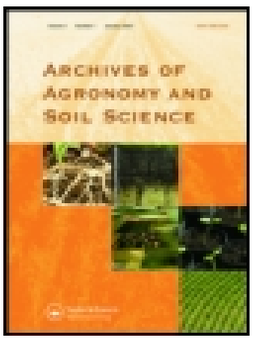

Archives of Agronomy and Soil Science

ISSN: 0365-0340 (Print) 1476-3567 (Online) Journal homepage: https://www.tandfonline.com/loi/gags20

Applying statistical methods to map soil organic carbon of agricultural lands in northeastern coastal areas of China

Zhenxing Bian, Xiaoyu Guo, Shuai Wang, Qianlai Zhuang, Xinxin Jin, Qiubing Wang & Shuhai Jia

To cite this article: Zhenxing Bian, Xiaoyu Guo, Shuai Wang, Qianlai Zhuang, Xinxin Jin, Qiubing Wang & Shuhai Jia (2019): Applying statistical methods to map soil organic carbon of agricultural lands in northeastern coastal areas of China, Archives of Agronomy and Soil Science, DOI: 10.1080/03650340.2019.1626983

To link to this article: https://doi.org/10.1080/03650340.2019.1626983

Accepted author version posted online: 31 May 2019.
Published online: 09 Jun 2019.

ARCHIVES OF AGRONOMY AND SOIL SCIENCE
https://doi.org/10.1080/03650340.2019.1626983

Applying statistical methods to map soil organic carbon of agricultural lands in northeastern coastal areas of China

Zhenxing Bian[a], Xiaoyu Guo[a], Shuai Wang[a,b], Qianlai Zhuang[b], Xinxin Jin[a,b], Qiubing Wang[a] and Shuhai Jia[a]

[a]College of Land and Environment, Shenyang Agricultural University, Shenyang, Liaoning Province, China; [b]Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, IN, USA

ABSTRACT

Soil organic carbon (SOC) is an important indicator to evaluate agricultural soil quality. Precise mapping SOC can help to facilitate soil and environmental management decisions. This study applied multiple stepwise regression (MSR), boosted regression trees (BRT) model, and boosted regression trees hybrid residuals kriging (BRTRK) to map SOC of agricultural lands in Wafangdian City, northeastern China. A 10-fold cross-validation procedure was used to evaluate the performance of the three models. The BRTRK method exhibited the best predictive performance and explained 78% of the total SOC variability. The distribution of SOC was mainly explained by elevation, followed by soil-adjusted vegetation index (SAVI), and topographic wetness index (TWI). We conclude that the BRTRK was the most accurate method in predicting spatial distribution of SOC. In addition, our study indicated that topographic variables as key factors to affect SOC should be considered in future SOC mapping.

ARTICLE HISTORY
Received 20 December 2018
Accepted 30 May 2019

KEYWORDS
Digital soil mapping; environmental variables; soil organic carbon; coastal areas

Introduction

Soil organic carbon (SOC) and its spatial distribution characteristics are important indicators of soil quality and soil health (Batjes 1996), which directly determine soil fertility and plant productivity (Guo et al. 2015). Crop systems are a potential carbon sink in their soils (Batjes 1996). However, minor changes in the large amount of SOC could greatly affect atmospheric CO_2 concentrations due to its sensitivity to climate changes and human activities (Lal 2004). Therefore, there has been an interest to understand spatial variations and controlling factors of SOC for soil quality, SOC accounting, and greenhouse gas emission quantification (Adhikari et al. 2014). Furthermore, accurate estimates of SOC are essential for analysing the regional-scale carbon balance of agroecosystems and the global carbon cycle (Baldock et al. 2012).

Wafangdian, which is located in the southwest of Liaoning Provence of northeastern China and rich in hydrothermal resources, has a long history of farming with various ways of farmland utilization (Wang et al. 2016). As a pioneer area of China's reform and opening up, its population grows persistently, urbanization and industrialization and urban infrastructure develops rapidly. Subsequently, farmlands are fragmented, and the landscape has a high spatial heterogeneity (Wang et al. 2016). This region becomes an ideal case study area to explore the changes in SOC content and its dominant factors.

CONTACT Shuai Wang shuaiwangsy@163.com College of Land and Environment, Shenyang Agricultural University, No. 120 Dongling Road, Shenhe District, Shenyang, Liaoning Province 110866, China; Qianlai Zhuang qzhuang@purdue.edu Department of Earth, Atmospheric, and Planetary Sciences, Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA

In agroecosystems, spatial distribution of SOC is affected by natural ecological processes influenced by a number of factors including climate, soil type, topography and land use. Thus, it is still a challenge to accurately map SOC at regional scales (Baldock et al. 2012). However, digital soil mapping (DSM) has been widely used to estimate soil properties including SOC using observational data and environmental variables (McBratney et al. 2003). A number of DSM techniques were constructed using soil-landscape models (Wang et al. 2018) to quantify the relationship between soil properties and easily accessible environment variables (Adhikari et al. 2014). Because the relationship between them is complicated and non-linear (Minasny et al. 2016); thus, machine learning techniques have been widely used in predicting soil properties (Wang et al. 2018).

However, an obvious drawback of these machine learning methods is that they only account for the relationships between SOC and environmental variables, but ignore the influences (spatial autocorrelation) of neighbouring observed data when predicting the spatial distribution of SOC (Guo et al. 2015). To overcome this shortcoming, a new approach of BRT hybrid residual kriging (BRTRK) was proposed to predict and map the spatial distribution pattern of SOC. Actually, BRTRK model can be considered as an extension of BRT model, which has an analogous design idea with the hybrid regression kriging approach.

The hybrid prediction methods have been applied in numerous scientific fields including meteorology, hydrology, remote sensing, environmental science, and soil science (Guo et al. 2015). Compared to the single machine learning method, the hybrid prediction methods are more powerful and efficient (Wang et al. 2016). Thus, the aim of this study is to apply the BRTRK to digitally map the spatial distribution of SOC in the northeastern coastal agroecosystems of China. The specific objectives are to (1) construct a hybrid prediction model; (2) identify environmental controls of SOC content; and (3) validate the performance of the model and analyse its potential applicability.

