Chin. Geogra. Sci. 2019 Vol. 29 No. 2 pp. 258–269
https://doi.org/10.1007/s11769-019-1020-8

Springer Science Press
www.springerlink.com/content/1002-0063

Combining Environmental Factors and Lab VNIR Spectral Data to Predict SOM by Geospatial Techniques

GUO Long[1], ZHANG Haitao[1], CHEN Yiyun[2], QIAN Jing[2]

(1. *College of Resources and Environment*, *Huazhong Agricultural University*, *Wuhan* 430070, *China*; 2. *School of Resource and Environment*, *Wuhan University*, *Wuhan* 430070, *China*)

Abstract: Soil organic matter (SOM) is an important parameter related to soil nutrient and miscellaneous ecosystem services. This paper attempts to improve the performance of traditional partial least square regression (PLSR) model by considering the spatial autocorrelation and soil forming factors. Surface soil samples (n = 180) were collected from Honghu City located in the middle of Jianghan Plain, China. The visible and near infrared (VNIR) spectra and six environmental factors (elevation, land use types, roughness, relief amplitude, enhanced vegetation index, and land surface water index) were used as the auxiliary variables to construct the multiple linear regression (MLR), PLSR and geographically weighted regression (GWR) models. Results showed that: 1) the VNIR spectra can increase about 39.62% prediction accuracy than the environmental factors in predicting SOM; 2) the comprehensive variables of VNIR spectra and the environmental factors can improve about 5.78% and 44.90% relative to soil spectral models and soil environmental models, respectively; 3) the spatial model (GWR) can improve about 3.28% accuracy than MLR and PLSR. Our results suggest that the combination of spectral reflectance and the environmental variables can be used as the suitable auxiliary variables in predicting SOM, and GWR is a promising model for predicting soil properties.

Keywords: visible near infrared spectral reflectance; environmental factors; spatial characteristics; partial least squares regression; geographically weighted regression

Citation: GUO Long, ZHANG Haitao, CHEN Yiyun, QIAN Jing, 2019. Combining Environmental Factors and Lab VNIR Spectral Data to Predict SOM by Geospatial Techniques. *Chinese Geographical Science*, 29(2): 258–269. https://doi.org/10.1007/s11769-019-1020-8

1 Introduction

Given the world's population and excessive pressures on land resources, spatial-temporal information of soil is highly valued in the sustainable management of soil resources (Hartemink et al., 2008). The formation, development, and movement of soil are affected by complex environmental and natural factors. This not only pose a great challenge on the modeling the spatial characteristics of soil properties but also provide an opportunity to the digital soil mapping by incorporating the related environmental covariates. Soil is three dimensional in character and heterogeneous with respect to the nutrient content and micro-nutrient content (Gupta, 2015). Soil organic matter (SOM) is one important component of soil, and it exerts positive effects on soil physical and chemical properties, as well as provide controlled ecosystem services (Schmidt et al., 2011).

Received date: 2017-06-09; accepted date: 2017-10-08

Foundation item: Under the auspices of the Natural Science Foundation of Hubei (No. 2018CFB372), the Fundamental Research Funds for the Central Universities (No. 2662016QD032), the Key Laboratory of Aquatic Plants and Watershed Ecology of Chinese Academy of Sciences (No. Y852721s04), the Chinese National Natural Science Foundation (No. 41371227), the National Undergraduate Innovation and Entrepreneurship Training Program (No. 201810504023, 201810504030)

Corresponding author: QIAN Jing. E-mail: 2011202050149@whu.edu.cn

Therefore, there has been an ongoing interest in the development of models for quick and cheap access to SOM data of both soil samples and unstamped sites.

It is widely recognized that a good soil dataset is a key factor to build an accurate soil prediction model and to evaluate the quality of its outputs (Lagacherie, 2008). Primary soil data information is collected from a traditional work flow, and it needs a large number of samples and numerous laboratory analyses to obtain physical-chemical and mineralogical properties of soils, as well as knowledge of their spatial variability in the environment. Thus, the collection of soil data has been a limiting factor which can severely slow the progress of soil prediction (Terra et al., 2015; Guo et al., 2018). The diffuse reflectance spectroscopy (DRS) technique is based on the detection of electromagnetic radiation (EMR) reflected at a characteristic wavelength without requiring direct contact between the sensor and the soil. It is less expensive and considerably faster than conventional analysis techniques and can be used with higher-density samples, thereby improving the characterization of an area (Viscarra Rossel and Hicks, 2015). In addition, the synchronous computational storage of DRS allows creating databases called 'spectral libraries' (Terra et al., 2015). This approach provides a useful way to develop and apply soil sensing techniques at different scales from the field (proximal sensing) to the orbital level (remote sensing), and provides support for constructing prediction models of soil properties (Shi et al., 2014). Spectral reflectance data from visible to near infrared (VNIR: 350 to 2500 nm) have been widely and effectively used in soil assessments (Terra et al., 2015). Many reports have shown that VNIR DRS can be successfully used to predict SOM because of its distinct absorption features over the VNIR regions caused by various chemical bonds, such as C–C, C–H, C–N, C=C, and O–H (Peon et al., 2017). The relationship between the reflectance spectra and the reference soil properties in the spectral library can be used as the empirical equations to predict the soil properties (Rossel and Webster, 2012). Several mature techniques have been used as empirical equations, including multiple linear regression (MLR), principal component regression (PCR), and partial least squares regression (PLSR) (Roudier et al., 2017). However, two important sets of information, i.e., soil covariates and their spatial characteristics, have been generally ignored when constructing soil spectral prediction models.

Many scholars have shown that soil properties have strong spatial heterogeneity and dependence, and these characteristics can be used to map the spatial distribution of soil properties by geostatistical models, such as ordinary kriging (OK), cokriging (COK) and simple kriging (SK) (Guo et al., 2017b). Based on the semivariable function and interpolation, geostatistical models can map soil properties by using a limited number of soil samples, thus serving as an economical and efficient way to provide better and more accurate information to soil researchers (Gaetan et al., 2010). At the same time, environmental factors, which have a greater influence on the soil properties, and soil-landscape relationships, which provide predictive tools and foundations of soil survey, are also important factors that need to be considered (Lagacherie, 2008). The soil covariates can be extracted from remote sensing images and soil maps, and these factors govern the soil chemical and physical information and the spatial information (Hartemink et al., 2008). Many scholars have shown that internal physical and chemical structures of soil can be influenced and changed by different natural conditions and environmental factors (Wang et al., 2013).

