Catena 210 (2022) 105897

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena

Anthropogenic controls over soil organic carbon distribution from the cultivated lands in Northeast China

Shuai Wang [a,b], Mingyi Zhou [a], Kabindra Adhikari [c], Qianlai Zhuang [d], Zhenxing Bian [a], Yan Wang [e], Xinxin Jin [a,*]

[a] College of Land and Environment, Shenyang Agricultural University, Shenyang, Liaoning Province 110866, China
[b] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[c] USDA-ARS, Grassland, Soil and Water Research Laboratory, Temple, TX 76502, USA
[d] Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, IN 47907, USA
[e] College of Tourism and Geography, Jiujiang University, Jiujiang 332005, China

ARTICLE INFO

Keywords:
Agroecosystem
Anthropogenic variable
Digital soil mapping
Soil organic carbon
Spatial variation

ABSTRACT

Both natural and anthropogenic variables affect soil C distribution and its pool, however studies about anthropogenic influence on soil C distribution are very limited in the literature. This study investigated anthropogenic effects on soil organic carbon (SOC) changes in the cultivated lands of Northeast China. A total of 196 topsoil samples (0–30 cm) were collected, and analyzed for SOC content, and 12 environmental variables (natural and anthropogenic) were selected as SOC predictors. Natural factors included elevation, slope gradient, slope aspect (SA), topographic wetness index (TWI), mean annual temperature, mean annual precipitation, and normalized difference vegetation index, while population (POP), gross domestic product (GDP), distance to the socioeconomic center, distance to roads, and reclamation period (PER) represented anthropogenic variables. Three different boosted-regression trees models with different combination of SOC predictors were constructed, and the model performance was evaluated with 10-fold cross-validation. We found that the model that included all predictors had the best performance, followed by the model with topography and climate variables, and the model with only anthropogenic variables. However, adding the anthropogenic variables in the model greatly improved its performance. Results showed that PER, POP and GDP were the key environmental variables affecting SOC content in the topsoil agroecosystems in Northeast China. This study suggests that anthropogenic variables should be selected as the main environmental variable in predicting of SOC content in agroecosystem with a higher human influence. We believe that the accurate prediction and mapping of SOC content in the topsoil agroecosystem will help formulate farmland soil management policies and promote soil carbon sequestration.

1. Introduction

Land use change and agricultural reclamation have altered carbon balance between soil and the atmosphere leading to more carbon dioxide emission into the atmosphere. Thus, an accurate estimation of SOC content is of great significance for regional as well as global ecosystem carbon balance. The digital soil mapping (DSM) models or algorithms used to predict SOC distribution, regression models have been widely used but use of advanced data mining or machine learning techniques has been gaining popularity quite recently (Minasny et al, 2013, Lamichhane et al, 2019).

There are many factors affecting the content and distribution of SOC in agroecosystems, and they include temperature, precipitation, land use and land cover types, and topography, among others (Minasny et al, 2013, Lamichhane et al, 2019, Adhikari et al, 2014). In recent years, human-induced factors or anthropogenic factors have also been recognized as important factors of SOC changes (Sanderman et al., 2017, Seto et al, 2012, Wang et al., 2020a). With rapid growth in population and economic development, the land use pattern has changed exponentially leading to great changes in SOC content and its distribution (Wang et al.,

* Corresponding author at: College of Land and Environment, Shenyang Agricultural University, No. 120 Dongling Road, Shenhe District, Shenyang, Liaoning Province 110866, China.
E-mail address: jinxinxin0218@syau.edu.cn (X. Jin).

https://doi.org/10.1016/j.catena.2021.105897
Received 15 March 2021; Received in revised form 14 November 2021; Accepted 22 November 2021

2016; Sanderman et al., 2017). Most of the previous SOC modeling studies focused on quantifying the effects of natural geographic environment change, land use and land cover, and farm management on SOC changes (Kumar et al., 2012; Conforti et al., 2016; Wang et al., 2019) However, there is not much research on the effect of anthropogenic factors such as population growth, and socio-economic development on SOC status changes. Wang et al (2020) identified population, gross domestic product, distance to the socio-economic center, and distance to the roads as main variables affecting the spatial variation of SOC in agroecosystems. In a separate study, Wang et al. (2019) made a strong recommendation to include cultivation history as an indicator of SOC changes in agroecosystems with a long cultivation history. A field study in the US also reported a positive influence of intensive farming on SOC levels (Adhikari and Hartemink, 2017).

The main aim of this study was to assess anthropogenic influence on SOC distribution in montane agroecosystems in Northeast China. Specific objectives included to: (1) identify potential anthropogenic variables influencing SOC distribution; (2) predict and map the SOC content with associated uncertainty; and (3) evaluate the predictive performance and potential model application.

2. Materials and methods

2.1. Study area

The study was conducted in the cultivated lands of Northeast Plain located in the middle of Northeastern China that includes Heilongjiang, Jilin and Liaoning provinces, and it covers an area of 787,300 km^2. The eastern, western and northern parts of the study area are surrounded by Changbai Mountain, Greater Khingan Range, and Xiaoxing'an Mountains. The terrain is high on these three sides with a low and open-wide plain facing south in the center. The altitude of the study area ranges from 0 to 2665 m above sea level, with an average altitude about 200 m (Fig. 1).

The study area consists of three major plains, namely, Songnen Plain, Liaohe Plain and Sanjiang Plain with a large area under cultivation (217,000 km^2; 20% of China s total cultivated land). The Northeast Plain has a long (>300 yrs) history of farming civilization and is the main farming area in northeastern China with frequent and intense human influence. Therefore, the Northeast Plain has been an ideal area to study anthropogenic influence on SOC distribution in agricultural ecosystems. Moreover, the study area is an important national grain base of China, and the main crops include rice, corn, soybean, potato, beet, sorghum, and temperate fruits and vegetable. The region has a temperate monsoon climate with four distinct seasons (Spring, Summer, Autumn, Winter). The annual precipitation increases from northwest to southeast, and gradually transit from semi-arid area, semi-humid to humid area. Mean annual precipitation (MAP) ranges from 350 mm to 1100 mm, mainly from July to August. Mean annual temperature (MAT) is between −4℃ and 11℃. According to the World Reference Base for Soil Resources (IUSS Working Group, 2006), the dominant soil types include Cambisols covering 46% of area Fluvisols (21% of area), and the rest are Anthrosols, Phaeozems, Gleysols, Histosols, and Andosols.

2.2. Sampling strategy and SOC determination

Due to the large area studied, intensive soil sampling was unrealistic. In order to accurately reflect spatial characteristics of SOC content of this typical agroecosystem, we used a purposive sampling strategy (Zhu et al., 2008) following a stratified simple random sampling principle, and identified 196 sample locations covering the entire study area. We first grouped four environmental variables namely, soil type, elevation, temperature, and precipitation into more homogenous sampling strata or landscape units using fuzzy c-means (FCM) clustering algorithms incorporated in the 'fanny' function of 'cluster' package in R (Pal et al., 2005). In the FCM algorithm, two parameters, namely cluster number and fuzzy parameter were set to be 34, and 2, respectively. Generally speaking, cluster number should be far less than the number of cluster samples, and ensure that it is at least>1. The fuzzy parameter controls the flexibility of the algorithm and is used to define the fuzziness of the