Materials and methods

Study area

Our study was conducted in Wafangdian City (39.33° to 40.12° N, 121.22° to 122.27° E), which is located in the southwest of Liaoning Provence, northeastern China. It covers a total area of 3,818 km^2, accounting for 71% of the study area with cultivated lands and garden plots. The study area has a warm temperate semi-humid continental monsoon climate with four clearly distinct seasons. The mean annual precipitation is 637 mm, and the mean annual temperature is 9.3 °C. Varied topographic conditions lead to the development of different soil types, including Anthrosols, Argosols, Cambosols, Halosols, and Primosols according to Chinese Soil Taxonomy. According to FAO-WRB classification system (2014), the main soil types are Anthrosols, Cambisols, Histosols, Leptisols and Luvisols. The main formation lithology is metamorphic rocks including quartzite, marble, slate and sort of mixed rocks. Land-use types mainly include woodland, orchard, cultivated land and grasslands.

Soil samples

Soil sampling was conducted on a 1.6 × 1.6 km grid covered the whole study area between August and September (growing season) in 2012 (Figure 1). A total of 1195 soil samples were collected from the topsoil depth (0–20 cm) excluding litter layer, if present. The coordinates and altitude of each sampling site are recorded by a handheld GPS. One kilogram soil samples were contained at each sample site for laboratory analysis. The samples were air-dried and then crushed and finally passed by a 2-mm sieve eliminating non-soil materials like gravel and plant roots. SOC content (g kg^{-1}) was determined by dry combustion using a Vario EL III elemental analyser (Elementar Analysensysteme GmbH, Hanau, Germany) (Wang et al. 2018).

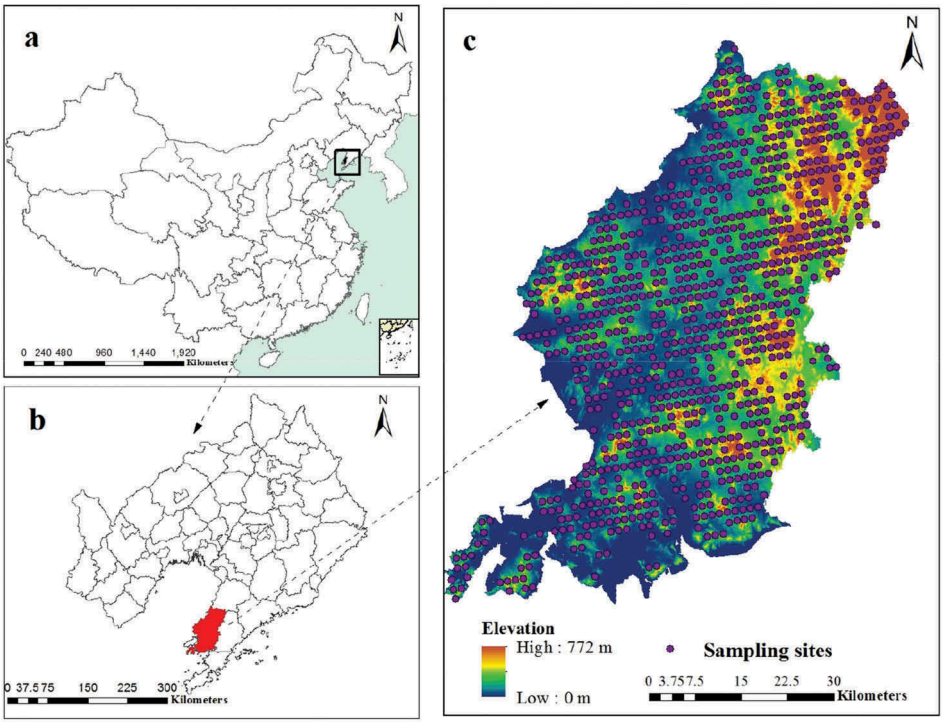

Figure 1. Soil sample locations overlaid on the digital elevation model of the study area (c) in Liaoning Province (b) of China (a).

Land-use and soil type data

Land-use data represents land-use and land-cover types in the study area, which are extracted from a land-use and land-cover map compiled at a cartographic scale of 1:200,000. The data were provided by National Science and Technology Infrastructure of China, Data Sharing Infrastructure of Earth System Science (http://www.geodata.cn). Land-use types were classified into cultivated land, grassland, woodland, and orchard according to the Second National Land Survey and Land Classification System (Ministry of land and resources, China, 2007).

Soil type data were obtained from the Second National Soil Survey of Liaoning Province conducted between 1979 and 1990. In addition, according to the Chinese Soil Taxonomic Classification, the soil types were classified into Anthrosols, Argosols, Cambosols, Halosols, and Primosols.

Environmental variables

Topographic variables

A 30 m resolution of DEM acquired from Shuttle Radar Topography Mission was used to derive topographic variables. Five topographic variables were selected including elevation, slope gradient, slope aspect, catchment area (CA), and topographic wetness index (TWI). Elevation, slope aspect, and slope gradient as the primary terrain attributes were directly derived from ArcGIS 10.1, and the corresponding secondary terrain attributes of TWI and CA were obtained by the System for Automated Geoscientific Analyses (SAGA, Hamburg, Germany) GIS software.

