


Evaluation of spectral pretreatments, spectral range, and regression methods for quantitative spectroscopic analysis of soil organic carbon composition

Hongzhang Kang, Huanhuan Gao & Wenjuan Yu

To cite this article: Hongzhang Kang, Huanhuan Gao & Wenjuan Yu (2017) Evaluation of spectral pretreatments, spectral range, and regression methods for quantitative spectroscopic analysis of soil organic carbon composition, Spectroscopy Letters, 50:3, 143-149, DOI: [10.1080/00387010.2017.1297956](https://doi.org/10.1080/00387010.2017.1297956)

To link to this article: <https://doi.org/10.1080/00387010.2017.1297956>

 View supplementary material 

 Accepted author version posted online: 09 Mar 2017.
Published online: 09 Mar 2017.

 Submit your article to this journal 

 Article views: 76

 View related articles 

 View Crossmark data 

Evaluation of spectral pretreatments, spectral range, and regression methods for quantitative spectroscopic analysis of soil organic carbon composition

Hongzhang Kang^a, Huanhuan Gao^a, and Wenjuan Yu^b

^aSchool of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China; ^bInstrumental Analysis Center, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Although there is an increasing interest in using infrared spectroscopy for the simple, rapid, and inexpensive prediction of soil organic carbon content, few studies have used this technique to measure organic carbon chemistry. In this paper, based on both near-infrared and mid-infrared diffuse reflectance spectroscopy, we compared the use of instrumentation, spectral pretreatment, and regression method for the prediction of three parameters related to organic carbon content, one related to isotopic composition, and five related to organic carbon chemistry. A total of 140 soil samples collected from seven oriental oak forest sites across East China were used as the data set for the calibration-validation procedure. Calibrations using sample set partitioning based on joint x-y distances method significantly outperformed those using Kennard-Stone method. Compared to models using linear method (i.e., partial least squares), those using non-linear regression method (i.e., support vector machines) greatly improved the prediction precision of the alkyl-to-O-alkyl ratio and performed slightly better for the other organic carbon chemical compositions. Instrumentation had a large effect as mid-infrared models had higher average prediction accuracies than near-infrared models. We finally proposed a model using second derivative preprocessing, joint x-y distances based sample set partitioning, mid-infrared spectra, and support vector machines regression to quantify organic carbon chemistry in this study. The results are helpful for the further study of soil composition measurement.

ARTICLE HISTORY

Received 21 September 2016
Accepted 18 February 2017

KEYWORDS

Infrared spectroscopy; partial least squares; soil organic carbon; support vector machines

Introduction

Soil organic carbon (SOC) is extremely important due to its large quantity stored in soil and its strong coupling to atmospheric CO₂.^[1] Although SOC is more commonly reported in terms of its gravimetric content (g C/g soil), knowledge on the chemical composition of SOC is also needed for estimates on C sequestration and decomposition processes.^[2] Nuclear magnetic resonance (NMR) spectroscopy is commonly used for characterizing SOC chemistry but is prohibitively expensive, time-consuming, and technically demanding.^[3] Stable isotope ratio mass spectrometry for measuring $\delta^{13}\text{C}$ value (an important indicator of SOC decomposition) has similar disadvantages. Thus, the use of the two techniques is currently beyond the reach of many labs.

Being able to quantify the nature of chemical bonds in soils, infrared (IR) spectroscopy offers a promising alternative as it is rapid, simple, and cost effective.^[4] As two commonly used infrared reflectance spectra involved in soil quantitative analysis, diffuse reflectance infrared Fourier transform spectroscopy (DRIFT) is usually used on dried/ground soils, and total attenuated reflectance spectroscopy (ATR) is applied to soil pastes.^[5,6] Many studies have used IR spectroscopy to quantitatively predict soil C content, and these findings have been reviewed.^[6,7] Nevertheless, only few studies have attempted to predict the percentages of four broad classes of NMR-derived carbon types using infrared spectra,^[2,3,8,9] and to our

knowledge, no study has utilized IR spectroscopy to quantify $\delta^{13}\text{C}$ in forest soils.

Over several decades, near-infrared (NIR) diffuse reflectance spectroscopy has been shown to be versatile for soil C determination. Yet some researches have demonstrated that for the analysis of soil C, mid-infrared (MIR) diffuse reflectance spectroscopy is often more accurate and produces more robust calibrations than NIR when analyzing ground, dry soils under laboratory conditions.^[7] Although both NIR^[2] and MIR^[3,8,9] have been used for predicting C species as revealed by NMR, none of these studies have made a direct comparison between the two instruments.

Since soil constituents interact in a complex way to produce a given spectrum, NIR and MIR spectra usually combine multivariate regression methods to develop models for measuring soil properties.^[10] There are two key points in the development of a prediction model: (i) spectral pretreatment and (ii) model development.^[11] Aiming at increasing variations in the signal, pretreatment of the spectral data before model development is viewed as a critical step, and a fine threshold needs to be found between removing noise and removing information. Only very few studies have referred to comparisons among different preprocessing methods.^[11] Model development is also of importance. The linear partial least square (PLS) has commonly been used in spectral calibration and validation; other nonlinear techniques (e.g., artificial neural

networks (ANN) and support vector machines (SVM)) have received much less attention. Although several studies have reported using SVM and ANN models for soil cation exchange capacity prediction^[12] and soil organic carbon measuring,^[13] there is no report of using SVM for the analysis of soil organic carbon composition as indicated by ¹³C NMR spectra.