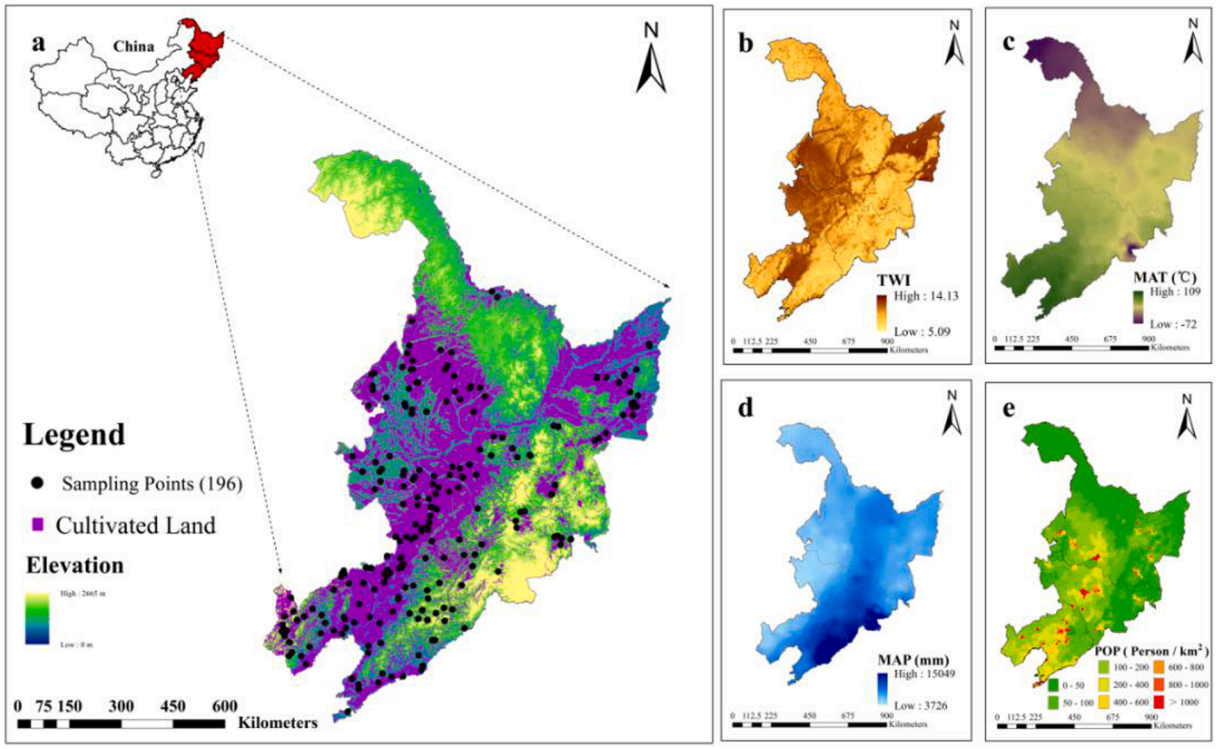

Fig. 1. Location of study area. (a) sampling sites overlaid on a 90-m resolution digital elevation model; (b) topographic wetness index (TWI) map; (c) mean annual temperature (MAT) map; (d) mean annual precipitation (MAP).map; and (e) population map.

whole data set. It is generally 2 by default. In this study, first initialize the membership matrix to calculate the cluster center or initialize the cluster center to iteratively calculate the value function. When it is less than a minimum or the difference between the previous and subsequent times is less than a minimum, stop updating the membership matrix, and finally the clustering result is determined to be 34. The FCM algorithms identified 34 clusters, landscape units hereafter, covering the study aera, and its map was converted into a raster of 90 m × 90 m grid resolution in ArcGIS. In each landscape unit, 5–8 sampling sites were randomly selected considering landscape position (peak, ridge, valley, and saddle) in the landscape making a total of 196 sampling points across the study area. The geographical coordinates of each sampling point were recorded by a handheld global positioning system. Soil sample at each sampling point was collected from the topsoil depth (0–30 cm), litters were removed if present, and the samples were dried and ground before sieving to obtain a fine earth fraction. The SOC content in the samples was determined by dry combustion method using a Vario EL III elemental analyzer (Elementar Analysensysteme GmbH, Hanau, Germany) in the Analysis and Testing Center of Shenyang Agricultural University, Shenyang, China.

2.3. Environmental variables

Twelve environmental variables representing anthropogenic, topography, climate, and biology, as described below, were used as predictors of SOC in the study area. Environmental variables were collected from different sources, and were converted to a raster of 90 m × 90 m resolution using nearest neighborhood method in ArcGIS (Manap et al., 2013).

2.3.1. Anthropogenic variables

Anthropogenic variables are human-induced or human-mediated variables that directly or indirectly impact SOC variations in agroecosystems; cultivation history or land reclamation period is one of the prime examples, among others (Wang et al 2019, Wang et al., 2020, Adhikari and Hartemink, 2017). Anthropogenic warming is found to be directly related to microbial-mediated SOC losses (García-Palacios et al., 2021), while temperature dependency of soil organic matter decomposition and its effect on SOC storage has long been studied (Kirschbaum, 1995). Further, impacts of increased population and urban expansion on carbon pool and biodiversity (Seto et al, 2012), and human-induced land use change has resulted in substantial losses of carbon from soils globally (Sanderman et al., 2017). In this study, we considered five anthropogenic variables namely, population (POP), gross domestic product (GDP), distance to the socioeconomic center or hotspots (DSE), distance to roads (DR), and reclamation period (PER) as SOC predictors. These specific variables reflect anthropogenic activities and were selected to reflect the driving forces of SOC content and distribution in agroecosystems. The GDP is one of the important indicators of socioeconomic development, and is a key to regional planning and resources management (Sokka et al., 2009). The POP, and GDP were spatialized to replace the traditional administrative statistical unit with spatial statistical unit (Vasenev et al., 2018), which brings great convenience for data sharing, and spatial analysis among multiple fields (Liu et al., 2005; Ling et al., 2006; Wang et al., 2020a). Firstly, 1-km grid data set of GDP, and POP were derived from the national statistical data. Using the multi-factor weight distribution method (Ling et al., 2006), the GDP, and POP data with the administrative region as the basic statistical unit were distributed to the grid unit, so as to realize the spatialization of GDP, and POP. Land use type, night light intensity, residential density and other factors closely related to anthropogenic economic activities were comprehensively considered during spatialization.

The POP and GDP variables attributed to the grassroots census only stayed at county level, which affected the accuracy of mapping. Therefore, the nearest distance from the field to the road, and to the economic center were considered as anthropogenic factors because the roads and economic center were the most active area of human influence. Detailed steps to generate these data are shown in Liu et al. (2006). Land use type, night brightness, residential density, socioeconomic center, and road network information were downloaded from the Institute of Geographic Sciences, Resource and Environment Cloud Data Platform, China (http://www.resdc.cn/) and were rasterized to a 1-km grid in ArcGIS. The variable PER reflects agricultural impact on agroecosystem and the data were obtained from Wang et al. (2019). PER was derived from the data of 300 years of farming or reclamation period following a factor correction and a human-land relationship test method. Seven PERs were identified-0–10 years, 10–30 years, 30–70 years, 70–120 years, 120–200 years, 200–300 years, and>300 years.

2.3.2. Topographic variables

Topography is one of the five soil forming factors and has been widely used in the spatial prediction of SOC (Yimer et al., 2006, Adhikari et al, 2014, Wang et al., 2016). In this study, we selected four topographic variables – elevation (ELE), slope gradient (SG), slope aspect (SA), and topographic wetness index (TWI), which were derived from the U.S. Geological Survey's 90-m digital elevation model (DEM). The ELE, SG, and SA were derived in ArcGIS software and TWI in System for Automated Geoscientific Analyses (Conrad et al., 2015)

2.3.3. Climatic variables

Climatic conditions affect input, decomposition, and transformation of C in soils (Jobbágy and Jackson, 2000), thus affecting overall SOC distribution. In this study, traditional climate variables like mean annual precipitation (MAP), and mean annual temperature (MAT) over thirty year period (1982–2010) were obtained from China Meteorological Data Service Center (http://data.cma.cn/en). The raster data with 1-km grid were generated by kriging interpolation using 673 weather stations across China. The kriged layer was then resampled to 90-m grid by using the nearest neighbor method.

2.3.4. Biological variable

For the biological variable, we used Normalized Difference Vegetation Index (NDVI) which is one of the mostly used biological variables as SOC predictor. NDVI reflects vegetation growth, and plant nutrition information (Wang et al., 2018; Wang et al., 2020b). A negative value indicates that the ground is either covered by clouds, water, or snow; 0 indicating rock or bare soil; and a positive value indicating vegetation coverage which increases with increasing chlorophyll content. We used Landsat-7 band 3 (0.63–0.69 μm) and band 4 (0.78–0.90 μm) representing vegetation growth and coverage to calculate the NDVI as follows:

$$\mathrm{NDVI} = (band\ 4 - band\ 3)/(band\ 4 + band\ 3) \quad (1)$$

Landsat-7 image was downloaded from the Resource and Environment Science and Data Center from July to September 2013 covering the study area with the cloud cover < 10%. The images were atmospherically corrected using the Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes atmospheric correction method (Perkins et al., 2005) in ENVI 5.1 software and were resampled to a grid resolution of 90-m by using the nearest neighbor method.