Remote sensing variables

Four variables were derived from the Landsat 5 Thematic Mapper (Landsat 5 TM). The data were acquired from the USGS (https://www.usgs.gov/) between July and September (growing season) in

2012. Vegetation variables were derived from RS data from the International Scientific and Technical Data Mirror Site, Computer Network Information Center, the Chinese Academy of Sciences (http://www.gscloud.cn) between July and September (growing season) in 2012, with cloud cover < 10%. From the RS data three primary vegetation attributes—the visible-red band 3 (B3, 0.63–0.69 μm), near-infrared band 4 (B4, 0.76–0.96 μm) and short-wave infrared band 5 (B5, 1.55–1.75 μm), and a secondary vegetation attributes—Soil Adjusted Vegetation Index (SAVI) was derived. B3, B4, B5 and SAVI were characterized as vegetation growth, coverage, biomass, vegetation cover and type (Adhikari et al. 2014; Yang et al. 2016). SAVI is defined as (Huete 1988):

$$SAVI = [(B4 - B3)(1 + S)]/(\mathrm{B4} + \mathrm{B3} + \mathrm{S}) \tag{1}$$

where 'S' is the soil condition factor, and its value is between 0 and 1. '0' and '1' represent two extreme cases of extremely high and very low vegetation coverage, respectively. Usually, 0.5 is chosen to reduce the background difference of soil and remove the influence of noise and sound.

Modelling approaches

Three modelling approaches were used in this study, including BRT, BRTSK and multiple stepwise regression (MSR). BRT was first proposed by Friedman (2001), which consisted of two statistical techniques of boosting and regression trees (Wang et al. 2016). BRT was a means of optimizing and regularizing the numerical predictions to achieve rapid and accurate prediction of the corresponding variables (Yang et al. 2016). The BRT model was implemented in R environment (R Development Core Team 2013) using R packages 'dismo' version 0.8–17 (Hijmans et al. 2013). Four parameters need to be defined, including the learning rate (LR), tree complexity (TC), bag fraction (BF) and the number of trees (NT). In order to achieve the best prediction performance, several sets of parameter combinations of LR (0.025, 0.05, 0.1, 0.15), TC (5, 8, 9, 10), BF (0.45–0.85) and NT (500, 1000, 1500, 2000) were tested. Finally, the optimal settings were 0.025, 9, 0.8, and 1500 for LR, TC, BF, and NT, respectively, based on the highest cross-validated method.

BRTRK was an extension of the BRT model. In BRTRK, the model residuals were obtained by subtracting the predicted values of BRT and the observed values, and then interpolated using ordinary Kriging (OK). The SOC in BRTRK is calculated:

$$SOC_{BRTSK}(i) = SOC_{BRT}(i) + \varepsilon_{OK}(i) \tag{2}$$

where SOC_{BRTSK} (i) is the final predicted SOC content at location i using the BRTRK method, SOC_{BRT} (i) represents the predicted value of the BRT model at location i, ε_{OK} (i) is the predicted value using OK model to interpolate the residual at location i.

To further demonstrate the robustness of the BRTRK model, we introduce a multiple stepwise regression model (MSR) and compare their predictive performance. As a classical approach, MSR has been widely used to predict the response to predictor variables and analyse their interactions among response and predictor variables (Ishii et al. 2014). This is an iterative process that continues until no explanatory variables are added or removed from the equation (Ishii et al. 2014), and the model only contains significant explanatory variables. The final MSR equation is:

$$\begin{aligned} SOC_{MSR}(i) = {} & 17.9825 + 0.0442Elevation(i) + 0.0021Slopeaspect(i) - \\ & 0.8079TWI(i) + 0.0051B5(i) + 1.6373SAVI(i) \end{aligned} \tag{3}$$

where SOC_{MSR} (i) is the predicted SOC content at location i using the MSR method.

Statistical analysis

Descriptive statistical analysis of soil properties and environmental variables was carried out using SPSS 22.0, including Pearson correlation coefficient, P values, Skewness, Kurtosis, variance inflation factors (VIF), and Kolmogorov–Smirnov test (K-S test). Pearson correlation coefficient was used to express the degree of linear correlation between the variables. P values were used to detect significant levels among variables. Skewness was a measure of the asymmetry of probability distribution of random variables and the degree of asymmetry relative to the mean value (Wang et al. 2016). Kurtosis was the characteristic number that characterizes the peak value of probability density distribution curve at the average value (Yang et al. 2016). VIF is the ratio of variance between explanatory variables with multiple collinearity and variance without multiple collinearity (Wang et al. 2018). K-S test is a test method to compare a frequency distribution f (x) with a theoretical distribution g (x) or two observations. Different from other methods such as t-test, K-S test does not need to know the distribution of data, so it can be regarded as a non-parametric test method (Adhikari et al. 2014).

Analysis of spatial autocorrelation and semi-variance of BRT model residuals

Residuals are considered the errors and represent the component of a model that could not be explained by the deterministic component (Guo et al. 2015). Ideally, residuals of a model should be identically and independently distributed. In fact, residuals might show spatial autocorrelations in some cases; and residuals could be added back to the deterministic component to further improve the model prediction accuracy through geostatistical analysis.

Spatial dependence as a special property of spatial data is different from general attribute data and is usually measured with spatial autocorrelations (Guo et al. 2015). At present, Moran's I index proposed by Moran in 1950 is the most commonly used spatial autocorrelation statistic. The variation of Moran's I is from −1 to 1, if the space is no autocorrelation, the value is approximately equal to 0. Value of Moran's I for residuals of BRT was calculated in ArcGIS 10.1, and the result was 0.245 ($p < 0.000$). Consequently, residuals were introduced to BRT model to further optimize the performance. In addition, the semi-variance analysis of BRT model residuals is carried out by using GS+7.0 statistical software (Gamma Design Software, Plainwell, MI). Linear, spherical, exponential and gauss model were compared to obtain the best fitting results of BRT residuals based on nugget, sill, and range values (Guo et al. 2015).

Model validation

Overall performance of MSR, BRT and BRTRK methods was evaluated using a 10-fold cross-validation method in the R version 3.2.2 (R Development Core Team 2013). Different measured and predicted levels of SOC content were calculated in four classical validation indices, i.e. absolute mean error (AME), mean error (ME), root-mean-square error (RMSE), and coefficient of determination (R^2). These indices were defined as below:

$$AME = \frac{1}{n}\sum_{i=1}^{n} |(P_i - O_i)| \tag{4}$$

$$ME = \frac{1}{n}\sum_{i=1}^{n} (P_i - O_i) \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (P_i - O_i)^2} \tag{6}$$

$$R^2 = \frac{\sum_{i=1}^{n} (P_i - \overline{O})^2}{\sum_{i=1}^{n} (O_i - \overline{O})^2} \quad (7)$$

where P_i, O_i and $\overline{O}$ are the predicted values, the observed values and the mean value of the observations at site *i, respectively*. *n* is the number of samples.

