Machine-learning-based quantitative estimation of soil organic carbon content by VIS/NIR spectroscopy

Jianli Ding[1,2], Aixia Yang[1,3], Jingzhe Wang[1,2], Vasit Sagan[4] and Danlin Yu[5,6]

[1] Key Laboratory of Smart City and Environment Modelling of Higher Education Institute, College of Resources and Environment Sciences, Xinjiang University, Urumqi, China
[2] Key Laboratory of Oasis Ecology, Xinjiang University, Urumqi, China
[3] College of Resources and Environment Science, Qinzhou University, Qinzhou, China
[4] Department of Earth and Atmospheric Sciences, Saint Louis University, St. Louis, MO, United States of America
[5] Department of Earth and Environmental Studies, Montclair State University, Montclair, NJ, United States of America
[6] School of Sociology and Population Studies, Renmin University of China, Beijing, China

ABSTRACT

Soil organic carbon (SOC) is an important soil property that has profound impact on soil quality and plant growth. With 140 soil samples collected from Ebinur Lake Wetland National Nature Reserve, Xinjiang Uyghur Autonomous Region of China, this research evaluated the feasibility of visible/near infrared (VIS/NIR) spectroscopy data (350–2,500 nm) and simulated EO-1 Hyperion data to estimate SOC in arid wetland regions. Three machine learning algorithms including Ant Colony Optimization-interval Partial Least Squares (ACO-iPLS), Recursive Feature Elimination-Support Vector Machine (RF-SVM), and Random Forest (RF) were employed to select spectral features and further estimate SOC. Results indicated that the feature wavelengths pertaining to SOC were mainly within the ranges of 745–910 nm and 1,911–2,254 nm. The combination of RF-SVM and first derivative pre-processing produced the highest estimation accuracy with the optimal values of R_t (correlation coefficient of testing set), $RMSE_t$ and RPD of 0.91, 0.27% and 2.41, respectively. The simulated EO-1 Hyperion data combined with Support Vector Machine (SVM) based recursive feature elimination algorithm produced the most accurate estimate of SOC content. For the testing set, R_t was 0.79, $RMSE_t$ was 0.19%, and RPD was 1.61. This practice provides an efficient, low-cost approach with potentially high accuracy to estimate SOC contents and hence supports better management and protection strategies for desert wetland ecosystems.

Subjects Ecosystem Science, Soil Science, Natural Resource Management, Environmental Impacts, Spatial and Geographic Information Science
Keywords Ebinur lake wetland, Desert wetland soil, Soil organic carbon, Machine learning

Submitted 15 May 2018
Accepted 10 September 2018
Published 17 October 2018

Corresponding author
Danlin Yu, yud@mail.montclair.edu

Academic editor
Richard Becker

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.5714**

INTRODUCTION

Wetlands account for a significant portion of global carbon stocks (*Hu et al., 2010*). According to the United Nations Environment Programme's (UNEP) World Conservation Monitoring Centre, the total area of wetlands is about 6% of the total land area globally. Carbon stocks within the wetlands accounts for 14% of the entire land ecosystems (*Foley*

et al., 2005). Due to its high carbon storage, any slight change in wetland carbon stocks might result in significant effect on global climate change (*Wang, Zhang & Haimiti, 2015*). For example, changes in wetland carbon stocks can increase carbon dioxide concentration and methane in the atmosphere, which might lead to more severe global warming (*Pott & Pott, 2004*).

Wetlands in arid and semi-arid regions play an important role as ecological barrier in desert ecosystems. Unlike wetlands in wet regions, arid wetlands are highly sensitive to human activities, and the restoration and rehabilitation of them often are extremely hard once degraded (*Zhao et al., 2009*). Therefore, inland wetlands in arid regions are small but important component that cannot be ignored, especially in global carbon cycle and balance of atmospheric greenhouse gases (GHG) studies (*Cole et al., 2007*; *Liu et al., 2010*). Over the past decades, environmental variables (e.g., mean annual precipitation and temperature), soil characteristics including texture and lithology, and the increasingly intensive human activities, like water diversion, reclamation, overgrazing, pollutant emissions have profoundly changed the arid region's wetland distribution and therefore the balance of carbon budget (*Ding & Yu, 2014*; *Thakur et al., 2012*). It is imperative to develop efficient, fast, and relatively accurate approaches to detect, monitor and predict soil organic carbon (SOC) content over large areas in arid regions (*Jaber & Al-Qinna, 2011*; *Stevens et al., 2010*; *Vohland et al., 2011*).

Traditional approaches measuring SOC content employ typical soil chemical analysis methods, mainly including dry combustion techniques (*Craft, Seneca & Broome, 1991*), chemical oxidation method (*West & Post, 2002*) and acid solution extraction (*Polglase, Jokela & Comerford, 1992*). Though those traditional approaches are relatively accurate and widely accepted, they require extensive lab work and often destroy the samples during processing, which renders repeating the lab work nearly impossible. On the other hand, recent studies have reported that the Visible/Near Infrared (VIS/NIR) spectroscopy is a rapid, cost-effective, quantitative and non-destructive technique which provides spectral information with large amount of data to monitor and detect soil quality and chemical components (*Kinoshita et al., 2012*). Due to the large amount of spectral information, scholars have attempted to establish a wide range of empirical models to seek in-depth understanding between various soil chemicals (such as organic phosphorous, organic carbon, among many others) and the reflective spectra obtained from the spectroscopy analysis (*Viscarra Rossel et al., 2006*). Among them, machine learning algorithms with their capability to relatively quickly and accurately analyze large amount of data, stand out to provide excellent opportunities for taking advantage of the spectral information (*Kuang, Tekin & Mouazen, 2015*; *Nawar & Mouazen, 2017*). The analysis can be done both within the laboratory and in the field.

SOC was one of most important controlling components of soil spectral features. With SOC content of 2% as a boundary, that is, when SOM content exceeded 2%, the SOC played a principal role in masking out the spectral features, while the SOC content was less than 2%, it became less effective (*Wang et al., 2017*). Many studies have used VIS/NIR spectroscopy to study, estimate and monitor SOC content, but mainly on Alfisols, Entisols, Ultisols (*Chang et al., 2001*; *Summers et al., 2011*; *Vasques, Grunwald & Harris,*

2010), and Mollisols (Araújo et al., 2015; Hong et al., 2018). In general, the conventional regression methods were sufficient for the spectral detection of soil types with higher SOC content. The composition, structure and sedimentary environment of wetland soils were extreme complex, especially in arid regions (Kayranli et al., 2010). Therefore, very few were conducted on wetland soils in arid regions. In addition, most of the empirical approaches focus primarily on applying multivariate linear regression, partial least square regression, or regression kriging to establish relationships between spectra and SOC (Dai et al., 2014; Guo et al., 2015; Liu, Zhang & Zhang, 2008; St. Luce et al., 2014). These approaches often suffer from autocorrelation, nonlinearity, or in some cases overestimation (Wang et al., 2017). On the other hand, machine learning approaches in recent years have gained momentum due to their relative flexibility in adapting (learning) the data structure prior to making any sensible prediction or simulation (McDowell et al., 2012; Peng et al., 2014; Viscarra Rossel & Behrens, 2010; Were et al., 2015).

Applications of various machine learning approaches, such as Support Vector Machine (SVM) and random forest (RF), have been attempted (Meng & Dennison, 2015; Nauman, Thompson & Rasmussen, 2014; Shi et al., 2013). SVM is a powerful calibration method based on the kernel learning methods, it could offer a possibility to train generalizable, nonlinear classifiers in high dimensional spaces using a small training set is that it offers a possibility to train generalizable, nonlinear classifiers in high dimensional spaces using a small training set (Mountrakis, Im & Ogole, 2011; Vapnik, 1999). Differing from existing linear and non-linear regression modeling methods, RF has acceptable predicting performance even if most independent variables are noise (Svetnik et al., 2003; Wang et al., 2018). For instance, Viscarra Rossel & Behrens (2010) and Were et al. (2015) collected soil samples from Kenya and Australia, and applied NIR spectroscopic analyses on the samples. Their study found that both support vector machine and random forest provide reasonably good estimation for SOC. Peng et al. (2014) also applied VIS/NIR spectroscopy with SVM to estimate SOC contents with samples from the middle and lower reaches of the Yangtze River, China. All these researches reported promising results in combining both machine learning approaches and VIS/NIR spectroscopic analysis. Studies on SOC measurement in arid wetlands predominantly employed traditional chemical analytical approaches (Anne et al., 2014; Cohen, Prenger & DeBusk, 2005; Wang et al., 2016; Wang, Zhang & Haimiti, 2015) and the use of machine learning algorithms with spectroscopy are limited.