2.4. Prediction model and uncertainty

We selected Boosted Regression trees (BRT) as a SOC prediction model. The BRT model as proposed by Friedman et al. (2000) is similar to other boosting models in that multiple models are trained and combined for prediction to boost model performance. The model consisted of two algorithms: regression trees and gradient boosting (Elith et al., 2008). The regression trees applies a set of predictive variables to analyze the response variable, and uses binary segmentation to fit the simple model to each split. The boosting algorithm uses the iterative

method to develop the final model, by gradually adding a tree to the model. BRT relied on random gradient propulsion through numerical optimization and regularization for more accurate and faster calculations (Pouteau et al., 2011). Compared with other data mining methods, BRT model had higher prediction accuracy and good interpretability (Ottoy et al., 2017).

We used "gbm.step" in 'dismo' R package (https://cran.r-project.org/web/packages/dismo/dismo.pdf) (Hijmans et al., 2013) to build the model in R language environment (R Development Core Team 2013). In the BRT model, four model parameters need to be set: Learning Rate (LR), Tree Complexity (TC), Bag Fraction (BF) and Number of trees (NT). The LR expresses the contribution made by each tree in the final fitting result (Pouteau et al., 2011), TC is the complexity of the tree, which is the maximum interaction level between predictive variables (Wang et al., 2018), BF represents the proportion of data used in the dataset (the more data the model used, the less random it is) (Elith et al., 2008). Although the BRT model could be used to avoid overfitting by expanding the model's operations, it was still necessary to set NT which can be determined based on the combination of LR, and TC (Wang et al., 2020a). A 10-fold cross-validation was used to optimize model parameter settings and to obtain the best predictive performance of the model. The final optimal values of LR, TC, BF, and NT were 0.025, 12, 0.60, and 2000, respectively. The model was then applied to predict .SOC distribution through study area.

Three different BRT models each with a specific group of predictors were developed for SOC predictions in the study area. The first model used anthropogenic variables only (MA), the second model used only topography, and climate variables (MB), and the third model used all variables, i.e., topography, climate, and anthropogenic variables (MC). For each model, a data matrix of 196 SOC measurements (log-transferred SOC) in rows and corresponding predictors values at measurement locations in columns was constructed. The model was built on these matrices and was iterated 100 times generating 100 maps; the average of 100 predictions was then considered as the final predicted map of SOC distribution. The log-transferred SOC values were back-transferred to original units for map display.

The uncertainty related to BRT prediction was derived as the standard deviation (SD) of 100 iterations of each model, and the map represented an indicator of SOC prediction uncertainty. The relative importance (RI) of variables were measured according to the number of times a variable was selected for modeling and weighted by the square improvement to each split and averaged across all trees (Wang et al., 2020a). The RI of each variable was then scaled so that the sum was added as a percentage to 100. The higher the percentage of the variable, the stronger the RI of the variable to the response.

2.5. Model evaluation

A 10-fold cross-validation technique was applied to evaluate the BRT model through multiple iterations using environmental variables, and measured SOC content at sampling locations. The prediction performance of each BRT model was tested by comparing the average predicted SOC value with the measured value with validation indices mean absolute error (MAE), root mean of squared error (RMSE), coefficient of determination (R^2), and Lin's concordance correlation coefficient (LCCC) (Lin, 1989) which were calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{3}$$

$$R^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{4}$$

$$\mathrm{LUCC} = \frac{2r\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\overline{X} + \overline{Y})^2} \tag{5}$$

where X_i and Y_i stands for predicted values and observed values, respectively; n is the number of samples, $\overline{X}$ and $\overline{Y}$ are the mean predicted and observed values ; r refers to the Person correlation coefficient between the predicted values and observed values , σ_X and σ_Y correspond to the standard deviation of the prediction set and observation set.

3. Results

3.1. Descriptive statistics

Statistical description of measured SOC content, and environmental variables are shown in Table 1. These statistics are restricted to the sampling sites (196 locations). Measured SOC content ranged from 6.30 g kg^{-1} to 67.30 g kg^{-1} and its average value was 24.60 g kg^{-1}. SOC data were positively skewed with a coefficient of variation of 43.9%.

The Pearson correlation coefficient between log-transformed SOC content and environmental variables is shown in Table 2. The SOC had positive correlations with ELE, SG, TWI, MAP, NDVI, and DR, with the correlation coefficients of 0.39, 0.29, 0.33, 0.36, 0.51 and 0.32, respectively. However, SOC was negatively correlated with MAT (r = -0.36), POP (r = -0.43), and anthropogenic variables such as GDP (r = -0.52), DSE (r = -0.34), and PER (r = -0.39), respectively. Measured SOC had a high correlation with all anthropogenic variables in our study area. It could be attributed to the fact that northeast China was China's main commercial grain production base, accounting for 20% of the country's total grain output. In order to ensure national food security, the Chinese government has been investing a lot to promote grain production every year. Moreover, the protective tillage and black land protection policies carried out in recent years had made the cultivated land more strongly affected by human interference.

3.2. Model performance

Summary statistics of the performance of MA (only anthropogenic variables), MB (only topography and climate variables), and MC (topography, climate, and anthropogenic variables) models to predict SOC content based on 100 iterations of the BRT model is showed in Table 3 and Fig. 2. The validation statistics indicated that MC had the best prediction performance, because it had the highest R^2 (0.74) and LUCC (0.78), and the lowest MAE (0.83) and RMSE (1.21), and it was followed by MA, and MB models. The MB model explained 53% of the SOC spatial variation in the region while the MC and MA models could explain 78%, and 42% of the variation. Results showed that adding anthropogenic variables significantly improved the prediction performance of the MC model. This suggested that SOC distribution in Northeast agroecosystems in China is greatly influenced by anthropogenic factors, and combining it with topographic, and climatic variables could improve SOC prediction performance. Overall, all three models showed a good performance in predicting SOC content, among which MC model with the full set of variables had the best prediction performance.

To further verify the importance of incorporating anthropogenic variables in SOC predictions, we generated a scatter plot between observed and predicted SOC values from the three BRT model (Fig. 3). The predicted value of MC model was closer to observed SOC, and its stability and accuracy (Table 3) were also higher, suggesting that the inclusion of anthropogenic variables greatly improved SOC predictions in northeastern agroecosystem in China.

Fig. 4 shows uncertainty map of MA, MB, and MC models, for which the corresponding mean values were 3.2 g kg^{-1}, 2.1 g kg^{-1} and 1.3 g kg^{-1}, respectively. Although all the three BRT models performed well, adding anthropogenic factors as SOC predictors reduced mean uncertainty of MC model compared to MB model.

Table 1
Summary statistics of measured SOC content, and environmental variables at 196 sampling sites.

Property	Unit	Min.	Mean	Max.	SD	Skewness	Kurtosis
SOC	g kg^{-1}	6.30	24.60	67.30	10.80	1.20	1.90
LnSOC	g kg^{-1}	1.84	3.13	4.21	0.67	0.23	2.14
ELE	m	0.80	186.70	652.70	129.20	1.10	1.50
SA	degree	0.00	163.40	345.80	102.50	0.30	−1.30
SG	degree	0.00	1.80	17.50	2.20	1.40	1.70
TWI	index	6.60	9.70	12.40	1.40	−0.30	−0.60
MAP	mm	410.20	594.90	1093.70	125.31	1.50	2.70
MAT	degree Celsius	−0.50	5.70	10.80	2.60	−0.10	−0.70
NDVI	index	0.17	0.40	0.56	0.09	−0.17	0.27
POP	Person / km^2	13.60	155.70	905.90	124.50	−1.30	2.40
GDP	10^4 yuan / km^2	35.00	747.20	12055.00	1179.50	−1.60	3.55
DSE	km	0.70	5.70	31.30	4.70	1.70	2.30
DR	km	0.10	0.90	3.40	0.70	0.50	1.10

Note: ELE, elevation; SA, slope aspect; SG, slope gradient; TWI, topographic wetness index; MAP, mean annual precipitation; MAT, mean annual temperature; NDVI, Normalized Difference Vegetation Index; POP, population , GDP, gross domestic product; DSE, distance to the socioeconomic center or hotspots; DR, and distance to roads.

Table 2
Pearson correlation coefficients between SOC content and environmental variables based on 196 samples.