Results and discussion

Descriptive statistics of SOC content

Descriptive statistics of the measured SOC content and values of environmental variables at sample sites are summarized using SPSS 22.0 (Table 1). Average SOC content at sites was 18.8 g kg^{-1}. The standard deviation (SD) of SOC was 6.3 g kg^{-1}. In addition, the data sets of SOC have a generalized skewed distribution based on the skewness coefficients of −0.3. In Dalian, Wang et al. (2018) found the average SOC content in the topsoil was 14.4 ± 4.5 g kg^{-1}. Distribution of SOC could be described well under generalized skewed distribution with the skewness coefficient of 0.87. In addition, because previous studies revealed that soil organic carbon was mainly stored in surface soils, the soil depth was limited to 0–20 cm in this study. In Liaoning Provence of China, Wang et al. (2017) found that 69% of SOC was stored in topsoil. This conclusion had been confirmed by Liu et al. (2012), indicating that 43% of the stocks of SOC are in the upper 30 cm. Furthermore, Adhikari et al. (2014) predicted that, in Denmark, 59% of SOC stock was in the upper 30 cm of soils.

Correlation coefficients between the measured values of SOC (g kg^{-1}) with the selected environmental variables are listed in Table 2. SOC was positively correlated with elevation (0.71), slope gradient (0.55), slope aspect (0.12), CA (0.23), and SAVI (0.23), but negatively correlated with TWI (−0.54) and B3 (−0.18). Correspondingly, B4 and B5 had a negative influence on SOC. To check the multicollinearity problems, variance inflation factors (VIF) were calculated for all environmental variables using SPSS 22.0 software. All environmental variables have VIF values less than 5.0, suggesting that multicollinearity does not exist in our studies.

Geostatistical analysis of model residuals

Typically, residuals contained the most important information in model construction. Assuming that the model was constructed sufficiently accurate, the residuals of the model were determined by the measurement error. Consequently, the information contained in the residuals was used to detect the performance of model and the stability of data (Wang et al. 2018). Ideally, the residuals of all the values in the model are identical. In reality, residuals usually have spatial autocorrelations. Therefore, in order to improve the prediction accuracy of model, residuals were usually

Table 1. Descriptive statistics of soil organic carbon (SOC) data, and environmental variables at the sampling sites.

Property	Parameter	Unit	Min.	Median	Mean	Max.	SD	CV	Skewness
Soil	SOC	g kg^{-1}	1.51	19.01	18.82	31.31	6.32	33.58	−0.31
Topography	Elevation	m	1.00	63.52	85.41	646.00	81.6	95.54	2.62
	Slope gradient	Degree	0.00	4.33	7.11	45.12	7.71	108.44	1.71
	Slope aspect	Degree	0.00	158.21	161.32	356.21	110.93	68.76	−0.01
	CA	m^2 m^{-1}	375.22	856.42	1225.61	12,137.52	1231.91	100.51	3.32
	TWI		1.43	4.43	4.81	9.01	1.82	37.84	1.03
Remote Sensing	B3	Digital number	0.00	87.01	92.42	255.00	55.92	60.51	0.52
	B4	Digital number	0.00	145.02	143.83	255.00	59.51	41.38	−0.23
	B5	Digital number	0.00	89.00	94.13	255.00	58.30	61.94	0.51
	SAVI		0.00	0.31	0.31	1.00	0.23	77.42	1.01

SOC, Soil organic carbon; CA, catchment area; TWI, topographic wetness index; B3, Landsat TM band 3; B4, Landsat TM band 4; B5, Landsat TM band 5; SAVI, soil-adjusted vegetation index; Min. minimum; Max. maximum; SD, standard deviation; CV, coefficient of variation.

Table 2. Pearson correlation analysis between TSN and environmental variables based on 1195 samples.

Property	SOC	Elevation	Slope gradient	Slope aspect	CA	TWI	B3	B4	B5
Elevation	0.71**								
Slope gradient	0.55**	0.68**							
Slope aspect	0.12**	0.13**	0.14**						
CA	0.23**	0.21**	0.17**	0.18**					
TWI	−0.54**	−0.56**	−0.73**	−0.38**	−0.30**				
B3	−0.18**	−0.20**	−0.17**	−0.03	−0.08*	0.16**			
B4	0.01	−0.03	−0.03	−0.05	0.01	0.02	0.39**		
B5	−0.06*	−0.10**	−0.07*	−0.03	−0.02	0.07*	0.70**	0.12**	
SAVI	0.26**	0.20**	0.19**	0.02	0.08*	−0.15**	−0.68**	−0.12**	−0.54**

*Correlation is significant at the 0.05 level.
**Correlation is significant at the 0.01 level.
SOC, Soil organic carbon; CA, catchment area; TWI, topographic wetness index; B3, Landsat TM band 3; B4, Landsat TM band 4; B5, Landsat TM band 5; SAVI, soil-adjusted vegetation index.

incorporated into the model by means of geostatistical analysis (Wang et al. 2018). The residual of BRT model range, skewness, and kurtosis was −7.88–8.76 g kg^{-1}, 0.079, and 0.158, respectively. And the residual of the model passed the K-S test, which was suitable for geostatistical interpolation.

Correspondingly when the value was less than 0, indicating the process was negatively spatially autocorrelated (Wang et al. 2018). Value of Moran's I for residuals of BRT was calculated in ArcGIS 10.1, and the result was 0.245 ($p < 0.000$). Consequently, residuals were introduced our model to further optimize the performance of the BRT model. Linear, spherical, exponential and gauss model were compared in GS+7.0 software to obtain the best fitting results of BRT residuals, the final exponential function with best fit, and the ratio of nugget/still only 0.103, representing a strong spatial dependence for the BRT residuals. This spatial correlation indicated that variability among samples is less likely to be caused by stochastic factors, and the residuals of the model can be further reduced using the BRTRK method.