At present, more machine learning algorithms have been proposed and used for variable selection, e.g., interval partial least squares (iPLS), ant colony optimization (ACO). The general success of combining VIS/NIR spectroscopy analysis and machine learning approaches in other regions calls for in-depth investigation on applying similar approaches in wetlands of arid regions. The combination of two algorithms could maximize the superiority of single method and overcome some faults, to a certain extent. In terms of the Ant Colony Optimization-interval Partial Least Squares regression (ACO-iPLS), it could exhibit certain advantage in distributed parallel calculation, information positive feedback and heuristic search ability (Huang et al., 2014; Zhu et al., 2018). The Support Vector Machine Recursive Feature Elimination (RF-SVM), an intelligent optimization method has demonstrated its outstanding performance and great potential for development in

solving many complex optimization problems (*Lin et al., 2011*). EO-1 Hyperion data with 242 spectral bands was the first satellite borne hyperspectral imaging spectrometer, which has been widely used in soil science, agricultural science, geological mapping, and accurate mapping (*Liu et al., 2009*). Due to the limitation of atmospheric influence and sensor observation conditions, it was difficult to obtain enough hyperspectral remote sensing imageries to meet the needs of different research fields. Nevertheless, the simulation of Hyperion data based on its spectral response function could serve as the effective substitute for research, when there were no available data. Hence, it is worthwhile to consider these coupling algorithms (RF-SVM and ACO-iPLS) and simulated remote sensed imageries as tools to develop estimating models in soil science.

Given these backgrounds and motivated by previous research, this study aims to combine VIS/NIR spectroscopy and machine learning approaches to quantitatively estimate SOC content of wetlands in arid regions. The study area, Ebinur Lake Wetland National Nature Reserve, located in the arid Northwestern China. The arid wetland ecosystem is very fragile due to the specific climatic conditions. Regional SOC is more sensitive to the climatic changes and human activities than in other areas. In addition, since the study area is located along the Silk Road Economic Belt, its ecological stability has profound significance on the sustainability of local economies as well as the entire Economic Belt (*Tan et al., 2018*; *Wang et al., 2018*; *Xu et al., 2017*). We collected 140 soil samples at various depth in 2012 and analyzed the samples with chemical analysis. In the meantime, we also obtained the spectra information through VIS/NIR spectroscopy. We choose three efficient machine learning approaches, namely, the ACO-iPLS, RF-SVM, RF, to extract feature wavelengths from the spectral data and simulated EO-1 Hyperion data, and further construct the adequately stable and reliable models for the SOC content in the arid wetland regions.

MATERIAL AND METHODS

Study area

Ebinur Lake wetland is a typical arid region lake wetland in Xinjiang Uyghur Autonomous Region, China (44°30′∼45°09′N, 82°36′∼83°50′E). The study area is a combination of lake, river and swamp wetlands, which is ideal for studying SOC contents in arid region (*Li, Zhao & Li, 2018*). The wetland is located in the northern slopes of the Tianshan Mountains, southwest of the Junggar Basin. The area is surrounded by mountains in the south, west and north, but connect to the Mutetaer desert in the east (Fig. 1). It is a designated eco-protection region, with a land area of 2670.85 km^2. The climate is typical temperate arid continental climate with limited annual precipitation (90.9 mm), but very high evaporation (3,400 mm). The annual average temperature is 8.3 °C (*Abuduwailil, Zhaoyong & Fengqing, 2015*). According to the World Reference Base for Soil Resources (WRB), local prevalent soil types are mainly Arenosols, Solonetz, and Solonchaks (*He et al., 2015*; *Wang et al., 2018*). The existence of various typical arid region soil types provides a good opportunity to test the proposed machine learning algorithms' effectiveness to monitor and evaluate SOC content.

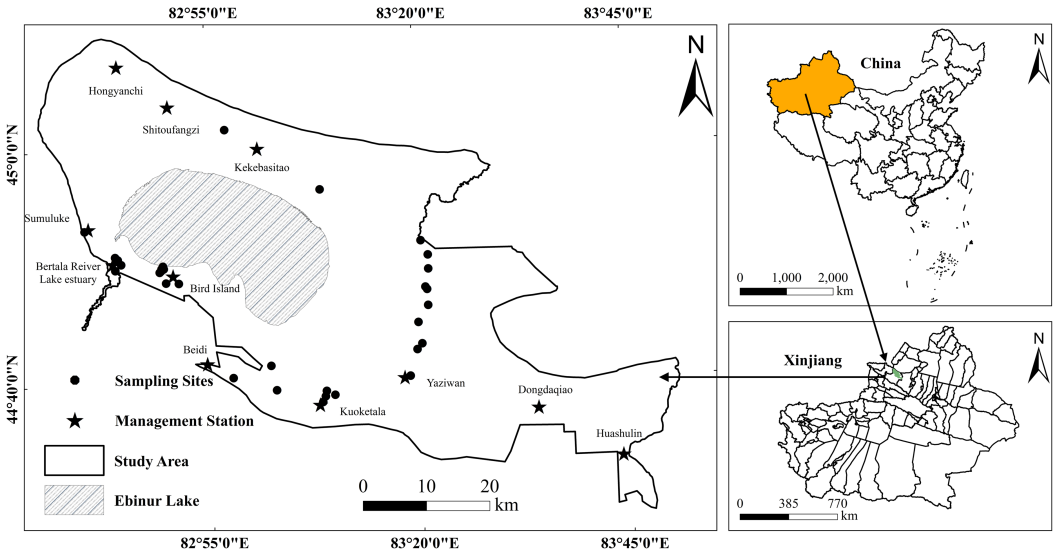

Figure 1 Study area and locations of sampling points. Vectorization by Jingzhe Wang.
Full-size DOI: 10.7717/peerj.5714/fig-1

Soil collection and chemical analysis

The soil samples were collected from a field trip to the Ebinur Lake Wetland National Nature Reserve in October 2012. The sampling sites were previously established for various soil properties monitoring purposes. They are located around relatively accessible locations in Kekebasitao Management Station, Yaziwan Management Station, Beidi Management Station, Bird Island, Bortala River Lake estuary, and the lower reaches of Kuitun River. There were in total 35 sampling sites (Fig. 1). At each sampling site, samples were collected at four vertical depths (5 cm, 20 cm, 40 cm, and 60 cm) and five evenly distributed points with a grid of 30×30 m (because the spatial resolution of Hyperion imagery is 30 m). The samples for the five points (at each depth) were then mixed evenly to represent the soil for that sampling site (at the specific depth). A total of (4×35) samples were collected and then brought to the laboratory for chemical measurements. All soil samples ($n = 140$) were sufficiently air-dried, ground and sieved through a two mm mesh to remove plant materials, residues, roots and stones. The potassium dichromate method was employed for the measurement of SOC content.

Spectral measurements and pre-processing

The reflectance spectra of all soil samples were measured in the laboratory via an ASD FieldSpec®3 portable spectroradiometer (Analytical Spectral Devices, Inc., St, Boulder, CO, USA) with a wavelength range of 350–2,500 nm. The spectral readings were interpolated to a 1 nm interval. Using recommendation by *Zhou et al. (2005)*, the spectra of all soil samples were measured in a dark room with a 50-W halogen lamp as the light source, which was positioned 0.5 m away from the soil sample, with a 25° zenith angle. The soil samples were put in a 12 cm diameter with 1.8 cm depth container evenly. The optical probe was installed about 0.1 m above the soil sample (*Shi et al., 2014*). Prior to the first scan, a standardized white Spectralon® panel with 99% reflectance was used to convert

radiance to reflectance. To eliminate random reflectance errors, 10 spectral measurements for each sample were taken and the average of these measurements was used as the final spectral reflectance.

High frequency random noises, baseline drifts, and scattering noises could affect spectral measurements. To remove the influence of these noises, Savitaky-Golay (SG) smoothing was implemented with a window size of 5 and polynomial order of 2 via Origin Pro software version 9.0. In general, the transformation of first order derivative was used for the enhancement of the spectral characteristics (*Savitzky & Golay, 1964*). In this study, the SG preprocessed spectral data was transformed into first order derivative (A′), the inversion of the first order derivative (1/A′), and logarithm transformation of the first order derivative (lg(A′)). In spectral analysis, they are effective pretreatments, which could eliminate the background noise to a certain extent, and enhance the spectral characteristics (*Wang et al., 2017*).

Model calibration, evaluation, and comparison

Considering the Euclidean distance of each sample, all 140 soil samples were separated into two equal parts (training set and testing set) using the Kennard-Stone (K-S) algorithm. Each set consists of 70 samples. To investigate the feasibility of using VIS/NIR to predict SOC content and select the most effective pre-processing methods, three machine learning approaches, i.e., Ant Colony Optimization-interval Partial Least Square (ACO-iPLS), Recursive Feature Elimination based on Support Vector Machine (RF-SVM), and Random Forest (RF), were applied for the reduction of inefficacious information and model construction.