Property	lnSOC	ELE	SA	SG	TWI	MAP	MAT	NDVI	POP	GDP	DSE	DR
ELE	0.39											
SA	−0.17	0.09										
SG	0.29	0.37	0.17									
TWI	0.33	−0.46	−0.33	−0.62								
MAP	0.47	−0.32	0.13	0.17	0.06							
MAT	−0.36	0.38	−0.14	0.22	0.10	0.25						
NDVI	0.41	0.23	−0.06	0.21	0.21	0.37	−0.17					
POP	−0.43	−0.23	−0.24	−0.16	−0.37	0.23	0.18	−0.27				
GDP	−0.52	−0.07	−0.09	−0.18	0.13	0.26	0.13	−0.13	0.43			
DSE	−0.34	−0.19	−0.11	−0.14*	0.11	0.07	0.08	0.11	−0.46	−0.33		
DR	0.32	0.18	−0.14	0.05	−0.16	−0.09	0.11	−0.16	−0.31	−0.29	0.35	
PER	−0.39	−0.21	0.23	−0.13	0.06	0.38	0.27	0.17	0.34	−0.16	0.21	0.32

Note: ELE, elevation; SA, slope aspect; SG, slope gradient; TWI, topographic wetness index; MAP, mean annual precipitation; MAT, mean annual temperature; NDVI, Normalized Difference Vegetation Index; POP, population, GDP, gross domestic product; DSE, distance to the socioeconomic center or hotspots; DR, and distance to roads; and PER, Reclamation period.

Table 3
Evaluation of prediction performances of MA (only anthropogenic variables), MB (only topography and climate variables), and MC (topography, climate, and anthropogenic variables) models for SOC content using BRT model with 100 iterations based on 196 samples.

Model	Index	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
MA	MAE	1.36	1.38	1.39	1.39	1.41	1.43
	RMSE	1.64	1.65	1.67	1.67	1.69	1.73
	R^2	0.39	0.40	0.42	0.42	0.44	0.45
	LUCC	0.60	0.61	0.63	0.64	0.64	0.65
MB	MAE	1.28	1.29	1.30	1.30	1.34	1.37
	RMSE	1.59	1.62	1.65	1.65	1.67	1.68
	R^2	0.49	0.50	0.51	0.52	0.54	0.55
	LUCC	0.69	0.70	0.72	0.72	0.74	0.74
MC	MAE	0.83	0.86	0.92	0.92	0.95	0.96
	RMSE	1.21	1.24	1.25	1.25	1.26	1.28
	R^2	0.74	0.76	0.76	0.78	0.79	0.81
	LUCC	0.78	0.81	0.88	0.88	0.89	0.91

Note: MAE, absolute prediction error; RMSE, root mean square error; R^2, coefficient of determination; and LCCC, Lin's concordance correlation coefficient.

3.3. Relative importance of environment variables

Three BRT models are iterated 100 times, and the average value of RI of environment variables in predicting SOC of each model is obtained respectively. The RI of each variable was then scaled so that the sum was added as a percentage to 100. The RI values of environmental variables in predicting SOC in the study area is displayed in Fig. 5. The variables showed different levels of importance in predicting SOC content of topsoil agroecosystem in Northeast China. Anthropogenic variables had the largest RI (50.6%), followed by topographic (23.02%), and climatic variables (20.07%). The biological variables showed the least importance (6.31%). Among the anthropogenic variables, PER had a highest influence (RI > 15%) and DSE had the lowest (RI 5.3%) of all. POP and GDP shared a comparable influence of about 11% RI.

3.4. Spatial distribution of SOC content

The predicted maps of SOC content in the study area from MA, MB, and MC models are shown in Fig. 6. The average SOC contents in the study area based on MA, MB, and MC models were 26.2 g kg^{-1}, 25.3 g kg^{-1}, and 25.8 g kg^{-1}, respectively. Overall, all three maps showed similar SOC distribution pattern in the study area, a gradually decreased SOC from northeast to southwest. To further demonstrate the difference between MB and MC model, we generated a difference map of SOC content predicted by MB, and MC models (Fig. 7). Although the average SOC difference between the maps was very low (0.6 g kg^{-1}), the predicted SOC by MB model in most of the study area was lower than that of MC model.

4. Discussion

4.1. Anthropogenic variables and SOC

Our results showed that adding anthropogenic variables in the SOC prediction model could improve model performance (Table 3), and results are consistent with previous findings (Wang et al., 2019; Vasenev

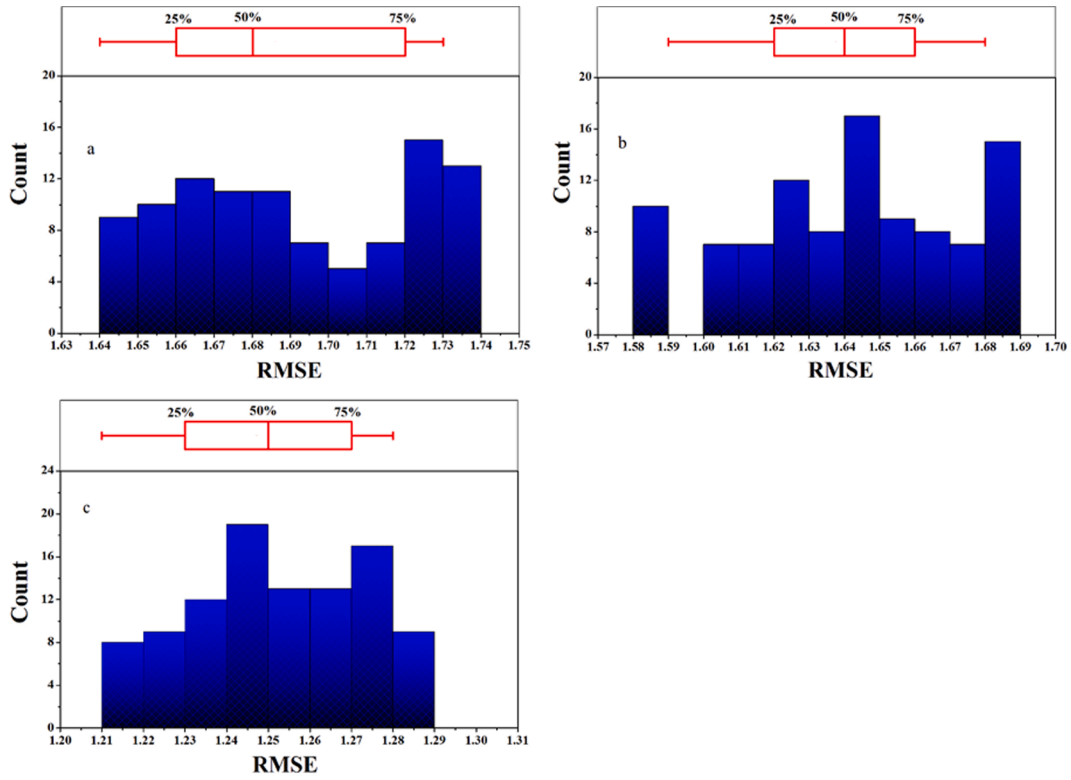

Fig. 2. RMSE distributions from BRT in predicting SOC based on 100 iterations. (a) MA model (only anthropogenic variables); (b) MB model (only topography and climate variables); (c) MC model (included all topography, climate, and anthropogenic variables).