Model performance and uncertainty

A 10-fold cross-validation procedure was used to evaluate the model performance. The model performance is summarized in Table 3 using statistics of AME, ME, RMSE, and R^2. The BRTRK method performed well to predict SOC content, better than using the BRT and using MSR. The MSR, BRT, and BRTRK methods explained 51%, 59%, and 78% of the total SOC variability, respectively. In order to further compare the predictive performance of the BRTRK model for different land use patterns and soil types, we calculated the model validation indicators (Table 4). Table 4 shows that BRTRK model had the best prediction performance in cultivated land and Cambosols, which can explain 81% and 85% of the spatial variation of SOC, respectively. This may be due to the widest distribution area of the two kinds of soil in the study area and the largest number of sampling sites (462 VS 674), respectively. However, the prediction performance on grassland and Primosols is poor, which is due to the minimum distribution area of the two types and the scarcity of sampling sites. The number of sampling points will affect the accuracy of model prediction (Wang et al. 2017).

Table 3. SOC prediction performance of the BRT, BRTRK and MSR methods.

Model	AME	ME	RMSE	R^2
BRT	3.44	3.22	4.14	0.59
BRTRK	2.79	2.14	3.60	0.78
MSR	3.51	3.32	4.18	0.51

SOC, soil organic carbon; BRT, boosted regression trees; BRTRK, boosted regression trees hybrid residuals kriging; MSR, multiple stepwise regression model; AME, absolute mean error; ME, mean error; RMSE, root-mean-square error; and R^2, model efficiency.

Table 4. Summary statistics of SOC prediction performance of BRTRK under different land-use patterns and soil type.

Name	Area (km^2)	Number	AME	ME	RMSE	R^2
Land-Use Patterns						
Woodland	904.5	412	3.01	2.31	3.31	0.72
Orchard	497.6	226	3.21	2.46	3.06	0.66
Cultivated land	1015.8	462	2.68	2.05	3.74	0.81
Grassland	208.1	95	3.38	2.59	2.84	0.62
Soil types						
Anthrosols	50.9	19	3.49	2.68	2.7	0.59
Argosols	1137.9	417	2.85	2.18	3.53	0.76
Cambosols	1839.8	674	2.54	1.95	3.92	0.85
Halosols	197.0	72	3.29	2.53	2.95	0.64
Primosols	36.3	13	3.6	2.76	2.56	0.55

Our prediction is also consistent with previous SOC mapping studies. Martin et al. (2011) developed a BRT model that explained 50–58% of the SOC variability in France. Using geographically weighted regression (GWR) and geographically weighted regression kriging (GWRK) approaches, Kumar et al. (2012) captured 23% and 36% of the SOC variability, respectively, in the state of Pennsylvania (PA), USA. Other machine learning tools and geostatistical methods such as random forest (RF) and ordinary kriging (OK) were also common in DSM (Wang et al. 2017). Typically, Guo et al. (2015) used three models of stepwise linear regression (SLR), RF, and RF plus residuals kriging (RFRK) to predict the spatial distribution patterns of soil organic matter (SOM) in a rubber plantation of Hainan Province, China, and RFRK model was proved to have the efficiently and steadily predictive performance. BRT model can flexibly deal with various data types, default values, outliers, and thus was widely applied to solve different scientific problems (Yang et al. 2016). Such complexity coupled with its microclimates would very often lead to the emergence and extension of a local ecological niche or pedogeomorphological units (Yang et al. 2016) with a diverse soil and vegetation types, making soil variability predictions more challenging. In such conditions, the BRT hybrid residuals kriging could be an effective method for SOC mapping.

In order to explore the uncertainty of the BRT model, a mean-standard -deviation (SD) map (Figure 2(a)) was generated through 100 iterations. The BRT model estimates have a small uncertainty with a mean SD of 0.58 g kg^{-1}. In addition, The R^2 is from 0.55 to 0.60, indicating the BRT model had stable prediction ability (Figure 2(b)). Similarly, the lower values of AME, ME, and RMSE, and higher R^2 also indicated that the BRTRK method was stable in predicting the SOC content. In

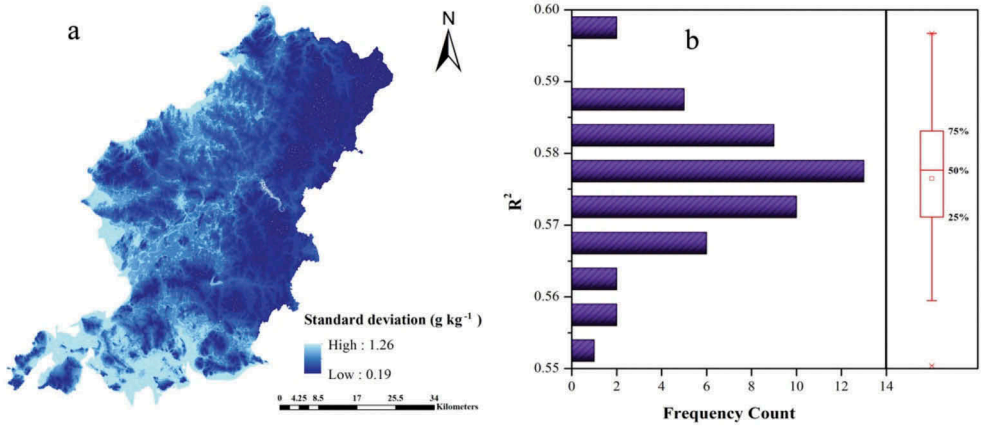

Figure 2. Standard deviation maps of the predicted topsoil organic carbon (g kg^{-1}) derived from 100 times boosted regression trees (BRT) models (a) and histogram showing the R^2 response to a number of model iterations for SOC (b).

addition, there also exist some uncertainties in this study such as sampling error, experimental error, and model error (Wang et al. 2016).

