Published April 26, 2018

Nutrient Management & Soil & Plant Analysis

Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library

Nuwan K. Wijewardane
Yufeng Ge*
Dep. of Biological Systems Engineering
Univ. of Nebraska
Lincoln, NE 68588

Skye Wills
Zamir Libohova
USDA–NRCS
National Soil Survey Center
100 Centennial Mall North Fed. Bdg.
Lincoln, NE 68508

Mid-infrared (MIR) reflectance spectroscopy is commonly studied as a rapid and nondestructive method for predictive soil analysis under laboratory conditions. The first objective of this paper is to report an MIR spectral library based on 20,000+ soil samples collected from the United States. The second objective is to assess, using partial least squares regression (PLSR) and artificial neural networks (ANN), the performance of the library to predict 12 physical and chemical soil properties: organic carbon (OC), inorganic carbon (IC), total carbon (TC), total nitrogen (TN), clay, silt, sand, Mehlich-3 extractable phosphorus (P), NH_4OAc extractable potassium (K), cation exchange capacity (CEC), total sulfur (TS), and pH. The third objective is to investigate whether the use of auxiliary variables of master horizon (HZ), taxonomic order (TAXON), and land use land cover (LULC) would improve MIR model performance. The results showed that OC, IC, TC, TN and TS were predicted most satisfactorily with R^2 > 0.95 and RPD (ratio of performance to deviation) > 5.5. Soil CEC, pH, clay, silt, and sand were also predicted satisfactorily with R^2 > 0.75 and RPD > 2.0. P and K were predicted poorly, with R^2 < 0.4 and RPD < 1.4. The ANN models generally outperformed PLSR models, except for clay, silt and sand. Using auxiliary variables (HZ, TAXON, and LULC) to develop stratified models generally improved model performance. The HZ-specific models showed the greatest improvements. Using an MIR spectral library for routine soil analysis would positively impact many modern applications where high spatial resolution, quantitative soil data are demanded.

Abbreviations: ANN, artificial neural network; CEC, cation exchange capacity; HZ, horizon; IC, inorganic carbon; LULC, land use land cover; MIR, mid-infrared; OC, organic carbon; PLSR, partial least squares regression; RMSE, root mean squared error; RPD, ratio of performance to deviation; TAXON, taxonomic order; TC, total carbon; TN, total nitrogen; TS, total sulfur; VisNIR, visible and near infrared.

Core Ideas

- **A mid-infrared spectral library containing 20,000+ samples was reported.**
- **Twelve soil physical and chemical properties were predicted with MIR spectra.**
- **ANN models performed better than PLSR for most soil properties.**
- **Horizon and taxonomic order as auxiliary variables improved prediction for PLSR.**
- **MIR library has the potential as an alternative to laboratory-based analysis for OC and IC.**

VisNIR (visible and near infrared, from 25000 to 4000 cm^{-1}) and MIR (mid-infrared, from 4000 to 400 cm^{-1}) are the two most commonly used spectral regions for soil analysis (Viscarra Rossel et al., 2006). Both techniques are rapid, nondestructive, and require only drying and grinding for sample preparation (which is already required for virtually all soil analyses in the laboratory). Multiple soil properties can be inferred from one spectral scan. These features make VisNIR and MIR highly desirable for many applications requiring high throughput analysis or in situ deployment.

Both VisNIR and MIR are vibrational spectroscopy, a type of spectroscopy involving the absorption of electromagnetic energy due to various vibrational modes of molecules (absorptions in the visible part caused by electronic transition). Fundamentals of these vibrational absorption bands appear in the MIR region. They are strong, distinct, and can be used to fingerprint specific chemical bonds associated with the bands. The VisNIR region is characterized by overtones and combinations of fundamental bands. They are generally weaker, overlapped, and

Soil Sci. Soc. Am. J. 82:722–731
doi:10.2136/sssaj2017.10.0361
Received 11 Oct. 2017.
Accepted 25 Feb. 2018.
*Corresponding author (yge2@unl.edu).

difficult to resolve for specific chemical constituents (Stenberg et al., 2010).

The performance of VisNIR versus MIR for soil analysis was compared in a number of papers, with an emphasis on soil carbon. McCarty et al. (2002) showed that MIR outperformed NIR for soil organic C and inorganic C modeling. Reeves et al. (2006) used MIR and NIR spectra to model organic C, inorganic C, and total C of soil samples from 10 States in the United States and obtained more accurate predictions with MIR. Similarly, Vohland et al. (2014) reported considerable better model performance with MIR compared with VisNIR for organic C, microbial biomass-C, hot water extractable C, and nitrogen. Ge et al. (2014b) demonstrated that, compared with VisNIR, MIR-ATR (attenuated total reflectance) models yielded much higher accuracy for inorganic C and clay content, and slightly higher accuracy for organic C. Henaka Arachchi et al. (2016) also showed improved predictions for organic C fractions of bulk soil samples using MIR compared with NIR. Bellon-Maurel and McBratney (2011) conducted a critical review on this topic and concluded that MIR is better than NIR in predicting C, with prediction errors generally 10 to 40% lower.

Although better modeling results are documented with MIR, VisNIR has a number of advantages that makes it more widely used in the soil research community. First, less expensive, portable, and off-the-shelf VisNIR instruments have been available for a long time. Second, acquiring VisNIR scans requires less sample preparation. Furthermore, VisNIR has been deployed in the field (Ackerson et al., 2017; Bricklemyer and Brown, 2010) and used for intact, field-moist soil samples (Ge et al., 2014a; Minasny et al., 2011).

Large VisNIR soil spectral libraries have been also developed and compiled. Different from local datasets which contain a few hundred samples covering field or watershed scales, large-scale VisNIR libraries cover regional, national or global scales with a few thousands to tens of thousands of soil VisNIR spectra and reference soil data (Brown et al., 2006;Viscarra Rossel et al., 2016; Wijewardane et al., 2016b). They can be accessed and used by researchers from different countries to model and predict soil properties with higher throughput and lower cost, compared with the traditional laboratory-based soil analysis (Wijewardane et al., 2016b). This becomes particularly relevant as there are increasing demands for densely sampled soil data in the space-time continuum for applications like precision agriculture, land resource management, and hydrological and ecological modeling (Schmugge et al., 2002; Viscarra Rossel et al., 2011; Viscarra Rossel and Bouma, 2016).

Literature is scanty on large-scale MIR soil spectral libraries. Viscarra Rossel et al. (2008) reported an MIR library from Australia containing ~1900 samples; and Terhoeven-Urselmans et al. (2010) reported an MIR database containing a diverse set of 971 samples from the International Soil Reference and Information Center. On the other hand, the superior performance of MIR models for many soil properties suggests the great potential of such MIR libraries for high throughput soil analysis under laboratory conditions. The MIR may not be quite suitable for in situ soil analysis, but it is conceivable that a soil MIR library can be a very powerful tool to supplement conventional laboratory analysis and increase throughput (Nocita et al., 2015). This paper intends to fill in this gap in the literature with the following three objectives: (i) introduce a compiled MIR soil spectral library comprising over 20,000 soil samples collected from the United States, (ii) evaluate the performance of the MIR library in predicting a number of common soil physical and chemical properties using partial least squares regression and artificial neural networks, and (iii) investigate the refinement of MIR models using auxiliary variables.