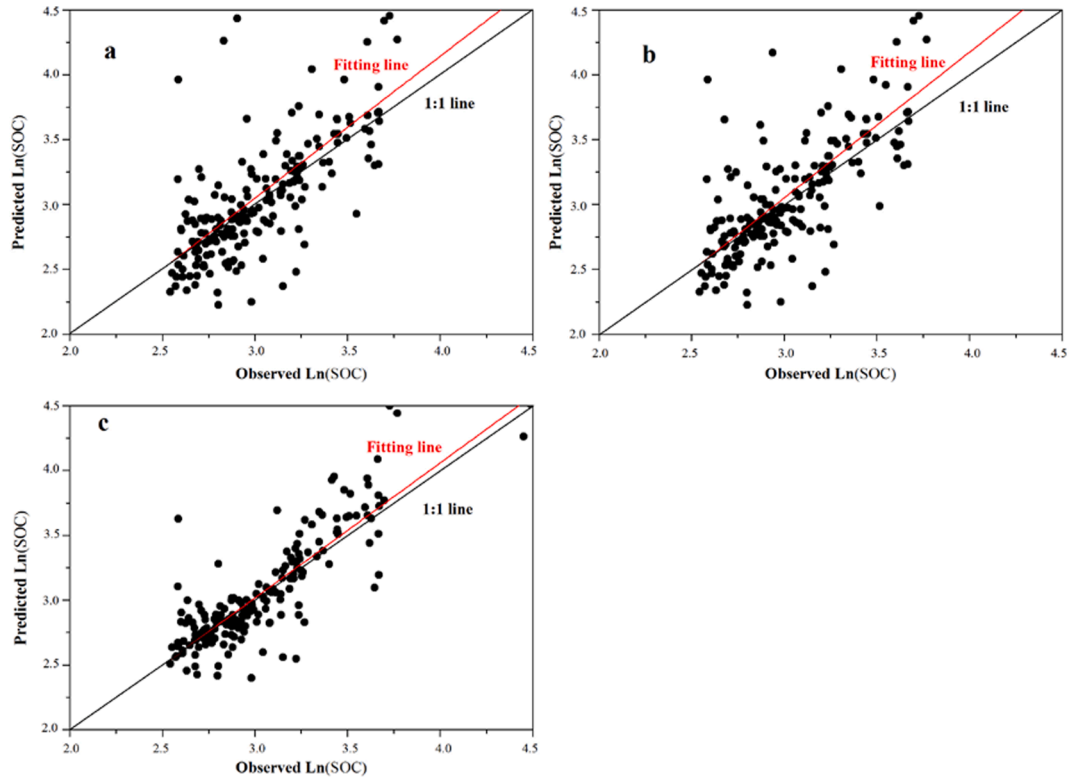

Fig. 3. Scatter plot between the observed SOC with its predicted values using the BRT model based on the 196 sampling point. (a) MA model (only anthropogenic variables); (b) MB model (only topography and climate variables); (c) MC model (included all topography, climate, and anthropogenic variables).

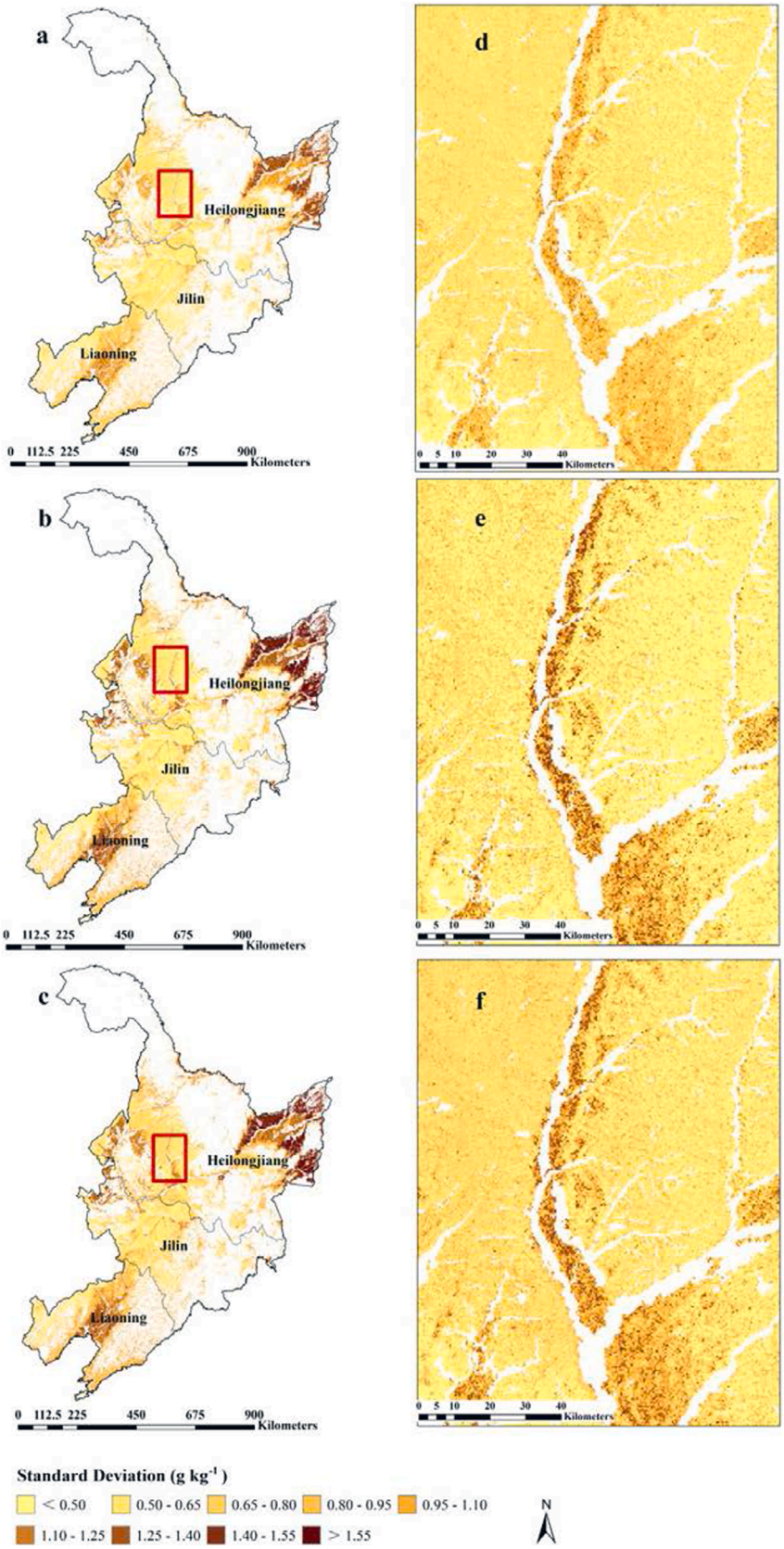

Fig. 4. Uncertainty map of SOC prediction carbon (g kg^{-1}). (a) MA model (only anthropogenic variables); (b) MB model (only topography and climate variables); (c) MC model (included all topography, climate, and anthropogenic variables); (d), (e) and (f) represent zoomed in areas in map (a), (b), and map (c), respectively. The areas in white are where no predictions were made.