Soil covariates can be used as inputs of spatial econometrical models, including MLR, geographically weighted regression model (GWR) and spatial auto-regression models (Lagacherie, 2008). A multitudinous number of regression models have been constructed and examined for the study of soil, and they can be classified into three groups based on their basic theories and input-output variables. The first group only considers spatial characteristics of soil properties, such as OK, inverse distance weighting model and others (Trangmar et al., 1985). The definition of the spatial weight and the suitable distance are key modeling components in the processing of soil properties prediction. The second group can be recognized as the multiple linear regression models, which only consider soil covariates and are based on non-spatial statistical models (Evrendilek et al., 2004; Zornoza et al., 2007). The important work of this group is to choose representative auxiliary variables and remove multicollinearity and random noise among them. The third group is about combining spatial characteristics and soil covariates to construct soil prediction models, including COK, GWR and other extensional models (regression kriging and

geographically weighted regression kriging model), which consider spatial autocorrelation of residuals of the prediction results (Guo et al., 2017a). Various prediction models of soil properties have been constructed in different regions of interest, and most of the publications have shown that the third class of models have better predictive accuracy and can be easily interpreted (Wang et al., 2013).

GWR, which is a local spatial linear regression method, can be used to test whether the model coefficients are non-constant over space, unlike traditional regression models (such as PLSR and MLR), which assume spatial stationary (Shekhar and Xiong, 2008). Thus, GWR will be used to construct the prediction models of SOM with soil covariates and spectral library in this paper. The overall goal of this study was to improve the performance of the traditional soil spectral models with the help of the spatial models and the environmental factors. The specific objectives were to: 1) compare the differences of the environmental factors and the VNIR spectra in predicting SOM; 2) analysis the spatial autocorrelation of SOM and its auxiliary variables, and explore the potential of them in constructing soil prediction models; and 3) explore one accurate and robust way to construct soil spectral models through combining the suitable models and the useful auxiliary information. Through analyzing the potential of environmental factors and spatial characteristics of SOM in constructing soil spectral models, this study could throw light on the precision digital soil mapping.

2 Materials and Methods

2.1 Study area

The study region is located at northwest of Honghu City, Hubei Province, China (29.87°N to 30.02°N, 113.11°E to 113.72°E). Honghu is representative of Jianghan Plain, which is an alluvial plain with the mean annual precipitation of 1150 mm and air temperature of 16℃, and it is one vital grain and cotton production area of China. Honghu has a subtropical humid monsoon climate including obvious features of continental climate, and it is a flat water land as the elevation is mainly distributed from 23 to 28 m and the mean value of slope is approximately 0.3°.

