Science of the Total Environment 669 (2019) 844–855

Contents lists available at ScienceDirect

Science of the Total Environment

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

Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms

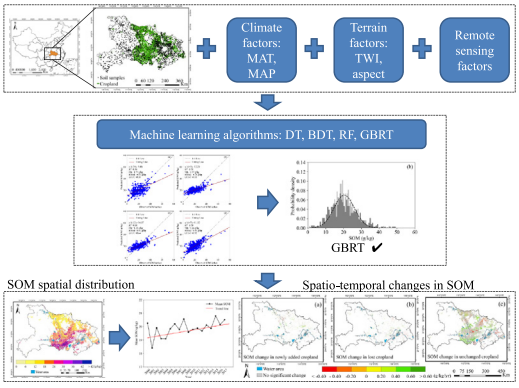
Di Chen [a,b,c,1], Naijie Chang [a,c,1], Jingfeng Xiao [c,*], Qingbo Zhou [a,b], Wenbin Wu [a,b,*]

[a] Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[b] Key Laboratory of Agricultural Remote Sensing, Ministry of Agriculture, Beijing 100081, China
[c] Earth Systems Research Center, Institute for the Study of Earth, Oceans, and Space, University of New Hampshire, Durham, NH, 03824, USA

HIGHLIGHTS

- GBRT was a better algorithm for spatially predicting and mapping SOM content in Hubei, China than DT, BDT, and RF.
- Remote sensing reflectance and vegetation indices were proved to be key factors for predicting SOM content.
- The SOM content in the topsoil in 2017 varied from 0.89 to 58.86 g/kg, with a mean value of 20.52 g/kg.
- The mean cropland SOM content of Hubei exhibited a slight increasing trend from 2000 to 2017.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 31 December 2018
Received in revised form 2 March 2019
Accepted 10 March 2019
Available online 11 March 2019

Editor: Jose Julio Ortega-Calvo

Keywords:
Digital soil mapping
Multi-year
Soil organic carbon
MODIS
Machine learning algorithms
Cropland

ABSTRACT

As an important indicator of soil quality, soil organic matter (SOM) significantly contributes to land productivity and ecosystem health. Accurately mapping SOM at regional scales is of critical importance for sustainable agriculture and soil utilization management and remains a grand challenge. Many studies used soil sampling data and machine learning algorithms to predict SOM at regional scales for a given year, while few studies mapped SOM for multiple years and examined its temporal dynamics. We compared the performance of four machine learning algorithms: decision tree (DT), bagging decision tree (BDT), random forest (RF), and gradient boosting regression trees (GBRT) in mapping SOM in Hubei province, China over the 18-year period from 2000 to 2017. Our results showed that RF and DT had the highest coefficient of determination (R^2) (0.61) and the lowest potential bias (9.48 g/kg), respectively, while GBRT had the lowest mean error (ME) (1.26 g/kg), root mean squared error (RMSE) (5.41 g/kg) and Lin's concordance correlation coefficient (LCCC) (0.72). The SOM map based on GBRT better captured the distribution of the soil sample data than that based on RF. The trained GBRT model and the spatially explicitly data on explanatory variables (e.g., climate, terrain, remote sensing) were used to predict SOM for each 500 m × 500 m grid cell in Hubei for the period from 2000 to 2017. Our results showed that the SOM content of cropland was relatively high in the southeast and relatively low in the north. The SOM content in the topsoil varied from 0.89 to 58.86 g/kg and was averaged at 20.52 g/kg. The mean cropland SOM content of the province exhibited an increasing trend from 2000 to 2017 with an increase of 0. 26 g/kg and a growth

* Corresponding authors.
E-mail addresses: j.xiao@unh.edu (J. Xiao), wuwenbin@caas.cn (W. Wu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scitotenv.2019.03.151

rate of 1.28%. Spatially, the SOM content increased in southern Hubei and decreased in central and northern parts of the province. A large portion of the areas with decreasing SOM content in northern Hubei was reclaimed cropland, while a large part of the high-quality cropland with rising SOM content in the east (~0.45 × 10^4 ha) was lost due to land use change (e.g., urbanization).

1. Introduction

Soil organic matter (SOM), reflecting soil quality and health (Zhang et al., 2006), is an important element of terrestrial ecosystems (Li et al., 2013) due to its great potential to affect the climate, food security, and agricultural sustainability (Were et al., 2015). As the major constituent and chemical measure of SOM, soil organic carbon (SOC) plays an important role in the global carbon cycle (Kumar and Lal, 2011; Yang et al., 2016). Quantifying the spatial distribution and temporal dynamics of SOM content is thus of great importance for informing climate policymaking and soil management (Meersmans et al., 2008), increasing food production (Taghizadeh-Mehrjardi et al., 2016) and providing essential benchmarking data for ecosystem models (Li et al., 2003). Moreover, accurate information on the spatio-temporal variations of SOM is useful for land use planning and other activities related to forestry, agriculture, environment protection and land degradation management (Li et al., 2013).