In the present study, based on the DRIFT spectra of both NIR and MIR, we experimented with the use of various spectral pretreatment methods as well as linear and non-linear regression methods for the prediction of the following soil properties: (i) elemental composition: total carbon (C) content, total nitrogen (N) content, the C/N ratio; (ii) isotopic composition: $\delta^{13}\text{C}$ value; (iii) chemical composition: the proportions of alkyl C, O-alkyl C, aromatic C, carbonyl C, and the alkyl-to-alkyl (A/O) ratio. By making comprehensive comparisons from the aspects of instrumentation, regression method, and preprocessing method, we tried to determine an optimal IR model for predicting both SOC content and composition in oriental oak forest soils of East China.

Experimental

Soil sample collection

The sampling procedure has been described in detail by Yu et al.^[14] In July 2014, seven oriental oak stands in East China were sampled along a latitudinal gradient, extending from Beijing (40.25°N, 117.12°E) to Jiangxi (29.09°N, 115.62°E) (Supplementary Fig. 1). At each site, mineral soils at four depths (0–2, 2–5, 5–10, and 10–20 cm) were sampled in five parallel transects with 10 m spacing between them, resulting in a total of 140 soil samples (7 sites \times 5 replicates \times 4 depths). The samples covered a wide range of soil types, textures, and colors as well as climatic conditions and thus presented a harsh test on the applicability of IR for the prediction of

several soil characteristics from different sites and under different climatic conditions. Prior to analysis, mineral soils were sieved (2 mm), air-dried, smashed, and screened through mesh size of 80 (0.18 mm). Each sample was divided into two parts, one for laboratory chemical analyses and one for optical measurements.

Elemental, isotopic, and ¹³C-NMR analyses

The isotopic composition, total carbon, and nitrogen of these samples were measured on a Vario EL III Elemental Analyzer coupled to an Isotopic Ratio Mass Spectrometer (Elementar Analysensysteme GmbH, Germany). C isotopic results were expressed in the δ -notation, as the ‰ variation from the standard reference material, Pee Dee Belemnite (PDB).

Two of the five replicates for each depth and site were used for NMR analysis. Prior to NMR analysis, the total 56 mineral soil samples (7 sites \times 2 replicates \times 4 depths) were treated with hydrofluoric acid (HF) to concentrate the organic matter and remove paramagnetic minerals.^[15] The CP-MAS ¹³C NMR spectra were acquired on a Bruker Avance 400 MHz NMR spectrometer, equipped with a 4 mm broadband CP-MAS probe. The detailed procedures for sample pretreatment and spectra acquisition could refer to Yu et al.^[14] The spectra were integrated into the following chemical shift regions: alkyl carbon (0–50 ppm), O-alkyl carbon (50–110 ppm), aromatic carbon (110–160 ppm), and carbonyl carbon (160–200 ppm).^[16]

Infrared spectral reflectance measurements

NIR spectra were acquired using a Nicolet 6700 spectrometer (Thermo Fisher Scientific Inc., MA, USA) equipped with a white-light source, a CaF₂ beam-splitter, an InGaAs detector, and an NIR diffuse reflectance accessory (Pike Technologies,

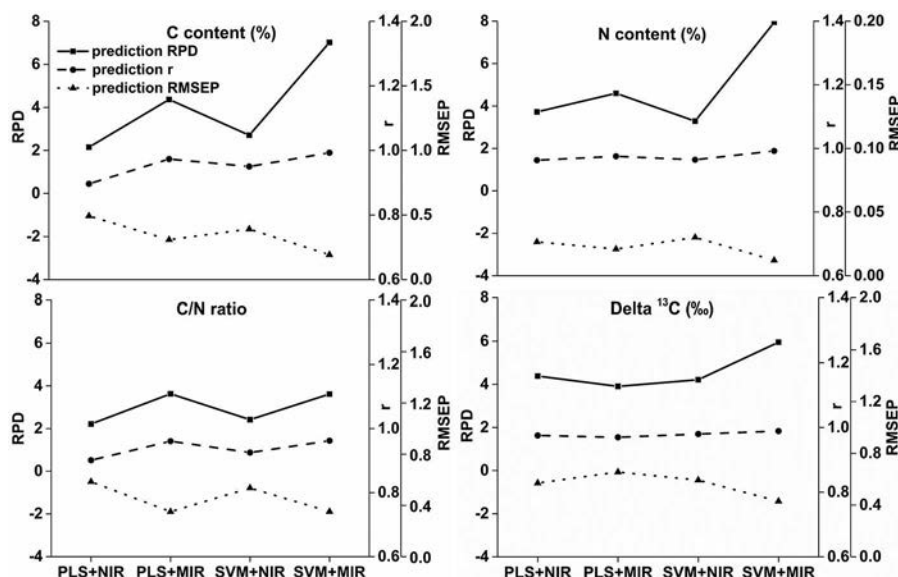


Figure 1. Comparison of performances of models developed with combinations of two instruments and two regression methods for total carbon (C) content, total nitrogen (N) content, the C/N ratio, and $\delta^{13}\text{C}$ value. Sample set partitioning based on x-y distances method was used. NIR, near-infrared spectra; MIR, mid-infrared spectra; PLS, partial least squares regression; SVM, support vector machines regression; r, correlation coefficient; RMSEP, root mean square error of prediction; RPD, ratio of performance to standard deviation.