Relative importance of environmental variables

Selected environmental variables presented different importance levels in the SOC prediction (Figure 3). The relative importance (RI) of each variable was obtained by iterating the BRT model estimation for 100 times. In the prediction of SOC, terrain-related variables (69% of RI) were considered as major explanatory variables, followed by vegetation-related variables (31% of RI). As one of the five soil forming factors, topography affects the soil moisture and temperature conditions (Jenny 1941), and controls water and energy flow regulating spatial distribution of soils at a landscape scale (Bonfatti et al. 2016). Throughout all terrain variables, elevation played a decisive character in predicting SOC. In a hilly region of central Iran, Tajik et al. (2012) used soil and topographic attributes to predict the activity of three soil enzymes using artificial neural networks (ANNs) and multiple linear regression (MLR) approaches, concluding that DSM can be applied to predict spatial distribution of soil enzymes at the hillslope scale. In addition, TWI and slope gradient were also recognized as important environment variables in our model (Figure 3). TWI represents potential features of topography and soil hydraulic characters, playing an important role in spatial distribution of SOC (Yang et al. 2016). Similarly, slope gradient impacts the movement and accumulation of surface water and mineral carbon loss in a landscape (Martin et al. 2014). In China and Denmark, a strikingly similar conclusion about slope gradient was drawn (Adhikari et al. 2014; Yang et al. 2016). However, a negative effect of slope gradient on SOC distribution was reported by Tsui et al. (2004) where slope gradient played a major role in determining prevailing land-use types. Slope aspect was also found as an important variable in

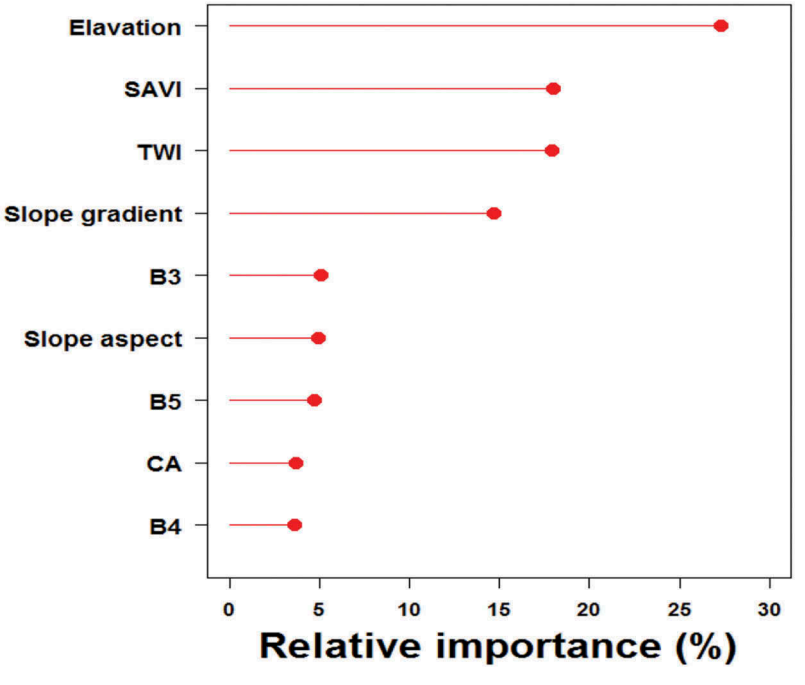

Figure 3. Relative importance (RI) of each predictor from 100 iterations using the boosted regression trees (BRT). CA: catchment area; TWI, topographic wetness index; B3, Landsat TM band 3; B4, Landsat TM band 4; B5, Landsat TM band 5; SAVI, soil adjust vegetation index.

predicting SOC. Bonfatti et al. (2016) considered slope aspect affecting vegetation community while impacting regional microclimate.

Vegetation is one of the main factors influencing SOC variability (Minasny et al. 2016). On the Qinghai Tibet Plateau, Yang et al. (2016) found that vegetation-related variables played an important role in mapping SOC. Curiously, our conclusion showed that vegetation variables had a weak correlation with topsoil SOC compared with topographic variables. Our analysis suggested that vegetation variables might be mediated by topographic variables, and there was usually better