Conventional field soil mapping techniques require a large amount of soil morphological data (e.g., soil texture) (Taghizadeh-Mehrjardi et al., 2014), while the collection of soil morphological data is usually time-consuming and costly. Moreover, conventional techniques have been criticized for their subjective and qualitative nature (Taghizadeh-Mehrjardi et al., 2016). With the development of machine learning methods and the increasing availability of remote sensing data streams, digital soil mapping (McBratney et al., 2003) is becoming a cost-effective solution to these problems (Camera et al., 2017). With digital soil mapping techniques, soil properties (e.g., SOM) can be quantitatively predicted by formulating relationships between field soil observations and readily measured environmental and ecological data (McBratney et al., 2003; Grimm et al., 2008). Advances in remote sensing during the last few decades have made it possible to obtain a large amount of spatially explicit information on a suite of ecological and environmental variables at regional to global scales at a relatively low cost (Doetterl et al., 2013). Remote sensing has become an important source of spatial data for predicting SOM at regional scales (Mulder et al., 2011; Were et al., 2015; Vågen et al., 2016).

In recent studies, the relationships between field soil observations and readily measured environmental data were built by machine learning algorithms which overcome the shortcomings of parametric and non-parametric statistical methods (Were et al., 2015) and have more potential for digital SOM mapping (Taghizadeh-Mehrjardi et al., 2016). A wide range of machine learning algorithms have been used to predict SOM at regional (Yang et al., 2016) and national (Li et al., 2013; Camera et al., 2017) scales. For example, Wang et al. (2017) examined the spatial distribution of SOC at five soil depth intervals in Liaoning, China using a boosted regression tree model. Wiesmeier et al. (2011) used the random forest (RF) method to estimate SOC stocks in the Xilin River basin in Inner Mongolia, China. However, because of the spatial variability in climate, soil properties, and land use management, no single algorithm is universally applicable (Kumar and Lal, 2011). Therefore, some researchers compared the performance of different algorithms for mapping SOM (Somarathna et al., 2016; Were et al., 2015). For example, Somarathna et al. (2016) compared multiple linear regression, regression tree model, and support vector regression (SVR) for SOC mapping in New South Wales, Australia. Were et al. (2015) used soil samples, climatic data, topographic data, and remote sensing data to estimate SOC stocks in the Eastern Mau Forest Reserve, Kenya, and compared the performance of SVR, artificial neural network, and RF models. However, most previous studies predicted soil properties (e.g., SOM) for a specific year. Quantifying the spatio-temporal variations of SOM over a multi-year period is critical for agricultural production and land management. To date, most studies on SOC dynamics were based on either ecosystem models like DNDC (Li et al., 2003; Wang et al., 2008; Zhang et al., 2017) or meta-analysis (i.e., analysis based on data from a large number of published studies) (Huang and Sun, 2006; Popeplau et al., 2011). These studies mainly focused on the period from the 1980s to the early 2000s. For example, Stockmann et al. (2015) used soil profile data to assess the global SOC change during the last five decades. However, soil profile data are not readily available in many regions or countries, particularly at decadal scales. Studies that directly use remote sensing data to map SOM at regional scales and to also analyze its trends at decadal scales are yet to emerge.

Here we compared four machine learning algorithms: decision tree (DT), bagging decision tree (BDT), RF, gradient boosting regression trees (GBRT) for predicting the SOM content in Hubei, one of the main grain production areas in China, and also mapped SOM at the provincial scale over the period 2000–2017. The surface reflectance and vegetation indices derived from the moderate resolution imaging spectroradiometer (MODIS), climatic variables and terrain factors were used as predictors in our study. Our objectives were to compare four different machine learning algorithms, map SOM at the regional scale, and examine the trends of predicted SOM from 2000 to 2017. Our results can inform agricultural land management and policymaking.

2. Study area and data

2.1. Study area

Hubei province, a transitional zone between northern and southern China, is located in Central China, approximately between 29° N and 33° N and between 108° E and 116° E (Fig. 1a). The region is dominated by a subtropical monsoon climate with four distinct seasons. The mean annual temperature is 15–17 °C; the coldest (3 °C) and hottest (29 °C) months are January and July, respectively. The rainy reason is from May to July. The mean annual precipitation varies spatially from 800 mm in the northwest to 1600 mm in the southeast. The province consists of an incomplete basin in the south-central region and mountains in the east, west and north (Fig. 1b). Mountainous, hilly and plain landscapes account for 56%, 24% and 20% of the total land area of the province (~18.59 × 10^6 ha), respectively. According to the Chinese Soil Taxonomy (Cooperative Research Group on Chinese Soil Taxonomy, 2001), paddy soil, fluvo-aquic soil and yellow-brown soil are the main soil types (50.4%, 19.0% and 14.5%), distributed in the plain, hilly and low mountain areas of the south central, northeastern and western parts of the province, respectively. The cropland area accounts for 28.6% (~5.32 × 10^6 ha) of the land area in the province, and the main crops are rice, oilseed rape, vegetables, cotton, wheat and corn. Hubei has always been one of the main grain production areas in China.