MATERIALS AND METHODS

Soil Samples and MIR Spectral Library

The MIR spectral library was compiled at USDA–NRCS National Soil Survey Center (NSSC). The library consisted of 20,153 soil samples. The soil samples were collected throughout the United States over the past 17 yr, belonged to various projects, and were consistently processed and analyzed following The Kellogg Soil Survey Laboratory (KSSL) protocols (Soil Survey Staff, 2014b).

Twelve soil properties in this study included organic carbon (OC), inorganic carbon (IC), total carbon (TC), total nitrogen (TN), percentages of clay, silt and sand, cation exchange capacity (CEC), Mehlich-3 extractable potassium (K), NH_4OAc extractable phosphorus (P), total sulfur (TS), and pH. Table 1 lists reference laboratory methods used to measure these soil properties (Soil Survey Staff, 2014b).

Each sample also had three auxiliary variables: Land use land cover (LULC, as described in Fry et al., 2011), Taxonomic orders (TAXON), and field described Master Horizons (HZ) (Soil Survey Staff, 2014a). These variables were later used to stratify samples for the refinement and improvement of spectral modeling. Table 2 shows the number of samples in each category. Nearly half of soil samples were from B horizons or Mollisols. In LULC, samples from agriculture lands (Ag Land) had the highest proportion while woody wetlands had the lowest proportion.

A Fourier-Transform Infrared spectrometer (FT-IR, Vertex 70, Bruker Optics) was used to acquire MIR spectra (in diffuse reflectance mode or DRIFT) of all soil samples. The instrument had an MCT (mercury cadmium telluride) detector cooled by

Table 1. Soil properties studied and the reference methods used for their lab analysis.

Soil property	Unit	Reference method
Organic C	%	Total carbon minus inorganic carbon
Inorganic C	%	HCl treatment/Manometric
Total C	%	Dry combustion
Total N	%	Dry combustion
Clay, silt and sand	%	Pipette method
CEC	$cmol_c\ kg^{-1}$	NH_4OAc/pH 7 extraction
Extractable K	$cmol_c\ kg^{-1}$	NH_4OAc/pH 7 extraction
Extractable P	$mg\ kg^{-1}$	Mehlich-3 extraction
Total S	%	Dry combustion
pH		1:1 water extraction

liquid nitrogen. The spectral resolution was set at 4 cm^{-1}. Air-dried and ground samples (<2 mm) were further ground to 180 μm and then loaded into a 96-well spot plate. The plate was coated by a layer of reflective anodized aluminum. The wells were 6 mm in diameter and 1.3 mm deep. For each soil sample, four subsamples were drawn and loaded into four wells on the spot plate; and the four scans were averaged to obtain the MIR spectrum of that sample. Care was given to ensure the surface of the sample in each well was flat (by using a cylindrical, flat-bottomed press tool). The MIR spectrum of each sample was then measured by the instrument in a sequential manner, and a reference spectrum was collected immediately before every soil sample by scanning an empty well as the reference. Both the sample and reference scans from each well were 32 co-added instantaneous scans. All spectra were first stored in the FT-IR spectrometer's OPUS software format and then converted to csv for processing and analysis.

Spectral Modeling with PLSR and ANN

Spectra collected by the instrument covered the spectral range from 7498 to 600 cm^{-1} (the lower end determined by MCT detector limit). However, for spectral analysis and modeling, only the range from 4000 to 600 cm^{-1} was used, which is commonly regarded as the MIR spectral region.

Spectra were preprocessed with an average window of 10 bands. This improved signal to noise ratio of the spectra, reduced the dimensionality of data for effective computation, and avoided model overfitting. The library was randomly split into two datasets as a calibration set (50%) used for model development, and a validation set (50%) for model validation. Models were calibrated using partial least squares regression (PLSR) and artificial neural network (ANN).

The PLSR is the most commonly used technique for chemometric modeling and the de facto standard method in soil spectroscopy. Similar to principal component analysis, PLSR reduces predictor variables to several synthetic variables known as "latent variables" while considering the response variable simultaneously. A linear model is then fitted between the latent variables and the response variable (Helland, 2004). Unlike ANN, PLSR modeling is less computationally demanding and more interpretable (Stenberg et al., 2010).

Table 2. Number of soil samples (in parentheses) in each category of field-described master horizon, taxonomic order, and land use land cover.†

Master horizon	Taxonomic order	Land use land cover
O (1058)	Alfisols (3115)	Ag Land (4258)
A (4275)	Andisols (945)	Deciduous Forest (1725)
E (389)	Aridisols (921)	Developed/Open Space (715)
B (7317)	Entisols (1503)	Evergreen Forest (3100)
C (2526)	Histosols (612)	Grassland/Pasture (3656)
	Inceptisols (2677)	Shrubland (2299)
	Mollisols (6948)	Woody Wetlands (655)
	Spodosols (1184)	Barren (101)
	Ultisols (1693)	Developed/High Intensity (84)
	Gelisols (215)	Developed/Med. Intensity (286)
	Oxisols (6)	Developed/Low Intensity (341)
	Vertisols (325)	Herbaceous Wetlands (110)
		Mixed Forest (399)
		Open Water (294)
		Perennial Ice/Snow (6)

† Categories having <500 samples were not used for model calibration and validation in stratified modeling.

The ANN is inspired by the networks of biological neurons, which have layers of nodes acting as nonlinear summing devices. These nodes are connected to input variables by weights which are adjusted iteratively in model calibration (Dayhoff and DeLeo, 2001). Back-propagation is one technique to adjust these weights to minimize the learning error by propagating the error back to the input layers (Rumelhart et al., 1985; Gallant, 1993). ANN is effective in scenarios where low signal-to-noise ratio is observed in data and interpretation is not one of the goals (Hastie et al., 2001).

For PLSR, models with the number of latent variables from 1 to 30 were considered (as the tuning parameter); and the optimum models were selected by number of latent variables which gave the first local minimum of root mean squared error of cross-validation ($RMSE_{CV}$) using 50-fold cross-validation. For ANN, a grid search with two tuning parameters (the number of nodes in the hidden layer from 3 to 25, and the decay of weight at each iteration set at 0.01, 0.1, and 0.3) was used to select the model with the lowest $RMSE_{CV}$ values.

Model performances were evaluated by calculating R^2 (coefficient of determination between predicted and reference values in the validation set), root mean squared error of validation ($RMSE_V$), ratio of performance to deviation (RPD), and ratio of performance to inter-quartile range (RPIQ, Bellon-Maurel et al., 2010). The RPIQ was included because the distributions of all soil properties deviated substantially from normal (Table 3).