ACO-iPLS

ACO-iPLS approach is a combination between principal component analysis based PLS and the meta heuristic optimization Ant Colony (ACO). PLS has been proven a robust and reliable approach in spectral quantitative research, primarily because of its advantages regarding dimension reduction and the synthesis and solving of multi-collinearity problems among independent variables. ACO, on the other hand, is an optimization algorithm that originates from the observation of the ants' food-seeking behavior in which each ant will leave certain amount of pheromone on the route to the food. A colony of ants would leave enough amount of pheromone to guide the colony to follow the optimal route (with the largest amount of pheromone) to the food. A combination of ACO and PLS algorithms seems to produce fairly useful information to select the most informative spectra or segments of spectra (*Huang et al., 2014*). Detailed steps are as follows:

(1) Initialization: set the size of the colony (k), for the m segments of spectra, the initial pheromones τ_i are all set to be 1.

$$\tau_i = 1, \quad i = 1, 2, \ldots, m. \tag{1}$$

(2) Determine the probability that one segment will be selected in the "route". The selection is done with the Roulette wheel method, namely, suppose that the time t, segment

i's pheromone is $\tau_i(t)$, then the probability that segment i will be selected is:

$$P_i(t) = \frac{\tau_i(t)}{\sum_{i=1}^{m} \tau_i(t)}, \quad i = 1, 2, \ldots, m. \tag{2}$$

(3) Target function: the prediction accuracy of a PLS model will be used as the target function, specifically, the inversion of the root mean squared error (RMSE) is used here:

$$F = Q/(1 + \mathrm{RMSE}) = \frac{Q}{1 + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}}. \tag{3}$$

where n is the number of samples, y_i is the actual value (SOC content in this study), and $\hat{y}_i$ is the predicted value, Q is a constant that represents the significant factor (significance level). Smaller RMSE indicates a better model.

(4) Update: assume $\rho \in (0, 1)$ is the information decaying rate (which can be selected based on empirical studies), then the pheromone of segment i will update as:

$$\tau_i(t+1) = (1 - \rho) \times \tau_i(t) + \rho \times F. \tag{4}$$

Following these simple steps, we will iterate steps 2–4. Once the iteration reaches certain amount, the algorithm will produce an optimal set of segments that contains the largest amount of pheromone (information), which also produces the least overall RMSE values.

The SOC contents were regarded as y_i. The soil spectra were divided in segments as the different number of routes. The initial number of ants was set to 50, maximum recursive attempts for the PLS to be 50, and maximum iteration to be 20. Based on the minimum RMSE, ρ was set to 0.53, and Q to be 0.01. The threshold of segment was set to 0.3. For the PLS model, the initial number of segments of spectra was set to 15.

RF-SVM

SVM is a machine learning algorithm that has recently attracted quite some attention in dealing with large amount of data. The basic principle of SVM is to use a kernel function that can maximally separate the different classes (*González Costa et al., 2017*; *Thissen et al., 2004*). By using non-linear kernels, SVM can essentially map the seemingly inseparable data points to higher dimension, to find the inherent structure of the data. Detailed steps follow:

(1) Set the training samples $X_0 = [x_1, x_2, x_3 \ldots x_i, \ldots, x_n]^T$, and the corresponding class labels: $Y = [y_1, y_2, y_3, \ldots y_i, \ldots, y_n]^T$. Each x_i $(i = 1, \ldots, n)$ is a vector containing the spectral information obtained from the VIS/NIR spectroscopic analysis. Each y_i is a measured SOC from chemical analysis.

(2) Initialize the feature subset vector $s = [1, 2, 3, \ldots, k]$. The initial k is the total number of wavelengths (k = 2,151, from 350 nm to 2,500 nm).

(3) Start the iteration: obtain a new training sample based on the remaining features: $X = X_0$ (:, s). The initial training sample includes all 2,151 wavelengths (features).

(4) Use the training sample in SVM and obtain the weight vector (w) for all the features.

(5) Set the ordering rule: $c = w^2$ (to consider only the magnitude of the weights instead of their signs).

(6) Eliminate the feature that has the lowest ordering score C, and then update the training sample.

(7) Repeat steps (3)–(6) until the set s is an empty set to obtain the final result with each feature (wavelength) being ordered.

During each iteration of the RF-SVM algorithm, the lowest ordered feature will be eliminated first, and the remaining features then be re-trained to obtain a new weight vector for next round of feature ordering. For individual features, RF-SVM might not produce optimal results. For groups of features, however, RF-SVM has the potential to produce the best supplementary combinations of features. In practice, the SOC contents obtained from chemical analysis were used as the class labels Y. The spectral information was used as the feature input for the support vector machine X_0. The algorithm was implemented with the libsvm-3.1 [FarutoUltimate3.1 Mcode] package based on MATLAB® software version R2015a (MathWorks, Inc., Natick, MA, USA). The optimization was done with a two-dimensional grid searching. With multiple trainings and experiments, we finally decided to use the epsilon-SVM model. The kernel function was selected as Sigmoid. The Gamma value and Eps were set to 0.0039 and 0.01, respectively, and the tuning parameter C was 1.

Random forest

Random forest is an ensemble learning technique developed by *Breiman (2001)* to improve the regression trees method (*Mutanga, Adam & Cho, 2012*). In RF regression, the procedure uses bootstrapping to repetitively generate subsamples from a training dataset and train a tree from each subsample. Averaging through these trees results in large gain in reduced variance. As in the previous section, predictors x represents the spectral data, while y represents the actual SOC content in this study, the Abhisheck Jaiantila's randomforest-matlab package was applied for the implement of random forest learning (*Liaw & Wiener, 2002*). In the present study, m, the number of sub-samples of the predictors was set to 500.

Model evaluation and comparison

All the above three algorithms were validated via 10 folds cross-validation. The cross-validation correlation coefficient (R_{cv}), and root mean square error of the cross validation (RMSE_{cv}) were used to optimize all the model parameters. Precision indices of R_{cv}, RMSE_{cv}, and residual prediction deviation (RPD) were also used to evaluate the performance of these algorithms. Higher R_{cv} and lower RMSE_{cv} indicate a more stable model. Similarly, for the testing set, higher the R_t, lower the RMSE_t suggest better performance. The RPD was used to assess the accuracy of the algorithm. If $\mathrm{RPD} \geq 2$, then the algorithm has very reliable prediction accuracy. If $1.4 \leq \mathrm{RPD} < 2$, then the prediction accuracy is acceptable. Only if $\mathrm{RPD} < 1.4$, the prediction accuracy is not acceptable (*Chang et al., 2001*). In addition, the deviation of the scatterplot with predicted and measured SOC content from the 1:1 diagonal line can also be used to evaluate a particular algorithm's prediction accuracy. Intuitively, higher accuracy prediction algorithm will have a better fit along the 1:1 line (*Luan et al., 2013*).

Table 1 Descriptive statistics of soil organic carbon in both training and testing sets.

Models	Sample size	Min/%	Max/%	Mean/%	St.dev/%
Training sets	70	0.02	2.97	0.51	0.64
Testing sets	70	0.01	3.42	0.40	0.65

Hyperion simulation

Despite advances in algorithm development, successful applications of satellite-based methods are limited due to the relative unavailability of sensors with both fine spectral and spatial resolution. The next generation multispectral and hyperspectral sensors such as NASA's Hyperspectral Infrared Imager (HyspIRI) attempt to address these issues with both increased spatial and spectral resolution but are not yet available. No Hyperion hyperspectral data were available for the study area during the study period. Considering Hyperion shares similar spectral and spatial characteristics with HyspIRI, we simulated Hyperion soil reflectance spectra by using the spectral response function shown in Eq. (5).

$$\rho_{Hyperion}(\lambda) = \frac{\int_{\lambda\min}^{\lambda\max} f(\lambda)\rho(\lambda)d(\lambda)}{\int_{\lambda\min}^{\lambda\max} f(\lambda)d(\lambda)} \quad (5)$$

where $\rho_{Hyperion}$ is the simulated Hyperion reflectance spectra of band λ; $f(\lambda)$ is the spectral response function of the simulated band λ; $\rho(\lambda)$ is the measured reflectance spectral at band λ; and $\lambda_{\min}$ and $\lambda_{\max}$ are the lower and upper bounds of measured reflectance spectra, respectively (*Maimaitiyiming, Miller & Ghulam, 2016*).

RESULTS

Descriptive statistical analysis of SOC content

The statistical characteristics of both the training and testing sets were shown in Table 1. The standard deviations between the training and testing sets were similar, and differences between the average value was reasonably small. The fact suggests that the selection of both data sets was representative. The spectral curves of soil samples with different SOC contents were illustrated in Fig. 2, overall spectral reflectance increases as the SOC contents decrease. The diagram shows that SOC contents of less than 1% and more than 2.5% correspond to the highest and lowest reflectance, respectively. The pattern of the spectral reflectance fluctuation remains similar regardless of the different SOC contents. Specifically, the reflectance tends to be low in the visible bands (350–780 nm), but high in the infrared bands (780–2,500 nm). The significant absorption features were in the range of 1,850–1,950 nm, and the spectral curves increased rapidly to the peak of 2,100 nm.