2.2. Soil samples

A provincial topsoil (0–20 cm) survey was conducted to sample across every natural village of Hubei province in 2017 (Fig. 1c). This provincial topsoil survey encompassed a large range of climate and soil

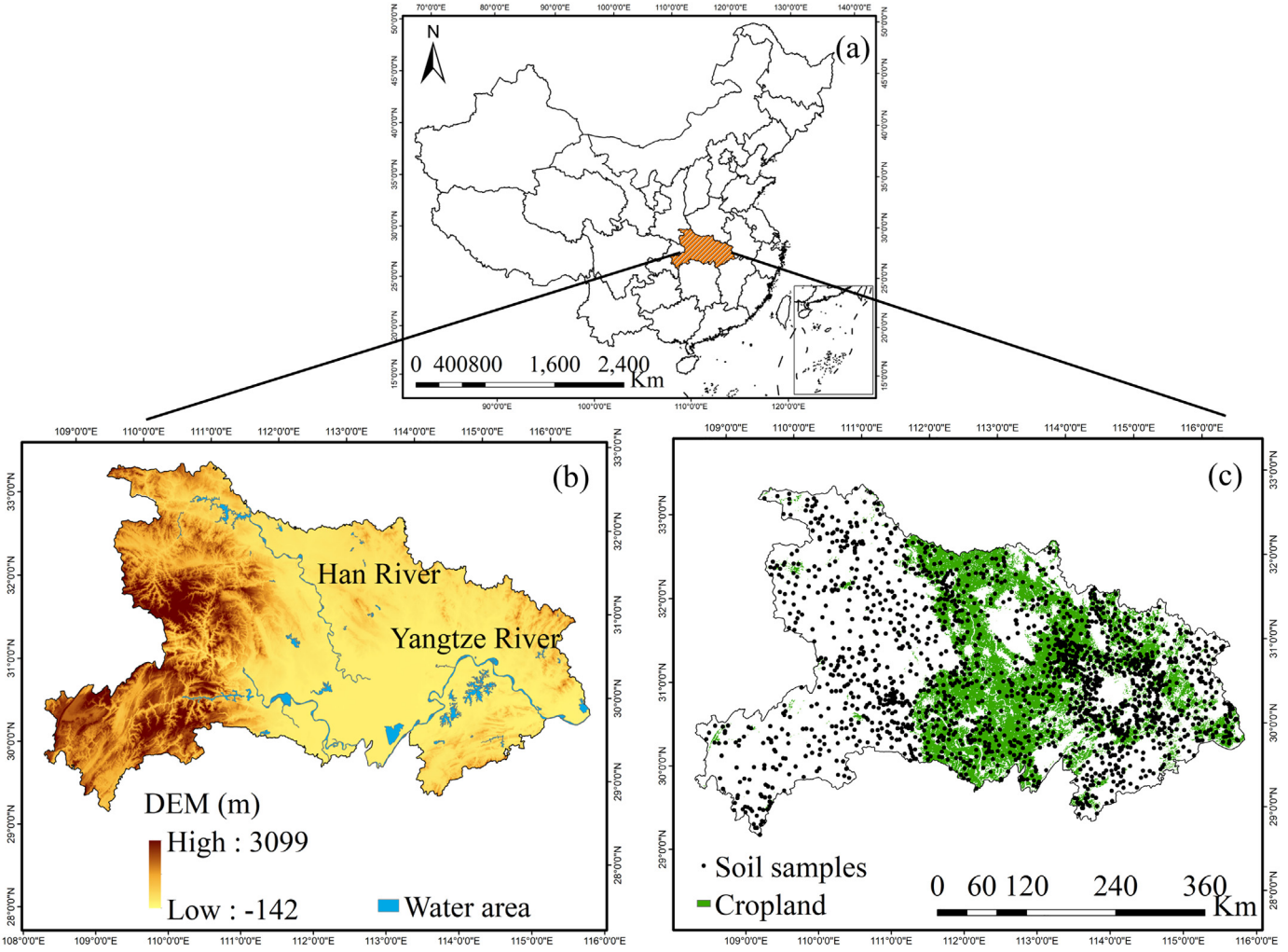


Fig. 1. Geographical location of the Hubei province (a), digital elevation model (DEM) (b), and distribution of cropland soil samples across the province (c).

types across 96 counties. At each site, soil samples were collected at five different locations randomly selected within a 10 m-radius area and the five samples were then mixed into one sample. These sites were all selected from croplands and their geographic locations were recorded with a global positioning system (GPS). In total, 1872 soil samples from 1872 natural villages were collected in the survey. All soil samples were air-dried and passed through a 0.25 mm mesh and were then analyzed for SOM using the Potassium dichromate volumetric method (Yeomans and Bremner, 1988). The measured SOM content in the topsoil varied from 1.52 to 71.69 g/kg, and the mean content was 20.29 g/kg. We randomly selected 70% (1310) of the soil samples as training data and used the remaining 30% (562) of the samples as validation data.