To investigate whether stratifying samples by the auxiliary variables would improve MIR modeling, we calibrated the MIR models according to the classes for each auxiliary variable. For example, when HZ was used as the auxiliary variable, we developed five HZ-specific PLSR models (that is, O-model, A-model, E-model, B-model, C-model; Table 1) using the calibration samples in each class. These HZ-specific models were then applied to the validation samples in their corresponding class. The R^2 and $RMSE_V$ were then calculated across all the validation samples and compared with the generic models (that is, without using the auxiliary variables). Note that for some classes, the number of samples was not enough to calibrate a robust PLSR or ANN model (e.g., there are only six Oxisol samples in the library). We set an arbitrary threshold of 500, meaning that if the number of samples in a class was less than 500, the specific model for that class was not built to avoid model overfitting.

Model calibrations were implemented in the supercomputer cluster at the Holland Computing Center of University of Nebraska-Lincoln with 64 2.1 GHz cores and 250 GB RAM. Data analysis was performed in the R environment (R Core Team, 2016) with the following packages: pls (Mevik et al., 2013) for PLSR, nnet (Venables and Ripley, 2002) for ANN, caret (Kuhn et al., 2015) as the modeling wrapper, and doParallel (Revolution Analytics and Weston, 2015) for parallel processing.

Table 3. Summary statistics of soil properties in the calibration and validation sets.

Data set	Statistic	Soil property†											
		OC	IC	TC	TN	Clay	Sand	Silt	CEC	K	P	Total S	pH
		%							$cmol_c\ kg^{-1}$		$mg\ kg^{-1}$	%	
Calibration	Minimum	0	0	0	0	0	0.2	0	0	0	0	0	2.32
	Median	0.64	0.13	0.97	0.08	21.11	33.6	39.3	14.97	0.33	6.03	0.01	6.23
	Mean	3.47	0.84	3.77	0.2	22.72	38.86	38.43	17.88	0.53	21.52	0.15	6.4
	Maximum	62.43	10.56	62.3	6.91	96.14	100	94.5	199.6	30	587.2	18.78	10.49
	Skewness	4	2.68	3.97	4.78	0.7	0.5	–0.05	4.35	12.95	5.82	12.9	0.13
	Kurtosis	18.48	11.84	18.43	31.61	3.26	2.08	2.31	34.69	319.97	55.27	188.77	2.11
Validation	Minimum	0	0	0	0	0	0.1	0	0	0	0	0	2.67
	Median	0.63	0.12	0.98	0.08	21.27	34.2	38.6	14.84	0.33	6.01	0.01	6.27
	Mean	3.13	0.85	3.42	0.19	22.63	39.4	37.98	17.78	0.53	22.71	0.21	6.42
	Maximum	57.17	12.62	57.17	5.23	90.82	100	94.2	377.1	32.33	595	25.24	10.23
	Skewness	4.35	2.84	4.32	5.03	0.7	0.48	–0.04	5.09	14.59	5.37	10.14	0.13
	Kurtosis	21.83	13.35	21.86	33.04	3.27	2.05	2.29	51.52	451.45	42.45	115.69	2.05
Total no. of samples		17298	7861	20102	20102	18952	18961	18961	17873	17877	4969	20102	18493

† OC, organic carbon; IC, inorganic carbon; TC, total carbon; TN, total nitrogen; CEC, cation exchange capacity.

RESULTS AND DISCUSSION

Soil Properties in the MIR Spectral Library

Table 3 gives summary statistics of the 12 soil properties in the calibration and validation sets. The mean and median values are comparable in both sets, indicating the split of data was balanced which was important for effective model evaluation. All soil properties appear to deviate substantially from normal distributions, as indicated by their Skewness and Kurtosis values. The extreme values in OC (>50%) are the samples of O horizons of Histosols or Spodosols in agricultural lands or forest areas. Samples with extreme IC values (>8%) are mainly from B horizons of Mollisols. Highly sandy samples (>90%) are from the C horizons of Entisols in south Michigan. Highly clayey samples (>80%) are from B horizons of Alfisols in Superior Lake Plain.

The wide range and non-normal distribution of properties are expected as samples in this database are from wide geographic areas covering a broad range of climate, parent material, land cover and management practices. In addition, the samples in the library do not represent all soil orders and horizons equally (Table 2), which also leads to the skewed distributions of soil properties. Previous studies have reported similar distributions (Brown et al., 2006; Wijewardane et al., 2016b) and applied log transformations (Askari et al., 2015; Mulder et al., 2016) to improve model performance. We did not employ variable transformation in this study, because the modeling method used (in particular ANN) and assessment metrics (RPIQ) would address this non-normality issue effectively (Bellon-Maurel et al., 2010).

Soil samples with high OC (37%), IC (10%), and clay (96%) are given as an example to show samples' MIR spectra (Fig. 1). Several strong absorption bands that can be associated with certain functional groups are identifiable. For the high clay sample, the absorption peak at 3620 cm^{-1} (labeled a in Fig. 1) is commonly seen for clay minerals (e.g., kaolinite, smectite, and illite); and the peak at 1645 cm^{-1} (labeled e in Fig. 1) is caused by H–O–H bonds of water in the clay lattice (Nguyen et al., 1991). For the high OC sample, C–H stretching bands from methyl and methylene groups of organic matter at 2920 cm^{-1} and 2850 cm^{-1} (labeled b and c in Fig. 1) are identified, along with the carbonyl C=O band of organic matter at 1750 cm^{-1} (labeled f in Fig. 1). Diagnostic bands of carbonates at 2510, 1800, 1415, 870, and 710 cm^{-1} can be identified in the high IC sample (all labeled d in Fig. 1). These characteristic bands, as well as the overall spectral shape, are in good agreement with the published soil MIR spectra (McCarty et al., 2002; Ge et al., 2014b).

Mid-Infrared Spectral Library Model Performance

The MIR modeling (PSLR and ANN) results (Table 4) indicate the target soil properties can be grouped into four classes in terms of prediction accuracy. The first group includes OC and IC. These two properties are predicted with the highest accuracy, with very high R^2 (0.99), very low bias, and high RPD (>11) (with ANN method for validation). The second group includes TC, TN, and TS. The models for these three properties also showed satisfactory validation statistics, with $R^2 = 0.97$, low bias and RPD values greater than 5.5 (again, with ANN method in validation). The third group includes clay, silt, sand, CEC, and pH, which are predicted with intermediate accuracy. Their validation R^2 values vary between 0.80 and 0.90, and RPDs vary between roughly 2.0 to 3.0 with the better of the two modeling approaches. The last group includes K and P, which are predicted with low accuracy (R^2 in general lower than 0.50). Biases of these models are small relative to $RMSE_V$, suggesting that lack-of-fit is the major source of error for these two properties. A visualization of six soil proprieties (OC, IC, clay, sand, CEC and pH) is given in Fig. 2 with scatterplots of MIR-predicted (with ANN) versus laboratory-measured values for the validation set and a 1:1 line for reference.