Comparing the results from this study with results reported in other regions (*Hong et al., 2018*; *Peng et al., 2014*; *Viscarra Rossel et al., 2006*), we observe that the spectral reflectance of soils in arid wetland regions were similar to those of agricultural soils, though in the range from 1,900 to 2,100 nm, spectral reflectance of arid wetland soil increases faster than that of the crop land.

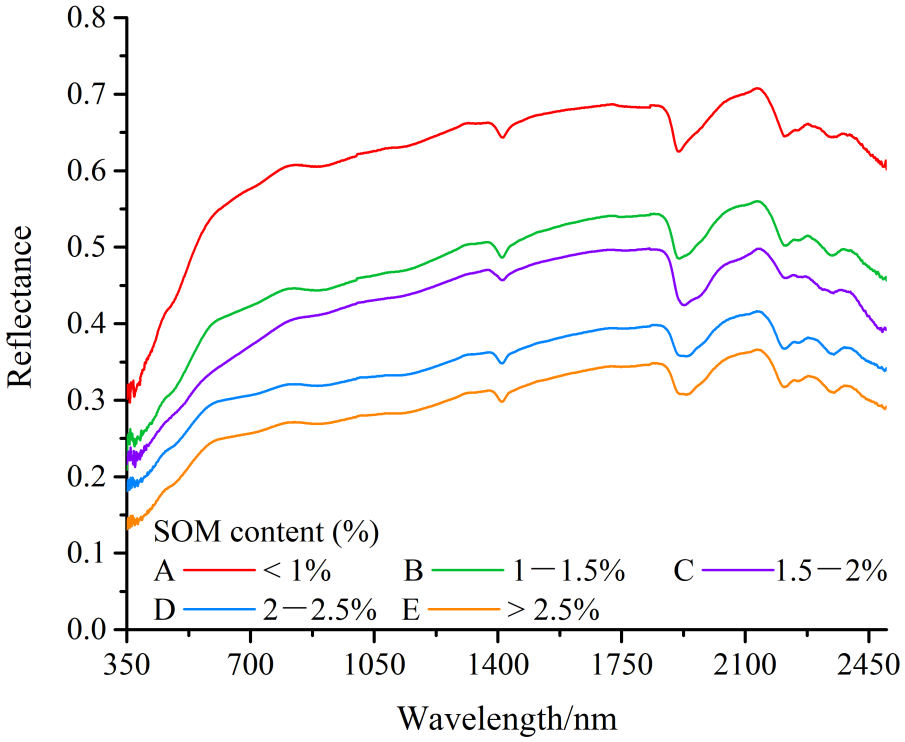

Figure 2 Spectral reflectance of different wetland soil organic carbon contents. (A) Arenosols, (B) Solonetz, (C) Solonetz, (D) Solonchaks, (E) Solonetz.

Full-size DOI: 10.7717/peerj.5714/fig-2

Results of ACO-iPLS algorithm

Although spectral data contains very large amount of data, not all of them are necessarily informative towards detecting SOC contents (*Stenberg et al., 2010*). If all wavelengths were applied in the construction of the models, some irrelevant spectral data could be included, and yields inferior estimation accuracies (*Liu, Zhang & Zhang, 2008*). Thus, the model based on informative spectra could outperform those using all wavelengths.

For that reason, the ACO-iPLS algorithm first selected the informative spectra segments (after pretreatments), and then relevant spectra segments were entered to the model to predict the SOC content. The selected wavelengths and the modeling results with ACO-iPLS were reported in Table 2. It was obvious that the original first order derivative transformation with selected wavelengths yielded the best model performance with the optimal R_{cv} and $RMSE_{cv}$ of 0.87 and 0.33%, respectively. For the testing set, R_t was 0.83, $RMSE_t$ was 0.40%, and the RPD was 1.63, indicating reasonable prediction. Other transformations yield rather low RPD scores (0.87 and 1.10, respectively).

Results of RF-SVM

We applied the RF-SVM algorithm to quantify the SOC contents using selected wavelengths via the three transformations. The selected wavelengths and modeling results were reported in Table 3. The result again suggested that the original first order derivative transformation yielded the best performance. R_{cv} is 0.97, $RMSE_{cv}$ is 0.16%. For the testing set, the R_t is 0.91,

Table 2 Selected feature wavelengths, training sets and testing sets results by ACO-iPLS method.

Pre-processing	Selected wavelengths	Training sets		Testing sets		
		R_{cv}	$RMSE_{cv}$	R_t	$RMSE_t$	RPD
A′	1,786∼1,929	0.86	0.33	0.83	0.40	1.63
1/A′	494∼638	0.64	0.57	0.76	0.74	0.87
lgA′	1,786∼1,929	0.73	0.50	0.82	0.59	1.10

Table 3 Selected feature wavelengths, training sets and testing sets results by RF-SVM method.

Pre-processing	Selected wavelengths	Training sets		Testing sets		
		R_{cv}	$RMSE_{cv}$	R_t	$RMSE_t$	RPD
A′	780, 1,911, 783, 779, 768, 759, 793, 794, 2,254, 910, 1,677, 1,912, 2,089, 745, 825, 2,088, 746, 2,090, 1,913, 1,751	0.97	0.16	0.91	0.27	2.41
1/A′	663, 1,836, 658, 2,431, 2,494, 618, 999, 746, 370, 2,475, 960, 510, 1,081, 443, 1,681, 1,123, 360, 793, 2,123, 2,476	0.99	0.03	0.84	0.34	1.88
lgA′	706, 736, 731, 1,943, 779, 721, 413, 510, 704, 397, 732, 1,944, 1,085, 2,091, 2,347, 881, 2,422, 1,966, 2,257, 2,111	0.99	0.03	0.81	0.45	1.44

Table 4 Selected feature wavelengths, training sets and testing sets results by RF method.

Pre-processing	Selected wavelengths	Training sets		Testing sets		
		R_{cv}	$RMSE_{cv}$	R_t	$RMSE_t$	RPD
A′	794, 740, 758, 713, 741, 821, 789, 766, 613, 682, 732, 776, 822, 720, 769, 746, 635, 733, 940, 668	0.98	0.15	0.92	0.33	1.98
1/A′	1,403, 1,402, 1,390, 1,399, 1,405, 1,404, 2,189, 2,196, 620, 2,176, 822, 2,192, 809, 2,177, 670, 632, 2,191, 1,388, 727, 2,315	0.98	0.14	0.83	0.43	1.51
lgA′	676, 633, 2,189, 2,202, 2,195, 675, 1,402, 2,183, 722, 632, 620, 703, 821, 2,205, 2,193, 689, 2,200, 646, 812, 714	0.98	0.14	0.90	0.41	1.58

$RMSE_t$ is 0.27%, and RPD is 2.41. The results indicated the RF-SVM algorithm recorded overall better model performance. Even for the other two transformations, the RPDs also reached 1.88 and 1.44, indicating that the algorithm could give reasonable prediction.

Results of RF

Based on the three transformations of spectral data, the selected feature bands and RF modeling results were reported in Table 4. Again, the original first order derivative transformation yields the best performance, R_{cv} was 0.98 and $RMSE_{cv}$ was 0.15%. For the testing set, the R_t is 0.92, $RMSE_t$ is 0.33%, and RPD is 1.98. The algorithm also yielded very good prediction. Even for the other two transformations, the RPDs are 1.51 and 1.58, indicating reasonably acceptable prediction accuracy.

From Tables 2–4, it was evident that the original first order derivative transformation of the spectral information yielded the most reasonable modeling performance among all three algorithms. Our discussion will then focus on modeling results based on this transformation only. Figures 3–5 are the results of selected feature wavelengths under the first order

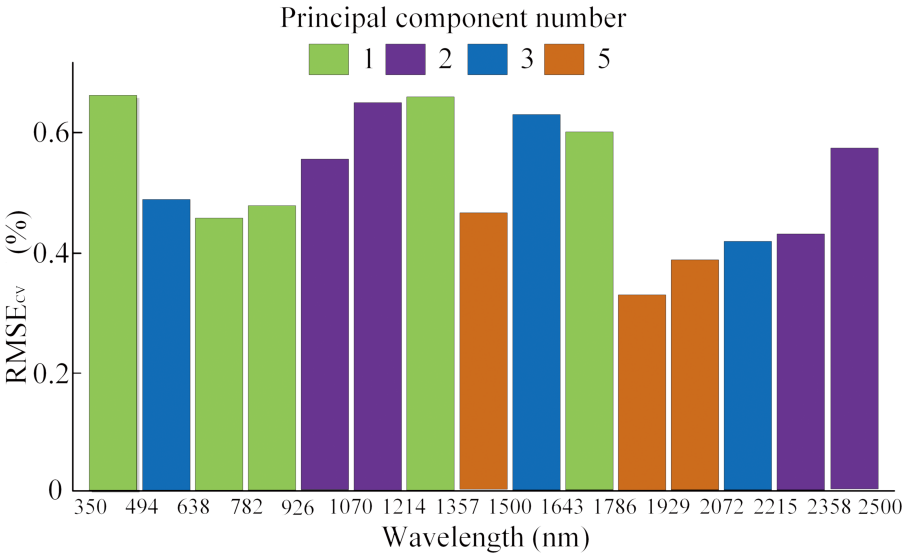

Figure 3 **Selected spectral interval by ACO-iPLS with first derivative spectra.**
Full-size DOI: 10.7717/peerj.5714/fig-3