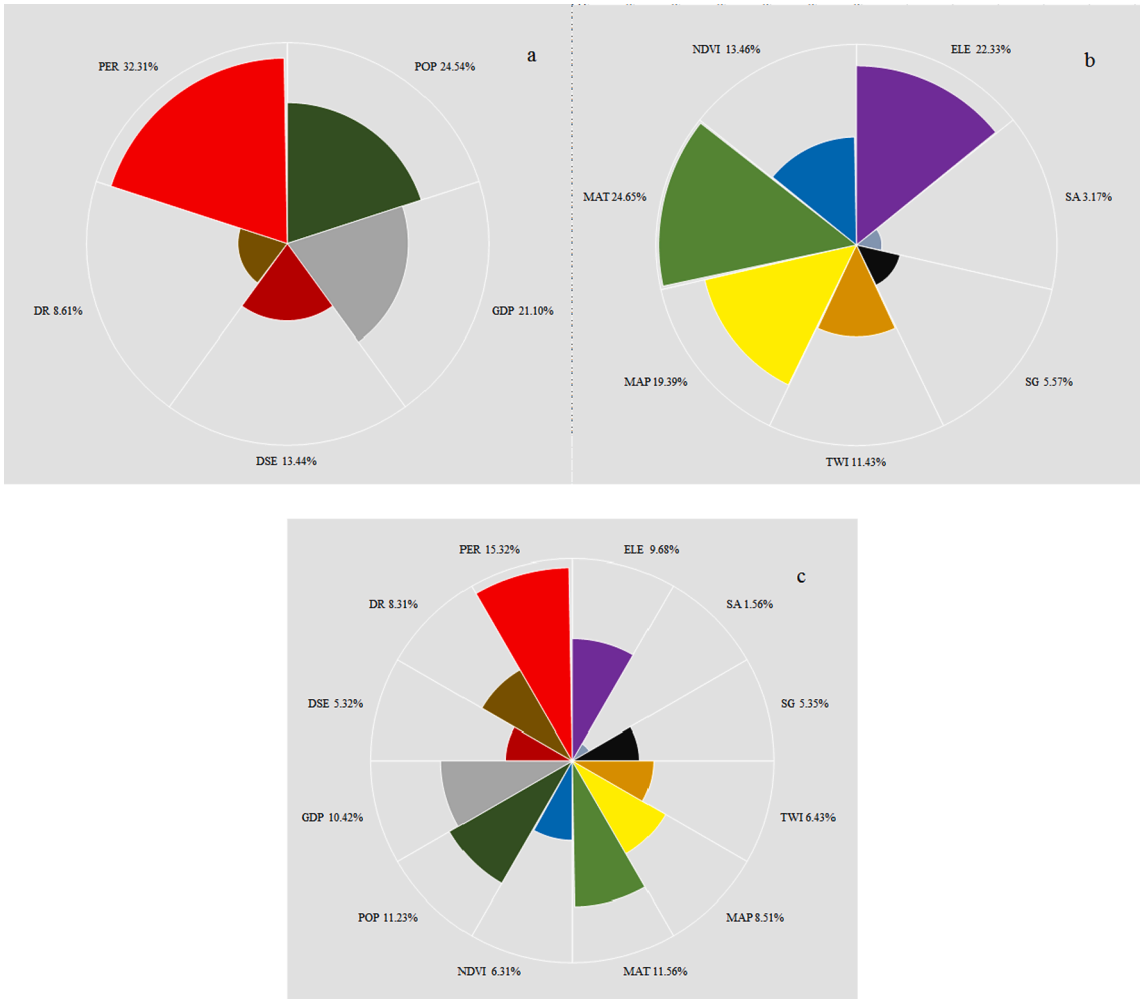

Fig. 5. Relative importance (RI) of SOC predictors in (a) MA model (only anthropogenic variables), (b) MB model (only topography and climate variables), and (c) MC model (included all topography, climate, and anthropogenic variables).

et al., 2018; Wang et al., 2020a). Wang et al. (2020a) found anthropogenic and related variables as important predictors of SOC affecting the SOC distribution in rapidly urbanized areas in Northeast China. It was also pointed out that there was a strong and a negative correlation between anthropogenic variables (POP and GDP) and SOC content in the topsoil. Increasing POP and economic development have led to an exponential increase in land use changes (Liu et al., 2016; Stumpf et al., 2018), resulting in large changes in soil carbon stocks, impacting the global climate. Wang et al. (2019) to include anthropogenic variables (POP and GDP) as potential SOC predictors to map SOC in areas with strong anthropogenic disturbance. Present study also reported the similar results. Few other studies, for example, e.g., Vasenev et al. (2014) and Wang et al. (2019) have also shown that anthropogenic related variables could impact SOC distribution in agroecosystem. Furthermore, variables like PER, GDP, and POP could reflect the input and cultivation level to a certain extent, which had an indirect impact on the SOC content in the topsoil (Vasenev et al., 2014, 2018; Wang et al., 2020a).

We found PER, GDP and POP as the most effective variables affecting SOC distribution the study area, and all variables represent anthropogenic influence in agroecosystems. This finding was consistent with Wang et al. (2019) who found PER as a key variable of SOC distribution in Norther eastern Chinese agroecosystems. Vasenev et al. (2014) introduced GDP, POP, DSE, and DR variables to predict topsoil SOC in a rapidly urbanized area in Russia, and concluded that these factors could significantly improve the prediction accuracy. These indices could be used as proxy indicators related to human-induced processes reflecting the anthropogenic disturbance on topsoil SOC levels in agroecosystem. In the cultivated soil, the effects of tillage and other agricultural activities would affect soil hydrology and microbiology, resulting in the destruction of physical protection layer and the exposure of organic matter to decomposition (Stoécio et al., 2019; Topa et al., 2021). The increase of soil microbial activity, soil respiration, organic carbon decomposition and mineralization rates could lead to the decrease of soil organic matter content in agricultural soils. In addition, soil erosion could be an important cause to SOC depletion due to mechanical removal. Road construction could often cause important soil erosion phenomena to occur in neighboring areas (Ren et al., 2018). This study showed that combining anthropogenic variables together with other environmental variables as SOC predictors significantly improves SOC prediction and we recommend considering such variables in future SOC modeling studies, especially in areas with a long history of anthropogenic influence.

4.2. *Spatial variation of SOC and associated predictors*

The spatial variation of SOC content predicted by three BRT models with different combination of environmental variables showed a comparable spatial distribution pattern or trend (Fig. 6). In general, SOC showed a decreasing trend from northeast to southwest, with the highest SOC content in the north of the study area, which was attributed to the black soil as the main soil type. It could be attributed to the high latitude

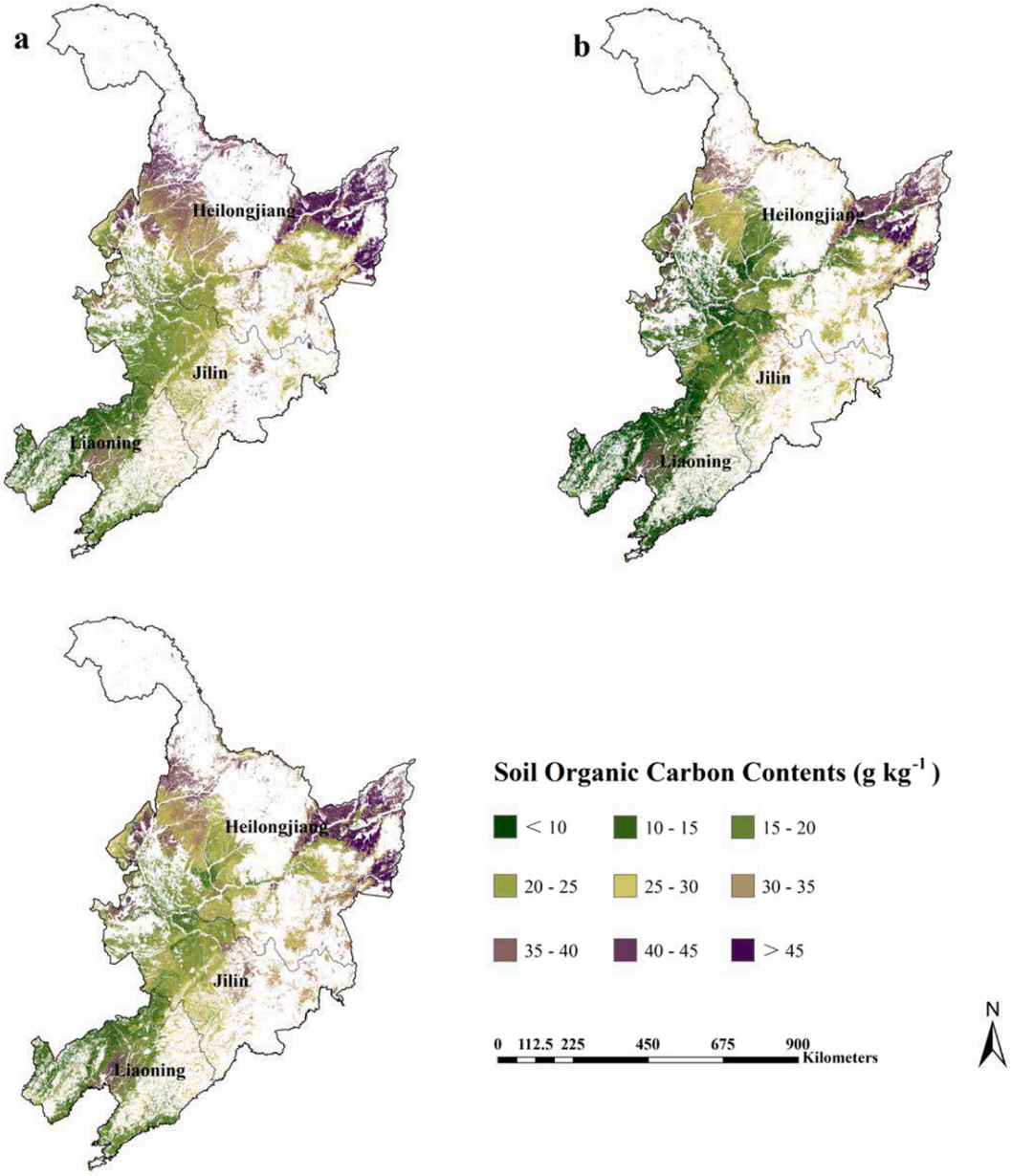

Fig. 6. Predicted maps of soil organic carbon (g kg^{-1}) distribution in the study area. Each map is an average of 100 predictions derived from 100 iteration of the BRT model. (a) MA model only anthropogenic variables; (b) MB model only topography and climate variables; (c) MC model included all predictors (topography, climate, and anthropogenic variables); (d), (e) and (f) represent zoomed in areas in map (a), (b), and map (c), respectively. The areas in white are where no predictions were made.

area in this region, so the long and cold winter had inhibited the soil microbial activity leading to the slow decomposition of organic matter and the accumulation of a large amount of humus in the upper part of the soil forming a deep black humus layer. However, with the increase of cultivated land reclamation time and human activities, the rate of SOC loss was accelerating rapidly in this area, which had been confirmed by many previous studies (Liu et al., 2006; Huang et al., 2010; Wang et al., 2019). The SOC content in southwest area was lower than northeast because it had the longest cultivated land reclamation period (>300 yrs). This area constitute the traditional agricultural area in China, which was deeply affected by human influence.