Soil OC, IC, TC, TN and TS contribute directly to chemical bonds of carbon-containing compounds in soil (namely, organic matter and carbonates). The high R^2 values for these properties can therefore be attributed to the specific strong absorption bands associated with these chemical bonds (Viscarra Rossel and Behrens, 2010). On the other hand, properties like K

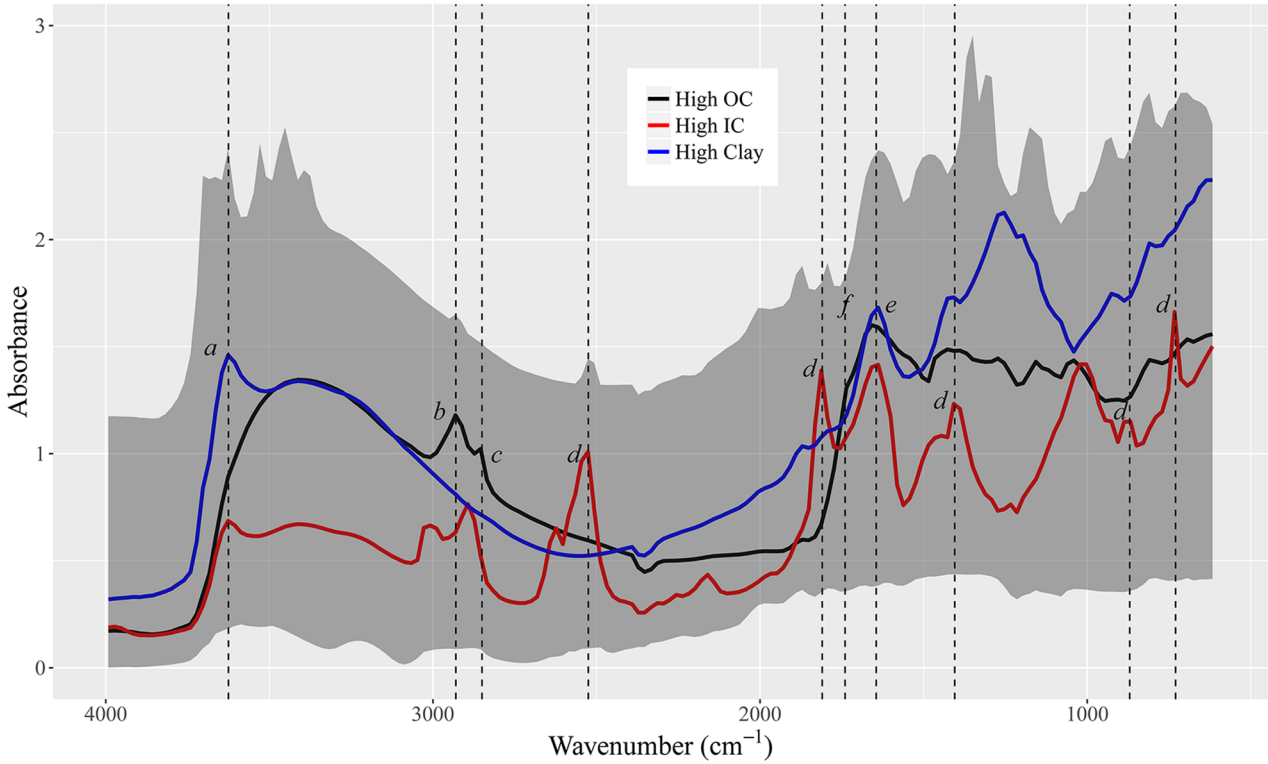

Fig. 1. Mid-infrared (MIR) spectra of selected samples of 37% organic carbon, 10% inorganic carbon, and 96% clay showing the common diagnostic absorption bands caused by organic matter (b, c, and f), carbonates (d), and clay minerals (a and e). The gray spectral envelope is the bounding maximum and minimum value of the MIR absorption of the entire library at each wavenumber.

Table 4. Cross validation and validation results of mid-infrared modeling for different soil properties using partial least squares regression (PLSR) and artificial neural network (ANN).

Soil property	Modeling method	Calibration set		Validation set				
		$RMSE_{CV}$†	R^2	$RMSE_V$‡	R^2	Bias	RPD§	RPIQ¶
Organic C (%)	PLSR	2.10	0.95	1.89	0.95	0.00	4.55	0.82
	ANN	1.00	0.99	0.75	0.99	–0.01	11.46	2.05
Inorganic C (%)	PLSR	0.24	0.97	0.26	0.97	–0.02	5.61	4.42
	ANN	0.15	0.98	0.13	0.99	0.00	11.23	8.98
Total C (%)	PLSR	2.02	0.95	1.90	0.95	0.00	4.44	1.08
	ANN	1.01	0.99	1.34	0.97	–0.01	6.32	1.54
Total N (%)	PLSR	0.15	0.86	0.15	0.86	0.00	2.69	0.83
	ANN	0.08	0.96	0.07	0.97	0.00	5.76	1.77
Clay (%)	PLSR	6.07	0.85	6.01	0.85	0.02	2.60	3.91
	ANN	5.98	0.86	7.61	0.77	–0.89	2.05	3.08
Silt (%)	PLSR	10.24	0.73	9.96	0.74	0.02	1.96	2.88
	ANN	10.59	0.70	10.09	0.73	0.38	1.93	2.84
Sand (%)	PLSR	12.22	0.82	12.07	0.82	–0.10	2.36	3.92
	ANN	13.20	0.78	14.25	0.75	–0.56	2.00	3.33
CEC ($cmol_c\ kg^{-1}$)	PLSR	6.33	0.86	6.50	0.86	–0.06	2.63	2.36
	ANN	5.72	0.88	5.58	0.90	0.01	3.09	2.75
K ($cmol_c\ kg^{-1}$)	PLSR	0.64	0.37	0.70	0.29	0.00	1.19	0.71
	ANN	0.53	0.51	0.51	0.48	0.02	1.33	0.97
P ($mg\ kg^{-1}$)	PLSR	37.26	0.20	45.17	0.14	–1.26	1.08	0.47
	ANN	37.40	0.19	44.64	0.16	–2.59	1.09	0.47
Total S (%)	PLSR	0.30	0.78	0.31	0.94	0.00	4.21	0.08
	ANN	0.19	0.91	0.23	0.97	0.01	5.85	0.11
pH	PLSR	0.57	0.80	0.57	0.80	0.00	2.24	3.84
	ANN	0.44	0.88	0.43	0.89	0.00	2.99	5.12

† Root mean squared error of cross validation.
‡ Root mean squared error of validation.
§ Ratio of performance to deviation.
¶ Ratio of performance to inter-quartile range.

and P cannot be assigned to particular MIR absorption bands, which suggests why these properties are not predicted as successfully as the first group.

In general, models calibrated with ANN outperformed those with PLSR (except for clay, silt and sand). The better results by ANN for majority of the soil properties are not surprising. The large soil library with samples from very different backgrounds (in terms of climate, parental material, for example), leads likely to complex and nonlinear relationship between soil properties and MIR spectra, which would be better modeled by ANN. The ANN approach may also be advantaged by modeling the soil properties in the original scale rather than log-transformed scale. These results are also in agreement with other VisNIR studies where nonlinear methods performed better than the linear method (i.e., PLSR) when modeling large datasets (Viscarra-Rossel and Behrens, 2010; Wijewardane et al., 2016b).

