters using absorbance spectra. The dashed line is the regression line, and the solid line is a 1:1 line. All y-axes represent predicted values. Residual prediction deviation (RPD) is the ratio of standard deviation and root mean squared error. Plot abbreviations are as follows: soil organic matter (SOM), Gram positive bacteria (GM+), Gram negative bacteria (GM-), fatty acid methyl ester (FAME), arbuscularmycorrhizal fungi (AMF), microbial biomass carbon (MBC), microbial biomass nitrogen (MBN)

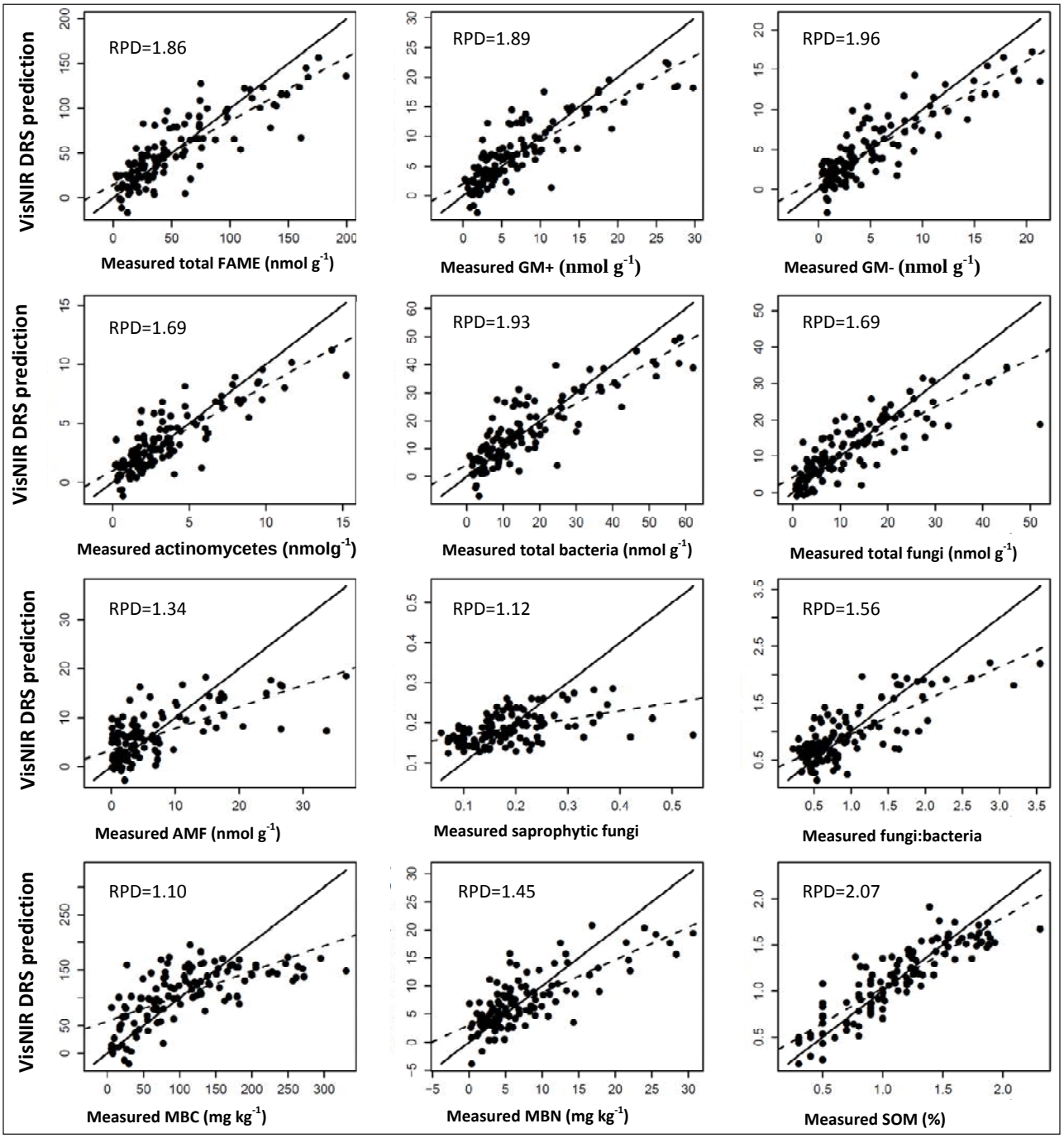

Fig. 3: Lab-measured vs. visible near infrared diffuse reflectance spectroscopy (VisNIR-DRS) partial least squares regression (PLS) predicted soil biological parameters using absorbancespectra. The dashed line is the regression line, and the solid line is a 1:1 line. All y-axes represent predicted values. Residual prediction deviation (RPD) is the ratio of standard deviation and root mean squared error. Plot abbreviations are as follows: soil organic matter (SOM), Gram positive bacteria (GM+), Gram negative bacteria (GM-), fatty acid methyl ester (FAME), arbuscularmycorrhizal fungi (AMF), microbial biomass carbon (MBC), microbial biomass nitrogen (MBN)

were satisfactorily predicted, exhibiting r^2>0.80 and RPD>2.5. Conversely, all variables except SOM (RPD=2.07) were poorly predicted by PLS models which had an RPD<2. Bias made a negligible contribution to the overall lack of cross-validation fit (<10% of MSE) for all models tested. The number of PLS latent factors (components) used to explain the prediction models were identical (10) for total FAME, GM+, GM-, actinomycetes, total bacteria, total fungi, and SOM. The ratio of RF: PLS and RF:PSR model results revealed some interesting trends, where the RF-MBC model substantially raised r^2 (417%) relative to the PLS-MBC model. Furthermore, RF-AMF produced roughly doubled

Table 2: Summary statistics of soil (n=123) parameters used in the study

Property	Minimum	Maximum	1st Quartile	Median	3rd Quartile	Mean	Variance (n-1)	Standard deviation (n-1)
Total FAME (nmol g^{-1})	2.55	199.25	19.76	35.53	73.42	51.83	1990.53	44.62
GM+ (nmol g^{-1})[a]	0.25	29.76	2.36	4.12	7.18	5.87	27.22	5.22
GM-(nmol g^{-1}) [b]	0.32	21.30	1.57	3.18	6.59	5.11	23.72	4.87
Actinomycetes (nmolg^{-1})	0.18	15.22	1.51	2.56	4.33	3.47	9.08	3.01
Total bacteria (nmol g^{-1})	0.82	62.02	5.45	10.04	18.52	14.35	163.98	12.81
Total fungi (nmol g^{-1})	0.43	52.06	4.09	9.21	16.42	11.85	98.25	9.91
AMF (nmol g^{-1}) [c]	0.09	36.72	1.23	3.42	7.16	6.03	54.29	7.37
Saprophytic fungi	0.06	0.54	0.13	0.19	0.25	0.20	0.01	0.09
Fungi: Bacteria	0.22	3.99	0.49	0.76	1.11	0.99	0.51	0.71
MBC (mg kg^{-1}) [d]	5.67	330.26	49.32	91.65	151.10	106.33	5405.59	73.52
MBN (mg kg^{-1}) [e]	0.09	30.62	3.24	5.39	9.08	7.33	42.17	6.49
Soil organic matter (%)	0.38	2.31	1.00	1.25	1.52	1.25	0.14	0.38

[a]GM+, Gram positive bacteria; [b]GM-, Gram negative bacteria; [c]AMF, Arbuscular mycorrhizal fungi; [d]MBC, Microbial biomass C; [e]MBN, Microbial biomass N.