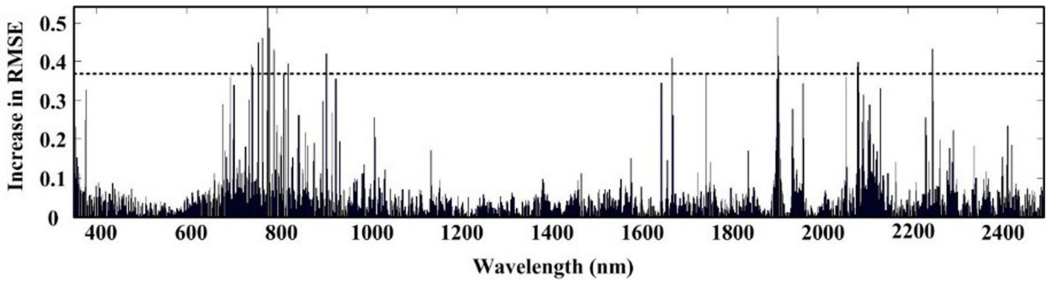

Figure 4 **Selected wavelengths by RF-SVM with first derivative spectra.**
Full-size DOI: 10.7717/peerj.5714/fig-4

derivative transformation. Table 5 lists the modeling performance using wavelengths after first order derivative transformation with both the training and testing sets.

The ACO-iPLS algorithm subdivided the entire wavelength (350–2,500 nm) into 15 segments (Fig. 3). The 11th segments (1,786–1,929 nm) yielded the best performance with the lowest RMSE value. From Fig. 4, based on RF-SVM algorithm, the optimal wavelengths were in the segments of 745–910 nm and 1,911–2,254 nm. Based on mean decrease in accuracy (Fig. 5), RF selected wavelengths ranging from 613 to 940 nm. Combing all three selection results, the optimal wavelengths that were most relevant to SOC content are located within segments of 745–910 nm and 1,911–2,254 nm for the arid Ebinur Lake wetland soils.

From Table 5, we learn that the RF-SVM produces the highest RPD ($\geq$2) among the three algorithms, followed by RF and ACO-iPLS. In addition, Fig. 6 shows scatterplots

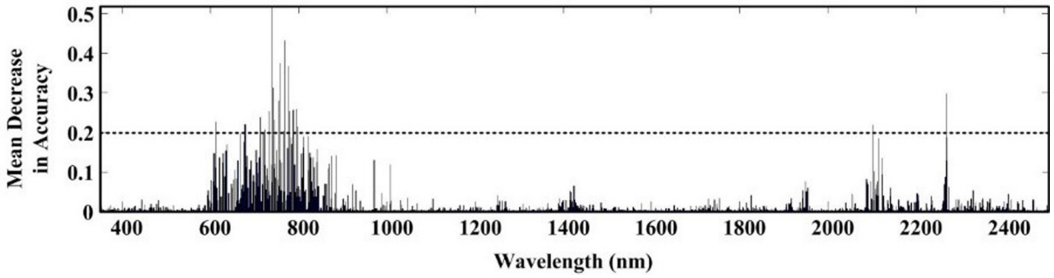

Figure 5 Selected wavelengths by RF with first derivative spectra.
Full-size DOI: 10.7717/peerj.5714/fig-5

Table 5 Comparison of the results by different models with first derivative spectra.

Modeling methods	Training sets		Testing sets		
	R_{cv}	$RMSE_{cv}$	R_t	$RMSE_t$	RPD
AOC-iPLS	0.86	0.33	0.83	0.40	1.63
RF	0.98	0.15	0.92	0.33	1.98
RF-SVM	0.97	0.16	0.91	0.27	2.41

of predicted and measured SOC contents. The slope for the RF-SVM model were well distributed on the 1:1 line indicating the best fit which further confirms the above observations.

Simulation application to satellite data

The results of all three algorithms performed on laboratory-derived spectra data showed that RF-SVM approach with the first derivative pre-processing produced the highest estimation accuracy. In this practice, we evaluated the feasibility of simulated Hyperion reflectance spectra to estimate SOC by using RF-SVM approach. The selected wavelengths and the modeling results were illustrated in Table 6. Using the RF-SVM algorithm, the optimal wavelengths were in the segments of 702–824 nm and 2,083–2,426 nm. And there was some difference between the feature bands of Hyperion and those of laboratory-derived data. It was observed that the original first order derivative transformation with selected wavelengths yielded higher R_{cv} (0.96) and $RMSE_{cv}$ of 0.23%. For the testing set, R_t was 0.79, and $RMSE_t$ was 0.19%, the RPD was 1.61 (Fig. 7). The results suggest that VIS/NIR bands of hyperspectral satellite data have a good potential for predicting wetland SOC content in arid areas.

DISCUSSION

Applying spectral techniques to evaluate, monitor and predict SOC content is an important approach, especially in arid regions where SOC is critical to soil quality yet soil sampling often is expensive and sometimes very hard for laboratory tests. Establishing an effective model to take advantage of the large amount of spectra information from VIS/NIR spectroscopy technology is of great interest in the study of SOC content in arid regions.

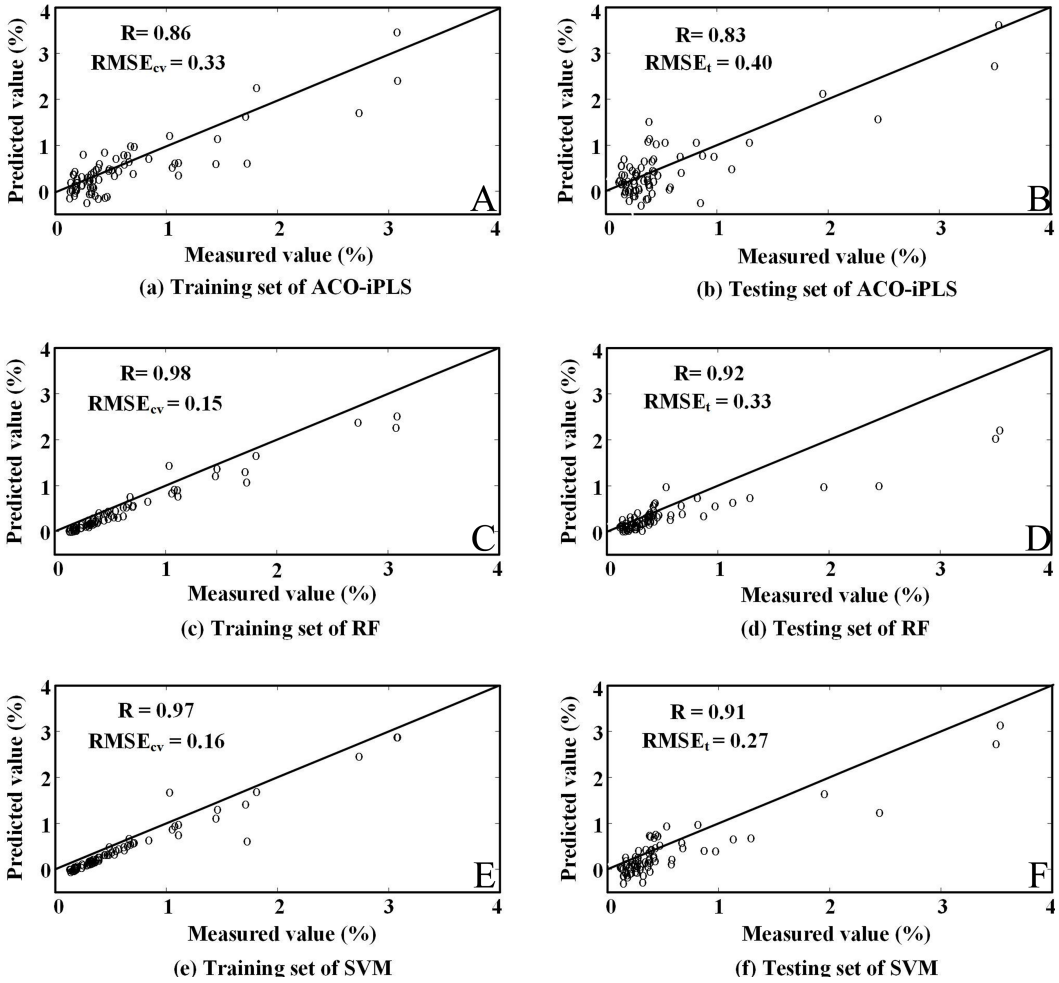

Figure 6 **Measured content and the values estimated by SVM model with simulated EO-1 Hyperion data.**

Full-size DOI: 10.7717/peerj.5714/fig-6

In the current study, we found that the original first order derivative transformation was able to retain the most useful information from spectral data among the three discussed transformations including original first order derivative, inversion of first order derivative, and logarithm of first order derivative. *Zornoza et al. (2008)* also found similar results in their study, with high coefficient of correlations of 0.95 and 0.98, which was consistent with our results.