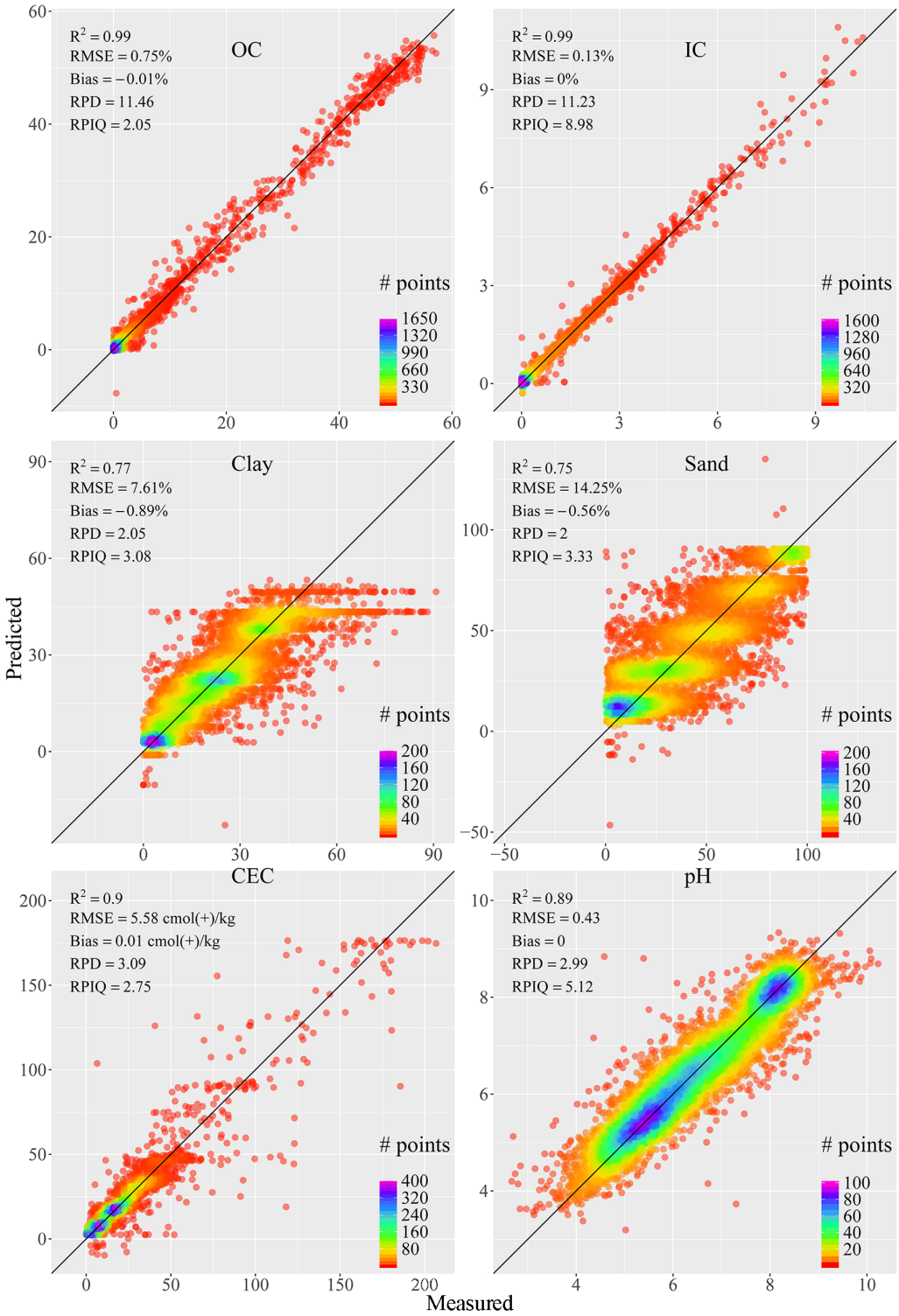

Fig. 2. Scatterplots of laboratory-measured versus mid-infrared-predicted organic carbon, inorganic carbon, clay, sand, cation exchange capacity (CEC), and pH using an artificial neural network.

While not tested in this study, other nonlinear chemometric methods such as random forests and support vector regression have also been widely and successfully used in modeling large soil spectral datasets (Wijewardane et al., 2016a).

One advantage of PLSR is that it allows the models (namely, regression coefficient in each wavenumber) to be plotted and examined while ANN yields black-box models. Wavenumbers with larger regression coefficients indicate a larger contribution to the final predicted values (Beebe and Kowalski, 1987), which in some cases could provide spectroscopic explanation of these models. Plots of wavenumber versus PLSR regression coefficient for the models of OC, IC, clay, sand, and CEC are shown in Fig. 3. Clearly, the diagnostic bands in Fig. 1 are also identified in these PLSR models. For example, C–H stretching at 2920 and 2850 cm^{-1} (again labeled b and c as in Fig. 1) appear to be significant in the OC model, together with the C=O stretching at 1750 cm^{-1} (labeled f). Similarly, all five bands associated with carbonates (labeled d) can be identified in the IC model; and the two bands associated with clay content (labeled a and e) appear in the clay model. It is also interesting to point out that, for the CEC model, both bands attributable to clay and OC are found. CEC is usually correlated with clay and OC, which likely explains the presence of clay and OC bands in the CEC model.

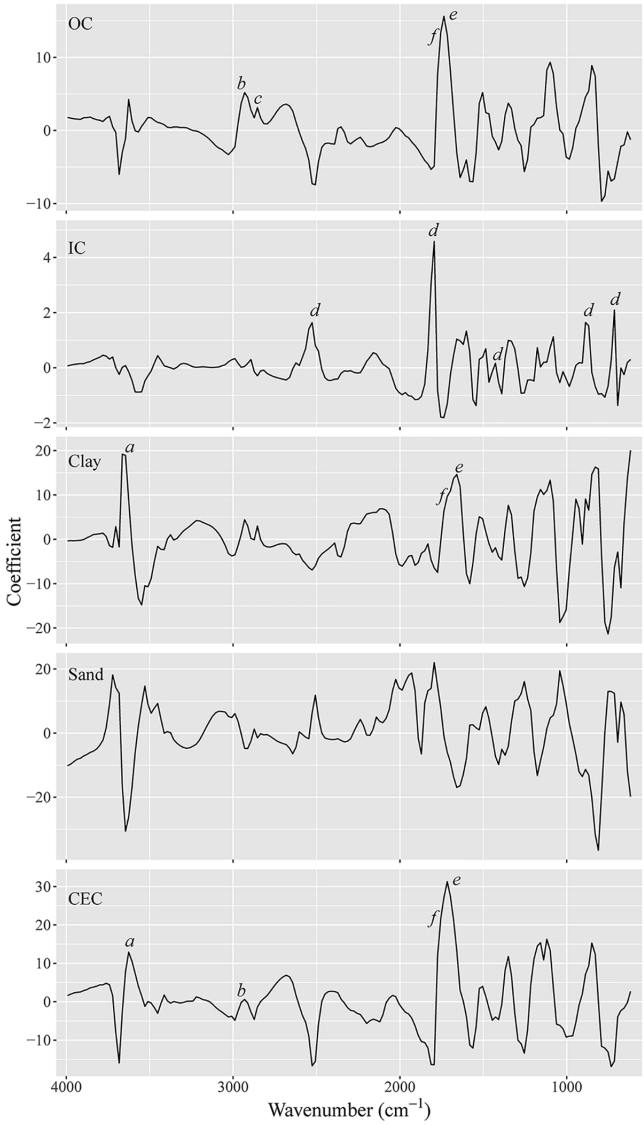

Fig. 3. Partial least squares regression model regression coefficients of organic carbon, inorganic carbon, clay, sand, and CEC. The large regression coefficients that can be associated with diagnostic absorption bands in soil mid-infrared spectra were labeled as in Fig. 1.