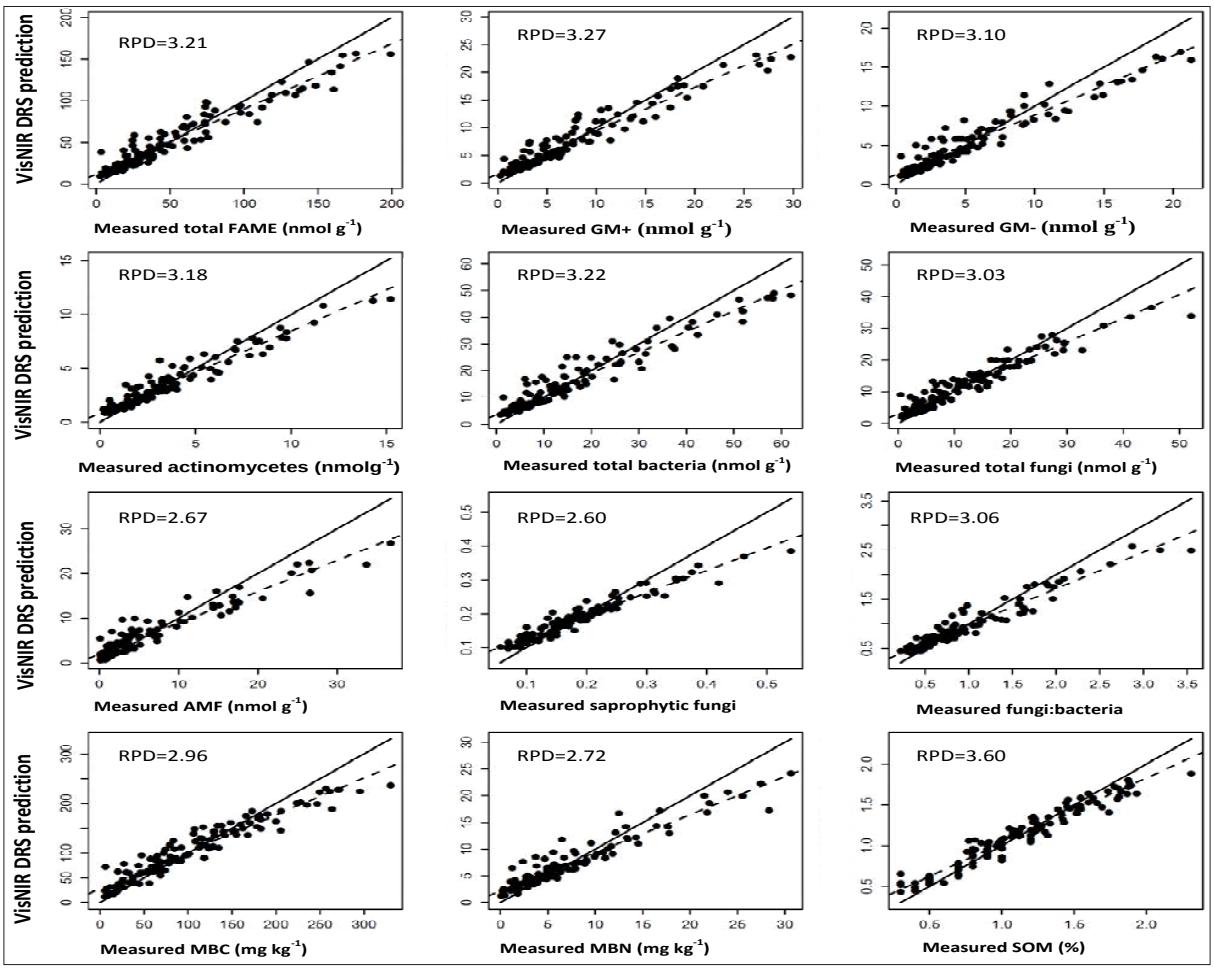

Fig. 4: Lab-measured vs. visible near infrared diffuse reflectance spectroscopy (VisNIR-DRS) random forest regression (RF) predicted soil biological parameters using absorbance spectra. The dashed line is the regression line, and the solid line is a 1:1 line. All y-axes represent predicted values. Residual prediction deviation (RPD) is the ratio of standard deviation and root mean squared error. Plot abbreviations are as follows: soil organic matter (SOM), Gram positive bacteria (GM+), Gram negative bacteria (GM-), fatty acid methyl ester (FAME), arbuscularmycorrhizal fungi (AMF), microbial biomass carbon (MBC), microbial biomass nitrogen (MBN)

Table 3: Correlation matrix (p-values) of measured soil biological properties. The last column exhibits r^2 values for simple linear regressions (SLR) of soil properties with soil organic matter

Variables	Total FAME	GM+	GM-	Actinomycetes	Total bacteria	Total fungi	AMF	Saprophytic fungi	Fungi: Bacteria	MBC	MBN	Soil organic matter	r^2 of SLR when Y= Soil organic matter
Total FAME (nmol g^{-1})	0												0.41
GM+ (nmol g^{-1})[a]	< 0.0001	0											0.42
GM-(nmol g^{-1})[b]	< 0.0001	< 0.0001	0										0.34
Actinomycetes (nmolg^{-1})	< 0.0001	< 0.0001	< 0.0001	0									0.37
Total bacteria (nmol g^{-1})	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0								0.40
Total fungi (nmol g^{-1})	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0							0.40
AMF (nmol g^{-1})[c]	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0						0.21
Saprophytic fungi	0.314	0.123	0.896	0.257	0.400	0.001	0.376	0					0.00
Fungi: Bacteria	0.870	0.030	0.076	0.023	0.036	0.001	0.379	< 0.0001	0				0.00
MBC (mg kg^{-1})[d]	0.000	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.118	0.002	0.045	< 0.0001	0			0.00
MBN (mg kg^{-1})[e]	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.253	0.053	< 0.0001	0		0.10
Soil organic matter (%)	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.600	0.742	0.484	0.001	0	-

[a]GM+, Gram positive bacteria; [b]GM-, Gram negative bacteria; [c]AMF, Arbuscular mycorrhizal fungi; [d]MBC, Microbial biomass C; [e]MBN, Microbial biomass N