2.3. Explanatory variables

SOM content is controlled by multiple environmental and ecological factors and their interactions (Li et al., 2013). Based on a review of the literature (e.g., Liu et al., 2010; Kumar and Lal, 2011; Li et al., 2013; Were et al., 2015; Song et al., 2016; Yang et al., 2016; Wang et al., 2017), a suite of 25 explanatory variables representing climatic, terrain and ecological factors were chosen for the prediction of SOM (Table 1). The spatially explicit data on these explanatory variables were obtained and resampled to 500-m spatial resolution to be consistent with the resolution of MODIS data. For each variable, the values were extracted for the 500 m × 500 m pixel in which each soil sample is located. The extracted values were used as the input for a feature importance evaluation. The 10 most important variables were selected as predictors for this study based on the feature relative importance evaluation (See more details in Section 3.1). Table 1 lists all the 25 explanatory variables and the 10 selected variables.

2.3.1. Climatic factors

Generally, climatic factors determine the broad patterns of SOM content (McLauchlan, 2006; Li et al., 2013). Mean annual temperature (MAT) and mean annual precipitation (MAP) of the study area over 36 years (1980–2015) were obtained from the Resource and Environment Data Cloud Platform, Chinese Academy of Sciences (http://www.resdc.cn/). The gridded MAT and MAP data in China were interpolated from daily meteorological observations of >2400 meteorological stations across the country. For each year during the period 2000–2016, we used the MAT and MAP over the 20-year period prior to this year. For example, in 2000, we used the MAT and MAP over the period 1980–1999 for 2000 and MAT and MAP over the period 1996–2015 for 2016. For 2017, we used the mean values of the previous 19 years (1997–2016). The climate data are at 1-km spatial resolution and were resampled to 500 m using the nearest neighbor method.

2.3.2. Terrain factors

Four terrain factors including elevation, slope, aspect, and topographic wetness index (TWI) (Beven and Kirkby, 1979) were derived from the Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) with 90-m resolution (https://gdex.cr.usgs.gov/gdex/).

Table 1
The list of all 25 explanatory variables and the 10 selected variables.

Explanatory variable	Acronym/abbreviation	Data source	Selected for final model
Mean annual temperature	MAT	http://www.resdc.cn/	Yes
Mean annual precipitation	MAP	http://www.resdc.cn/	Yes
Elevation	DEM	http://www.resdc.cn/	No
Slope	slope	Calculated from DEM	No
Aspect	aspect	Calculated from DEM	Yes
Topographic wetness index	TWI	Calculated from DEM	Yes
mean MODIS red band 1	b1_mean	https://earthdata.nasa.gov	Yes
Mean MODIS near infrared band 2	b2_mean	https://earthdata.nasa.gov	Yes
Mean MODIS blue band 3	b3_mean	https://earthdata.nasa.gov	Yes
Mean MODIS green band 4	b4_mean	https://earthdata.nasa.gov	No
Mean MODIS mid infrared band 5	b5_mean	https://earthdata.nasa.gov	No
Mean MODIS shortwave infrared1 band 6	b6_mean	https://earthdata.nasa.gov	No
Mean MODIS shortwave infrared2 band 7	b7_mean	https://earthdata.nasa.gov	No
Mean normalized difference vegetation index	NDVI_mean	Calculated from MOD09A1	No
Maximum normalized difference vegetation index	NDVI_max	Calculated from MOD09A1	No
Mean enhanced vegetation index	EVI_mean	Calculated from MOD09A1	Yes
Maximum enhanced vegetation index	EVI_max	Calculated from MOD09A1	No
Mean ratio vegetation Index	RVI_mean	Calculated from MOD09A1	No
Maximum ratio vegetation index	RVI_max	Calculated from MOD09A1	Yes
Mean difference vegetation index	DVI_mean	Calculated from MOD09A1	Yes
Maximum difference vegetation index	DVI_max	Calculated from MOD09A1	No
Mean soil-adjusted vegetation index	SAVI_mean	Calculated from MOD09A1	No
Maximum soil-adjusted vegetation index	SAVI_max	Calculated from MOD09A1	No
Mean normalized difference water index	NDWI_mean	Calculated from MOD09A1	No
Maximum normalized difference water index	NDWI_max	Calculated from MOD09A1	No

The SRTM DEM dataset is perhaps the most widely used ancillary data source related to various soil properties and soil classes (McBratney et al., 2003). For areas with large terrain variations, elevation is an important factor affecting the magnitude and distribution of SOM (Yang et al., 2016). Slope and aspect impact soil erosion and are closely related to the spatial variation of SOM (Odeh et al., 1995). TWI is considered as a good indicator of SOM and soil moisture at different landscapes, and shows a significant correlation with the distribution of SOM (Zhang et al., 2012). TWI was calculated based on the following equation:

$$TWI = \ln\left(\frac{\alpha}{\tan\beta}\right) \quad (1)$$

where α is the specific catchment area (SCA), and tan β is the local slope gradient. SCA indicates the potential flow accumulation of a specific location, and tan β reflects the local drainage potential (Beven and Kirkby, 1979).