The massive information obtained from hyperspectral sensors often requires effective algorithms to produce the best and most robust prediction. The selection of feature spectra is hence a critical step in applying VIS/NIR spectroscopy technique. From Tables 2–4, we could see that the approaches of RF-SVM and RF were able to select 20 wavelength segments, and the ACO-iPLS selected a large chunk of segments that produce the most desired results. Comparing the results of the three algorithms, we observed that the relevant wavelengths concentrated in the range of 745–910 nm and 1,911–2,254 nm. Results obtained in the

Table 6 Selected feature wavelengths and training sets and testing sets results by RF-SVM method with simulated EO-1 Hyperion data.

Modeling methods	Selected wavelengths	Training sets		Testing sets		
		R_{cv}	$RMSE_{cv}$	R_t	$RMSE_t$	RPD
RF-SVM	824, 813, 2,194, 2,426, 2,093, 702, 2,083, 712, 2,436, 803, 2,174, 2,214, 2,163, 2,416, 2,103, 722, 2,133, 1,810, 1,669, 2,123	0.96	0.23	0.79	0.19	1.61

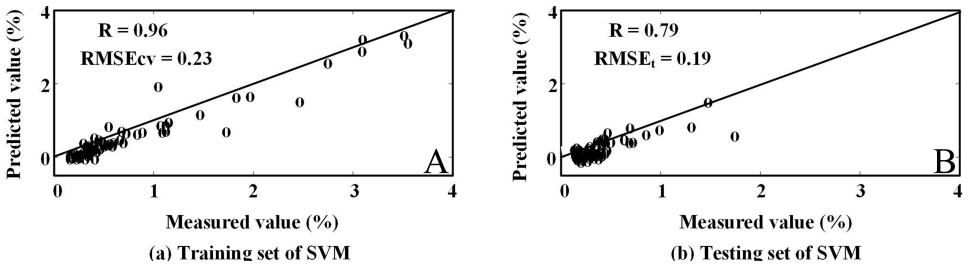

Figure 7 Comparison of the measured content and the values estimated by different models.
Full-size DOI: 10.7717/peerj.5714/fig-7

current study were not in accord with previous research (400–800 nm, 1,030–1,080 nm and 2,250–2,340 nm), this could be attributed to the difference among different soil types (*Wang et al., 2016*).

In addition, our experiment with the three algorithms also indicated that different algorithms yield different results. Among the three algorithms, the RF-SVM produced the best results, followed by RF algorithm, while the ACO-iPLS algorithm performs least ideally. We argue that the ACO-iPLS algorithm is not necessarily an inappropriate algorithm as the results are still acceptable. However, ACO-iPLS algorithm, a bionic approach attempting to mimic the ants' intelligence, could yield effective optimization if the amount of information was set to appropriate size. On the other hand, if the amount of information was massive, the complexity of the information might lead to local optimal to dominate the optimization process. Though by further fine-dividing the segments, we might be able to produce better results, we are still at risk of being stuck at potential local optimal with the ACO-iPLS algorithm. Random forest algorithm, on the other hand, uses multiple trees (the forest) to produce enhanced learning outcomes. Because of its flexibility, it often produces the best results with the training set since it can distinguish between the useful information and inevitable noises existed within the training set. The problem, however, as with many tree-related algorithms, is that it can easily slide to over-fitting and might be impacted by the very skewed dataset which is the case for our dataset. The results from the current study support this argument. The $RMSE_{cv}$ of the random forest algorithm for the training set was 0.15, lowest among all three algorithms. While the $RMSE_t$ was higher than that of SVM (0.33 versus 0.27, Fig. 6), indicating possible over-fitting issues for the random forest algorithm. SVM is a structural empirical risk model, the parameters of the decision function are determined by empirical analysis. Since the goal of the algorithm when training the parameters was to minimize risks, it allows for some errors during fitting while assigning

certain penalty to such errors (by adjusting the tuning parameter, C). This also agrees logically with the fact that there shall be inevitable noises in the training data. Our analysis suggested that the SVM produced results were better. The RPD of the model reached 2.41 while those of RF and ACO-iPLS were 1.98 and 1.63, respectively. The results were in line with *Were et al. (2015)* and *Viscarra Rossel & Behrens (2010)* who used similar algorithms to study SOC contents in Kenya and Australia.

The results of simulated Hyperion spectral analysis showed that the feature wavelengths pertaining to SOC were mainly located around the ranges of 702–824 nm and 2,083–2,426 nm. This finding differs from previously reported work by *Morgan et al. (2009)* that identified the feature wavelengths between 610 and 650 nm for SOC. It is worth noting that the 702–824 nm wavelengths overlap with the findings by *Ji, Viscarra Rossel & Shi (2015)* who identified the absorption feature at the wavelengths from 600 to 800 nm for SOC. Although the combination of laboratory-derived reflectance with RF-SVM produced slightly better estimation than the simulated Hyperion reflectance spectra, the advantage of the simulated approach is evident, which is critical for the potential uses of the planned hyperspectral sensors soon to be available.

In fact, the simulated hyperspectral was constructed based on the specified spectral response function and field-derived spectral data (*Jin et al., 2017*; *Liu et al., 2009*). However, the obtained continuous spectrum information of every pixel was affected by geographical atmospheric, meteorological, and lighting variations, e.g., cloud cover, precipitation, pixel purity, and temporal-spatial resolution of ground target (*Hill, 2013*; *Zhou et al., 2013*). Hence, there was a difference between the actual VIS/NIR reflectance and simulated satellite data, the quantitative estimation models were only theoretically valid. To further improve the applicable capability of the established model, the prediction accuracy, combination of various simulated and actual satellite sensor data will be analyzed in the future study. Various spaceborne hyperspectral data or imaging spectroscopy will be available, e.g., ESA's PROBA, Chinese GaoFen-4, and upcoming Germany's DLR's EnMAP, which is very helpful for achievement of the quantitative analysis on remote sensing (*Liu et al., 2018*). Therefore, actual satellite data should be used in future studies to evaluate the SOM estimation model. We did not consider the effect of the spatial resolution of remote sensing imagery on the estimation accuracy of SOC content. In terms of the actual satellite data, the different radiometric correction and atmospheric correction approaches could result in the changes of spectra of targets. Additionally, the actual satellite data and the simulated satellite data differ because of the measurement errors, signal noises, and atmospheric environment (*Maimaitiyiming, Miller & Ghulam, 2016*). The accuracy and detection limits of estimations could be affected by these mentioned factors. In future study, more intelligent algorithms will be applied to overcome the scale differences in both spectral and spatial dimension of actual and simulated data.

This study clearly suggests that VIS/NIR spectroscopy is an effective method to detect wetland SOC content of soils in arid regions. Our work provides a comprehensive evaluation of models and algorithms for their power to identify relevant feature wavelengths to estimate SOC. Such endeavor is of critical and practical importance. Increasing population and intensive human activities have put ever increasing pressure on wetland in arid

regions. Changes in the SOC content in fragile ecosystems can be drastic even with slight increase of human activities (*Câmara et al., 2016*; *Smith et al., 2016*). Such changes could have significant impact on local climatic and ecological systems, and even contribute to large-scale carbon equilibrium (*Prasad et al., 2016*). The proposed VIS/NIR spectroscopic approach plus the relatively mature classification and prediction models provide effective means to the local ecological and environmental management authorities.

CONCLUSION

In this study, the first order derivative transformation provides the best predictive power among proposed wavelength transformation strategies (A′, 1/A′, and lg(A′)). All three algorithms consistently suggest that the wavelength segments of 745–910 nm and 1,911–2,254 nm are the most effective spectral regions to detect SOC content. Among the three models, SVM based recursive feature elimination algorithm produces the best overall results for both the training and testing datasets with an RPD of 2.41. The other two approaches, namely, ACO-iPLS and RF also produce reasonably well results following SVM. In addition, the simulated EO-1 Hyperion data combined with SVM based recursive feature elimination algorithm produces high accuracy of estimating SOC content with an RPD of 1.61. The RF-SVM algorithm identified the wavelength segments of 702–824 nm and 2,083–2,426 nm as the most effective spectral regions to detect SOC content at the satellite level. Overall, the simulated Hyperion data have a great potential for predicting wetland SOC content in arid regions. The proposed combination of VIS/NIR spectroscopy technique and SVM based recursive feature elimination algorithm provides a fast, economic, and robust approach to monitor, detect, and predict SOC contents in the arid and semi-arid region wetlands.

ACKNOWLEDGEMENTS

We are especially grateful to the anonymous reviewers and editors for appraising our manuscript and for offering instructive comments.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Natural Science Foundation of China (No. 41771470 and No. U1603241). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 41771470, U1603241.