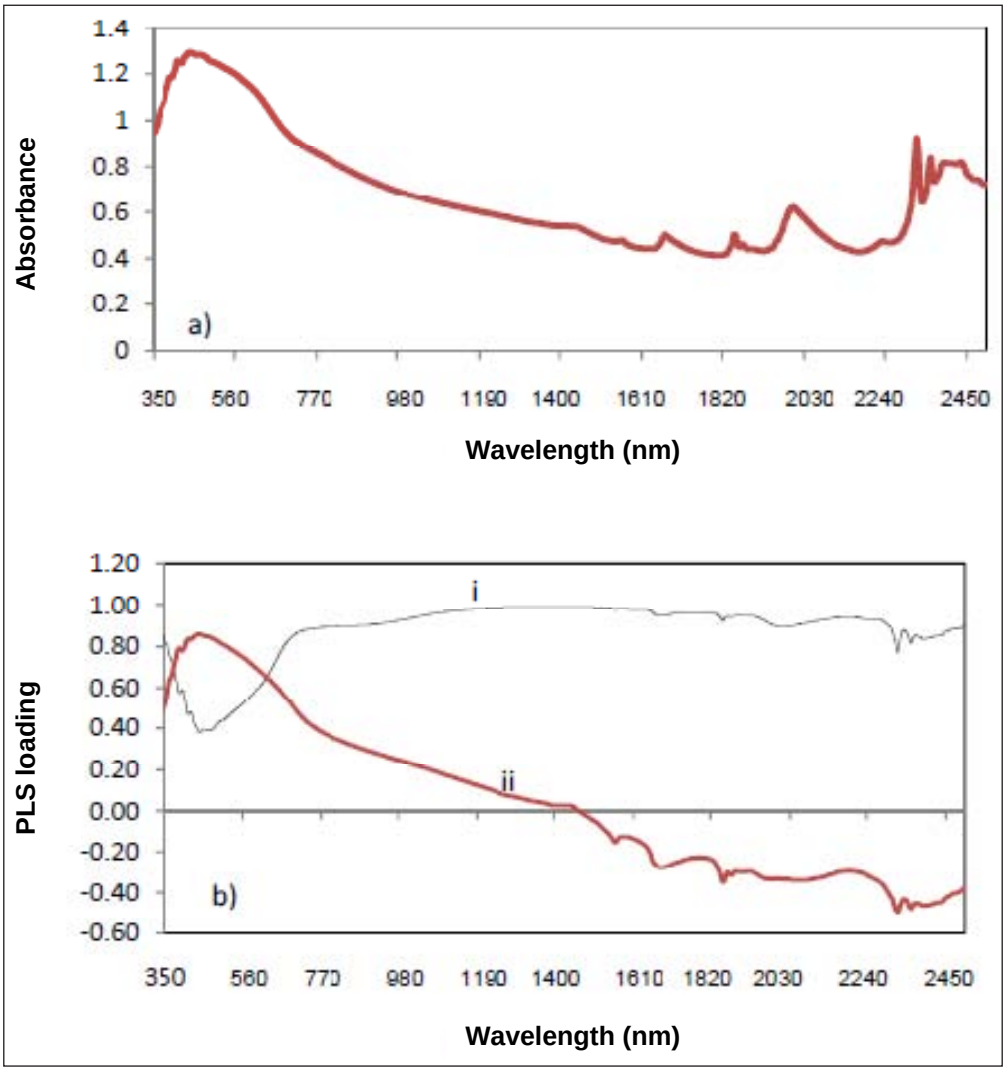

Fig. 5: Plots showing a) the average absorbance spectrum across a 10-band window of one randomly selected sample, and b) the first twopartial least squares (PLS) factor loading weight vectors (i and ii) centered on zero

the coefficient of determination (95% increment) while halving the RMSEcv (2.75nmol g^{-1}) relative to the PLS-AMF model (5.48 nmol g^{-1}). Among other variables, the RF model for saprophytic fungi produced 325% and 123% greater r^2 values and 132% and 103% greater RPDs from PLS and PSR counterparts, respectively. These increments were accompanied with a 67% and 72% decrease in RMSEcv for PSR and PLS, respectively.

Qualitative spectral analysis

The average absorbance spectrum across a 10-band window of one randomly selected sample was characterized by high absorbance values in the visible range (Fig. 5a), indicating the VisNIR sensitivity to soil color. The shape of the spectrum resembles that of the NIR absorbance spectra reported by Rinnan and Rinnan (2007) except with an almost muted 1450 nm peak and two slight but prominent positive peaks at ~1700 (aromatics, $2\upsilon_1$)nm and ~2000 nm (amides, $3\upsilon_1$) (where, υ_i= fundamental mode). The dip near 1940 nm in the spectra could also be suggestive of water ($3\upsilon_1$)or amide N-H.

Given that PLS loading weight vectors can be interpreted as correlations between the variables (wavelength) and the components of interest (soil biological parameter),we plotted the first two loading weight vectors to qualitatively characterize ABS spectra (Fig. 5b) (Chakraborty *et al.* 2013). Haaland and Thomas (1988) revealed that the first loading weight vector indicates a first-order approximation to the "pure" component spectrum and can be useful for making assignments of

Table 4: Summary model statistics obtained for soil biological parameters by three different multivariate methods using absorbance spectra. Also ratio of random forest (RF) results with penalized spline regression (PSR) and partial least squares (PLS) regression results are shown. For example, a value of 1.11 for PSR means the r^2 of RF is about 11% higher than PSR

Property[a]	Model[b]	LF[c]	r^2	RMSEcv[d]	RPD[e]	Bias	r^2	RMSEcv	RPD
			Leave-one-out-cross validation				*Ratio with RF results*		
Total FAME (nmol g^{-1})	PSR	—	0.81	19.20	2.32	-6.62×10^{-12}	1.11	0.72	1.38
	RF	—	0.90	13.88	3.21	0.541	1.00	1.00	1.00
	PLS	10	0.71	23.87	1.86	8.79×10^{-13}	1.26	0.58	1.72
GM+ (nmol g^{-1})	PSR	—	0.85	2.49	2.63	2.81×10^{-13}	1.07	0.80	1.24
	RF	—	0.91	2.00	3.27	0.06	1.00	1.00	1.00
	PLS	10	0.72	3.45	1.89	9.32×10^{-15}	1.26	0.57	1.73
GM-(nmol g^{-1})	PSR	—	0.85	1.88	2.58	-4.06×10^{-13}	1.05	0.83	1.20
	RF	—	0.90	1.57	3.10	0.037	1.00	1.00	1.00
	PLS	10	0.74	2.47	1.96	2.81×10^{-13}	1.21	0.63	1.58
Actinomycetes (nmol g^{-1})	PSR	—	0.83	1.25	2.42	4.14×10^{-13}	1.08	0.75	1.31
	RF	—	0.90	0.94	3.18	0.012	1.00	1.00	1.00
	PLS	10	0.68	1.62	1.69	9.24×10^{-15}	1.32	0.58	1.88
Total bacteria (nmol g^{-1})	PSR	—	0.86	5.26	2.70	-3.34×10^{-13}	1.04	0.84	1.19
	RF	—	0.90	4.44	3.22	0.088	1.00	1.00	1.00
	PLS	10	0.73	7.33	1.93	-8.79×10^{-14}	1.23	0.60	1.66
Total fungi (nmol g^{-1})	PSR	—	0.72	5.24	1.89	8.31×10^{-13}	1.23	0.63	1.60
	RF	—	0.89	3.32	3.03	0.120	1.00	1.00	1.00
	PLS	10	0.65	5.86	1.69	1.88×10^{-13}	1.36	0.56	1.79
AMF (nmol g^{-1})	PSR	—	0.67	4.22	1.74	4.16×10^{-13}	1.26	0.65	1.53
	RF	—	0.86	2.75	2.67	0.125	1.00	1.00	1.00
	PLS	4	0.44	5.48	1.34	-2.32×10^{-14}	1.95	0.50	1.99
Saprophytic fungi	PSR	—	0.38	0.06	1.28	3.28×10^{-15}	2.23	0.33	2.03
	RF	—	0.85	0.02	2.60	0.001	1.00	1.00	1.00
	PLS	6	0.20	0.07	1.12	-2.71×10^{-16}	4.25	0.28	2.32
Fungi: Bacteria	PSR	—	0.68	0.34	1.78	4.06×10^{-14}	1.30	0.58	1.71
	RF	—	0.89	0.20	3.06	0.004	1.00	1.00	1.00
	PLS	9	0.59	0.39	1.56	8.78×10^{-15}	1.50	0.51	1.96
MBC (mg kg^{-1})	PSR	—	0.22	64.87	1.13	-2.04×10^{-12}	4.00	0.38	2.61
	RF	—	0.88	25.01	2.96	1.02	1.00	1.00	1.00
	PLS	7	0.17	66.93	1.10	-1.88×10^{-12}	5.17	0.37	2.69
MBN (mg kg^{-1})	PSR	—	0.62	3.91	1.64	-4.10×10^{-14}	1.38	0.60	1.65
	RF	—	0.86	2.35	2.72	0.123	1.00	1.00	1.00
	PLS	7	0.52	4.40	1.45	1.10×10^{-13}	1.65	0.53	1.87
Soil organic matter (%)	PSR	—	0.86	0.16	2.67	-3.01×10^{-14}	1.06	0.75	1.34
	RF	—	0.92	0.12	3.60	-0.0003	1.00	1.00	1.00
	PLS	10	0.76	0.21	2.07	1.10×10^{-14}	1.21	0.57	1.73

[a]GM+, Gram positive bacteria; GM-, Gram negative bacteria; AMF, arbuscular myccorhizal fungi; MBC, microbial biomass C; MBN, microbial biomass N; [b]RF, random forest; PLS, partial least squares regression; PSR, penalized spline regression; [c]LF, PLS latent factor; [d]RMSEcv, root mean squared error of cross-validation; [e]RPD, residual prediction deviation.

spectral bands that may be important in the analysis. Accordingly,the positive and negative peaks are associated with the component of interest and interfering components, respectively (Viscarra Rossel *et al.* 2006). The first-factor loading weights showed a positive contribution for the whole VisNIR range with minor interfering positive peaking to varying magnitudes at ~2300 to 2400 nm which could arise from methyles ($3\upsilon_1$) and carbohydrates ($4\upsilon_1$) (Viscarra Rossel and Beherens, 2010). The