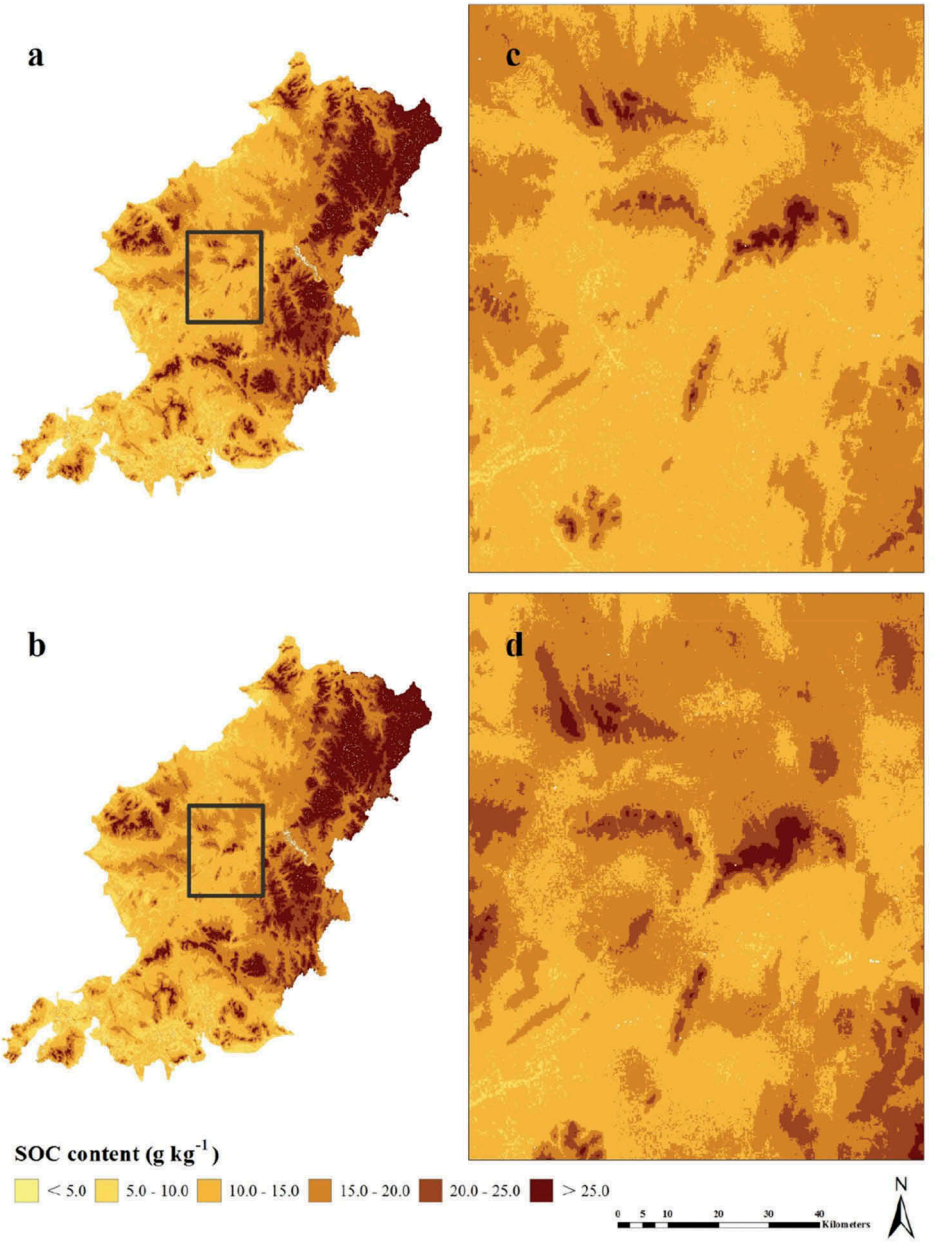

Figure 4. Distribution of topsoil organic carbon (g kg^{-1}) derived from boosted regression tree (BRT) model (a) and BRT hybrid residuals kriging (BRTRK) (b); c), and d) small areas outlined with black colour in left large areas for showing detail information.

vegetation cover at high altitudes in our study areas. Of all the vegetation variables, SAVI and B3 were the crucial variables in SOC prediction in hilly dominated areas (Wang et al. 2016). The simultaneous effect of B4 and B5 could reflect the present situation of land use, thus different land-use patterns have different levels of SOC content. Similar to Adhikari and Hartemink (2015), our study indicated that SAVI, B3 and B4 were major predictor variables in predicting the spatial distribution of SOC. In addition, the higher resolution RS data better characterize the spatial heterogeneity of vegetation types in the study area, especially in the region with dense vegetation cover, which help increase our prediction accuracy.

Spatial distribution of SOC content

Average predicted SOC contents were 17.68 g kg^{-1} using the BRTRK method and 17.60 g kg^{-1} using the BRT (Figure 4). The northeastern mountain areas were estimated with the highest SOC content. Elevation was the main predictors in this area (Figures 1 and 4). Generally, high altitude areas usually have better vegetation coverage, which corresponds to the increase of plant litter returned to the soil, thus increasing SOC (Figure 5). The effect of elevation on SOC has been demonstrated in several researches (Martin et al. 2014; Adhikari and Hartemink 2015). Wang et al. (2016) reported that SOC significantly increased with elevation. Differences in elevation gradients might have affected the input and loss of soil carbon mainly through indirect measures such as its influence on precipitation and temperature (Lal 2004; Martin et al. 2014).

The spatial distribution of SOC content indicates that the woodland had the highest content, followed by orchard, grassland and cultivated land, which is consistent with previous studies (Tsui et al. 2004; Tajik et al. 2012). In all land-use types, soil under cultivated land had the lowest SOC content (16.2 g kg^{-1}), confirmed by Yang et al. (2016). Argosols had the highest SOC content (18.1 g kg^{-1}). Most of the cultivated lands are on Cambisol (1839.8 km^2). There was a spatial link between soil types and land-use patterns with SOC distribution in the region dominated by agroecosystems.

Conclusion

This study used the BRTRK method to map spatial distribution of SOC in the northeastern coastal areas of China. The results indicated that BRTRK was effective in predicting spatial

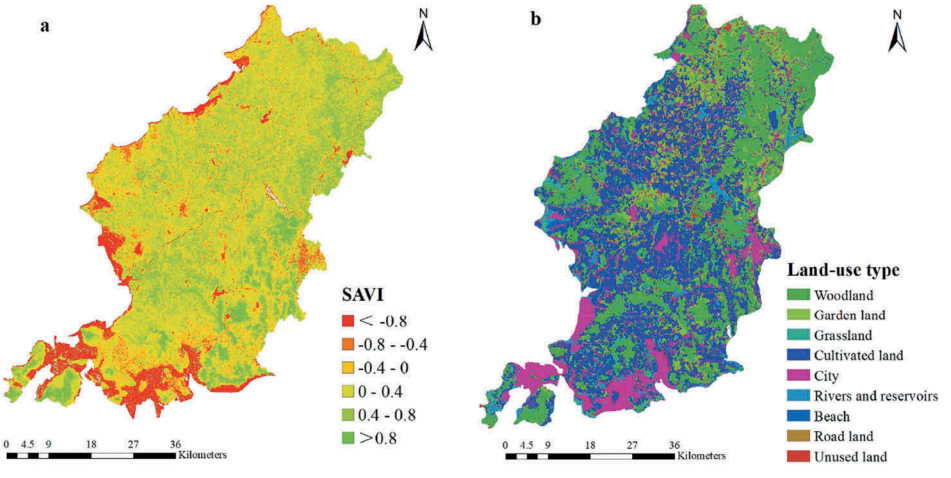

Figure 5. Spatial distribution map of SAVI and (a) and Land-use map (b).

distribution SOC with low AME, ME, and RMSE and high R^2 explaining 78% of the SOC variability. The northeastern areas had higher SOC levels than anywhere else. We found that topographic variables were the main factors for the spatial distribution of SOC in the northeastern coastal area of China. Consequently, topographic variables such as elevation shall be considered in future SOC mapping in the coastal hill areas in China. Woodlands and Argosols have the highest SOC content. The predicted SOC distribution is valuable information for soil conservation, environment protection, and agricultural production planning in the northeastern coastal agroecosystems of China.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Natural Science Foundation Project of Shenyang Agricultural University [880418058].