2.3.3. MODIS time-series data

We used the MODIS/Terra Surface Reflectance Collection 6 product (MOD09A1 V006) with 8-day composites (https://earthdata.nasa.gov). The MOD09A1 V6 product provides an estimate of the surface spectral reflectance of Terra MODIS bands 1–7 at 500 m resolution, and has already been cloud screened, atmospherically corrected, and transformed to a standard Sinusoidal projection. In our study, we used the annual mean reflectance values of the seven MODIS bands: b_1_mean (band 1, red), b_2_mean (band 2, near infrared), b_3_mean (band3, blue), b_4_mean (band 4, green), b_5_mean (band 5, mid-infrared), b_6_mean (band 6, shortwave infrared) and b_7_mean (band 7, shortwave infrared). The MODIS surface reflectance product was also used to calculate remote sensing indices: normalized difference vegetation index (NDVI) (Rouse Jr. et al., 1974), enhanced vegetation index (EVI) (Huete et al., 2002), ratio vegetation index (RVI) (Jordan, 1969), difference vegetation index (DVI) (Richardson and Weigand, 1977), soil-adjusted vegetation index (SAVI) (Huete, 1988), and normalized difference water index (NDWI) (Gao, 1996). The mean and maximum values of the six remote sensing indices were calculated as predictors: mean and maximum NDVI (NDVI_mean and NDVI_max), mean and maximum EVI (EVI_mean and EVI_max), mean and maximum RVI (RVI_mean and RVI_max), mean and maximum DVI (DVI_mean and DVI_max), mean and maximum SAVI (SAVI_mean and SAVI_max), and mean and maximum NDWI (NDWI_mean and NDWI_max). For each 8-day interval from 2000 to 2017, four MODIS tiles covering the entire Hubei region were obtained from the NASA EARTHDATA (https://earthdata.nasa.gov) and were mosaicked and clipped to the extent of the study area. NDVI was calculated as follows:

$$NDVI = \frac{b_2 - b_1}{b_2 + b_1} \quad (2)$$

where b_1 is the MODIS red band and b_2 is the MODIS near infrared band. EVI was calculated based on the following equation:

$$EVI = 2.5 \times \frac{b_2 - b_1}{b_2 + 6b_1 - 7.5b_3 + 1} \quad (3)$$

where b_3 is the MODIS blue band. RVI, DVI, and SAVI were calculated as follows:

$$RVI = \frac{b_2}{b_1} \quad (4)$$

$$DVI = b_2 - b_1 \quad (5)$$

$$SAVI = \frac{b_2 - b_1}{b_2 + b_1 + L}(1 + L) \quad (6)$$

where L is equal to 0.5. NDWI was calculated based on the following equation:

$$NDWI = \frac{b_2 - b_5}{b_2 + b_5} \quad (7)$$

where b_5 is the MODIS mid infrared band. For each year, the maximum and mean values of these remote sensing indices used in the study are the maximum and mean values of the year.

We also used the MODIS Collection 6 Land Cover Type product (MCD12Q1). The MCD12Q1 product provides global maps of land cover type at 500 m spatial resolution annually based on MODIS bands

1–7 and the EVI using a decision tree classification algorithm (Friedl et al., 2010). We used the MCD12Q1 product to identify all the cropland pixels in Hubei province for each year over the period from 2001 to 2016.

3. Methodology

3.1. Feature relative importance evaluation

A feature relative importance evaluation was used to assess the relative importance of each feature (i.e., explanatory variable) in predicting the target variable - SOM. The relative rank (i.e., depth) of a feature used as a decision node in a tree was calculated to measure the predictability of each variable. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples that a feature contributes to can thus be used as an estimate of its relative importance. The fraction of samples that a feature contributes to is combined with the decrease in impurity from splitting them to create a normalized estimate of the predictive power of that feature. The averaging of the estimates of the predictive ability over several randomized trees can reduce the variance of the estimate. This is known as the mean decrease in impurity, or MDI. The readers are referred to Louppe (2014) for more information on MDI and feature importance evaluation.

3.2. Machine learning approaches

DT is a non-parametric supervised learning method used for both classification and regression (Tien Bui et al., 2012) by creating a model that predicts the value of a target variable using simple decision rules inferred from data features. DT learns from training data to approximate a sine curve with a set of if-then-else decision rules. Deeper trees have more complex decision rules and better fitting but also more likely to overfit. In the DT regression, the average value of the output variable of a sample in leaf nodes is treated as the prediction result. The GridSearch Cross-validation (GridSearchCV) method (Bergstra and Bengio, 2012) which exhaustively searches over specified parameter values and finds the optimized one over a parameter grid was used to adjust the following two parameters: the maximum depth of the tree (max_depth) and the minimum number of samples required to be at a leaf node (min_samples_leaf). In this research, the optimal max_depth was set to 10, and min_samples_leaf was set to 1.