WI, USA); spectra were acquired over 10000–4000 cm^{-1} with a resolution of 4 cm^{-1} . MIR diffuse reflectance spectra were also acquired using the Nicolet 6700 spectrometer, equipped with a KBr beam-splitter, a MCT detector, and a MIR diffuse reflectance accessory (Pike Technologies, WI, USA); spectra were acquired over 4000–650 cm^{-1} with a resolution of 4 cm^{-1} . Approximately 200 mg of ground, air-dried soil was placed into an 11 mm stainless steel cup, and the surface was smoothed. A total of 64 scans were acquired and averaged to produce a reflectance spectrum for each individual soil sample.

Chemometric analyses—spectral pretreatment and sample set partitioning

In order to decrease the noise and enhance possible spectral features linked to the property studied, the raw NIR and MIR reflectance spectra were corrected for effects of atmospheric CO_2 and water vapor, converted into absorbance spectra (log transform of the inverse of reflectance) and automatically baseline-corrected (obtained by polynomial fitting) by Omnic software Version 8.2 (Thermo Fisher Scientific, Waltham, MA, USA). The preprocessed spectra were then imported into Matlab R2009a (The MathWorks, Natick, MA, USA) and subjected to gap derivation;^[17] the segment length in data points over which the derivative was taken and the segment length over which the function was smoothed were both set to 7,^[18] and both first derivative and second derivative were tried. Multiplicative scatter correction (MSC) was also tried.^[19]

The total number of spectra (140 for elemental and isotopic analyses and 56 for NMR analyses) was divided into two sets of 75% and 25%. The former set was used to establish the calibration models, and the latter one was used for independent validation of the established models. The method for separating calibration and validation sets was also of importance, and we compared Kennard-Stone (KS) method^[20] and sample set partitioning based on joint x-y distances (SPXY) method.^[21] Four combinations of treatments were compared in our study: first derivative + SPXY, MSC + first derivative + SPXY, second derivative + SPXY, and first derivative + KS. Each pretreatment was then calibrated to SOC with a multivariate model.

Chemometric analyses—partial least-squares regression and prediction

Two different calibration techniques were used, namely, PLS and SVM. The linear PLS is commonly used for quantitative spectral analysis and projects predictors (X spectra) and response (Y laboratory data) into a low-dimensional space (i.e., a set of orthogonal variables called latent variables that maximize the covariance between X - and Y -scores).^[22] In order to avoid the problem of overfitting, a critical step in the algorithm is the determination of the appropriate number of latent variables. This was usually determined by minimizing the value of root mean square error of cross validation (RMSECV) by K -fold cross-validation.^[23] In our study, both the maximum number of latent variables and K were set to 20.

Chemometric analyses—support vector machines regression and prediction

SVM has been developed to solve nonlinear regression problems with a limited number of observations. In support vector regression, the input x is first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear model is constructed in this feature space. We selected the radial basis function (RBF) kernel due to computational convenience. Thus, the performance of SVM for regression depends on the combination of the following factors: regularization parameter C , ϵ of ϵ -insensitive loss function, and the width γ of the radial basis function. A systematic grid search was carried out to select the proper values for the C , ϵ , and γ ; the set of values with the best five-fold cross-validation performance is selected for further analysis.^[24] Matlab R2009a was used for all calculations. The PLS and SVM models were implemented with a toolbox developed by Hong-Dong Li et al.^[25] and a package developed by Chih-Chung Chang et al.,^[26] respectively. In both of two models, the data were mean-centered before model training to avoid inconsistent dimension.

Chemometric analyses—performance measures

Three performance criteria, correlation coefficient (r), root mean square error of prediction (RMSEP), and ratio of performance to standard deviation (RPD), were used in this study to assess the goodness of fit of the models. The r is a statistical measure of how well the predicted data is close to the observed data, and a perfect model will have a value of 1. The RMSEP can provide a balanced evaluation of the goodness of fit of the model as it is more sensitive to the larger relative errors caused by the low value, and the perfect model will have a value of 0. RPD is also commonly used to evaluate calibration accuracy. Viscarra Rossel et al. classified RPD values as follows: RPD < 1.0 indicates very poor predictions; RPD between 1.0 and 1.4 indicates poor predictions; RPD between 1.4 and 1.8 indicates fair predictions; RPD between 1.8 and 2.0 indicates good predictions where quantitative predictions are possible; RPD between 2.0 and 2.5 indicates very good predictions; and RPD > 2.5 indicates excellent predictions.^[27]

Results and discussion

Elemental and isotopic composition of SOC

Table 1 listed some statistical descriptions of soil carbon properties using regular laboratory analyses. Figure 1 presented the correlation coefficient (r), root mean square error of prediction (RMSEP), and ratio of performance to standard deviation (RPD) for total carbon (C) content, total nitrogen (N) content, the C/N ratio and $\delta^{13}\text{C}$ value when prediction models were developed with four combinations of instruments (NIR, MIR) and regression methods (PLS, SVM) by using the same preprocessing method (first derivative + SPXY). The results showed that four models all had good performances in predicting the above four parameters in mineral soils. The average r , RPD, and RMSEP of the four prediction models were 0.95, 4.06, and 0.35% for C, 0.97, 4.89, and 0.02% for N, 0.94, 2.97, and 0.46 for the C/N ratio and 0.98, 4.62, and

Table 1. Statistical description of the observed soil carbon properties using conventional laboratory analyses. Soils were collected from seven oriental oak forest sites along a 1500-km latitudinal gradient in East China.