Stratified Modeling with Auxiliary Variables of Horizon, Taxon, and Land Use

Validation results of each soil property with stratified modeling using auxiliary variables of HZ, TAXON, and LULC are given in Supplemental Tables S1 through S12. Because there are many stratified MIR models to be compared, it is not easy to generalize trends or patterns regarding model performance. However, it is evident that, for the soil proprieties which are predicted satisfactorily with the generic model, those are also predicted satisfactorily with stratified models. K and P are still the two properties showing the poorest overall performance in stratified models; even though for K, some stratified models have improved substantially (for example, the Mollisols-K model using TAXON as the stratifying criterion and the Developed/Open Space-K model using LULC as the stratifying criterion using ANN). The good performance of these stratified models also suggests the usefulness of the large-scale MIR library on predicting local-scale datasets, and alleviates the concern that the good performance of generic models for some soil properties is inflated by the great range of values in the library.

To compare the effectiveness of the three auxiliary variables for sample stratification, Fig. 4 is produced by pooling all samples from different stratified models and calculating the overall $RMSE_V$ across the strata for the six selected soil proper-

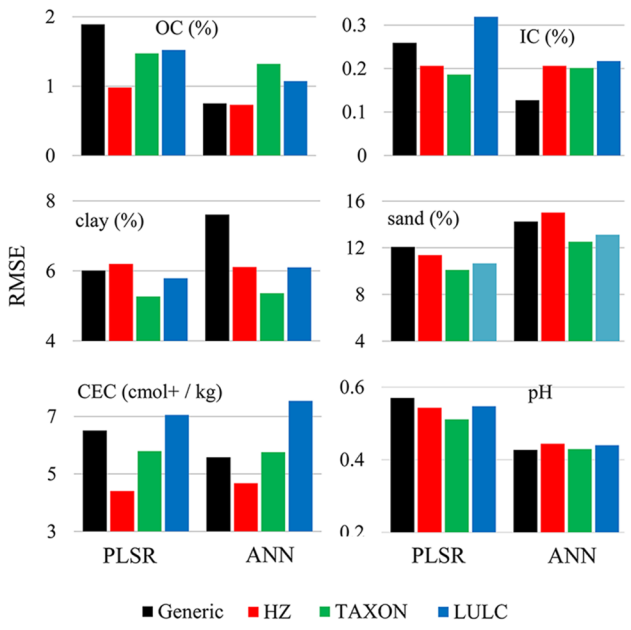

Fig. 4. Comparison of generic modeling with specific modeling using master horizon (HZ), taxonomic order (TAXON), and land use land cover (LULC) as the stratifying variables. Both modeling approaches of partial least squares regression (PLSR) and artificial neural networks (ANN) are compared. The root mean squared error of validation ($RMSE_V$) was calculated by pooling the samples from all strata.

ties. Among the three variables, HZ or TAXON appears to be more effective in stratifying the samples to improve the PLSR model performance, compared with LULC, particularly for OC, IC and CEC. For example, $RMSE_V$ of the generic OC model was 1.89%. When HZ, TAXON and LULC were used for sample stratification, $RMSE_V$ was improved to 0.98, 1.47, and 1.52%, respectively. Similarly, HZ and TAXON also reduced $RMSE_V$ for IC, clay, sand, CEC and pH, with the exception that HZ stratification increased $RMSE_V$ of clay slightly. The use of LULC, however, led to mixed results. It reduced $RMSE_V$ for OC, clay, sand and pH, but increased $RMSE_V$ for clay and CEC.

The higher effectiveness of HZ or TAXON compared with LULC as stratifying variables for MIR PLSR model improvement can be attributed to the fact that HZ and TAXON are associated with the pedogenic processes and intrinsic physical properties of soils. Using either variable to stratify the library could lead to sample classes which have spectrally and compositionally more similar samples, and result in better MIR models within each class. LULC, on the other hand, is more related to management aspect of soils that are applied over diverse soils and soil properties. Therefore, LULC may not be as effective as HZ or TAXON in stratifying samples. This finding is also consistent with another study conducted by Wijewardane et al. (2016b) where sample stratification with HZ and TAXON improved the performance of a VisNIR spectra library for OC and TC prediction.

In contrast to PLSR, using auxiliary variables does not seem to improve the performance of ANN models for these soil properties. Clay is the only soil property that shows significant decrease in $RMSE_V$ with all three auxiliary variables when compared with the generic ANN model. This improvement, nevertheless, is discounted when the high $RMSE_V$ of the generic clay ANN model is considered. One possible explanation is that as a nonlinear modeling approach, ANN can effectively account for the nonlinearity issue (for instance, by accommodating different associations between the target soil properties and MIR spectra among different soil horizons or orders) in the library. This resulted in already good ANN models, leaving only limited room for improvement with sample stratification.

When combining modeling approaches and stratification, it seems that for OC, IC, and pH, generic ANN models show better performances compared with PLSR. For clay and sand, using TAXON-specific PLSR give the best result. For CEC, HZ-specific PLSR models give the highest overall prediction accuracy.

Stratification can also be done with two or more auxiliary variables, which potentially can form more uniform calibration sets and further improve model performance. As an example, we stratified the MIR library with "HZ = A and LULC = Ag Land". This process selected shallow layer, surface samples from lands for agricultural use, which are the target of many precision agriculture applications (Ge et al., 2007). The modeling result (Table 5) show that, with this stratified subset, $RMSE_V$ for OC is 0.60 and 0.52% for PLSR and ANN models, respectively. $RMSE_V$ for clay is 5.01 and 4.33% for PLSR and ANN models, respectively. Compared with their respective generic models (Table 4), $RMSE_V$ for these two soil properties are much lower. Some applications, such as generating site-specific fertilization or irrigation management zones based on grid soil sampling, can be supported and expedited by the MIR library.

Table 5. Validation results of mid-infrared modeling (partial least squares regression [PLSR] and artificial neural network [ANN]) for the subset selected from the library using "HZ = A and LULC = Ag Land".

Soil property	Modeling method	Validation statistic				
		$RMSE_V$†	R^2	Bias	RPD‡	RPIQ§
Organic C (%)	PLSR	0.60	0.92	0.05	3.45	2.78
	ANN	0.52	0.94	0.01	4.00	3.22
Clay (%)	PLSR	5.01	0.85	–0.47	2.49	3.55
	ANN	4.33	0.88	–0.73	2.87	4.11

† Root mean squared error of validation.