Table 5: Confusion matrix showing classification of soil biological properties using linear discriminant analysis (LDA). The weighted kappa coefficients are also given (n=123). Q1, Q2, Q3, and Q4 represent 1st quartile, 2nd quartile, 3rd quartile, and 4th quartile of the soil properties, respectively. Highlighted cells represent numbers of correctly classified samples

	Total FAME (κ = 0.89)				GM+ (κ = 0.95)[a]				GM- (κ = 0.93)[b]			
	Measured Q1	Measured Q2	Measured Q3	Measured Q4	Measured Q1	Measured Q2	Measured Q3	Measured Q4	Measured Q1	Measured Q2	Measured Q3	Measured Q4
Predicted Q1	30	2	1	0	31	0	0	0	31	0	0	0
Predicted Q2	1	26	3	0	1	30	0	0	2	28	1	0
Predicted Q3	2	2	27	1	1	2	28	0	1	1	28	0
Predicted Q4	0	0	0	28	0	0	1	29	0	0	3	28
	Overall misclassification			10%	Overall misclassification			4%	Overall misclassification			6.5%

	Actinomycetes (κ = 0.93)				Total bacteria (κ = 0.92)				Total fungi (κ = 0.91)			
	Measured Q1	Measured Q2	Measured Q3	Measured Q4	Measured Q1	Measured Q2	Measured Q3	Measured Q4	Measured Q1	Measured Q2	Measured Q3	Measured Q4
Predicted Q1	29	2	0	0	31	0	0	0	28	3	0	0
Predicted Q2	3	26	1	0	2	28	1	0	3	26	2	0
Predicted Q3	1	0	29	0	1	1	28	0	0	2	27	1
Predicted Q4	0	0	1	30	1	0	2	28	0	0	0	31
	Overall misclassification			6.5%	Overall misclassification			6.5%	Overall misclassification			8%
	AMF (κ = 0.95)[c]				Saprophytic fungi (κ = 0.8)				Fungi: bacteria (κ = 0.91)			
	Measured Q1	Measured Q2	Measured Q3	Measured Q4	Measured Q1	Measured Q2	Measured Q3	Measured Q4	Measured Q1	Measured Q2	Measured Q3	Measured Q4
Predicted Q1	31	0	0	0	28	2	1	1	30	2	1	0
Predicted Q2	1	28	1	1	5	27	1	1	1	29	0	0
Predicted Q3	0	0	30	0	4	3	22	1	1	2	26	0
Predicted Q4	0	0	1	30	1	1	2	23	1	0	1	29
	Overall misclassification			3%	Overall misclassification			18.5%	Overall misclassification			7%
	MBC (κ = 0.90)[d]				MBN (κ = 0.90)[e]				Soil organic matter (κ = 1)			
	Measured Q1	Measured Q2	Measured Q3	Measured Q4	Measured Q1	Measured Q2	Measured Q3	Measured Q4	Measured Q1	Measured Q2	Measured Q3	Measured Q4
Predicted Q1	28	2	1	0	28	3	0	0	32	0	0	0
Predicted Q2	0	30	1	0	2	28	2	0	0	30	0	0
Predicted Q3	0	1	25	4	0	0	26	1	0	0	30	0
Predicted Q4	0	0	1	30	0	0	1	30	0	0	0	31
	Overall misclassification			8%	Overall misclassification			9%	Overall misclassification			0%

[a]*GM+, Gram positive bacteria;* [b]*GM-, Gram negative bacteria;* [c]*AMF, Arbuscular mycorrhizal fungi;* [d]*MBC, Microbial biomass C;* [e]*MBN, Microbial biomass N.*

broad shoulder at 2137 nm was perhaps due to polysaccharides, such as cellulose (Chakraborty *et al.* 2013), which are part of the hard-to-decompose organic C. Conversely, the second-factor loading weights exhibited pronounced positive contributions for wave bands between 350 and ~1450 nm, possibly arising from the combined effect of electronic transitions, hydroxyls (930 nm, $3\upsilon_1$; 700 nm, $4\upsilon_1$; 1400 nm, $2\upsilon_1$), aromatics (1100 nm, $3\upsilon_1$; 825 nm; $4\upsilon_1$), carboxylic acids (1449 nm, $4\upsilon_1$), amines (1000 nm, $3\upsilon_1$; 751 nm, $4\upsilon_1$), and alkyl asymmetric-symmetric doublets (1138 nm, $3\upsilon_3$; 1170 nm, $3\upsilon_1$; 853 nm,$4\upsilon_3$; 877 nm, $4\upsilon_1$) (Viscarra Rossel and Beherens, 2010).

Notwithstanding the high dimensionality of the spectral data (215 spectral channels from 350 to 2500 nm at 10-nm intervals), the first three linear discriminants explained up to 90% of the spectral variance for each variable. Separate pairwise linear discriminant plots (F1 vs.F2) indicating four classes of each parameter were used to distinguish soil ABS spectra and identify spectral similarities within a single class for each parameter (Fig. 6). Almost all plots exhibited ''class-clustered'' structure. Fig. 6*l* confirmed that SOM classes were very well discriminated on the F1 axis which explained roughly three quarters (~72%) of the overall spectral variance. No obvious outlier among the samples was seen. Except for class 4 (Q4), the plots for total FAME (Fig. 6a), GM+ (Fig. 6b), actinomycetes (Fig. 6d), total bacteria (Fig. 6e), total fungi (Fig. 6f), and MBN (Fig. 6k) showed some subtle content-wise separations (Q1 followed by Q2 and Q3) along F2 axes. Nevertheless, some overlapping among higher classes was apparent. Besides, as light separation of first three classes of MBC was apparent along the F1 axis with some overlapping between Q3 and Q4 (Fig. 6j). Interestingly, panels for total FAME, GM+, GM-, actinomycetes, total bacteria, saprophytic fungi, and fungi: bacteria showed separation of Q4 from the other three classes along F1. The LDA confusion matrix summarized the reclassification of the observations (Table 5), and allowed quick visualization of the % of misclassified observations. It is noteworthy that the LDA classification closely followed results of visual F1 vs. F2 plot inspections (Table 5). Indeed, for SOM class classification, LDA was 100% accurate. For total FAME, GM+, GM-, actinomycetes, total bacteria, total fungi, AMF, saprophytic fungi, fungi: bacteria, MBC, and MBN, LDA correctly classified all but 12,5, 8, 8, 8, 11, 4, 23, 9, 10, and 9 samples, respectively. We obtained highest κ (linear weighting) for SOM (1). However, it should be noted that κ-coefficient of agreement values were specific to a given data distribution and thus cannot be directly compared across different datasets (Brown *et al.* 2006).

VisNIR DRS for soil microbial characterization

Our results strongly converged with previous studies in that GM+, actinomycetes, total bacteria, total fungi, MBC, MBN, and SOM are well correlated with soil reflectance (Chodak *et al.* 2007; Rinnan and Rinnan 2007; Zornoza *et al.* 2008).Even though fundamental vibration of organic molecules occurs in the mid-IR region (Williams and Norris, 1987), relatively muted absorption features of their overtones and combination bands arising from stretching and bending of N–H, C–H, and C=O groups present in soil biological fractions were identified by hyperspectral DRS in our study (Ben-Dorand Banin, 1995). Further, the relatively higher light-penetrability arising from high extinction coefficients, less sample handling, and lesser specificity requirements gives an edge to VisNIR DRS over MIR instruments which have been mainly restricted to laboratory analysis (Reeves and Smith, 2009). However, one must use caution while interpreting soil VisNIR spectra, since not only was the spectrum (Fig. 5a) encumbered by the abundance or combination of overtone bands, but these bands were broad and perhaps overlapping. Often the location of the overtone bands was shifted slightly from the exact expected location since real molecules do not act totally harmonically (Bishop *et al.* 1994). Hence, while the target microbial group is present only in trivial quantity as in the cases of several soil biological fractions, assigning band specificity in soil VisNIR spectra poses stern difficulties and thus warrants multivariate modeling.