BDT is a bagging (ensemble) decision tree method that can repeatedly build multiple decision trees and then vote for a better prediction integrating these decision tree models. A bagging regression is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregates their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can reduce the variance of a black-box estimator (e.g., a decision tree) by introducing randomization into its construction procedure and then making an ensemble out of it. This randomization is reflected in the random repeated sampling of training samples. The process of adjusting parameters in BDT was similar to that in DT. In this research, BDT had the same parameter settings based on by the GridSearchCV method as DT.

RF is an algorithm that integrates multiple trees through ensemble learning (Breiman, 2001). Its basic unit is the decision tree, and multiple decision trees are combined to reduce the risk of over-fitting. The training process of multiple decision trees is parallel, while the decision trees are slightly different due to the random process added to the algorithm. It is worth mentioning that the random process of feature selection is different between RF and BDT. RF contains random selection of features, but BDT does not. The predicted value of the RF is the weighted mean of target variables of leaf nodes. The RF uses the Out-Of-Bag (OOB) data to provide reliable error estimates (Breiman, 1996; Breiman, 2001). In RF modeling, the following three parameters need to be specified: the number of trees in the forest (n_estimators), the minimum number of samples required to be at a leaf node (min_samples_leaf), and the maximum depth of the tree (max_depth). These parameters were set to 200, 1, and 40, respectively, by the GridSearchCV method.

Unlike bagging, boosting is a forward, stage-wise procedure in which tree models are iteratively fitted to the training data to increase attention to the poor observations by the existing collection of trees and to minimize the loss function (Elith et al., 2008). GBRT is a generalization of boosting to arbitrary differentiable loss functions. GBRT builds an additive model in a forward stage-wise fashion, and it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. The stochastic gradient boosting procedure can reduce overfitting and improve model performance (Friedman, 2002). GBRT can handle data of mixed type naturally and have a powerful predictability. Because of robust loss functions, GBRT has a good robustness to outliers in the output space. In GBRT modeling, six parameters are user defined: loss function (loss), the number of boosting stages to perform (n_estimators), the minimum number of samples required to split an internal node (min_samples_split), the minimum number of samples required to be at a leaf node (min_samples_leaf), learning rate (learning_rate), maximum depth of the individual regression estimators (max_depth). The GridSearchCV method was used to adjust these parameters: the optimal loss = 'huber', n_estimators = 200, min_samples_split = 3, min_samples_leaf = 1, learning_rate = 0.3, and max_depth = 10.

3.3. Model training and validation

We randomly selected 70% (1310) of the soil samples as training data to model the relationships between the SOM observations and the predictors. The DT, BDT, RF and GBRT algorithms were used for developing predictive SOM models. The remaining 30% (562) of the soil samples were used for validation. The coefficient of determination (R^2), root mean squared error (RMSE), mean error (ME) and Lin's concordance correlation coefficient (LCCC) were calculated using the 10-fold cross-validation procedure to evaluate the predictive performance of the predictive models. These statistical measures were calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(o_i - \overline{o})^2} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(o_i - p_i)^2}{n}} \tag{9}$$

$$ME = \sum_{i=1}^{n}\frac{o_i - p_i}{\mathrm{n}} \tag{10}$$

$$LCCC = \frac{2r\sigma_o\sigma_p}{\sigma_o^2 + \sigma_p^2 + (\overline{o} - \overline{p})^2} \tag{11}$$

where p_i is the SOM value of the validation sample i predicted by each model; o_i is the observed value of SOM as the reference; $\overline{p}$ and $\overline{o}$ are the means of the predicted SOM and observed SOM, respectively; n is the number of validation samples; σ_p and σ_o are the variances of predicted and observed values, respectively; and r is the Pearson correlation coefficient between the predicted and observed values.

3.4. Spatial prediction and data analysis

The 10 most important explanatory variables were selected through the feature importance evaluation and were used to train the four machine learning models. Then, we applied the trained DT, BDT, RF and GBRT models to predict SOM for each 500 m × 500 m pixel in Hubei in 2017. We compared the results of the four models and identified

the best performing model for the prediction of SOM. Finally, we used the best model to predict SOM for each 500 m × 500 m pixel in Hubei province for each year over the period from 2000 to 2017.

We calculated the transition matrix, a quantitative measure of the conversion between different land use types, of the MODIS C6 cropland products in 2001 and 2016, and also examined the spatial pattern of cropland in Hubei province from 2001 to 2016. For each pixel, we then analyzed the linear trend of SOM over the 18-year period by regressing SOM as a function of time on a per-pixel basis as follows:

$$y = a + bt \tag{11}$$

where y is SOM, t is time (year), and a and b are the intercept and slope, respectively. A positive slope indicates an increasing SOM trend, while a negative slope indicates a decreasing SOM trend. The SOM trend map can better reflect the temporal change of SOM and its spatial pattern in the study area.

Finally, we used the cropland change map to stratify the SOM trend map and then examined the changes of SOM in the newly added cropland, the lost cropland and the unchanged cropland, separately, during the 18-year period.