Competing Interests

Danlin Yu is an Academic Editor for PeerJ.

Author Contributions

- Jianli Ding conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.
- Aixia Yang conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, approved the final draft.
- Jingzhe Wang performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Vasit Sagan performed the experiments.
- Danlin Yu conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data are provided in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.5714#supplemental-information.

REFERENCES

Abuduwailil J, Zhaoyong Z, Fengqing J. 2015. Evaluation of the pollution and human health risks posed by heavy metals in the atmospheric dust in Ebinur Basin in Northwest China. *Environmental Science and Pollution Research* **22**:14018–14031 DOI 10.1007/s11356-015-4625-1.

Anne NJP, Abd-Elrahman AH, Lewis DB, Hewitt NA. 2014. Modeling soil parameters using hyperspectral image reflectance in subtropical coastal wetlands. *International Journal of Applied Earth Observation and Geoinformation* **33**:47–56 DOI 10.1016/j.jag.2014.04.007.

Araújo SR, Söderström M, Eriksson J, Isendahl C, Stenborg P, Demattê JM. 2015. Determining soil properties in Amazonian Dark Earths by reflectance spectroscopy. *Geoderma* **237–238**:308–317 DOI 10.1016/j.geoderma.2014.09.014.

Breiman L. 2001. Random forests. *Machine Learning* **45**:5–32 DOI 10.1023/a:1010933404324.

Câmara FAC, Carvalho LLF, Ferreira AAS, Nico E. 2016. Land-use type effects on soil organic carbon and microbial properties in a semi-arid region of Northeast Brazil. *Land Degradation & Development* **27**:171–178 DOI 10.1002/ldr.2282.

Chang C-W, Laird DA, Mausbach MJ, Hurburgh CR. 2001. Near-infrared reflectance spectroscopy—principal components regression analyses of soil properties. *Soil Science Society of America Journal* **65**:480–490 DOI 10.2136/sssaj2001.652480x.

Cohen MJ, Prenger JP, DeBusk WF. 2005. Visible-near infrared reflectance spectroscopy for rapid, nondestructive assessment of wetland soil quality. *Journal of Environmental Quality* **34**:1422–1434 DOI 10.2134/jeq2004.0353.

Cole JJ, Prairie YT, Caraco NF, McDowell WH, Tranvik LJ, Striegl RG, Duarte CM, Kortelainen P, Downing JA, Middelburg JJ, Melack J. 2007. Plumbing the global carbon cycle: integrating Inland waters into the terrestrial carbon budget. *Ecosystems* **10**:172–185 DOI 10.1007/s10021-006-9013-8.

Craft CB, Seneca ED, Broome SW. 1991. Loss on ignition and kjeldahl digestion for estimating organic carbon and total nitrogen in estuarine marsh soils: calibration with dry combustion. *Estuaries* **14**:175–179 DOI 10.2307/1351691.

Dai F, Zhou Q, Lv Z, Wang X, Liu G. 2014. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. *Ecological Indicators* **45**:184–194 DOI 10.1016/j.ecolind.2014.04.003.

Ding J, Yu D. 2014. Monitoring and evaluating spatial variability of soil salinity in dry and wet seasons in the Werigan–Kuqa Oasis, China, using remote sensing and electromagnetic induction instruments. *Geoderma* **235–236**:316–322 DOI 10.1016/j.geoderma.2014.07.028.

Foley JA, DeFries R, Asner GP, Barford C, Bonan G, Carpenter SR, Chapin FS, Coe MT, Daily GC, Gibbs HK, Helkowski JH, Holloway T, Howard EA, Kucharik CJ, Monfreda C, Patz JA, Prentice IC, Ramankutty N, Snyder PK. 2005. Global consequences of land use. *Science* **309**:570–574 DOI 10.1126/science.1111772.

González Costa JJ, Reigosa MJ, Matías JM, Covelo EF. 2017. Soil Cd, Cr, Cu, Ni, Pb and Zn sorption and retention models using SVM: variable selection and competitive model. *Science of the Total Environment* **593–594**:508–522 DOI 10.1016/j.scitotenv.2017.03.195.

Guo P-T, Li M-F, Luo W, Tang Q-F, Liu Z-W, Lin Z-M. 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. *Geoderma* **237–238**:49–59 DOI 10.1016/j.geoderma.2014.08.009.

He X, Lv G, Qin L, Chang S, Yang M, Yang J, Yang X. 2015. Effects of simulated nitrogen deposition on soil respiration in a populus euphratica community in the Ebinur lake area, a desert ecosystem of Northwestern China. *PLOS ONE* **10**:e0137827 DOI 10.1371/journal.pone.0137827.

Hill MJ. 2013. Vegetation index suites as indicators of vegetation state in grassland and savanna: an analysis with simulated SENTINEL 2 data for a North American transect. *Remote Sensing of Environment* **137**:94–111 DOI 10.1016/j.rse.2013.06.004.

Hong Y, Yu L, Chen Y, Liu Y, Liu Y, Liu Y, Cheng H. 2018. Prediction of soil organic matter by VIS–NIR spectroscopy using normalized soil moisture index as a proxy of soil moisture. *Remote Sensing* **10**:Article 28 DOI 10.3390/rs10010028.

Hu M-H, Yuan J-H, Yang X-E, He Z-L. 2010. Effects of temperature on purification of eutrophic water by floating eco-island system. *Acta Ecologica Sinica* **30**:310–318 DOI 10.1016/j.chnaes.2010.06.009.

Huang X, Zou X, Zhao J, Shi J, Zhang X, Mel H. 2014. Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models. *Food Chemistry* **164**:536–543 DOI 10.1016/j.foodchem.2014.05.072.

Jaber SM, Al-Qinna MI. 2011. Soil organic carbon modeling and mapping in a semi-arid environment using thematic mapper data. *Photogrammetric Engineering and Remote Sensing* **77**:709–719 DOI 10.14358/pers.77.7.709.

Ji W, Viscarra Rossel RA, Shi Z. 2015. Accounting for the effects of water and the environment on proximally sensed vis—NIR soil spectra and their calibrations. *European Journal of Soil Science* **66**:555–565 DOI 10.1111/ejss.12239.

Jin X, Song K, Du J, Liu H, Wen Z. 2017. Comparison of different satellite bands and vegetation indices for estimation of soil organic matter based on simulated spectral configuration. *Agricultural and Forest Meteorology* **244–245**:57–71 DOI 10.1016/j.agrformet.2017.05.018.

Kayranli B, Scholz M, Mustafa A, Hedmark Å. 2010. Carbon storage and fluxes within freshwater wetlands: a critical review. *Wetlands* **30**:111–124 DOI 10.1007/s13157-009-0003-4.

Kinoshita R, Moebius-Clune BN, Van Es HM, Hively WD, Bilgilis AV. 2012. Strategies for soil quality assessment using visible and near-infrared reflectance spectroscopy in a Western Kenya chronosequence. *Soil Science Society of America Journal* **76**:1776–1788 DOI 10.2136/sssaj2011.0307.

Kuang B, Tekin Y, Mouazen AM. 2015. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil and Tillage Research* **146**:243–252 DOI 10.1016/j.still.2014.11.002.

Li Y, Zhao M, Li F. 2018. Soil respiration in typical plant communities in the wetland surrounding the high-salinity Ebinur Lake. *Frontiers of Earth Science* **12(3)**:611–624 DOI 10.1007/s11707-018-0687-y.

Liaw A, Wiener M. 2002. Classification and regression by random forest. *R News* **2**:18–22.

Lin X, Wang Q, Yin P, Tang L, Tan Y, Li H, Yan K, Xu G. 2011. A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics* **7**:549–558 DOI 10.1007/s11306-011-0274-7.

Liu L, Feng J, Rivard B, Xu X, Zhou J, Han L, Yang J, Ren G. 2018. Mapping alteration using imagery from the Tiangong-1 hyperspectral spaceborne system: example for the Jintanzi gold province, China. *International Journal of Applied Earth Observation and Geoinformation* **64**:275–286 DOI 10.1016/j.jag.2017.03.013.

Liu W, Su Y, Yang R, Wang X, Yang X. 2010. Land use effects on soil organic carbon, nitrogen and salinity in saline-alkaline wetland. *Sciences in Cold and Arid Regions* **2**:263–270.

Liu H, Zhang Y, Zhang B. 2008. Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China. *Environmental Monitoring and Assessment* **154**:Article 147 DOI 10.1007/s10661-008-0385-4.

Liu B, Zhang L, Zhang X, Zhang B, Tong Q. 2009. Simulation of EO-1 hyperion data from ALI multispectral data based on the spectral reconstruction approach. *Sensors* **9**:3090–3108 DOI 10.3390/s90403090.

Luan FL, Zhang XL, Xiong HG, Zhang F, Wang F. 2013. Comparative analysis of soil organic matter content based on different hyperspectral inversion models. *Spectroscopy and Spectral Analysis* **33**:196–200.