From LDA plots and classification results, it can be concluded that VisNIR sensing offers both the power of LDA as a means of discriminating the spectra of various classes of soil biological parameters and exhibiting a clustering tendency for content-wise similarity with great sensitivity. Overlapping among higher classes, Q2 and Q3 suggested that spectral diversity was relatively low in these two classes. A misclassification rate between 18.5 and 0%is an

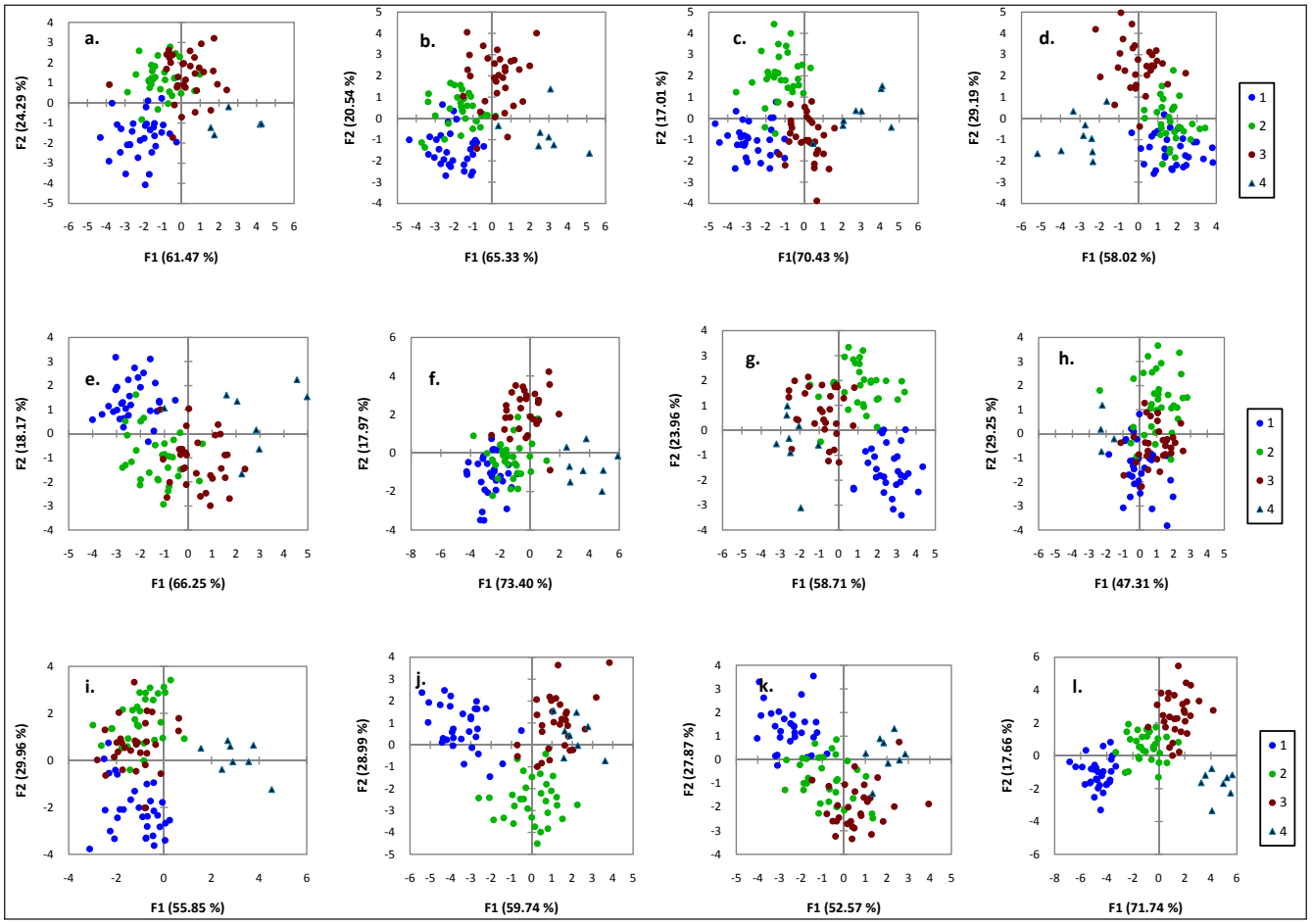

Fig. 6: Plots representing the observations on the first two linear discriminant analysis (LDA)factor axes (F1 and F2) for **(a)** total fatty acid methyl ester (FAME), **(b)** Gram positive bacteria (GM+), **(c)** Gram negative bacteria (GM), **(d)** actinomycetes, **(e)** total bacteria, **(f)** total fungi, **(g)** arbuscularmycorrhizal fungi (AMF), **(h)** saprophytic fungi, **(i)** fungi: bacteria, **(j)** microbial biomass carbon (MBC), **(k)** microbial biomass nitrogen (MBN), and **(l)** soil organic matter (SOM). The classes 1 (blue circle), 2 (green circle), 3 (brown circle), and 4 (black triangles) represent 1st quartile, 2nd quartile, 3rd quartile, and 4th quartile of each property, respectively

encouraging and realistic estimate of the ability for VisNIR spectroscopy to classify soil biological parameters. Nonetheless, it is noteworthy that the lack of high-intensity spectral bands to some extent constrained the utility of qualitative analysis.

Potential of advanced chemometric modeling

Although infrared spectroscopy has been extensively developed in agriculture in the last 40 years, this research area has been experiencing a boom over the last 10 years. Johnson *et al.* (2003) reported the first bacterial groupings based on their soil reflectance properties, similar to bacterial intergenic transcribed spacer analysis. In our study, the robustness and stability of the total RF-FAME model corroborated the findings of Whittaker *et al.* (2003) who successfully measured and discriminated food borne bacterial mixtures of FAMEs using an attenuated total reflection (ATR)-Fourier transform infrared (FTIR) spectroscopic and multivariate analysis. Our study, however, provided unique information of alternative RF and PSR models to improve the predictability of soil biological parameters. The better performance of PSR can be attributed to its stability and flexibility; more so than other parametric PLS approaches since the shape of the functional relationship between covariates and the dependent variable (soil biological parameter, in this study) was managed by the data (Marx and Eilers, 1999). The improvement of model generalization capability by implementing PSR followed the same trend reported elsewhere for other soil and compost properties (Chakraborty *et al.* 2012, 2014). Nonetheless, some of the underestimations of PLS, RF and PSR models could be due to the relative scarcity of observations

at the higher ends of the property scales, as previously outlined by Brown *et al.* (2006). Our RF model results for MBC exceeded those from Palmborg and Nordgren (1993), Chang *et al.* (2001), and Zornoza *et al.* (2008). The latter produced an PLS RMSEcv value of 142 mg kg^{-1} while using 393 air-dried soil samples and is thus unacceptable for mineral soils. Chang *et al.* (2001) reported moderate MBC model generalizing capability with an r^2 of 0.60 and an RPD of 1.10. Although, Couteaux *et al.* (2003) achieved excellent prediction for MBC (r^2=0.95, RPD=4.42) while investigating 204 air-dried forest Oh and A1 layers, their results were biased by the formation of subpopulations in the linear regression (Terhoeven-Urselmans *et al.* 2008). Our study also produced better predictabilities for GM+ (RPD=3.27) and GM- (RPD=3.10) than those of Zornoza *et al.* (2008). Furthermore, our prediction for MBN was better than that of Ludwig *et al.* (2002) who predicted MBN for 120 air-dried forest soils with a coefficient of determination of 0.7. The best predictability generated by RF can be supplemented by: (i) automatic identification of the best predictors, (ii) no need for data transformation or rescaling, (iii) resistance to outliers, (iv) more accurate results than a single tree, and most importantly (v) even growing a large number of RF trees does not create a risk of over-fitting (Breiman, 2001). Notably, on a sample containing several hundred thousand rows, a single RF tree can be built within a minute using a Pentium IV processor.