‡ Ratio of performance to deviation.

§ Ratio of performance to inter-quartile range.

Implications for the Use of Mid-Infrared Spectra

The results in this study show that MIR-based prediction can be accurate (validation $R^2 \geqslant 0.97$, RPD > 5.5) for OC, IC, TC, TN and TS over a range of different soils in the United States. Given this level of accuracy, the MIR library can be used to predict the properties of new soil samples with higher confidence, and in certain situations, replace wet-chemistry laboratory-based analysis to increase speed or reduce cost (Janik et al., 1998; Nocita et al., 2015). Recently, there have been emerging and cross-cutting fields where spatially dense and low-cost soil data are increasingly demanded, such as precision agriculture, soil process modeling, and soil carbon inventory and change detection at different scales. In this context, MIR-based soil analysis could be a solution to this problem of soil data scarcity in these modern applications. The MIR models can also be used as a rapid screening tool for laboratory quality control and quality assurance protocols (i.e., in addition to wet-chemistry analysis). For these reasons, MIR libraries would be attractive to commercial or noncommercial (such as those affiliated with governments and universities) soil laboratories, with great potential to reduce operating costs and increase analysis speed.

The use of the MIR spectral library in this study will also face several challenges. The first challenge comes from the fact that MIR models are derived from data-driven approaches and empirical in nature. If new samples to be analyzed fall outside the regime of the calibration set (for instance, the target soil properties outside the prediction range), it is likely that the prediction will be unduly poor. This is a key reason that hinders the practical use of MIR. Further studies are needed especially in understanding the intrinsic relationships between MIR wavelengths and soil properties if MIR would be used as the only measurement method of soil properties.

A second challenge is the computational resources needed for calibrating the prediction models. For the majority of soil spectroscopy studies, modeling is done on a personal computer with a few hundred soil samples. As an MIR spectral library can

have tens of thousands of soil samples, model calibration can become cumbersome and time consuming, especially with data mining techniques such as ANN. In this study, computation was done on a high-performance computing cluster, which reduced the time requirement of model calibration significantly. It is worth to note that ANN modeling with our dataset cannot be done on a personal computer with RAM of 16GB (simply because the memory was not enough). Therefore it is essential to have adequate computational resources to support MIR spectral libraries for model calibration, library updates (to include new samples), and recalibration.

Another challenge of MIR spectral libraries involves with the approaches for model calibration. Should all samples in the library be used, or only a smaller subset of the library be selected adaptively, to develop the calibration? Using all samples in a large spectral library will lead to large-scale, generic models. They encompass more variations in soil properties and MIR spectra, but also give higher prediction uncertainties. Conversely, a smaller subset will result in small-scale models applicable to local samples but maybe with better prediction accuracy.

The results of this study show that, if nonlinear modeling techniques such as ANN are used, good generic models involving all library samples could be developed. On the other hand, the method like PLSR can benefit from using auxiliary variables (such as HZ and TAXON) to select a subset of library samples for model calibration. However, these auxiliary variables may not be readily available for some parts of the world where field soil surveying programs are not well developed or established, rendering sample stratification impractical. Alternatively, spectral information itself can be used to select a subset of library samples to improve prediction accuracy (Ramirez-Lopez et al., 2013). Techniques such as spiking and extra-weighted spiking (Guerrero et al., 2014; Wetterlind and Stenberg, 2010) are also demonstrated to augment the spectral library and improve local predictions. It is not easy to predict what approach would work best. Many factors, such as size of the spectral library, availability of computational resources, target application and accuracy, and frequency of update with the arrival of new samples, would affect the approach being used. With the experience gained from this study, it is recommended to calibrate stratified models based on the HZ or TAXON to improve prediction if PLSR is used. However, if the computational demand can be met, nonlinear global models could provide improved accuracies.

CONCLUSIONS

The goal of this study is to predict an array of 12 soil physical and chemical properties from a national soil MIR spectral library comprising 20000+ samples from the United States. Two modeling techniques, namely, PLSR and ANN are employed and compared; and three auxiliary variables (HZ, TAXON and LULC) are used to explore the strategy of sample stratification for model improvement. The main conclusions drawn are as follows.

- Soil properties OC, IC, TC, TN, and TS can be predicted with MIR spectra satisfactorily, followed by CEC, pH, clay, silt, and sand. The prediction of soil K and P is poor.

- The ANN models generally outperform PLSR models, except for clay, silt, and sand.

- The use of auxiliary variables to develop stratified MIR models improves prediction performance for most soil properties and strata. Stratification appears more effective for PLSR than ANN.

- Among the three auxiliary variables, HZ and TAXON appear more effective as the stratifying criteria to improve MIR prediction on the soil properties than LULC.

- Stratifying the MIR library can be done on two or more auxiliary variables, which lead to even better model performance.

ACKNOWLEDGMENTS

Funding for this work was provided by USDA–NIFA–AFRI foundational program (Award No. 2017-67021-26248) and USDA–NRCS (Award No. 68-7482-14-517). The authors would also like to thank Richard Ferguson and Scarlet Bailey for their assistance in retrieving the samples and relevant datasets from National Soil Survey Center's databases.

SUPPLEMENTAL MATERIAL

Supplemental material is available with the online version of this manuscript. These materials contain the results of stratified modeling of 12 soil properties using auxiliary variables of master horizon, taxonomic order, and land use land cover.

REFERENCES

Ackerson, J.P., C.L.S. Morgan, and Y. Ge. 2017. Penetrometer-mounted VisNIR spectroscopy: Application of EPO-PLS to in situ VisNIR spectra. Geoderma 286:131–138. doi:10.1016/j.geoderma.2016.10.018

Askari, M.S., J. Cui, S.M. O'Rourke, and N.M. Holden. 2015. Evaluation of soil structural quality using VIS–NIR spectra. Soil Tillage Res. 146:108–117. doi:10.1016/j.still.2014.03.006

Beebe, K.R., and B.R. Kowalski. 1987. An introduction to multivariate calibration and analysis. Anal. Chem. 59:1007A–1017A. doi:10.1021/ac00144a725

Bellon-Maurel, V., E. Fernandez-Ahumada, B. Palagos, J.-M. Roger, and A.B. McBratney. 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC. Trends Analyt. Chem. 29:1073–1081. doi:10.1016/j.trac.2010.05.006

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

Bricklemyer, R.S., and D.J. Brown. 2010. On-the-go VisNIR: Potential and limitations for mapping soil clay and organic carbon. Comput. Electron. Agric. 70:209–216. doi:10.1016/j.compag.2009.10.006

Brown, D.J., K.D. Shepherd, M.G. Walsh, M. Dewayne Mays, and T.G. Reinsch. 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132:273–290. doi:10.1016/j.geoderma.2005.04.025

Dayhoff, J.E., and J.M. DeLeo. 2001. Artificial neural networks. Cancer 91:1615–1635.

Fry, J.A., G. Xian, S. Jin, J.A. Dewitz, C.G. Homer, Y. Limin, C.A. Barnes, N.D. Herold, and J.D. Wickham. 2011. Completion of the 2006 national land cover database for the conterminous United States. Photogramm. Eng. Remote Sensing 77:858–864.