In the MC model (all variable prediction models), PER was the most important variable with a relative importance of 15.3%. This finding was similar to the previous research results of Vasenev et al. (2018), Wang et al. (2019), and Wang et al., (2020a) who reported PER as one of the main variables affecting SOC change in cultivated land. Overall, PER had some "positive" effects on topsoil SOC in the short term (Zhang et al., 2016), but in the long-term or with longer reclamation period, a negative impact on SOC could be expected, which increased mineralization rate (and related carbon dioxide emissions) (Vasenev et al., 2014; Li et al., 2018; Wang et al., 2019; Wang et al., 2020b). Therefore, the positive or negative impact of PER on SOC should be evaluated according to the PER period (short-term, medium-term, or long-term).

Temperature and precipitation were considered to be the key climate variables affecting SOC spatial variability (Baldock et al., 2012;, Adhikari et al., 2020), and similar findings were obtained in this study. In the MC model, MAT had a higher RI than MAP indicating a bigger role of MAT in SOC distribution in Northeast Chinese agroecosystems. Temperature and precipitation affected SOC content mainly by affecting crop productivity, and litter decomposition rate (Jobbágy and Jackson, 2000; Lal, 2004; Reyes Rojas et al., 2018). Topographic variables were widely used in predicting SOC research, especially in the area with large

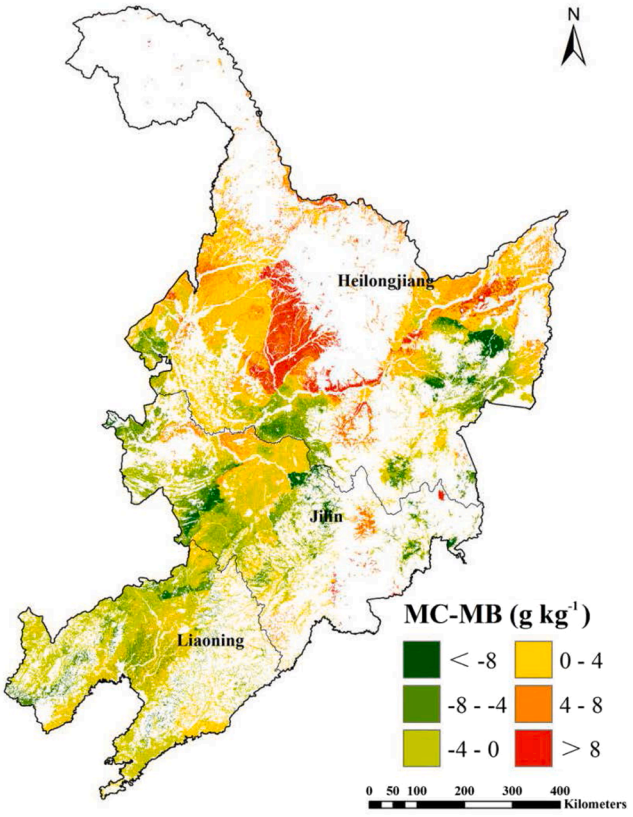

Fig. 7. Maps showing the SOC differences between MB (only topography and climate variables) and MC (topography, climate, and anthropogenic variables) models. The areas in white are where no predictions were made.

topographic variation (Chaminade, 2005; Yimer et al., 2006; Conforti et al., 2016; Guo et al., 2019; Fissore et al., 2017; Adhikari et al., 2020). Among all topographic variables, elevation was the most important variable, followed by TWI, SG, and SA. Elevation had been proved to be the most effective covariate of SOC in previous studies (e.g., Jobbágy and Jackson, 2000, Chaminade, 2005; Yimer et al., 2006; Wang et al., 2019; Wang et al., 2020a; Zhang et al., 2021). Elevation was closely related to the spatial distribution pattern of SOC, as shown in the Figs. 1, and 6. In the small terrain of montane ecosystems, elevation affected microclimate, thus affecting SOC distribution (Wang et al., 2016).

4.3. Model uncertainty

Although most uncertainties associated with SOC prediction and mapping were accounted for, there were other sources of uncertainties as well, that were not evaluated. For example, the reclamation period data was sorted out from the historical data, and a unified value was assigned to the county as the smallest unit, which could influence model accuracy. The GDP, and POP data used in this study were obtained from statistical yearbooks but the way they were primarily collected might not be error free. Same could be true with climatic, topographic, and biological variables. Other possible error sources could be associated with the clustering of environmental variables into sampling units, soil sampling strategy, and GIS operations itself. However, these sources of error are inevitable in DSM, and we did not assess such errors as it was beyond the scope of this study.

5. Conclusions

Three BRT models were used to predict SOC spatial distribution in topsoil (0–30 cm) of agroecosystem in Northeast China using a wide range of environmental variables as SOC predictors. The MC model that included anthropogenic variables as SOC predictors greatly improved the prediction performance compared to the MA and MB models that excluded such variables. MC model had a higher R^2, and LCCC, and lower MAE, and RMSE compared to MA and MB. The average SOC content predicted by MC model was 25.8 g kg^{-1}, and the model could explain 78% variations in SOC measurements. Among the anthropogenic variables, PER, GDP, and POP were the most important environmental variables affecting SOC distribution in the study area. These variables directly reflect SOC footprints of anthropogenic influence in soils. Therefore, future SOC mapping research, especially in the areas with rapid economic development, anthropogenic variables should be considered as potential SOC predictors. We believe that our SOC content map will have a positive impact on land use decision-making and agricultural management in the study region.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are thankful to the editor and anonymous reviewers for their constructive comments on the manuscript. This work was supported by the China Postdoctoral Science Foundation (Grant No. 2019M660782), National Science and Technology Basic Resources Survey Program of China (Grant No. 2019FY101300), and Young Scientific and Technological Talents Project of Liaoning Province (Grant No. LSNQN201910 and LSNQN201914). USDA is an equal opportunity provider and employer.

References:

Adhikari, K., Hartemink, A.E., 2017. Soil organic carbon increases under intensive agriculture in the central sands, Wisconsin, USA. Geoderma Regional 10, 115–125.
Adhikari, K., Hartemink, A.E., Minasny, B., Kheir, R.B., Greve, M.B., Greve, M.H., 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS One 9 (8), e105519.
Adhikari, K., Mishra, U., Owens, P.R., Libohova, Z., Smith, D.R., 2020. Importance and strength of environmental controllers of soil organic carbon changes with scale. Geoderma 375, 1–13.
Baldock, J.A., 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.
Chaminade, G., 2005. Topography, soil carbon-nitrogen ratio and vegetation in boreal coniferous forests at the landscape level. A Master of Science Thesis in Soil Sciences at the Department of Forest Soils at the Swedish University of Agricultural Sciences.
Conforti, M., Lucà, F., Scarciglia, F., Matteucci, G., Buttafuoco, G., 2016. Soil carbon stock in relation to soil properties and landscape position in a forest ecosystem of southern Italy (Calabria region). Catena 144, 23–33.
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8, 1991–2007.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.
Fissore, C., Dalzell, B.J., Berhe, A.A., Voegtle, M., Evans, M., Wu, A., 2017. Influence of topography on soil organic carbon dynamics in a southern california grassland. Catena 149, 140–149.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28, 337–407.
García-Palacios, P., Crowther, T.W., Dacal, M., Hartley, I.P., Bradford, M.A., 2021. Author Correction: Evidence for large microbial-mediated losses of soil carbon under anthropogenic warming. Nat Rev Earth Environ 2, 585.
Guo, Z., Adhikari, K., Chellasamy, M., Greve, M.B., Owens, P.R., Greve, M.H., 2019. Selection of terrain attributes and its scale dependency on soil organic carbon prediction. Geoderma 340, 303–312.
Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., 2013. Dismo: Species Distribution Modeling. packages/dismo/vignettes/sdm.pdf R Package Version 8–17. http ://www.idg.pl/mirrors/CRAN/web/.
Huang, Y., Sun, W., Zhang, W., Yongqiang, Y.U., Yanhua, S.U., Song, C., 2010. Marshland conversion to cropland in northeast china from 1950 to 2000 reduced the greenhouse effect. Global Change Biology 16 (2), 680–695.