Sample processing and limitations

In this study one obvious question was: is it possible to directly analyze soil biological parameters in fresh samples by VisNIR DRS? This question was critical as we attempted to develop modeling alternatives which could ultimately lead to *in situ* measurements. Given that, VisNIR DRS can sense the changes in the matrix materials canned, particularly moisture (as it relates to O-H bonding and color) (Bishop *et al.* 1994; Zhu *et al.* 2010), bringing the soil samples to standard water content (field capacity) prior to scanning is critical for obtaining consistent results. In the present study, soil samples were collected during a drought and thus were almost air-dried. Despite that, it cannot be excluded that even in low soil moisture conditions, there was still remaining water adsorbed on the surface areas of clay minerals (e.g., hygroscopic water) and organic matter in equilibrium with atmospheric water vapor. Interestingly, even pre-treatment methods like quick-freezing and freeze-drying were unable to remove the water completely from the layer minerals of soil (Terhoeven-Urselmans *et al.* 2008). However, these minor variations perhaps did not cause much difference in the NIR spectra, as previously identified by Minasny *et al.* (2011). Our results, however, only showed the applicability of predictive models under low moisture conditions as previously demonstrated by other researchers (Rinnan and Rinnan, 2007; Zoronoza *et al.* 2008). Under laboratory controlled settings or sample collection following drought (like in the present study)and/or water logged conditions,soil can be scanned under uniform moisture content. However, maintaining homogeneous water content in the field during *in-situ* scanning is not easy. The RF-MBC model reported an RPD of 2.96 which is interesting since very few studies thus far using air-dried sample produced acceptable MBC predictability (Chodak *et al.* 2003; Zoronoza *et al.* 2008). Perhaps dry conditions are common occurrences in these soils and thus microbial communities have adapted to survive under drier conditions and their response to drought is different than that observed when moist soils are air dried. Despite that, establishing the best sample pretreatment (field-moist or air-dried) was beyond the scope of this study and requires further investigations. Hence, the question of "how best to develop calibrations when moisture is present" is one of the big issues that remains to be properly answered about calibrations for soil biological properties.

Important wavelengths selected by multivariate algorithms

The present study could not identify high resemblance between the regression coefficients of the soil biological parameters and SOM. Indeed, the correlation matrix even suggested absence of any correlation between several variables and SOM (Table 3). While plotting the significant wavelengths of all three models, it was observed that calibrations were carried out independently, with diverse spectral regions implied in each property (Fig. 7). Whilst correlation between SOM and VisNIR absorbance data was developed in the

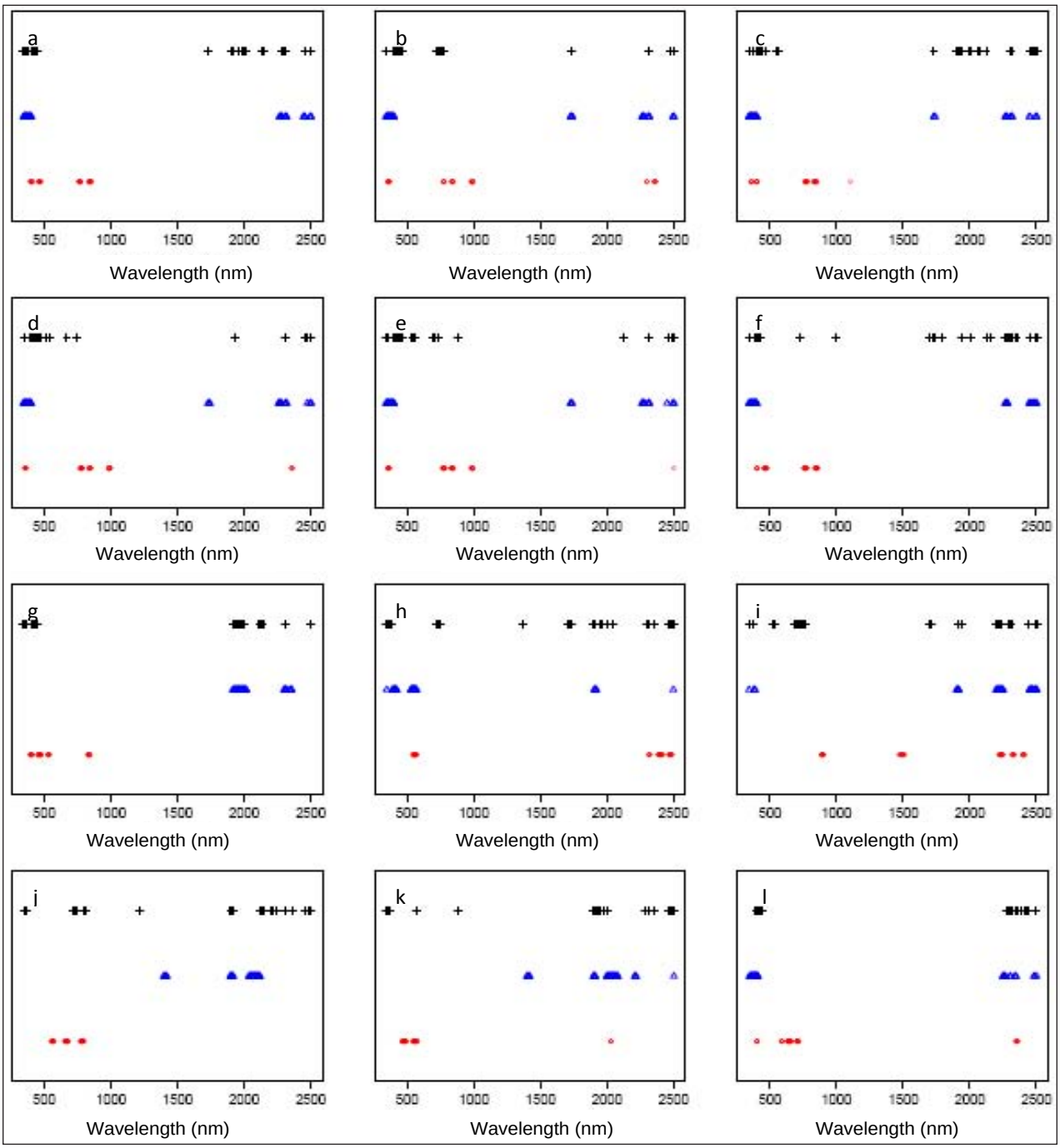

Fig. 7: Significant wavelengths used in the partial least squares (PLS) (black), random forest (RF) (blue), and penalized spline regression (PSR) (red) models for **(a)** total fatty acid methyl ester (FAME), **(b)** Gram positive bacteria (GM+), **(c)** gram negative bacteria (GM-), **(d)** actinomycetes, **(e)** total bacteria, **(f)** total fungi, **(g)** arbuscularmycorrhizal fungi (AMF), **(h)** saprophytic fungi, **(i)** fungi: bacteria, **(j)** microbial biomass carbon (MBC), **(k)** microbial biomass nitrogen (MBN), and **(l)** soil organic matter (SOM)

ranges 400-500 nm and 2300-2500 nm, other ranges were perceptible for other biological properties. Regions like ~2200-2500 nm were used in all PLS and RF and most PSR calibrations, suggesting that this region has deviations in functional groups bound to biological properties, which ultimately translated intovariations in the biological parameter concentrations. Note that, this range comes under the best NIR range of 1650-2500 nm for characterizing organic carbon compounds (Hummel *et al.* 2001; Lee

et al. 2009; Morgan *et al.* 2009). Furthermore, while comparing the $r^2$from simple linear regression of biological properties with SOM, and the values of r^2 obtained from respective RF models with VisNIR spectra, the former were consistently lower than the latter, implying that RF's accuracy cannot be explained by direct correlation with SOM (Table 3). In fact, three variables (saprophytic fungi, fungi: bacteria, and MBC) produced an r^2 of 0. Thus, the postulations of Cohen *et al.* (2005) and Rinnan and Rinnan (2007), that good predictions of soil biological parameters that are present in trivial quantities could be the consequence of high correlations with total soil organic matter quantity, may not be generalized.