	Soil properties	$n_{\text{calib}}/n_{\text{valid}}^a$	Mean	Range	Standard deviation
Elemental characteristics	C (%)	105/35	3.0	0.6 ~ 8.0	1.6
	N (%)	105/35	0.24	0.07 ~ 0.55	0.11
	C/N ratio	105/35	12.1	9.3 ~ 15.8	1.4
Isotopic characteristics	Delta ¹³ C (‰)	105/35	-23.6	-27.6 ~ -13.1	2.8
NMR characteristics	Alkyl C (%)	42/14	27.7	24.0 ~ 32.8	2.1
	O-alkyl C (%)	42/14	44.8	40.7 ~ 52.2	2.1
	Aromatic C (%)	42/14	15.0	12.8 ~ 18.6	1.3
	Carbonyl C (%)	42/14	12.5	10.6 ~ 14.9	0.9
	Alkyl/O-alkyl C	42/14	0.62	0.46 ~ 0.81	0.07

^a $n_{\text{calib}}/n_{\text{valid}}$ show the number of samples used in the spectral calibration and validation, respectively.

0.56‰ for $\delta^{13}\text{C}$. Infrared analysis is well suited for SOC analysis because of its sensitivity to the C-H, C-O, and C-N functional groups that dominate in organic matter,^[28] and this is most likely responsible for the good ability of infrared spectroscopy to quantify SOC content documented in our manuscript and many other literatures.^[29,30] Although IR spectra seem to contain no direct information related to ^{13}C due to their low sensitivity, a possible explanation of our successful $\delta^{13}\text{C}$ prediction is that the spectroscopy is based on the same changes in C composition that the $\delta^{13}\text{C}$ differences are used to determine or are based on.^[31] Specifically, the model combining MIR with SVM had the highest prediction accuracy (highest RPD and r , and lowest RMSEP) for all the four parameters (Fig. 1 and Table 2).

Chemical composition of SOC

Table 3 presented the r , RPD, and RMSEP for the five parameters related to SOC chemical composition when models were developed with 16 combinations of 4 different preprocessing methods, 2 instrumentations, and 2 regression methods. Compared to the above characteristics, chemical composition of SOC was more difficult to be accurately predicted by IR spectra, especially for carbonyl C, as indicated by lower r and RPD in some of the models listed in Table 3. In spite of the fact that the performance of a model was dependent on instrument, pretreatment, and regression, we tried to find some general trends by making the following comparisons and decide an optimal model for prediction of the chemical composition of SOC under oriental oak in this study.

Effect of spectral preprocessing methods on SOC chemical composition

The results in Table 3 showed that the effect of pretreatment on the prediction accuracy of SOC composition varied with

instrument and regression method. These results are consistent with the literatures.^[11,32] This situation makes the interpretation of the effect of each method difficult, since for a given parameter (i.e., carbonyl C), a method (i.e., first derivative + SPXY) working well in a specific model (i.e., PLS + MIR) did not perform well in others (i.e., SVM + NIR). However, MSC + first derivative + SPXY method generally performed better compared to the other three preprocessing methods (average values of the four models for $r = 0.84\text{--}0.92$, RMSEP = 0.41%–0.92% for the percentages of four NMR C groups and 0.02 for the alkyl-to-O-alkyl ratio, and RPD = 1.77–2.53), while first derivative + KS method usually gave the lowest prediction precision (average values of the four models for $r = 0.40\text{--}0.80$, RMSEP = 0.65%–1.50% for the proportions of four NMR C types and 0.02 for the alkyl-to-O-alkyl ratio, and RPD = 0.68–1.80). Our results suggested that sample set partitioning method played a key role in the prediction model, which have been largely ignored. The SPXY method employs a partitioning algorithm that takes into account the variability in both x and y -spaces, and the multidimensional space may be covered more effectively in comparison with partitioning schemes based on x -information alone (such as the Kennard–Stone (KS) algorithm) or random sampling (RS).^[21] This might be a possible reason why SPXY method was much more suitable for SOC chemical composition prediction than KS method in our study.

Effect of regression methods on SOC chemical composition

As far as we know, it is the first time that a non-linear regression method (i.e., SVM) has been used to predict the chemical composition of SOC as indicated by NMR; naturally, we compared the performances of traditional PLS models with SVM models. As demonstrated in Table 3, the choice of the regression method was also of importance. By averaging the

Table 2. Performance parameters for the predictions of various soil carbon attributes using mid-infrared spectra + support vector machines regression models. Soils were collected from seven oriental oak forest sites along a 1500-km latitudinal gradient in East China, and some basic soil properties could refer to Table 1. Sample set partitioning based on x - y distances method was used. C , regularization parameter; γ , width of the radial basis function; r , correlation coefficient; RMSEP, root mean square error of prediction; RPD, ratio of performance to standard deviation.

	Soil properties	Spectral pretreatment	Model C , γ	r	RMSEP	RPD
Basic characteristics	C (%)	1st derivative	2.00, 0.0010	0.99	0.19	7.02
	N (%)	1st derivative	2.00, 0.0010	0.99	0.01	7.95
	C/N ratio	1st derivative	5.66, 0.0010	0.96	0.35	3.62
Isotopic characteristics	Delta ¹³ C (‰)	1st derivative	4.00, 0.0010	0.99	0.43	5.96
NMR characteristics	Alkyl C (%)	2nd derivative	0.71, 0.0010	0.86	0.95	1.92
	O-alkyl C (%)	2nd derivative	1.00, 0.0014	0.92	0.45	2.62
	Aromatic C (%)	2nd derivative	4.00, 0.0010	0.93	0.43	2.67
	Carbonyl C (%)	2nd derivative	1.41, 0.0055	0.95	0.45	2.03
	Alkyl/O-alkyl C	2nd derivative	0.71, 0.0020	0.95	0.03	2.30