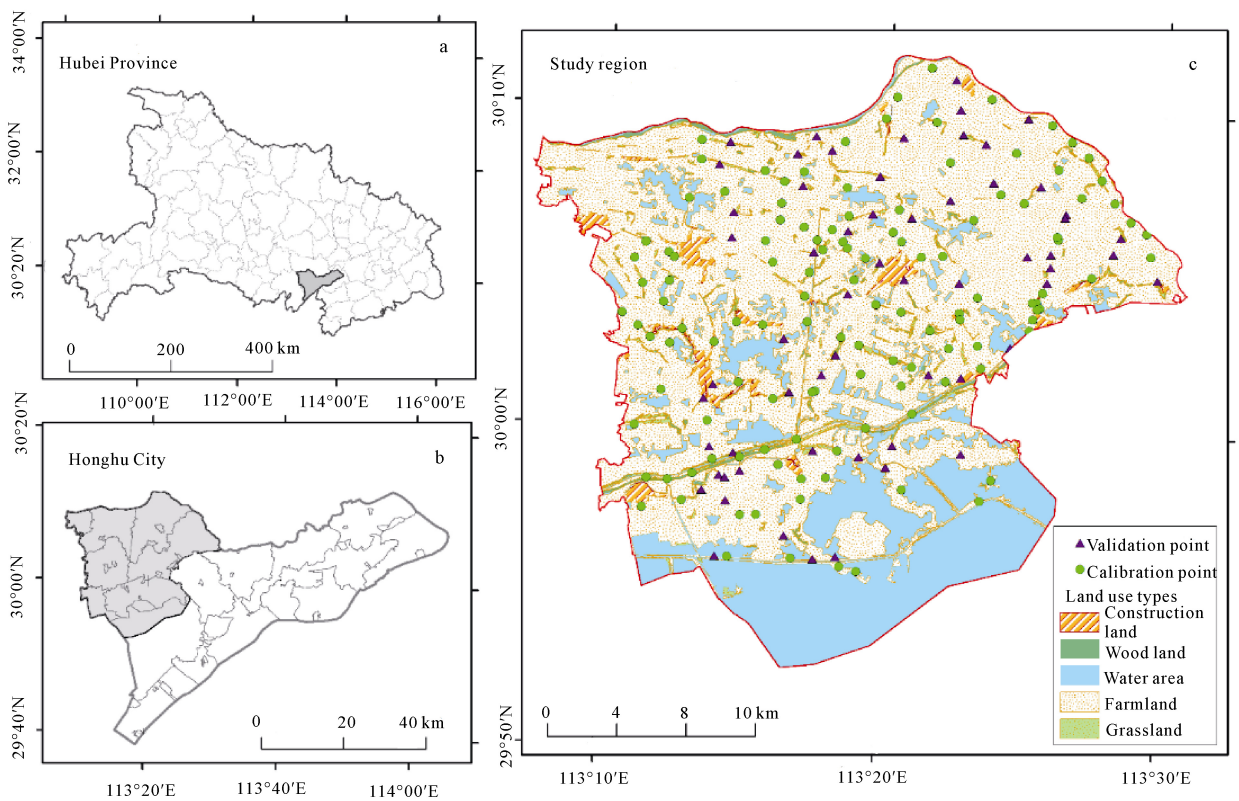

Fig. 1 Spatial distribution of soil sampling plots at the study region

2.2 Soil samples collection

The surface soil samples (n = 180) were collected in July 2014 by random sampling, with a minimum distance between two soil samples greater than 100 m. For each soil sample, a representative soil sample was collected from five mixed soil subsamples taken within a square of 1 m^2. The soil samples were first air-dried in the laboratory at 20℃–25℃ for 14 days. A porcelain mortar was used to break down large aggregates of the soil samples, and a 0.25 mm stainless steel sieve was used to choose the soil granule. The potassium dichromate method was used to measure the SOM contents (Nelson and Sommers, 1974). The soils were found to be diverse and composed of various classes defined by the World Reference Base of Soil Resources, including dystrochrept, typical haplaquept, hapludalf and eutroboralf (FAO, 1998). The spatial distribution of the sampling sites is shown in Fig. 1.

2.3 Soil auxiliary variables

2.3.1 VNIR spectrum collection and pre-processing

An ASD FieldSpec3 portable spectral radiometer was used to measure the spectral reflectance of soils in the VNIR (350–2500 nm) range. The operation procedure of the VNIR spectrum collection can reference our published papers (Guo et al., 2017b). A Matlab toolbox (ROBPCA) can detect and eliminate the outliers from the original datasets (Hubert et al., 2005). First, the reflectance spectra were reduced to 400–2350 nm to eliminate the noise at spectral edges. The Savitzky-Golay smoothing (SG) method with a moving window of 15 nm was used to smooth the reflectance curves. And then, the 1st derivative was used to transform the spectral reflectance, and the data were detrended (Detrend) to remove the linear trends. Finally, standard normal variate (SNV) was used as one typical example of scatter-corrective method to remove undesired scatter or particle-size information from spectra reflectance to some extent (Bendini et al., 2007). After preprocessing, the VNIR spectral reflectance will be used as auxiliary variables to construct the soil prediction models by PLSR and GWR.

2.3.2 Soil covariates collection and pre-processing

In this paper, multiple soil covariates were selected to construct the soil prediction models, including 5items of terrain variables (elevation, slope, aspect, roughness and relief amplitude (RDLS)) generated from the Global Digital Elevation Model Version 2 (GDEM V2), 3 artificial influence factors i.e., the index of land use types (ILUT), nearest distance to road (NDR) and nearest distance to construction land (NDC), derived from the global land cover mapping at 30 m resolution (GlobalLand30, http://www.globallandcover.com/GLC30Download/index.aspx), and 8 vegetation indexes, i.e., normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), atmospherically resistant vegetation index (ARVI), soil-adjusted vegetation index (SAVI), modified soil-adjusted vegetation index (MSAVI), normalized difference water index (NDWI), land surface water index (LSWI), and modification of normalized difference water index (MNDWI), calculated by the Landsat 8 OLI image (LC81230392014278LGN00, stripe number: 123, line number: 39, time: 06 October 2014). The definition of ILUT can reference the research of Zhang et al. (2013). Euclidean distance was used to measure the nearest distance from soil sample points to road and construction. Lastly, Pearson's correlation, variance inflation factor (VIF) and stepwise linear regression were used to reduce the dimensionality and remove multicollinearity of these auxiliary variables. Then, 6 factors (Elevation, ILUT, LSWI, EVI, RDLS, and NDC) were chosen as the auxiliary variables of the MLR and GWR prediction models. To ensure that the units of the SOM and its auxiliary variables were the same, the Zero-Mean normalization method was used to transform the original datasets of the auxiliary factor, such as the environmental factors and the spectral factors (Rai et al., 2005), and the transformed dataset will be used to construct the SOM prediction models.

2.4 Prediction methods

A total of 180 soil samples were separated into a calibration dataset (123, 2/3) and a validation dataset (61, 1/3). The environmental factors and the VNIR spectral data were used as the auxiliary variables, and then synthetically or separately construct the prediction models of SOM by PLSR and GWR. The environmental factors included elevation, ILUT, LSWI, EVI, RDLS, and NDC. After pre-processing of the auxiliary data, MLR1 and GWR1 only used the environmental factors, and PLSR2 and GWR2 only used the VNIR data, and MLR3 and GWR3 used the combination of the environmental factors and the VNIR data. At last, the degree of fitting (R^2), the root mean square error (RMSE), and improve-

ment rates (IR) were used to check the prediction accuracy of different models, and then recommend the suitable modeling strategy. There were many published papers have introduced the basic theories of MLR, PLSR and GWR, and more additional information can reference these papers (Liu et al., 2015; Guo et al., 2017b).

2.5 Evaluation indexes

In this paper, R^2, *RMSE* and *IR* were used to evaluate the performance of different prediction models: the coefficient of determination of calibration (R^2C) and the root mean square error of calibration (*RMSE*C) were used to evaluate these models' modeling ability, and the coefficient of determination of prediction (R^2P) and the root mean square error of prediction (*RMSE*P) were used to evaluate the prediction performance of these models. Also, the IR index between different models were calculated by the *RMSE*P. These functions can be described as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i = \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i = \overline{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

$$IR = \frac{RMSE\mathrm{P} - RMSE\mathrm{P}_{basg}}{RMSE\mathrm{P}_{basg}} \quad (3)$$

where n is the number of samples, y_i and $\hat{y}_i$ are the measured and the predicted SOM for the sample i, respectively, and $\overline{y}$ is the mean value of the measured values. $RMSE\mathrm{P}_{basg}$ is the *RMSE*P of the base model.