Hypothetical sensor development

Improvement of models with variable selection is obvious and should be generalized. Considering that the model with fewer regressors involves fewer optical filters and detectors in a hypothetical sensor, a simpler model would always simplify the sensor's configuration,reduce its weight, and make it more robust (Ge *et al.* 2007). The visual inspection of Fig. 7 revealed that the number of significant 10-nm averaged wavebands for PLS was far greater than RF and PSR counterparts. All these wavebands were almost consistently distributed along the entire VisNIR spectral range, implying that several optical filter-detector pairs with different central wavelengths and a uniform bandwidth of 10 nm would be needed. Further, an extra circuit block to combine these detectors' output signals into four to ten synthetic signals representing the PLS latent factors would also be required (Table 4), unambiguously complicating the sensor development. Similar observations were made by Ge *et al.* (2007) while comparing the PLS based sensors with a DWT based hypothetical sensor. In the present study, the lesser significant RF regressors would open new opportunities for designing a low-cost VisNIR spectrometer from laboratory toon-the-go sensors, dedicated to soil biological measurement, which would drastically reduce measurement costs.

CONCLUSION

Summarily, we have demonstrated the premise of using a VisNIR–RF model as a viable alternative to the VisNIR–PLS model for rapid and low-cost estimation of various biological properties as an addition to the standard methods for soil biological analysis. The study was intended for testing the capability of VisNIR-RF or VisNIR-PSR viability instead of making a lab-grade predictive model. While the RF model remained superior to the other two models evaluated, it was not exhaustive and should be explored further under different soil and management conditions before drawing a stronger conclusion. The next step will be the validation of these models on independent samples. Auxiliary soil properties that can be estimated rapidly and easily (pH, electrical conductivity, etc.) may improve these predictive models when combined with the soil spectra. More improvement could be achieved by increasing sample numbers and mapping the regressors with further refined algorithms like DWT. Clearly, more fundamental investigations as to how total FAME and taxonomic biomarkers and other parameters influence optical properties are warranted. Our study showed good potential as an impetus toward future VisNIR–RF-based soil studies. Spectral scattering properties of soil are highly complex; real-time soil biological characterization is expected to be complex as well. Our future research will be directed toward developing a general model, so that exact spectral features linked with each soil biological and biochemical property can be identified and modeled as appropriate, reflecting different soil compositions.

ABBREVIATIONS

SOM: Soil organic matter; **GM+:** Gram positive bacteria; **GM-:** Gram negative bacteria; **FAME:** Fatty acid methyl ester; **AMF:** Arbuscularmycorrhizal fungi; **MBC:** Microbial biomass carbon; **MBN:** Microbial biomass nitrogen; **PLS:** Partial least squares regression; **PSR:** Penalized spline regression; **RPD:** Residual prediction deviation; **RF:** Random forest; **VisNIRDRS:** Visible near-infrared diffuse reflectance spectroscopy.

ACKNOWLEDGEMENTS

The authors are grateful for financial support from the BL Allen Endowment in Pedology at Texas Tech University. The authors also thank Jon Cotton for his assistance in sample analysis. This project was supported by Agriculture and Food Research

Initiative Competitive Grant Program no. TEXW-2011-03783 from the USDA National Institute of Food and Agriculture.

REFERENCES

Acosta-Martinez, V., Zobeck, T.M.and Allen, V. 2004. Soil microbial, chemical and physical properties in continuous cotton and integrated crop–livestock systems. *Soil Science Society of America Journal,* **68**(6): 1875-1884.

Bellon-Maurel, V. and McBratney, A. 2011. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – critical review and research perspectives. *Soil Biology & Biochemistry,* **43**: 1398–1410.

Ben-Dor, E. and Banin, A. 1995. Near infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Science Society of America Journal,* **59**: 364–372.

Bishop, J.L., Pieters, C.M. and Edwards, J.O. 1994. Infrared spectroscopic analyses on the nature of water in montmorillonite. *Clays & Clay Minerals,* **42**: 702-716.

Breiman, L. 2001. Random forests. *Machine Learning,* **45**: 5–32.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M. and Reinsch, T.G. 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma,* **132**(3-4): 273-290.

Cécillon, L., Cassagne, N., Czarnes, S., Gros, R. and Brun, J.J. 2008. Variable selection in near infrared spectra for the biological characterization of soil and earthworm casts. *Soil Biology & Biochemistry,* **40**: 1975–1979.

Chakraborty, S., Das, B.S., Ali, N., Li, B., Sarathjith, M.C., Majumdar, K. and Ray, D.P. 2014. Rapid estimation of compost enzymatic activity by spectral analysis method combined with machine learning. *Waste Management,* **34**: 623-631.

Chakraborty, S., Weindorf, D.C., Ali, N., Li, B., Ge, Y. and Darilek, J.L. 2013. Spectral data mining for rapid measurement of organic matter in unsieved moist compost. *Applied Optics,* **52**: B82–B92.

Chakraborty, S., Weindorf, D.C., Morgan, C.L.S., Ge, Y., Galbraith, J., Li, B. and Kahlon, C.S. 2010. Rapid identification of oil contaminated soils using visible near-infrared diffuse reflectance spectroscopy. *Journal of Environmental Quality,* **39**: 1378–1387.

Chakraborty, S., Weindorf, D.C., Zhu, Y., Li, B., Morgan, C.L.S., Ge, Y. and Galbraith, J. 2012. Spectral reflectance variability from soil physicochemical properties in oil contaminated soils. *Geoderma,* 177-178, 80-89.

Chang, C., Laird, D.A., Mausbach, M.J. and Hurburgh, C.R. 2001. Near infrared reflectance spectroscopy: principal components regression analysis of soil properties. *Soil Science Society of America Journal,* **65**: 480–490.

Chodak, M., Khanna, P. and Beese, F. 2003. Hot water extractable C and N in relation to microbiological properties of soils under beech forests. *Biology and Fertility of Soils,* **39**: 123–130.

Chodak, M., Ludwig, B., Khanna, P. and Beese, F. 2002. Use of near infrared spectroscopy to determine biological and chemical characteristics of organic layers under spruce and beech stands. *Journal of Plant Nutrition and Soil Science,* **165**: 27–33.

Chodak, M., Niklińska, M. and Beese, F. 2007. Near-infrared spectroscopy for analysis of chemical and microbiological properties of forest soil organic horizons in a heavy-metal-polluted area. *Biology and Fertility of Soils,* **44**: 171-180.

Cohen, M.J., Prenger, J.P. and DeBusk, W.F. 2005. Visible-Near infrared reflectance spectroscopy for rapid, non-destructive assessment of wetland soil quality. *Journal of Environmental Quality,* **34**: 1422–1434

Coleman, D.C. and Whitman, W.B. 2005. Linking species richness, biodiversity and ecosystem function in soil systems. *Pedobiologia,* **49**(6): 479-497.

Couteaux, M.M., Berg, B. and Rovira, P. 2003. Near infrared reflectance spectroscopy for determination of organic matter fractions including microbial biomass in coniferous forest soils. *Soil Biology & Biochemistry,* **35**: 1587–1600.

Curtis, T.P. and Sloan, W.T. 2005. Exploring microbial diversity--A vast below. *Science,* **309**(5739): 1331-1333.

Drenovsky, R.E., Elliott, G.N., Graham, K.J. and Scow, K.M. 2004. Comparison of phospholipid fatty acid (PLFA) and total soil fatty acid methyl esters (TSFAME) for characterizing soil microbial communities. *Soil Biology & Biochemistry,* **36**(11): 1793-1800.

Fernandes, M.F., Saxena, J. and Dick, R.P. 2013. Comparison of Whole-Cell Fatty Acid (MIDI) or Phospholipid Fatty Acid (PLFA) Extractants as Biomarkers to Profile Soil Microbial Communities. *Microbial Ecology,* **66**(1): 145-157.