References

Adhikari K, Hartemink AE. 2015. Digital mapping of topsoil carbon content and changes in the Driftless Area of Wisconsin, China. Soil Sci Soc Am J. 79(1):155–164. doi:10.2136/sssaj2014.09.0392.

Adhikari K, Hartemink AE, Minasny B, Kheir RB, Greve MB, Greve MH. 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS One 9:e105519. doi:10.1371/journal.pone.0105519.

Baldock JA,Wheeler I,McKenzieN,McBratenyA.2012.Soils and climate change: potential impacts on carbon stocks and greenhouse gas emissions, and future research for australian agriculture.Crop Pasture Sci.63(3):269-283.

Batjes NH. 1996. Total carbon and nitrogen in the soils of the world. Eur J Soil Sci. 47:151–163. doi:10.1111/j.1365-2389.1996.tb01386.x.

Bonfatti BR, Hartemink AE, Giasson E, Tornquist CG, Adhikari K. 2016. Digital mapping of soil carbon in a viticultural region of Southern Brazil. Geoderma 261:204–221. doi:10.1016/j.geoderma.2015.07.016.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Ann Statist. 29:1189–1232. doi:10.1214/aos/1013203451.

Guo PT, Li MF, Luo W, Tang QF, Liu ZW, Lin ZM. 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma 237:49–59. doi:10.1016/j.geoderma.2014.08.009.

Hijmans RJ, Phillips S, Leathwick J, Elith J. 2013. Dismo: species distribution modeling. R package version 0.8-17.*The R Foundation for Statistical Computing, Vienna.*

Huete AR. 1988. A soil-adjusted vegetation index (SAVI). Remote Sens Environ. 25(3):295–309. doi:10.1016/0034-4257(88)90106-X.

Ishii Y, Murakami J, Sasaki K, Tsukahara M, Wakamatsu K. 2014. Efficient folding/assembly in Chinese hamster ovary cells is critical for high quality (low aggregate content) of secreted trastuzumab as well as for high production: stepwise multivariate regression analyses. J Biosci Bioeng. 118(2):223–230. doi:10.1016/j.jbiosc.2014.01.013.

Jenny H. 1941. Factors of soil formation. New York: McGraw-Hill.

Kumar S, Lal R, Liu D. 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189:627–634. doi:10.1016/j.geoderma.2012.05.022.

Lal R. 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304:1623–1627. doi:10.1126/science.1097396.

Liu W, Chen S, Qin X, Baumann F, Scholten T, Zhou Z, Sun W, Zhang T, Ren J, Qin D. 2012. Storage, patterns, and control of soil organic carbon and nitrogen in the northeastern margin of the Qinghai–tibetan Plateau. Environ Res Lett. 7(3):035401.

Martin MP, Orton TG, Lacarce E, Meersmans J, Saby NPA, Paroissien JB, Jolivet C, Boulonne L, Arrouays D. 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. Geoderma 223–225:97–107. doi:10.1016/j.geoderma.2014.01.005.

Martin MP, Wattenbach M, Smith P, Meersmans J, Jolivet C, Boulonne L, Arrouays D. 2011. Spatial distribution of soil organic carbon stocks in France. Biogeosciences 8:1053–1065. doi:10.5194/bg-8-1053-2011.

McBratney AB, Mendonça Santos ML, Minasny B. 2003. On digital soil mapping. Geoderma 117:3–52. doi:10.1016/S0016-7061(03)00223-4.

Minasny B, Setiawan BI, Arif C, Saptomo SK, Chadirin Y. 2016. Digital mapping for cost-effective and accurate prediction of the depth and carbon stocks in Indonesian peatlands. Geoderma 272:20–31. doi:10.1016/j.geoderma.2016.02.026.

R Development Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Tajik S, Ayoubi S, Nourbakhsh F. 2012. Prediction of soil enzymes activity by digital terrain analysis: comparing artificial neural network and multiple linear regression models. Environ Eng Sci. 29(8):798–806. doi:10.1089/ees.2011.0313.

Tsui CC, Chen ZS, Hsieh CF. 2004. Relationships between soil properties and slope position in a lowland rain forest of southern Taiwan. Geoderma123(1–2):131–142. doi:10.1016/j.geoderma.2004.01.031.

Wang S, Adhikari K, Wang Q, Jin X, Li H. 2018. Role of environmental variables in the spatial distribution of soil carbon (C), nitrogen (N), and C: N ratio from the northeastern coastal agroecosystems in China. Ecol Indic. 84:263–272. doi:10.1016/j.ecolind.2017.08.046.

Wang S, Wang Q, Adhikari K, Jia S, Jin X, Liu H. 2016. Spatial-temporal changes of soil organic carbon content in wafangdian, China. Sustainability8(11):1154. doi:10.3390/su8111154.

Wang S, Zhuang Q, Wang Q, Jin X, Han C. 2017. Mapping stocks of soil organic carbon and soil total nitrogen in Liaoning Province of China. Geoderma 305:250–263. doi:10.1016/j.geoderma.2017.05.048.

Yang RM, Zhang GL, Liu F, Lu YY, Yang F, Yang F, Li DC. 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecol Indic. 60:870–878. doi:10.1016/j.ecolind.2015.08.036.