3 Results

3.1 Basic statistics of SOM and its auxiliary variables

The distribution of SOM along with some of the predictors was described using classical descriptive statistics (Table 1). SOM values ranged from 17.34 to 50.50 g/kg, and the mean value was 30.12 g/kg and the median value was 28.96 g/kg. The standard variance was 6.69 g/kg and the coefficient of variation (*CV*) was 22.21%, suggesting that SOM had a moderate spatial variation based on Wilding (1985). Six variables were selected as the predictors to construct the SOM prediction models. Table 1 shows the basic statistics of the other predictors. The mean values of elevation, ILUT, LSWI, EVI, RDLS and NDC were 27.47 m, 3.06, 0.08, 0.12, 9.28 and 2390.67 m, respectively. The *CV*s of these auxiliary variables indicated that all of these predictors had a different degree of spatial variation across the study region. NDC had the greatest variation (75.71%) and ILUT had the least variation (13.35%). More detailed basic statistics of SOM and its environmental factors can be found in Table 1.

3.2 Relationship between SOM content and spectral reflectance

Fig. 2 shows the raw reflectance and the transformed reflectance curves of the soil samples. The original spectral reflectance ranged from 0.08 to 0.55, and the reflectance curves had an increasing trend over the range of 410 to 1300 nm and later fluctuated between1300 and 2300 nm (Fig. 2a), consistent with the typical characteristics of soil spectra. The raw spectral reflectance had an apparent absorption near 1350, 1850, and 2300 nm, which were strengthened in the

Table 1 The basic statistics of the soil organic matter and its auxiliary variables

Variable	Range	Minimum	Mean	Maximum	Median	Standard variance	*CV* (%)
SOM (g/kg)	33.15	17.34	30.12	50.50	28.96	6.69	22.21
Elevation (m)	44.00	5.00	27.47	49.00	28.00	8.49	30.91
ILUT	2.48	1.35	3.06	3.82	3.18	0.41	13.35
LSWI	0.19	–0.02	0.08	0.17	0.08	0.04	49.92
EVI	0.19	0.02	0.12	0.21	0.12	0.04	29.07
RDLS	18.00	2.00	9.28	20.00	9.00	3.28	35.34
NDC (m)	7336.17	0.00	2390.67	7336.17	2295.18	1810.01	75.71

Notes: SOM: soil organic matter, ILUT: the index of land use type, LSWI: land surface water index, EVI: enhanced vegetation index, NDC: the nearest distance to construction land, RDLS: roughness and relief amplitude, *CV*: coefficient of variance

transformed spectral reflectance (Figs. 2a and 2b), related to soil water (Brown et al., 2006). Pre-processing of the spectrum can remove the drifting and extend the range of the spectral reflectance (from 0.09–0.55 to –1.50–1.60), which can also strength the variation of these absorption wavelengths that have a relationship with SOM, enhancing the robustness and reliability of the prediction models. The raw spectral reflectance had an obvious increasing trend with decreased SOM values in Fig. 3a, and there was an empirical relationship between the SOM and the reflectance, which can be quantified by the Pearson's correlation coefficients. The coefficients between the SOM values and the raw spectral reflectance in different wavelengths ranged from –0.53 to –0.26 (Fig. 3a). The lowest values were observed at approximately 600 nm (–0.53) and the higher values were observed at 400 nm (–0.26) and 2150 nm (–0.37). Relative to the relationship between the SOM and the transformed spectral reflectance, the Pearson's correlation coefficients ranged from –0.50 to 0.50. Strong relationships between the spectral reflectance and the SOM content were observed at 500 nm (–0.50), 700 nm (0.50), 1300 nm (–0.10) and from 1850 nm (–0.45) to 1950 nm (0.45) (Brown et al., 2006).

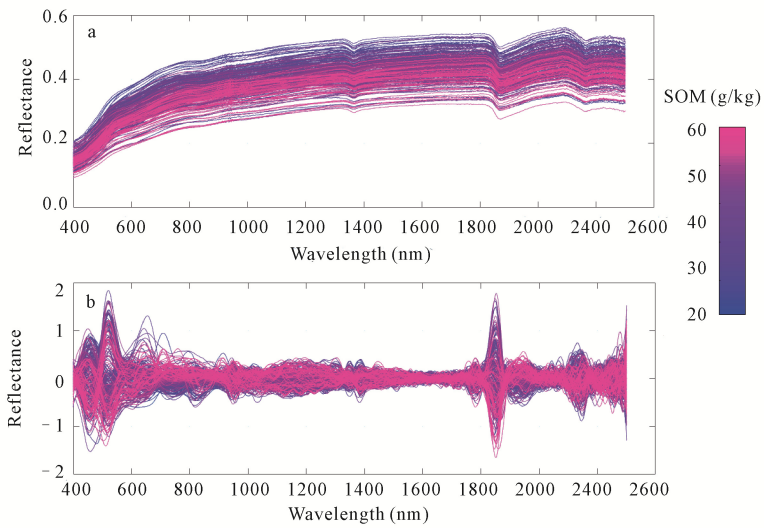

Fig. 2 The raw spectral reflectance (a) and the transformed spectral reflectance (b) of different SOM values for the wavelength from 400 to 2300 nm

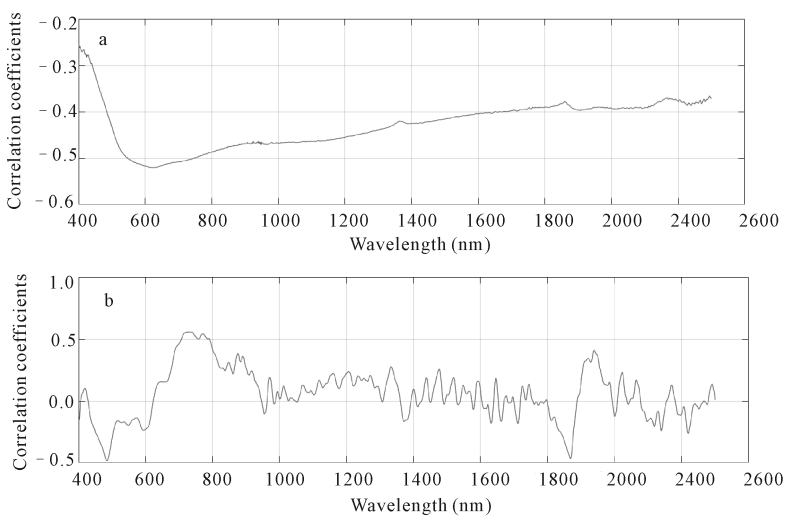

Fig. 3 Pearson's correlation coefficients between spectral reflectance and SOM. (a), original spectral reflectance; (b), spectral reflectance after pre-processing