Gallant, S.I. 1993. Neural network learning and expert systems. MIT press, Cambridge, MA.

Ge, Y., C.L.S. Morgan, and J.P. Ackerson. 2014a. VisNIR spectra of dried ground soils predict properties of soils scanned moist and intact. Geoderma 221-222:61–69. doi:10.1016/j.geoderma.2014.01.011

Ge, Y., J.A. Thomasson, C.L. Morgan, and S.W. Searcy. 2007. VNIR diffuse reflectance spectroscopy for agricultural soil property determination based on regression-kriging. Trans. ASABE 50:1081–1092. doi:10.13031/2013.23122

Ge, Y., J.A. Thomasson, and C.L.S. Morgan. 2014b. Mid-infrared attenuated total reflectance spectroscopy for soil carbon and particle size determination. Geoderma 213:57–63. doi:10.1016/j.geoderma.2013.07.017

Guerrero, C., B. Stenberg, J. Wetterlind, R.A. Viscarra Rossel, F.T. Maestre, A.M. Mouazen, R. Zornoza, J.D. Ruiz-Sinoga, and B. Kuang. 2014. Assessment of soil organic carbon at local scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking subset. Eur. J. Soil Sci. 65:248–263. doi:10.1111/ejss.12129

Hastie, T., R. Tibshirani, and J. Friedman. 2001. The elements of statistical learning. Springer-Verlag, New York. doi:10.1007/978-0-387-21606-5

Helland, I. 2004. Partial least squares regression In: S. Kotz, C.B. Read, N. Balakrishnan, B. Vidakovic, and N.L. Johnson, Encyclopedia of statistical sciences. John Wiley & Sons, New York. doi:10.1002/0471667196.ess6004

Henaka Arachchi, M.P.N.K., D.J. Field, and A.B. McBratney. 2016. Quantification of soil carbon from bulk soil samples to predict the aggregate-carbon fractions within using near- and mid-infrared spectroscopic techniques. Geoderma 267:207–214. doi:10.1016/j.geoderma.2015.12.030.

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

Kuhn, M. , J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, et al. 2015. caret: Classification and regression training. R Foundation for Statistical Computing, Vienna, Austria. https://CRAN.R-project.org/package=caret (Verified 20 Mar. 2018).

McCarty, G.W., J.B. Reeves, V.B. Reeves, R.F. Follett, and J.M. Kimble. 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Sci. Soc. Am. J. 66:640–646.

Mevik, B.-H., Wehrens, R., Liland, K.H., 2013. pls: Partial least squares and principal component regression.

Minasny, B., A.B. McBratney, V. Bellon-Maurel, J.-M. Roger, A. Gobrecht, L. Ferrand, and S. Joalland. 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. Geoderma 167–168:118–124. doi:10.1016/j.geoderma.2011.09.008

Mulder, V.L., M. Lacoste, A.C. Richer-de-Forges, M.P. Martin, and D. Arrouays. 2016. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263:16–34. doi:10.1016/j.geoderma.2015.08.035

Nguyen, T.T., L.J. Janik, and M. Raupach. 1991. Diffuse reflectance infrared fourier transform (DRIFT) spectroscopy in soil studies. Soil Res. 29:49–67. doi:10.1071/SR9910049

Nocita, M., A. Stevens, B. van Wesemael, M. Aitkenhead, M. Bachmann, B. Barthès, E. Ben Dor, D.J. Brown, M. Clairotte, and A. Csorba. 2015. Chapter four-soil spectroscopy: An alternative to wet chemistry for soil monitoring. Adv. Agron. 132:139–159. doi:10.1016/bs.agron.2015.02.002

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramirez-Lopez, L., T. Behrens, K. Schmidt, A. Stevens, J.A.M. Dematte, and T. Scholten. 2013. The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets. Geoderma 195-196:268–279. doi:10.1016/j.geoderma.2012.12.014

Reeves, J.B., III, R.F. Follett, G.W. McCarty, and J.M. Kimble. 2006. Can near or mid- infrared diffuse reflectance spectroscopy be used to determine soil carbon pools? Commun. Soil Sci. Plant Anal. 37:2307–2325. doi:10.1080/00103620600819461

Revolution Analytics, and S. Weston, 2015. doParallel: Foreach parallel adaptor for the 'parallel' package. CRAN, R Foundation for Statistical Computing, Vienna, Austria.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams. 1985. Learning internal representations by error propagation. DTIC Document. Defense Technical Information Center, Fort Belvoir, VA. doi:10.21236/ADA164453

Schmugge, T.J., W.P. Kustas, J.C. Ritchie, T.J. Jackson, and A. Rango. 2002. Remote sensing in hydrology. Adv. Water Resour. 25:1367–1385. doi:10.1016/S0309-1708(02)00065-9

Soil Survey Staff. 2014a. Soil Taxonomy: A basic system of soil classification for making and interpreting soil surveys. 12th ed. USDA–NRCS, Washington, DC.

Soil Survey Staff. 2014b. Kellogg soil survey laboratory methods manual. Soil Survey Investigations Report No. 42, Version 5.0. USDA–NRCS, Washington, DC.

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

Terhoeven-Urselmans, T., T.-G. Vagen, O. Spaargaren, and K.D. Shepherd. 2010. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. Soil Sci. Soc. Am. J. 74:1792–1799. doi:10.2136/sssaj2009.0218.

Venables, W.N., and B.D. Ripley. 2002. Modern applied statistics with S. 4th ed. Springer, New York. doi:10.1007/978-0-387-21706-2

Viscarra Rossel, R.A., V.I. Adamchuk, K.A. Sudduth, N.J. McKenzie, and C. Lobsey. 2011. Proximal soil sensing: An effective approach for soil measurements in space and time. Adv. Agron. 113:243–291. doi:10.1016/B978-0-12-386473-4.00005-1

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

Viscarra Rossel, R., and J. Bouma. 2016. Soil sensing: A new paradigm for agriculture. Agric. Syst. 148:71–74. doi:10.1016/j.agsy.2016.07.001

Viscarra Rossel, R.A., T. Behrens, E. Ben-Dor, D.J. Brown, J.A.M. Demattê, K.D. Shepherd, et al. 2016. A global spectral library to characterize the world's soil. Earth-Science Rev. 155:198–230. doi:10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R.A., Y.S. Jeon, I.O.A. Odeh, and A.B. McBratney. 2008. Using a legacy soil sample to develop a mid-IR spectral library. Aus. J. Soil Res. 46:1–16. doi:10.1071/SR07099

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

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

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

Wijewardane, N.K., Y. Ge, and C.L.S. Morgan. 2016a. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. Geoderma 267:92–101. doi:10.1016/j.geoderma.2015.12.014

Wijewardane, N.K., Y. Ge, S. Wills, and T. Loecke. 2016b. Prediction of soil carbon in the conterminous United States: Visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project. Soil Sci. Soc. Am. J. 80:973–982. doi:10.2136/sssaj2016.02.0052