Table 3. Effects of preprocessing methods, instruments, and regression methods on the prediction of five parameters related to soil carbon chemical composition. Soils were collected from seven oriental oak forest sites along a 1500-km latitudinal gradient in East China, and some basic soil properties could refer to Table 1. P1, first derivative preprocessing + sample set partitioning based on joint x-y distances method; P2, multiplicative scatter correction followed by first derivative preprocessing + sample set partitioning based on joint x-y distances method; P3, second derivative preprocessing + sample set partitioning based on joint x-y distances method; P4, first derivative preprocessing + sample set partitioning based on Kennard-Stone method; NIR, near-infrared spectra; MIR, mid-infrared spectra; PLS, partial least squares regression; SVM, support vector machines regression; n, the number of latent variables in the partial least squares regression models; r, correlation coefficient; RMSEP, root mean square error of prediction; RPD, ratio of performance to standard deviation.

Model		Alkyl C (%)				O-alkyl C (%)				Aromatic C (%)				Carbonyl C (%)				Alkyl/O-alkyl C			
		n	r	RPD	RMSEP	n	r	RPD	RMSEP	n	r	RPD	RMSEP	n	r	RPD	RMSEP	n	r	RPD	RMSEP
Separate	NIR+P1 + PLS	7	0.90	1.75	1.06	7	0.75	1.07	0.94	7	0.93	2.51	0.45	8	0.59	0.97	0.74	8	0.68	1.22	0.03
	NIR+P2 + PLS	9	0.86	1.80	0.70	6	0.89	2.17	0.64	8	0.94	2.87	0.37	8	0.92	2.38	0.32	8	0.87	1.84	0.03
	NIR+P3 + PLS	6	0.59	1.15	0.86	5	0.78	1.57	0.86	6	0.80	1.58	0.62	7	0.37	0.80	0.68	4	0.76	1.31	0.03
	NIR+P4 + PLS	7	0.64	0.68	1.51	6	0.32	0.82	1.57	8	0.69	1.36	0.81	8	0.32	1.00	0.62	7	0.44	0.52	0.06
	MIR+P1 + PLS	6	0.83	1.69	1.15	10	0.87	1.35	0.89	7	0.92	1.85	0.62	9	0.88	2.22	0.44	7	0.66	1.35	0.03
	MIR+P2 + PLS	6	0.76	1.44	1.35	10	0.87	1.56	0.84	6	0.94	2.64	0.43	6	0.85	1.94	0.48	7	0.71	1.43	0.03
	MIR+P3 + PLS	8	0.93	2.12	0.85	7	0.76	1.40	0.85	7	0.93	2.26	0.51	8	0.91	2.43	0.37	6	0.88	2.18	0.03
	MIR+P4 + PLS	4	0.44	0.84	1.46	8	0.63	1.12	1.48	6	0.88	2.16	0.63	7	0.66	1.20	0.69	4	0.40	0.82	0.05
	NIR+P1 + SVM	/	0.91	2.02	0.91	/	0.76	1.51	0.67	/	0.91	2.26	0.50	/	0.64	1.30	0.55	/	0.84	1.83	0.02
	NIR+P2 + SVM	/	0.86	1.86	0.65	/	0.93	2.00	0.69	/	0.88	2.02	0.52	/	0.92	2.08	0.36	/	0.94	2.69	0.02
	NIR+P3 + SVM	/	0.74	1.23	0.81	/	0.74	1.43	0.94	/	0.83	1.67	0.59	/	0.34	0.99	0.64	/	0.93	1.89	0.02
	NIR+P4 + SVM	/	0.72	0.77	1.34	/	0.17	0.79	1.63	/	0.74	1.42	0.77	/	0.45	1.14	0.54	/	0.50	0.59	0.05
	MIR+P1 + SVM	/	0.88	1.98	0.98	/	0.88	2.18	0.55	/	0.90	2.07	0.56	/	0.93	2.22	0.44	/	0.83	1.79	0.02
	MIR+P2 + SVM	/	0.88	1.98	0.98	/	0.93	2.56	0.51	/	0.93	2.60	0.43	/	0.85	1.92	0.48	/	0.85	1.94	0.02
MIR+P3 + SVM	/	0.86	1.92	0.95	/	0.92	2.62	0.45	/	0.93	2.67	0.43	/	0.95	2.03	0.45	/	0.95	2.30	0.03	
MIR+P4 + SVM	/	0.52	0.84	1.45	/	0.72	1.25	1.33	/	0.90	2.26	0.60	/	0.52	1.12	0.74	/	0.25	0.78	0.05	
Preprocessing method	P1 ^a	/	0.88	1.86	1.03	/	0.81	1.53	0.76	/	0.92	2.17	0.53	/	0.76	1.68	0.54	/	0.75	1.55	0.02
	P2 ^a	/	0.84	1.77	0.92	/	0.90	2.07	0.67	/	0.92	2.53	0.44	/	0.89	2.08	0.41	/	0.84	1.98	0.02
	P3 ^a	/	0.78	1.60	0.87	/	0.80	1.75	0.78	/	0.87	2.05	0.54	/	0.64	1.56	0.54	/	0.88	1.92	0.02
	P4 ^a	/	0.58	0.78	1.44	/	0.46	1.00	1.50	/	0.80	1.80	0.70	/	0.49	1.12	0.65	/	0.40	0.68	0.05
Regression method	PLS ^b	/	0.81	1.66	0.99	/	0.82	1.52	0.83	/	0.91	2.29	0.50	/	0.75	1.79	0.50	/	0.76	1.55	0.03
	SVM ^b	/	0.85	1.83	0.88	/	0.86	2.05	0.63	/	0.90	2.22	0.51	/	0.77	1.76	0.49	/	0.89	2.07	0.02
Instrument	NIR ^c	/	0.81	1.63	0.83	/	0.81	1.62	0.79	/	0.88	2.15	0.51	/	0.63	1.42	0.55	/	0.84	1.79	0.02
	MIR ^c	/	0.86	1.85	1.04	/	0.87	1.94	0.68	/	0.93	2.35	0.50	/	0.90	2.13	0.44	/	0.81	1.83	0.03
Regression method and instrument	PLS+NIR ^d	/	0.78	1.57	0.87	/	0.81	1.60	0.81	/	0.89	2.32	0.48	/	0.62	1.38	0.58	/	0.77	1.45	0.03
	SVM+NIR ^d	/	0.83	1.70	0.79	/	0.81	1.65	0.77	/	0.87	1.98	0.54	/	0.63	1.46	0.52	/	0.90	2.14	0.02
	PLS+MIR ^d	/	0.84	1.75	1.12	/	0.83	1.43	0.86	/	0.93	2.25	0.52	/	0.88	2.20	0.43	/	0.75	1.65	0.03
	SVM+MIR ^d	/	0.87	1.96	0.97	/	0.91	2.45	0.50	/	0.92	2.45	0.47	/	0.91	2.06	0.45	/	0.88	2.01	0.02