3.3 Prediction models of SOM by MLR and GWR

3.3.1 SOM prediction with environmental variables

The parameters of SOM prediction models by MLR and GWR, including the coefficients and identification indexes, are shown in Table 2. In MLR, the descending order of the absolute mean values of these auxiliary variables were LSWI (–2.90), Elevation (–1.06), ILUT (1.05), EVI (–0.92), RDLS (–0.74), NDC (0.55) (higher values indicate greater influence on SOM). LSWI had the greatest influence among all of the factors, whereas NDC had the smallest influence. Meanwhile, the coefficients of Elevation, LSWI, EVI and RDLS were less than 0, suggesting a negative relationship between them and SOM, while the other variables had positive relationships. The results of VIF also showed that no multicollinearity existed among these soil covariates, as all VIF values were less than 7.5. Because GWR is a local spatial model, the coefficients of the explanatory variables were not constant and varied with the change of geographical locations, and these can show the detailed degree of influence of environmental factors on SOM in different geographical locations.

The descending order of the absolute mean values of the coefficients was LSWI (–2.70), ILUT (1.14), Eleavation (–1.09), NDC (0.93), EVI (–0.90), and RDLS (–0.44). LSWI had the strongest influence and RDLS had the smallest influence among all factors. In Table 2, the range field showed the difference between the minimum and the maximum coefficients of a variable, and the standard error (Std) and *CV* field showed the degree of variation of the coefficients of a variable. EVI had the greatest values of range (2.06), Std (0.64) and *CV* (71.78%), suggesting that EVI had the greatest difference in the estimated coefficients over geographical locations. Elevation had the smallest values of range (0.55), Std (0.14) and *CV* (12.46%), suggesting that the coefficients of elevation were similar across the study region. The descending order of the degrees of variation of the coefficients of auxiliary variables was EVI, NDC, LSWI, RDLS, ILUT, and Elevation. Thus, the coefficients estimated by GWR directly reflected the detailed relationships between environmental factors and SOM in different geographical locations.

3.3.2 SOM prediction with spectral reflectance

PLSR and GWR were used to construct the prediction models of SOM by using the spectral reflectance, and the parameters are shown in Table 3. The first four PCs of the spectral reflectance were chosen as the auxiliary variables, since the accumulating contribution rate was 80.85% and the eigenvalues were bigger than 1. The coefficients of intercept, PC1, PC2, PC3 and PC4, were 30.73, 3.49, 2.90, 2.92, and 2.26, respectively. PC1 had the biggest influence on SOM among all PCs. Because the datasets had been transformed, a *p*-value and VIF value equal to 0 and a *t*-value greater than 1 would suggest a statistically significant role of the PCs in the PLSR model. The PCs also had different influences on the SOM in different geographical locations based on the coefficients in GWR. The descending order of the mean values was PC1 (3.57), PC3 (3.01), PC2 (2.92) and PC4 (2.33). PC1 had the strongest influence on SOM, and PC4 had the smallest influence. Meanwhile, coefficients of PC3 had the greatest degree of variation (*CV* = 10.61%), and coefficients of PC1 had the smallest variation (*CV* = 5.10%). All of these PCs had a positive relationship with SOM since their coefficients were positive.

Table 2 Parameters of multiple linear regression model (MLR) and geographically weighted regression (GWR) constructed by the environmental variables

Variables	MLR				GWR					
	Coefficients	*t*	*P*	VIF	Min	Mean	Max	*SE*	Range	*CV* (%)
Intercept	30.73	58.40	0	–	30.59	30.93	31.14	0.19	0.55	0.62
Elevation	–1.06	–1.93	0.06	1.09	–1.38	–1.09	–0.83	0.14	0.55	12.46
ILUT	1.05	1.72	0.09	1.33	0.70	1.14	1.51	0.26	0.81	22.53
LSWI	–2.90	–4.87	0	1.27	–3.14	–2.70	–2.08	0.25	1.06	9.32
EVI	–0.92	–1.43	0.16	1.49	–1.91	–0.90	0.16	0.64	2.06	71.78
NDC	0.55	1.02	0.31	1.04	0.24	0.93	1.65	0.39	1.41	41.97
RDLS	–0.74	–1.34	0.18	1.09	–0.74	–0.44	–0.04	0.16	0.70	36.54

Notes: VIF: variance inflation factor; *SE*: standard error, *CV*: coefficient of variation; ILUT: the index of land use type, LSWI: land surface water index, EVI: enhanced vegetation index, NDC: the nearest distance to construction land, RDLS: roughness and relief amplitude

Table 3 Parameters of Partial least squares regression model (PLSR) and Geographically weighted regression model (GWR) constructed by the spectral reflectance

Variables	PLSR						GWR					
	Coefficients	*t*	*P*	VIF	ACR	EIG	Min	Mean	Max	*SE*	Range	*CV* (%)
Intercept	30.73	96.29	0	0	–	–	30.60	30.74	30.84	0.06	0.24	0.21
PC1	3.49	10.89	0	0	52.24	48.30	3.22	3.57	3.84	0.18	0.62	5.10
PC2	2.90	9.04	0	0	72.80	15.40	2.56	2.92	3.25	0.21	0.70	7.19
PC3	2.92	9.11	0	0	78.81	4.35	2.44	3.01	3.32	0.32	0.87	10.61
PC4	2.26	7.04	0	0	80.85	3.57	2.11	2.33	2.58	0.15	0.47	6.51

Notes: *SE*: standard error, VIF: variance inflation factor, ACR: Accumulating Contribution rates, EIG: Eigenvalues, *CV*: coefficient of variation, PC: principal component

3.3.3 SOM prediction with the combination of spectral and environmental variables

The next step involved combining the spectral reflectance and environmental variables to construct the prediction models of SOM by MLR and GWR. The parameters of MLR and GWR are shown in Table 4. The absolute values of the coefficient of these auxiliary variables in descending order in MLR were PC2 (3.14), PC1 (2.76), PC3 (2.66), PC4 (2.29), ILUT (1.38), Elevation (–0.85), LSWI (–0.64), EVI (–0.39), RDLS (–0.32) and NDC (–0.11). All of the VIF values were less than 7.0, suggesting that there was no multicollinearity among these auxiliary variables. The descending order of the absolute mean values of these coefficients in GWR was same with MLR.