4. Results

4.1. Model evaluation

The feature relative importance evaluation was needed for DT, RF and GBRT but not for BDT, and therefore we evaluated the relative importance of the input variables for the first three models. Prior to the relative importance evaluation, we used a total of 25 explanatory variables to train the DT, BDT, RF and GBRT models and assessed their performance using the validation data. The R^2 of the DT, BDT, RF and GBRT models were 0.33, 0.49, 0.61 and 0.59, respectively. The relative importance of a predictor varied with model. We selected the 10 most important variables (MAT, MAP, aspect, TWI, b_1_mean, b_2_mean, b_3_mean, EVI_mean, RVI_max, and DVI_mean) as the final predictors for models training (Fig. 2). The R^2 values of the DT, BDT, RF and GBRT models based on the 10 most important variables were 0.36, 0.51, 0.61 and 0.58, respectively, and their predictive performance was almost identical with that of the models based on all 25 variables.

Fig. 3 shows the performance of the DT, BDT, RF and GBRT models for predicting SOM. When ME and RMSE approached zero, predictions became increasingly optimal. The GBRT model had the lowest ME (1.26 g/kg) and RMSE (5.41 g/kg); the RF model had slightly higher ME (1.36 g/kg) and RMSE (5.63 g/kg) values; the DT model had the highest ME (1.99 g/kg) and RMSE (6.81 g/kg) values. The GBRT model had the highest LCCC (0.72), followed by the RF model (0.64); the BDT model had the lowest LCCC (0.53). The DT model had the lowest potential bias (9.48 g/kg), followed by the GBRT model (11.02 g/kg); the RF model had the highest potential bias (14.07 g/kg).

The comparison of the three statistical measures (ME, RMSE and LCCC) among the models above indicated that the predictive ability of the RF model was close to that of the GBRT model, and therefore we compared the histogram in SOM between the soil samples and the SOM map predicted by each model (Fig. 4). The peak (~20 k/kg) of the SOM distribution of the soil samples coincided with the peak of the distribution of the SOM predicted by GBRT, while the SOM distribution by RF had a much higher peak. In addition, the RF model had much