^athe average results of two instruments combining two regression methods.

^bthe average results of three preprocessing methods (P1-P3) combining two instruments.

^cthe average results of three preprocessing methods (P1-P3) combining two regression methods.

^dthe average results of three preprocessing methods (P1-P3).

results of two instruments and three pretreatment methods (the results of first derivative + KS method were removed due to the poor performances), both PLS and SVM models had the ability to predict SOC composition within acceptable limits. SVM models were slightly better than PLS models, with r ranging from 0.77 to 0.90 for SVM vs. 0.75–0.91 for PLS, and RPD ranging from 1.76 to 2.22 for SVM vs. 1.52–2.29 for PLS. Specifically, the average prediction precision of the alkyl-to-O-alkyl ratio (an important index of SOC stabilization) was greatly improved by SVM (r = 0.89 and RPD = 2.07) compared to PLS (r = 0.76 and RPD = 1.55). Although no literature has compared the performances of PLS against SVM for prediction of SOC chemical composition, research has showed that nonlinear responses such as SVM perform better when compared to the traditional PLS for soil analysis.^[33] This might partly be due to the good abilities of SVM to generalize and to deal with sparse data.^[34]

Effect of instruments on SOC chemical composition

Still shown by Table 3, the effect of instruments on the prediction precision is larger than regression methods. For all parameters except the alkyl-to-O-alkyl ratio, MIR instrument gave significantly better results than NIR instrument. Averaging results of the six combinations of two regression

methods and three pretreatment methods, r varied from 0.81 to 0.93 for MIR and from 0.63 to 0.88 for NIR, and RPD ranged between 1.83 and 2.35 for MIR and between 1.42 and 2.15 for NIR. Specifically, the average prediction precision of the percentage of carbonyl C was dramatically improved by MIR (r = 0.90 and RPD = 2.13) compared to NIR (r = 0.63 and RPD = 1.42). For second derivative + SPXY method, MIR units gave better results than NIR units for all the five parameters. Many papers have demonstrated that for the determination of soil C content, MIR often produces more robust calibrations than NIR when analyzing ground, dry soils under laboratory conditions;^[7,28] MIR also performed better than NIR in predicting SOC chemical composition in our study. Peaks in the MIR are frequently better resolved and much more intense and contain better information related to SOC (such as alkyl C (wavenumbers 3000–2800 cm⁻¹), C-O-C from polysaccharides (1080–1060 cm⁻¹), aromatic carbonyl bands (1700 and 1510 cm⁻¹), and aromatic C=C stretching vibrations (at around 1610 cm⁻¹), and C=O from carboxylic acids, aldehydes, and ketones (1700–1640 cm⁻¹)), while frequencies in the NIR are generally overtones and combination bands from the fundamental vibrations occurring in the MIR.^[8,28,30,35,36] This might partly explain why MIR analysis substantially outperformed NIR when analyzing SOC.