However, this order was uncertain at specific geographical locations given that the minimum and maximum values of these coefficients were different. Thus, the spatial information of the soil sample points plays an important role in estimating the degree of influence of different auxiliary variables in a specific location. The PCs had a greater influence on SOM than the environmental factors, because the spectral reflectance can directly reflect the physical and chemical structure of soil, whereas the environmental factors take a long time to influence and change the soil structures. The *CV*s of PCs were smaller than the *CV*s of the environmental factors. The main reason was that the spatial variation of soil was influenced by the complexity and variation of the environmental factors, so it can hardly be represented completely by a limited set of environmental factors. In contrast, the spectral reflectance was able to significantly explain the physical and chemical structures of soil that led to the spatial variation of the spectral reflectance. Additionally, the result of *CV* showed that the geographical locations had a smaller influence on

Table 4 The parameters of Multiple linear regression model (MLR) and Geographically weighted regression model (GWR) constructed by the environmental variables and the spectral reflectance

Variables	MLR					GWR					
	Coefficients	*SE*	*t*	*P*	VIF	Min	Mean	Max	*SE*	Range	*CV* (%)
Intercept	30.73	0.30	102.70	0	–	30.61	30.70	30.76	0.03	0.15	0.10
PC1	2.76	0.36	7.71	0	1.42	2.68	2.93	3.19	0.12	0.51	4.03
PC2	3.14	0.34	9.28	0	1.27	2.89	3.15	3.44	0.18	0.55	5.67
PC3	2.66	0.35	7.65	0	1.34	2.39	2.66	3.04	0.18	0.65	6.93
PC4	2.29	0.31	7.42	0	1.06	1.90	2.33	2.72	0.27	0.82	11.36
Elevation	–0.85	0.32	–2.66	0.01	1.14	–1.11	–0.81	–0.52	0.18	0.59	22.45
ILUT	1.38	0.38	3.61	0	1.62	0.90	1.26	1.53	0.14	0.62	11.21
LSWI	–0.64	0.39	–1.64	0.11	1.69	–0.91	–0.70	–0.56	0.09	0.35	12.72
EVI	–0.39	0.37	–1.05	0.29	1.54	–0.65	–0.36	0.03	0.23	0.68	63.17
NDC	–0.11	0.31	–0.34	0.74	1.10	–0.47	–0.26	–0.12	0.09	0.35	34.19
RDLS	–0.32	0.32	–1.00	0.32	1.15	–0.59	–0.37	–0.13	0.11	0.46	30.76

Notes: *SE*: standard error; VIF: variance inflation factor; *CV*: coefficient of variation; PC: principal component; ILUT: the index of land use type, LSWI: land surface water index, EVI: enhanced vegetation index, NDC: the nearest distance to construction land, RDLS: roughness and relief amplitude

the spectral reflectance than on the environmental factors. This also suggests that the prediction model using both the environmental factors and the spectral reflectance are promising tools to predict the soil properties, as it not only ensures the robustness and reliability of the prediction models but also fully considers the spatial characteristics of soil properties.

3.4 Model validation and evaluation

Three series of prediction models were built, respectively, by the environmental factors, spectral reflectance and their integration data. First, only the environmental factors were taken as the explanatory variables to construct the MLR1 and GWR1 for the prediction of SOM, and the modeling and predictive abilities of these models were lower than that of other prediction models in Table 5. Next, the spectral reflectance was used as the explanatory variables to construct PLSR2 and GWR2, and the modeling abilities of PLSR2 and GWR2 increased by 38.68% and 40.57% compared to MLR1 based on the values of IR. At last, the spectral reflectance and environmental factors were combined to construct MLR3 and GWR3, and these models had a great improvement compared to previous models based on the values of *RMSE* and R^2.

This showed that all of the environmental factors and the spectral reflectance play an important role in predicting SOM. While, the relationship between the complex environmental factors and soil properties cannot be easily represented by a single linear regression model at different geographical locations, and the environmental factors may confuse the predicted ability of the prediction models. When we compared the performance of these models in inter-class, the spatial characteristics of soil properties played important roles to improve the modeling abilities (IR equals approximately 3%), since GWR is a local spatial regression model which considers the spatial dependence of the soil samples in different geographical locations. The comprehensive analysis suggested that there were two main factors that influenced the modeling and predicted abilities of the prediction models: one was the auxiliary variables, and the other was the spatial characteristics of the soil samples. Thus, combining the spectral reflectance and suitable environmental factors together to construct the local spatial models was a valid and efficient strategy, as demonstrated in this paper.