IUSS Working Group, 2014. World Reference Base for Soil Resources 2014 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. FAO, Rome.
Jobbágy, E.G., Jackson, R.B., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10, 423–436.
Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189, 627–634.
Kirschbaum, M.U.F., 1995. The temperature dependence of soil organic matter decomposition, and the effect of global warming on soil organic c storage - sciencedirect. Soil Biology Biochemistry 27 (6), 753–760.
Lal, R., 2004. Soil C sequestration impacts on Global Climatic Change and Food Security. Science 304 (5677), 1623–1627.
Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review - sciencedirect. Geoderma 352, 395–413.
Li, J., Yang, W., Li, Q., et al., 2018. Effect of reclamation on soil organic carbon pools in coastal areas of eastern China. Front. Earth Sci. 12, 339–348.
Lin, L., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268.
Liu, D., Wang, Z., Zhang, B., Song, K., Duan, H., 2006. Spatial distribution of soil organic carbon and analysis of related factors in croplands of the black soil region, northeast china. Agriculture Ecosystems & Environment 113 (1–4), 73–81.
Liu, X., Li, T., Zhang, S., Jia, Y., Li, Y., Xu, X., 2016. The role of land use, construction and road on terrestrial carbon stocks in a newly urbanized area of western Chengdu. China. Lands. Urban Plan. 147, 88–95.
Ling, Y.I., Xiong, L.Y., Yang, X.H., 2006. Method of Pixelizing GDP Data Based on the GIS. J. Gansu Sci.
Liu, H., Jiang, D., Yang, X., et al., 2005. Spatialization Approach to 1km Grid GDP Supported by Remote Sensing. Geo-information Sci. 7 (2), 120–123.
Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital mapping of soil carbon. Advances in Agronomy. 118. Academic Press, pp. 1-47.
Manap, M.A., Sulaiman, W.N.A., Ramli, M.F., et al., 2013. A knowledge-driven GIS modeling technique for groundwater potential mapping at the Upper Langat Basin, Malaysia. Arab J Geosci 6, 1621–1637.
Ottoy, S., Vos, B.D., Sindayihebura, A., Hermy, M., Orshoven, J.V., 2017. Assessing soil organic carbon stocks under current and potential forest cover using digital soil mapping and spatial generalisation. Ecological Indicators 77, 139–150.
Pal, N., Pal, K., Keller, J.M., Bezdek, J.C., 2005. A possibilistic fuzzy c-means clustering algorithm. IEEE Transactions on Fuzzy Systems 13 (4), 517–530.
Pouteau, R., Rambal, S., Ratte, J.P., Gogé, F., Joffre, R., Winkel, T., 2011. Downscaling MODIS-derived maps using GIS and boosted regression trees: The case of frost occurrence over the arid Andean highlands of Bolivia. Remote Sens. Environ. 115, 117–129.
Perkins, T., Adlergolden, S., Matthew, M., Berk, A., Anderson, G., Gardner, J., 2005. Retrieval of atmospheric properties from hyper and multispectral imagery with the FLAASH atmospheric correction algorithm. International Society for Optics and Photonics, Orlando, USA.
R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2013, Vienna, Austria. Available online: https://www.rproject.org/ (accessed on 7 March 2013).
Ren, CJ, Wang, YD, Deng, & Zhao, et al. (2018). Differential soil microbial community responses to the linkage of soil organic carbon fractions with respiration across land-use changes. FOREST ECOL MANAG, 2018,409(-), 170-178.
Reyes Rojas, L.A., Adhikari, K., Ventura, S.J., 2018. Projecting soil organic carbon distribution in central Chile under future climate scenarios. J. Environ. Qual. 47 (4), 735–745.
Sanderman, J., Hengl, T., Fiske, G.J., 2017. Soil carbon debt of 12,000 years of human land use. Proc. National Academy Sci. 114 (36), 9575–9580.
Sokka, T., Kautiainen, H., Pincus, T., Toloza, S., Da, R.C.P.G., Lazovskis, J., et al., 2009. Disparities in rheumatoid arthritis disease activity according to gross domestic product in 25 countries in the quest-ra database. Ann. Rheum. Dis. 68 (11), 1666–1672.
Stoécio, Malta, Ferreira, Maia, Adriana, & Tamie, et al. 2019. Combined effect of intercropping and minimum tillage on soil carbon sequestration and organic matter pools in the semiarid region of brazil. Soil Research, 57(3), 266-275.
Stumpf, F., Keller, A., Schmidt, K., Mayr, A., Gubler, A., Schaepman, M., 2018. Spatio-temporal land use dynamics and soil organic carbon in Swiss agroecosystems. Agric. Ecosyst. Environ. 258, 129–142.
Seto, K.C., Reenberg, A., Boone, C.G., Fragkias, M., Haase, D., Langanke, T., et al., 2012. Urban land teleconnections and sustainability. Proc. Nat. Academy Sci. USA 109 (20), 7687–7692.
Topa, D., Cara, I.G., Jitreanu, G., 2021. Long term impact of different tillage systems on carbon pools and stocks, soil bulk density, aggregation and nutrients: a field meta-analysis. Catena 199 (2), 105102.
Vasenev, V.I., Stoorvogel, J.J., Leemans, R., Valentini, R., Hajiaghayeva, R.A., 2018. Projection of urban expansion and related changes in soil carbon stocks in the Moscow Region. J. Clean. Prod. 170, 902–914.
Vasenev, V.I., Stoorvogel, J.J., Vasenev, I.I., Valentini, R., 2014. How to map soil organic carbon stocks in highly urbanized regions? Geoderma 226, 103–115.
Wang, S., Adhikari, K., Zhuang, Q., Gu, H., Jin, X., 2020a. Impacts of urbanization on soil organic carbon stocks in the northeast coastal agricultural areas of china. Sci. Total Environ. 721, 137814.
Wang, S., Zhuang, Q., Jia, S., Jin, X., Wang, Q., 2018. Spatial variations of soil organic carbon stocks in a coastal hilly area of China. Geoderma 314, 8–19.
Wang, S., Wang, Q., Adhikari, K., Jia, S., Jin, X., Liu, H., 2016. Spatial-temporal changes of soil organic carbon content in Wafangdian. China. Sustainability 8, 1154.
Wang, Y., Wang, S., Adhikari, K., Wang, Q., Sui, Y., Xin, G., 2019. Effect of cultivation history on soil organic carbon status of arable land in northeastern china. Geoderma 342, 55–64.
Wang, S., Gao, J., Zhuang, Q., Lu, Y., Jin, X., 2020b. Multispectral remote sensing data are effective and robust in mapping regional forest soil organic carbon stocks in a northeast forest region in china. Remote Sensing 393.
Yimer, F., Ledin, S., Abdelkadir, A., 2006. Soil organic carbon and total nitrogen stocks as affected by topographic aspect and vegetation in the Bale Mountains, Ethiopia. Geoderma 135, 335–344.
Zhang, Y., Ai, J., Sun, Q., Li, Z., Hou, L., Song, L., Tang, G., Li, L., Shao, G., 2021. Soil organic carbon and total nitrogen stocks as affected by vegetation types and altitude across the mountainous regions in the Yunnan Province, south-western China. Catena 196, 104872.
Zhang, F., Li, C., Wang, Z., Li, X., 2016. Long-term effects of management history on carbon dynamics in agricultural soils in Northwest China. Environmental Earth Sciences 75 (1), 1–9.
Zhu, A.X., Yang, L., Li, B., Qin, C., English, E., Burt, J.E., Zhou, C.H., 2008. Purposive sampling for digital soil mapping for areas with limited data. In Digital Soil Mapping with Limited Data; Hartemink, A.E., McBratney, A.B., de Lourdes, M.-S., Eds.; Springer: Amsterdam, The Netherlands, Volume 8, pp. 233-245.