Maimaitiyiming M, Miller AJ, Ghulam A. 2016. Discriminating spectral signatures among and within two closely related grapevine species. *Photogrammetric Engineering and Remote Sensing* **82**:51–62 DOI 10.14358/PERS.82.2.51.

McDowell ML, Bruland GL, Deenik JL, Grunwald S, Knox NM. 2012. Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. *Geoderma* **189–190**:312–320 DOI 10.1016/j.geoderma.2012.06.009.

Meng R, Dennison PE. 2015. Spectroscopic analysis of green, desiccated and dead tamarisk canopies. *Photogrammetric Engineering and Remote Sensing* **81**:199–207 DOI 10.14358/pers.81.3.199-207.

Morgan CLS, Waiser TH, Brown DJ, Hallmark CT. 2009. Simulated *in situ* characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. *Geoderma* **151**:249–256 DOI 10.1016/j.geoderma.2009.04.010.

Mountrakis G, Im J, Ogole C. 2011. Support vector machines in remote sensing: a review. *ISPRS Journal of Photogrammetry and Remote Sensing* **66**:247–259 DOI 10.1016/j.isprsjprs.2010.11.001.

Mutanga O, Adam E, Cho MA. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation* **18**:399–406 DOI 10.1016/j.jag.2012.03.012.

Nauman TW, Thompson JA, Rasmussen C. 2014. Semi-automated disaggregation of a conventional soil map using knowledge driven data mining and random forests in the sonoran desert, USA. *Photogrammetric Engineering and Remote Sensing* **80**:353–366 DOI 10.14358/pers.80.4.353.

Nawar S, Mouazen AM. 2017. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *CATENA* **151**:118–129 DOI 10.1016/j.catena.2016.12.014.

Peng X, Shi T, Song A, Chen Y, Gao W. 2014. Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. *Remote Sensing* **6**:2699–2717 DOI 10.3390/rs6042699.

Polglase PJ, Jokela EJ, Comerford NB. 1992. Phosphorus, nitrogen, and carbon fractions in litter and soil of Southern pine plantations. *Soil Science Society of America Journal* **56**:566–573 DOI 10.2136/sssaj1992.03615995005600020036x.

Pott A, Pott VJ. 2004. Features and conservation of the Brazilian Pantanal wetland. *Wetlands Ecology and Management* **12**:547–552 DOI 10.1007/s11273-005-1754-1.

Prasad JVNS, Rao CS, Srinivas K, Jyothi CN, Venkateswarlu B, Ramachandrappa BK, Dhanapal GN, Ravichandra K, Mishra PK. 2016. Effect of ten years of reduced tillage and recycling of organic matter on crop yields, soil organic carbon and its fractions in Alfisols of semi arid tropics of southern India. *Soil and Tillage Research* **156**:131–139 DOI 10.1016/j.still.2015.10.013.

Savitzky A, Golay MJE. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**:1627–1639 DOI 10.1021/ac60214a047.

Shi T, Cui L, Wang J, Fei T, Chen Y, Wu G. 2013. Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. *Plant and Soil* **366**:363–375 DOI 10.1007/s11104-012-1436-8.

Shi Z, Wang Q, Peng J, Ji W, Liu H, Li X, Viscarra Rossel RA. 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Science China Earth Sciences* **57**:1671–1680 DOI 10.1007/s11430-013-4808-x.

Smith P, House JI, Bustamante M, Sobocká J, Harper R, Pan G, West PC, Clark JM, Adhya T, Rumpel C, Paustian K, Kuikman P, Cotrufo MF, Elliott JA, McDowell R, Griffiths RI, Asakawa S, Bondeau A, Jain AK, Meersmans J, Pugh TA. 2016. Global change pressures on soils from land use and management. *Global Change Biology* **22**:1008–1028 DOI 10.1111/gcb.13068.

St. Luce M, Ziadi N, Zebarth BJ, Grant CA, Tremblay GF, Gregorich EG. 2014. Rapid determination of soil organic matter quality indicators using visible near infrared reflectance spectroscopy. *Geoderma* **232–234**:449–458 DOI 10.1016/j.geoderma.2014.05.023.

Stenberg B, Viscarra Rossel RA, Mouazen AM, Wetterlind J. 2010. Chapter five—visible and near infrared spectroscopy in soil science. In: Sparks DL, ed. *Advances in agronomy*. Burlington: Academic Press, 163–215.

Stevens A, Udelhoven T, Denis A, Tychon B, Lioy R, Hoffmann L, Van Wesemael B. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* **158**:32–45 DOI 10.1016/j.geoderma.2009.11.032.

Summers D, Lewis M, Ostendorf B, Chittleborough D. 2011. Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. *Ecological Indicators* **11**:123–131 DOI 10.1016/j.ecolind.2009.05.001.

Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* **43**:1947–1958 DOI 10.1021/ci034160g.

Tan C, Guo B, Kuang H, Yang H, Ma M. 2018. Lake area changes and their influence on factors in arid and semi-arid regions along the silk road. *Remote Sensing* **10**:Article 595 DOI 10.3390/rs10040595.

Thakur JK, Srivastava PK, Singh SK, Vekerdy Z. 2012. Ecological monitoring of wetlands in semi-arid region of Konya closed Basin, Turkey. *Regional Environmental Change* **12**:133–144 DOI 10.1007/s10113-011-0241-x.

Thissen U, Pepers M, Üstün B, Melssen WJ, Buydens LMC. 2004. Comparing support vector machines to PLS for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems* **73**:169–179 DOI 10.1016/j.chemolab.2004.01.002.

Vapnik VN. 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks* **10**:988–999 DOI 10.1109/72.788640.

Vasques GM, Grunwald S, Harris WG. 2010. Spectroscopic models of soil organic carbon in Florida, USA All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. *Journal of Environmental Quality* **39**:923–934 DOI 10.2134/jeq2009.0314.

Viscarra Rossel RA, Behrens T. 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 RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO. 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, Besold J, Hill J, Fründ H-C. 2011. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma* **166**:198–205 DOI 10.1016/j.geoderma.2011.08.001.

Wang J, Ding J, Abulimiti A, Cai L. 2018. Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS–NIR) spectroscopy, Ebinur Lake Wetland, Northwest China. *PeerJ* **6**:e4703 DOI 10.7717/peerj.4703.

Wang J, Ding J, Zhang D, Liu W, Wang F, Tashpolat N. 2017. Desert soil clay content estimation using reflectance spectroscopy preprocessed by fractional derivative. *PLOS ONE* **12**:e0184836 DOI 10.1371/journal.pone.0184836.

Wang P, Ma Y, Wang X, Jiang H, Liu H, Ran W, Shen Q. 2016. Spectral exploration of calcium accumulation in organic matter in gray desert soil from Northwest China. *PLOS ONE* **11**:e0145054 DOI 10.1371/journal.pone.0145054.

Wang Y, Zhang L, Haimiti Y. 2015. Study on spatial variability of soil nutrients in Ebinur Lake Wetlands in China. *Journal of Coastal Research* **73**:59–63 DOI 10.2112/si73-011.1.

Were K, Bui DT, Dick ØB, Singh BR. 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators* **52**:394–403 DOI 10.1016/j.ecolind.2014.12.028.

West TO, Post WM. 2002. Soil organic carbon sequestration rates by Tillage and crop rotation. *Soil Science Society of America Journal* **66**:1930–1946 DOI 10.2136/sssaj2002.1930.

Xu L, Fan X, Wang W, Xu L, Duan Y, Shi R. 2017. Renewable and sustainable energy of Xinjiang and development strategy of node areas in the ''Silk Road Economic Belt''. *Renewable and Sustainable Energy Reviews* **79**:274–285 DOI 10.1016/j.rser.2017.05.031.

Zhao R, Chen Y, Zhou H, Li Y, Qian Y, Zhang L. 2009. Assessment of wetland fragmentation in the Tarim River basin, western China. *Environmental Geology* **57**:455–464 DOI 10.1007/s00254-008-1316-y.

Zhou Z, Zhou K, Hou X, Luo H. 2005. Arc/Spark optical emission spectrometry: principles, instrumentation, and recent applications. *Applied Spectroscopy Reviews* **40**:165–185 DOI 10.1081/ASR-200052001.

Zhou L, Xu B, Ma W, Zhao B, Li L, Huai H. 2013. Evaluation of hyperspectral multi-band indices to estimate chlorophyll-A concentration using field spectral measurements and satellite data in Dianshan Lake, China. *Water* **5**:525–539 DOI 10.3390/w5020525.

Zhu Y, Chen X, Wang S, Liang S, Chen C. 2018. Simultaneous measurement of contents of liquirtin and glycyrrhizic acid in liquorice based on near infrared spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **196**:209–214 DOI 10.1016/j.saa.2018.02.021.

Zornoza R, Guerrero C, Mataix-Solera J, Scow KM, Arcenegui V, Mataix-Beneyto J. 2008. Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils. *Soil Biology and Biochemistry* **40**:1923–1930 DOI 10.1016/j.soilbio.2008.04.003.