4 Discussion

SOM plays an important role in the precision agriculture and the carbon cycle of the ecosystem. It is one meaningful and valuable research to quickly obtain the SOM content and soil carbon source. In nature conditions, the environmental factors and human activities can influence the pedogenesis, development and degradation of soil. Thus, the soil environmental factors always be used as the auxiliary variables in the prediction of soil properties and the digital soil mapping. Due to the spatial variability and uncertain of SOM, previous studies indicated that the performance of prediction models was not very good when only using environmental factors as the auxiliary variables (Wang et al., 2013; Jaber and Al-Qinna, 2015; Kumar, 2015). To improve the quality of the prediction result, the VNIR spectral data were integrated into the prediction models. VNIR can respond to the chemical and physical structures of the soil. The distinct absorption features over the VNIR regions have

Table 5 The modeling and predicted abilities of prediction models based on evaluation indexes

Auxiliary variables	Models	R^2C	*RMSE*C	R^2P	*RMSE*P	Improvement rates (%)	
						Inter-class	Global
Environmental factors	MLR1	0.301	5.654	0.315	5.316	–	–
	GWR1	0.345	5.484	0.255	5.801	3.01	3.01
Spectral reflectance	PLSR2	0.738	3.467	0.611	3.980	–	38.68
	GWR2	0.754	3.36	0.668	3.717	3.09	40.57
Comprehensive factors	MLR3	0.782	3.146	0.635	3.950	–	44.36
	GWR3	0.800	3.028	0.641	3.981	3.75	46.44

Notes: PLSR: partial least squares regression model, MLR: multiple linear regression model, GWR: geographically weighted regression model, R^2C: the coefficient of determination of calibration, R^2P: the coefficient of determination of prediction, *RMSE*C*:* the root mean square error of calibration model, *RMSE*P*:* the root mean square error of prediction, Inter-class: the comparison within the same auxiliary variables, Global: the comparison to the *RMSE*P of MLR1

a strong relationship with the organic matter, minerals and hydrogen groups in soil. Many studies have shown the feasibility of VNIR data in predicting soil properties with a non-destructive and cost-efficient alternative (Al-Asadi and Mouazen, 2014; Cambou et al., 2016). Our study showed the combination of the environmental factors and spectral reflectance can improve the prediction accuracy of SOM relative to separately use them.

Additionally, the spatial heterogeneity and dependence are important theoretical bases to construct the prediction models of SOM. Many scholars have proved that spatial dependence and heterogeneity exist in soil properties and environmental factors, and successfully used this theoretical knowledge to improve the performances of prediction models. Zhang et al. (2011) demonstrated the performance of GWR relative to OK, IDW and MLR in predicting SOC based on land cover, rainfall and soil type in Ireland. Kumar et al. (2013) showed the prediction accuracy of GWR was better than MLR. The main reason was that GWR had a better ability to capture the spatial characteristics of SOC compared to other global statistical techniques.

The spectral reflectance can reflect the physical and chemical structure of soil, and the soil properties have strong spatial dependence especially at local and fine scales. Thus, we inferred that the spectral reflectance of soil and its transformational PCs also contained spatial autocorrelation. Conforti et al. (2015) considered the independent and identical distribution of the predicted residuals by using PLSR combined with a linear mixed effect model (LMEM) based on the laboratory-based soil VNIR spectra. Ge et al. (2007) used the regression-kriging (RK) method to consider the spatial dependence of spectral models based on the lab soil reflectance spectra. Additionally, many studies have indicated that the spatial dependence also existed in the spectral reflectance, and the prediction accuracy can be improved by the local spatial weighted regression models (Guo et al., 2017b). In our paper, the spatial characteristics of environmental factors and spectral reflectance were separately used to construct the prediction models based on the GWR model. Our results highlighted the importance of spatial heterogeneity and dependence of soil properties in constructing the prediction models.

Also, many challenges and problems exist in the process of modeling. The pre-processing of environmental factors and VNIR data is one important task to ensure the quality and accuracy of soil spectral models. It is one hard work to obtain the VNIR data of the continuous or high density soil samples. The relationships between the VNIR data and various environmental factors should be further explored.

5 Conclusions

Six environmental factors (the index of land use type, land surface water index, enhanced vegetation index, the nearest distance to construction land and the roughness and relief amplitude) chosen from 16 candidate environmental factors were used as the environmental variables. The first four PCs of the spectral reflectance extracted from the partial least square regression (PLSR) were chosen as the spectral variables. The six environmental factors and four PCs were respectively used to construct the prediction models of multiple linear regression (MLR1), PLSR2, geographically weighted regression (GWR1) and GWR2, also the environmental and spectral variables were together used to construct MLR3 and GWR3.

(1) When the environmental and spectral variables were respectively used to construct the prediction model, the spectral reflectance can obtain the better prediction results than the environmental factors, and the prediction accuracy can be improved about 40%.

(2) The local spatial weighted regression model of GWR considered the spatial characteristics of SOM and auxiliary variables data in the prediction of soil properties, and GWR can improve about 3.00% prediction accuracy relative to the global linear regression models (MLR and PLSR).

(3) When using the fusion data of environmental factors and VNIR data as the auxiliary variables, the prediction models achieved the best prediction accuracy than other models.

References

Al-Asadi R A, Mouazen A M, 2014. Combining frequency domain reflectometry and visible and near infrared spectroscopy for assessment of soil bulk density. *Soil & Tillage Research*, 135: 60–70. doi: 10.1016/j.still.2013.09.002

Bendini A, Cerretani L, Di Virgilio F et al., 2007. In process monitoring in industrial olive mill by means of FT-NIR.

European Journal of Lipid Science and Technology, 109(5): 498–504. doi: 10.1002/ejlt.200700001

Brown D J, Shepherd K D, Walsh M G et al., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, 132(3): 273–290. doi: 10.1016/j.geoderma.2005.04.025

Cambou A, Cardinael R, Kouakoua E et al., 2016. Prediction of soil organic carbon stock using visible and near infrared reflectance spectroscopy (VNIRS) in the field. *Geoderma*, 261: 151–159. doi: 10.1016/j.geoderma.2015.07.007

Conforti M, Castrignano A, Robustelli G et al., 2015. Laboratory-based Vis-NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content. *Catena*, 124: 60–67. doi: 10.1016/j.catena.2014.09.004

Evrendilek F, Celik I, Kilic S, 2004. Changes in soil organic carbon and other physical soil properties along adjacent Mediterranean forest, grassland, and cropland ecosystems in Turkey. *Journal of Arid Environments*, 59(4): 743–752. doi: 10.1016/j.jaridenv.2004.03.002

FAO, 1998. *World Reference Base for Soil Resources*. Rome: Food and Agriculture Organization of the United Nations.

Gaetan C, Guyon X, Bleakley K, 2010. *Spatial Statistics and Modeling*. Springer, 90.