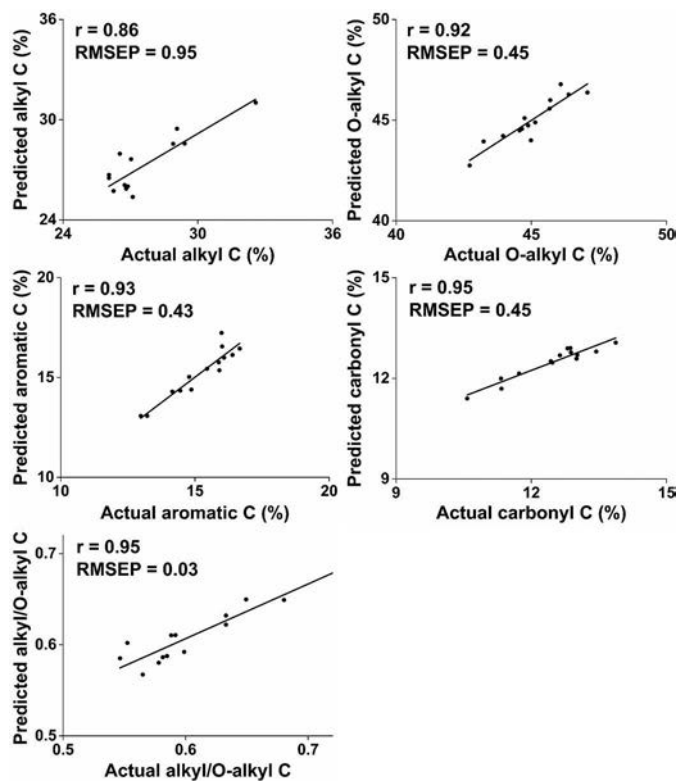


Figure 2. Predicted and actual values for the contents of alkyl C, O-alkyl C, aromatic C, carbonyl C, and alkyl/O-alkyl C using the mid-infrared spectra + second derivative preprocessing + sample set partitioning based on joint x-y distances method + support vector machines regression model. *r*, correlation coefficient; RMSEP, root mean square error of prediction.

Performances of mid-infrared spectra + support vector machines regression model for SOC chemical composition prediction

Given the above three comparisons, we tried to propose a best model for the five parameters related to SOC chemical composition in this study. By averaging the results of three pretreatment methods, SVM + MIR model showed the best predictive performance ($r = 0.87$ – 0.92 , $RMSEP = 0.45\%$ – 0.97% for the percentages of four functional groups and $RPD = 1.96$ – 2.45) (Table 3). Comprehensively, we chose second derivative + SPXY + MIR + SVM model to predict the composition of soil organic matter ($r = 0.86$, $RMSEP = 0.95\%$, and $RPD = 1.92$ for the percentage of alkyl C; $r = 0.92$, $RMSEP = 0.45\%$, and $RPD = 2.62$ for the percentage of O-alkyl C; $r = 0.93$, $RMSEP = 0.43\%$, and $RPD = 2.67$ for the percentage of aromatic C; $r = 0.95$, $RMSEP = 0.45\%$, and $RPD = 2.03$ for the percentage of carbonyl C; $r = 0.95$, $RMSEP = 0.03$, and $RPD = 2.30$ for the alkyl-to-O-alkyl ratio) (Table 2, Fig. 2). The first derivative + SPXY + MIR + SVM model also had the highest prediction precisions for SOC content and $\delta^{13}C$ (Fig. 1). Given the overall advantages of MIR over NIR and SVM over PLS, it was not surprising that models combining SVM with MIR outperformed the other three models in prediction of SOC content and composition.

Conclusion

Our results indicated that in addition to C content, N content, and the C/N ratio, the rarely reported $\delta^{13}C$ could also be

excellently predicted by both NIR and MIR spectra. We were the first to use a non-linear regression method (i.e., SVM) to predict SOC composition as indicated by ^{13}C NMR spectra. Our results showed that differences in IR spectral regions, pre-processing methods, and regression methods all have large impacts on final results, and only an iterative process can help in the development of the best models. Overall, FTIR data, especially when MIR data are developed with SVM algorithm, is a useful tool to predict both SOC content and composition in oriental oak forest ecosystems in East China. The results may provide some implications for the further study of soil composition measurement.

Funding

This work was financially supported by the National Natural Science Foundation of China (31270491), the National Key Research and Development Program of China (2016YFD0600206), SJTU Agri-X funding (Agri-X2015007) and SJTU SMC-C.

References

- [1] Trumbore, S. E.; Chadwick, O. A.; Amundson, R. Rapid exchange between soil carbon and atmospheric carbon dioxide driven by temperature change. *Science* **1996**, *272*, 393–396.
- [2] Terhoeven-Urselmans, T.; Michel, K.; Helfrich, M.; Flessa, H.; Ludwig, B. Near-infrared spectroscopy can predict the composition of organic matter in soil and litter. *Journal of Plant Nutrition and Soil Science-Zeitschrift Fur Pflanzenernahrung Und Bodenkunde* **2006**, *169*, 168–174.
- [3] Forouzangohar, M.; Baldock, J. A.; Smernik, R. J.; Hawke, B.; Bennett, L. T. Mid-infrared spectra predict nuclear magnetic resonance spectra of soil carbon. *Geoderma* **2015**, *247*, 65–72.
- [4] Mouazen, A. M.; Kuang, B.; De Baerdemaeker, J.; Ramon, H. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma* **2010**, *158*, 23–31.
- [5] Du, C.; Zhou, J. Evaluation of soil fertility using infrared spectroscopy: a review. *Environmental Chemistry Letters* **2009**, *7*, 97–113.
- [6] Bellon-Maurel, V.; McBratney, A. 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* **2011**, *43*, 1398–1410.
- [7] Reeves, J. B. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? *Geoderma* **2010**, *158*, 3–14.
- [8] Leifeld, J. Application of diffuse reflectance FT-IR spectroscopy and partial least-squares regression to predict NMR properties of soil organic matter. *European Journal of Soil Science* **2006**, *57*, 846–857.
- [9] Ludwig, B.; Nitschke, R.; Terhoeven-Urselmans, T.; Michel, K.; Flessa, H. Use of mid-infrared spectroscopy in the diffuse-reflectance mode for the prediction of the composition of organic matter in soil and litter. *Journal of Plant Nutrition and Soil Science-Zeitschrift Fur Pflanzenernahrung Und Bodenkunde* **2008**, *171*, 384–391.
- [10] Cecillon, L.; Barthes, B. G.; Gomez, C.; Ertlen, D.; Genot, V.; Hedde, M.; Stevens, A.; Brun, J. J. Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). *European Journal of Soil Science* **2009**, *60*, 770–784.
- [11] Igne, B.; Reeves, J. B.; McCarty, G.; Hively, W. D.; Lund, E.; Hurburgh, C. R. Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils. *Journal of near Infrared Spectroscopy* **2010**, *18*, 167–176.