Ge, Y., Morgan, C.L.S., Thomasson, J.A. and Waiser, T. 2007. A new perspective to near infrared reflectance spectroscopy: a wavelet approach. *Transaction of the ASABE.,* **50**: 303–311.

Haaland, D.M. and Thomas, E.V. 1988. Partial least-squares methods for spectral analyses: 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry,* **60**: 1193–1202.

Hummel, J.W., Sudduth, K.A. and Hollinger, S.E. 2001. Soil moisture and organic matter prediction of surface and subsurface soils using an NIR soil sensor. *Computers and Electronics in Agriculture,* **32**(2): 149-165.

Hunt, G.R. and Salisbury, J.W. 1970. Visible and near-infrared spectra of minerals and rocks: I. Silicate minerals. *Modern Geology,* **1**: 283–300.

Islam, K., Stingh, B. and McBratney, 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Australian Journal of Soil Research,* **41**: 1101–1114.

Jenkinson, D.S. 1988. Determination of microbial biomass carbon and nitrogen in soil. In Wilson, J.R. (Ed.), Advances in nitrogen cycling in agricultural ecosystems. CAB International, Wallingford, pp. 368–386.

Johnson, M.J., Lee, K.Y. and Scow, K.M. 2003. DNA fingerprinting reveals links among agricultural crops,

soil properties, and the composition of soil microbial communities. *Geoderma,* **114**: 279–303.

Lee, K.S., Lee, D.H., Sudduth, K.A., Chung, S.O., Kitchen, N.R. and Drummond, S.T. 2009. Wavelength identification and diffuse reflectance estimation for surface and profile soil properties. *Transactions of the ASABE.*, **52**(3): 683-695.

Ludwig, B., Khanna, P., Bauhus, J. and Hopmans, P. 2002. Near infrared spectroscopy of forest soils to determine chemical and biological properties related to soil sustainability. *Forest Ecology and Management*, **171**: 121–132.

Malley, D.F., Yesmin, L. and Eilers, R.G. 2002. Rapid analysis of hog manure and manure amended soils using near-infrared spectroscopy. *Soil Science Society of America Journal*, **66**: 1677–1686.

Martens, H.A. and Dardenne, P. 1998. Validation and verification of regression in smalldata sets. *Chemometrics and Intelligent Laboratory Systems,* **44**: 99–121.

Marx, B.D. and Eilers, P.H.C. 1999. Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics,* **41**: 1-13.

Millard, P. and Singh, B.K. 2010. Does grassland vegetation drive soil microbial diversity? *Nutrient Cycling in Agroecosystems,* **88**(2): 147-158.

Minasny, B., McBratney, A., Bellon-maurel, V., Roger, J.M., Gobrecht, A., Ferrand, L. and Joalland, S. 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma,* 167-168, 118-124.

Morgan, C.L.S., Waiser, T.H., Brown, D.J. and Hallmark, C.T. 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. *Geoderma,* **151**(3-4), 249-256.

Nelson, D.W. and Sommers, L.E. 1996. Total carbon, organic carbon and organic matter. In: Sparks, D.L. (Ed.), Methods of soil analysis. Part 3. Chemical methods. ASA and SSSA, Madison, WI.

Palmborg, C. and Nordgren, A. 1993. Modelling microbial activity and biomass in forest soil with substrate quality measured using near infrared reflectance spectroscopy. *Soil Biology & Biochemistry,* **12**: 1713–1718.

Pietikainen, J. and Fritze, H. 1995. Clear-cutting and prescribed burning in coniferous forest: comparison of effects on soil fungal and total microbial biomass, respiration activity and nitrification. *Soil Biology & Biochemistry,* **27**: 101–109.

R Development Core Team, 2008. R: a language and environment for statistical computing. Avaliable online with updates at http://www.cran.r-project.org. R Foundation for Statistical Computing, Vienna, Austria. (Verified 18th July 2014).

Reeves III, J.B., McCarty, G.W. and Meisinger, J.J. 2000. Near infrared reflectance spectroscopy for the determination of biological activity in agricultural soils. *Journal of Near Infrared Spectroscopy,* **8**: 161–170.

Reeves III, J.B. and Smith, D.B., 2009. The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America. *Applied Geochemistry,* **24**(8): 1472-1481.

Rinnan, R. and Rinnan, A. 2007. Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of artic soil. *Soil Biology & Biochemistry,* **39**: 1664–1673.

Saeys, W., Mouazen, A.M. and Ramon, H. 2005. Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. *Biosystems Engineering,* **91**: 393–402.

Schutter, M.E. and Dick, R.P. 2000. Comparison of fatty acid methyl ester (FAME) methods for characterizing microbial communities. *Soil Science Society of America Journal*, **64**(5): 1659-1668.

Soil Survey Staff, 2012. Official soil series descriptions. Available at soils.usda.gov/technical/classification/osd/index.html. NRCS, Washington, DC. (Verified 11 July 2014).

Terhoeven-Urselmans, T., Schmidt, H., Georg Joergensen, R. and Ludwig, B. 2008. Usefulness of near-infrared spectroscopy to determine biological and chemical soil properties: Importance of sample pre-treatment. *Soil Biology & Biochemistry,* **40**(5): 1178-1188.

Thompson, W.D. and Walter, S.D. 1988. A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology,* **41**: 949–958.

Vance, E.D., Brookes, P.C. and Jenkinson, D.S. 1987. An extraction method for measuring soil microbial biomass C. *Soil Biology and Biochemistry,* **19**(6), 703-707.

Vasques, G.M., Grunwald, S. and Sickman, J.O. 2009. Modeling of soil organic carbon fractions using visible-near-infrared spectroscopy. *Soil Science Society of America Journal,* **73**: 176–184.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J. and Skjemstad, J.O. 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma,* **131**(1-2): 59-75.

Viscarra Rossel, R.A. and Beherens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma,* **158**: 46–54.

White, D.C., Stair, J.O. and Ringelberg, D.B. 1996. Quantitative comparisons of in situ microbial biodiversity by signature biomarker analysis. *Journal of Industrial Microbiology,* **17**(3-4): 185-196.

Whittaker, P., Mossoba, M.M., Al-Khaldi, S., Fry, F.S., Dunkel, V.C., Tall, B.D. and Yurawecz, M.P. 2003. Identification of foodborne bacteria by infrared spectroscopy using cellular fatty acid methyl esters. *Journal of Microbiological Methods,* **55**(3): 709-716.

Williams, P. and Norris, K. 1987. Near-infrared Technology in the Agricultural and Food Industries. Am. Assoc. Cereal Chem., St. Paul, MN.

Wu, J., Joergensen, R.G., Pommerening, B., Chaussod, R. and Brookes, P.C. 1990. Measurement of soil microbial biomass C by fumigation-extraction-an automated procedure. *Soil Biology & Biochemistry,* **22**(8): 1167-1169.

Zelles, L., Bai, Q.Y., Rackwitz, R., Chadwick, D. and Beese, F. 1995. Determination of phospholipid-and lipopolysaccharide-derived fatty acids as an estimate of microbial biomass and community structures in soils. *Biology and Fertility of Soils,* **19**(2-3): 115-123.

Zelles, L. 1999. Fatty acid patterns of phospholipids and lipopolysaccharides in the characterisation of microbial communities in soil: A review. *Biology and Fertility of Soils,* **29**(2): 111-129.

Zhu, Y., Weindorf, D.C., Chakraborty, S., Haggard, B., Johnson, S. and Bakr, N. 2010. Characterizing surface soil water with field portable diffuse reflectance spectroscopy. *Journal of Hydrology,* **391**: 133-140.

Zornoza, R., Guerrero, C., Mataix-Solera, J., Scow, K.M., Arcenegui, V. and Mataix-Beneyto, J. 2008. Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils. *Soil Biology & Biochemistry,* **40**(7): 1923-1930.