Ge Y, Thomasson J A, Morgan C L et al., 2007. VNIR diffuse reflectance spectroscopy for agricultural soil property determination based on regression-kriging. *Transactions of the Asabe*, 50(3): 1081–1092. doi: 10.13031/2013.23122

Guo L, Chen Y, Shi T et al., 2017a. Exploring the role of the spatial characteristics of visible and near-infrared reflectance in predicting soil organic carbon density. *ISPRS International Journal of Geo-Information*, 6(10): 308. doi: 10.3390/ ijgi6100308

Guo L, Linderman M, Shi T et al., 2018. Exploring the sensitivity of sampling density in digital mapping of soil organic carbon and its application in soil sampling. *Remote Sensing*, 10(6): 888. doi: 10.3390/rs10060888

Guo L, Zhao C, Zhang H et al., 2017b. Comparisons of spatial and non-spatial models for predicting soil carbon content based on visible and near-infrared spectral technology. *Geoderma*, 285: 280–292. doi: 10.1016/j.geoderma.2016.10.010

Gupta D D, 2015. Soils as launching pad for healthy society and humannity-reality and not myth. *International Journal Environmental & Agricultural Science,* 1(2): 37–45.

Hartemink A E, McBratney A, de Lourdes M M, 2008. *Digital Soil Mapping with Limited Data*. Springer Science & Business Media, 250–251.

Hubert M, Rousseeuw P J, Vanden Branden K, 2005. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1): 64–79. doi: 10.1198/004017004000000563

Jaber S M, Al-Qinna M I, 2015. Global and local modeling of soil organic carbon using Thematic Mapper data in a semi-arid environment. *Arabian Journal of Geosciences*, 8(5): 3159–3169. doi: 10.1007/s12517-014-1370-6

Kumar S, 2015. Estimating spatial distribution of soil organic carbon for the Midwestern United States using historical database. *Chemosphere*, 127: 49–57. doi: 10.1016/j.chemosphere.2014.12.027

Kumar S, Lal R, Liu D S et al., 2013. Estimating the spatial distribution of organic carbon density for the soils of Ohio, USA. *Journal of Geographical Sciences*, 23(2): 280–296. doi: 10.1007/s11442-013-1010-1

Lagacherie P, 2008. *Digital Soil Mapping*: *A State of the Art*. Springer, 3–14.

Liu Y, Guo L, Jiang Q et al., 2015. Comparing geospatial techniques to predict SOC stocks. *Soil and Tillage Research*, 148: 46–58. doi: 10.1016/j.still.2014.12.002

Mouazen A, Kuang B, De Baerdemaeker J et al., 2010. 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*, 158(1): 23–31.

Peon J, Fernandez S, Recondo C et al., 2017. Evaluation of the spectral characteristics of five hyperspectral and multispectral sensors for soil organic carbon estimation in burned areas. *International Journal of Wildland Fire*, 26(3): 230–239. doi: 10.1071/wf16122

Rai P, Majumdar G, DasGupta S et al., 2005. Prediction of the viscosity of clarified fruit juice using artificial neural network: a combined effect of concentration and temperature. *Journal of Food Engineering*, 68(4): 527–533. doi: 10.1016/j.jfoodeng.2004.07.003

Rossel R A V, Webster R, 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. *European Journal of Soil Science*, 63(6): 848–860. doi: 10.1111/j.1365-2389.2012.01495.x

Roudier P, Hedley C B, Lobsey C R et al., 2017. Evaluation of two methods to eliminate the effect of water from soil vis–NIR spectra for predictions of organic carbon. *Geoderma*, 296: 98–107. doi: https://doi.org/10.1016/j.geoderma.2017.02.014

Schmidt M W, Torn M S, Abiven S et al., 2011. Persistence of soil organic matter as an ecosystem property. *Nature*, 478(7367): 49–56. doi: 10.1038/nature10386

Shekhar S, Xiong H, 2008. *Encyclopedia of GIS*. Springer Science & Business Media, 60–61.

Shi Z, Wang Q, Peng J et al., 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Science China Earth Sciences*, 57(7): 1671–1680. doi: 10.1007/s11430-013-4808-x

Terra F S, Demattê J A M, Viscarra Rossel R A, 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis-NIR and mid-IR reflectance data. *Geoderma*, 255–256: 81–93. doi: 10.1016/j.geoderma.2015.04.017

Trangmar B B, Yost R S, Uehara G, 1985. Application of geostatistics to spatial studies of soil properties. *Advances in agronomy*, 38(1): 45–94. doi: 10.1016/S0065-2113(08)60673-2

Viscarra Rossel R A, Hicks W S, 2015. Soil organic carbon and its fractions estimated by visible-near infrared transfer functions. *European Journal of Soil Science*, 66(3): 438–450. doi: 10.1111/ejss.12237

Wang K, Zhang C, Li W, 2013. Predictive mapping of soil total

nitrogen at a regional scale: a comparison between geographically weighted regression and cokriging. *Applied Geography*, 42: 73–85. doi: 10.1016/j.apgeog.2013.04.002

Wilding L, 1985. Spatial variability: its documentation, accommodation and implication to soil surveys. Soil spatial variability. Workshop.

Zhang C, Tang Y, Xu X et al., 2011. Towards spatial geochemical modelling: use of geographically weighted regression for mapping soil organic carbon contents in Ireland. *Applied Geochemistry*, 26(7): 1239–1248. doi: 10.1016/j.apgeochem.2011.04.014

Zhang Haitao, Guo Long, Chen Jiaying et al., 2013. Modeling of spatial distributions of farmland density and its temporal change using geographically weighted regression model. *Chinese Geographical Sci*ence, 24 (2): 191–204. doi: 10.1007/s 11769-013-0631-8

Zornoza R, Mataix-Solera J, Guerrero C et al., 2007. Evaluation of soil quality using multiple lineal regression based on physical, chemical and biochemical properties. *Science of the Total Environment*, 378(1): 233–237. doi: 10.1016/j.scitotenv.2007.01.052