- [12] Jafarzadeh, A. A.; Pal, M.; Servati, M.; FazeliFard, M. H.; Ghorbani, M. A. Comparative analysis of support vector machine and artificial neural network models for soil cation exchange capacity prediction. *International Journal of Environmental Science and Technology* **2016**, *13*, 87–96.
- [13] Stevens, A.; Udelhoven, T.; Denis, A.; Tychon, B.; Lioy, R.; Hoffmann, L.; van Wesemael, B. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* **2010**, *158*, 32–45.
- [14] Yu, W.; Fahey, T. J.; Kang, H.; Zhou, P. Surface soil organic carbon in temperate and subtropical oriental oak stands of East China. *Canadian Journal of Forest Research* **2016**, *46*, 621–628.
- [15] Simpson, M. J.; Simpson, A. J. The chemical ecology of soil organic matter molecular constituents. *Journal of Chemical Ecology* **2012**, *38*, 768–784.
- [16] Clemente, J. S.; Gregorich, E. G.; Simpson, A. J.; Kumar, R.; Courtier-Murias, D.; Simpson, M. J. Comparison of nuclear magnetic resonance methods for the analysis of organic matter composition from soil density and particle fractions. *Environmental Chemistry* **2012**, *9*, 97–107.
- [17] Norris, K. H.; Williams, P. C. Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size. *Cereal Chemistry* **1984**, *61*, 158–165.
- [18] Ludwig, B.; Khanna, P. K.; Bauhus, J.; Hopmans, P. Near infrared spectroscopy of forest soils to determine chemical and biological properties related to soil sustainability. *Forest Ecology and Management* **2002**, *171*, 121–132.
- [19] Geladi, P.; Macdougall, D.; Martens, H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy* **1985**, *39*, 491–500.
- [20] Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **2012**, *11*, 137–148.
- [21] Galvao, R. K. H.; Araujo, M. C. U.; Jose, G. E.; Pontes, M. J. C.; Silva, E. C.; Saldanha, T. C. B. A method for calibration and validation subset partitioning. *Talanta* **2005**, *67*, 736–740.
- [22] Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109–130.
- [23] Liu, Q. Z.; Sung, A. H.; Chen, Z. X.; Liu, J. Z.; Huang, X. D.; Deng, Y. P. Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. *Plos One* **2009**, *4*, e8250.
- [24] Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1257–1266.
- [25] Li, H.; Xu, Q.; Liang, Y. libPLS: An integrated library for partial least squares regression and discriminant analysis. *Peerj Preprints* **2014**, *2*, e190v1.
- [26] Chang, C. C.; Lin, C. J. LIBSVM: A library for support vector machines. *Acm Transactions on Intelligent Systems and Technology* **2011**, *2*, 27.
- [27] Viscarra Rossel, R. A.; Mcglynn, R. N.; Mcbratney, A. B. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. *Geoderma* **2006**, *137*, 70–82.
- [28] Soriano-Disla, J. M.; Janik, L. J.; Viscarra Rossel, R. A.; Macdonald, L. M.; McLaughlin, M. J. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews* **2014**, *49*, 139–186.
- [29] Viscarra Rossel, R. A.; Walvoort, D. J.J.; McBratney, A. B.; Janik, L. J.; Skjemstad, J. O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006**, *131*, 59–75.
- [30] McCarty, G. W.; Reeves, J. B.; Reeves, V. B.; Follett, R. F.; Kimble, J. M. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Science Society of America Journal* **2002**, *66*, 640–646.
- [31] Reeves Iii, J. B.; Follett, R. F.; McCarty, G. W.; Kimble, J. M. Can Near or Midinfrared and near-infrared diffuse reflectance used to Determine Soil Carbon Pools? *Communications in Soil Science and Plant Analysis* **2006**, *37*, 2307–2325.
- [32] Fernandez-Cabanas, V. M.; Garrido-Varo, A.; Perez-Marin, D.; Dardenne, P. Evaluation of pretreatment strategies for near-infrared spectroscopy calibration development of unground and ground compound feedingstuffs. *Applied Spectroscopy* **2006**, *60*, 17–23.
- [33] Viscarra Rossel, R. A.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54.
- [34] Venkoba Rao, B.; Gopalakrishna, S. J. Hardgrove grindability index prediction using support vector regression. *International Journal of Mineral Processing* **2009**, *91*, 55–59.
- [35] Grinand, C.; Barthes, B. G.; Brunet, D.; Kouakoua, E.; Arrouays, D.; Jolivet, C.; Caria, G.; Bernoux, M. Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). *European Journal of Soil Science* **2012**, *63*, 141–151.
- [36] Margenot, A. J.; Calderón, F. J.; Bowles, T. M.; Parikh, S. J.; Jackson, L. E. Soil organic matter functional group composition in relation to organic carbon, nitrogen, and phosphorus fractions in organically managed tomato fields. *Soil Science Society of America Journal* **2015**, *79*, 772–782.