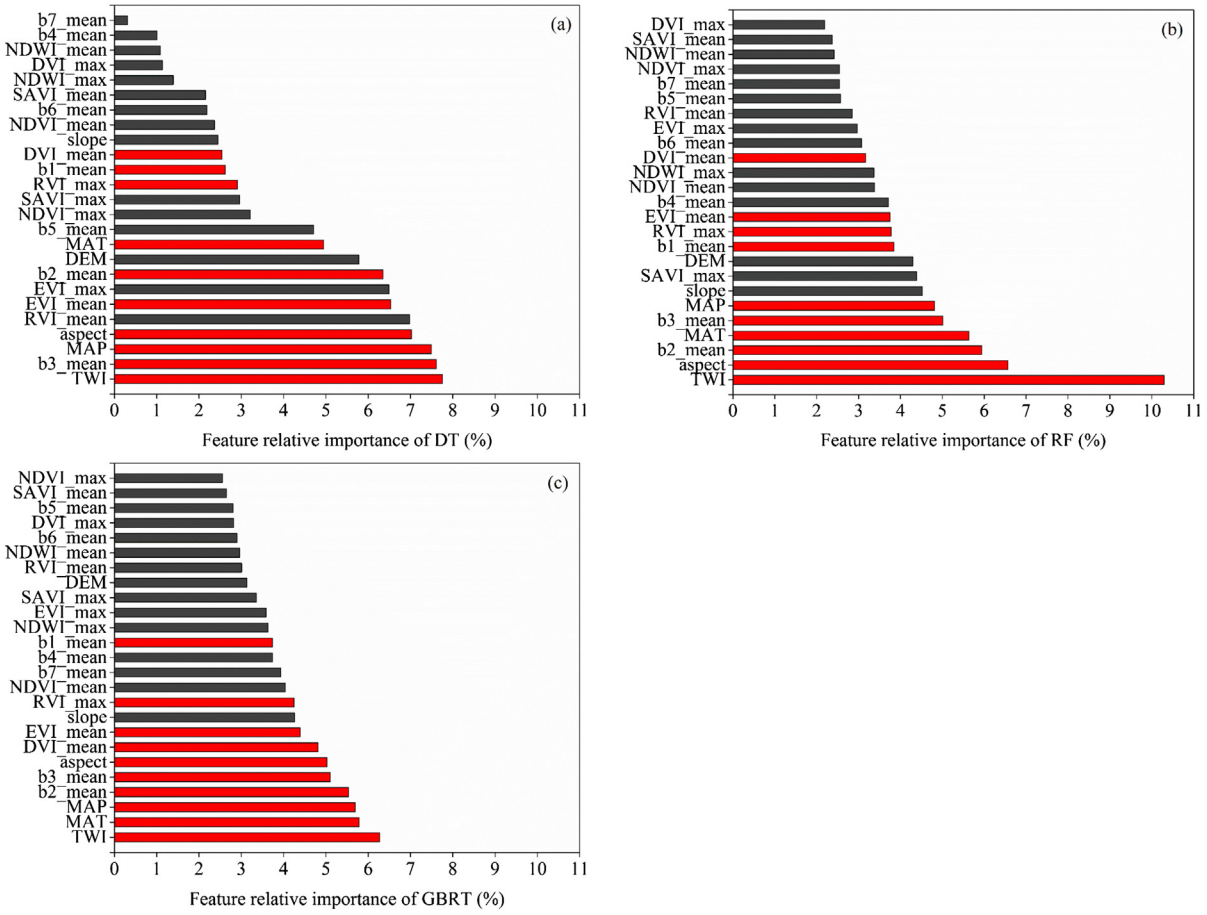


Fig. 2. Feature relative importance evaluation of DT (a), RF (b) and GBRT (c). The length of the bars indicates the relative importance of the explanatory variables, the red bar indicates the 10 variables that are important for all three models.

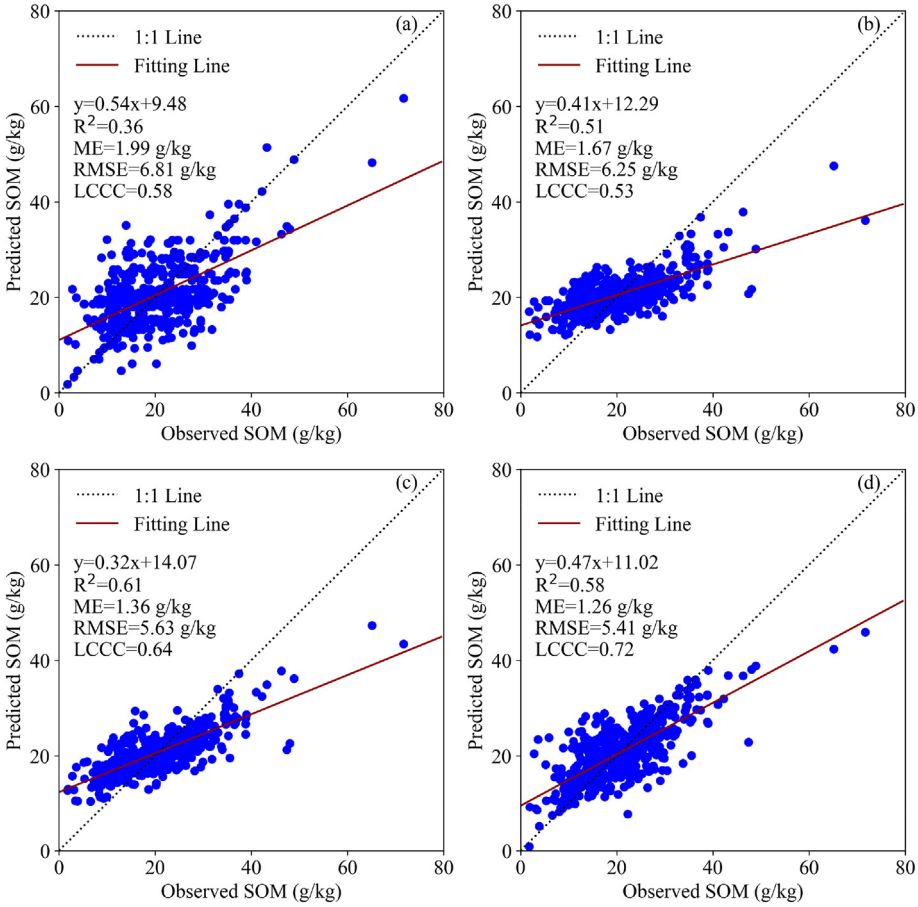


Fig. 3. Scatter plots of predicted SOM versus observed SOM from validation data for the DT (a), BDT (b), RF (c) and GBRT (d) models. The linear regression equations and statistical measures (R^2, ME, RMSE and LCCC) between observed SOM and predicted SOM are provided.

lower probability density in SOM at both lower (<15 g/kg) and higher (>25 g/kg) end of the range than the original soil samples, while the GBRT model had similar probability density to the original soil samples. Our results showed that the SOM map based on the GBRT model better captured the SOM distribution of the original soil samples than that based on the RF model. Therefore, compared with other machine learning approaches, GBRT had the best performance in capturing the magnitude and spatial distribution of SOM in Hubei province.

4.2. Spatial distribution of predicted SOM

As shown in Section 4.1, we found that among all the models, GBRT had the highest accuracy performance in predicting SOM in Hubei province. Therefore, we used the GBRT model to predict SOM for each 500 m × 500 m cropland pixel in the province for each year over the period 2000–2017. The changes in the explanatory variables derived from remote sensing data were used to reflect the changes in SOM content. For each year, we also used the mean values of temperature and precipitation over the 20 years prior to this year. The topography was assumed to remain constant during the study period. We then examined the magnitude and distribution of predicted SOM in 2017 (Fig. 5). The predicted SOM content of cropland was relatively high in the south of the province and was relatively low in the north. Spatially, the lower the latitude, the higher the predicted SOM content. The predicted SOM content in the topsoil varied from -0.89 to 58.86 g/kg and was averaged at 20.52 g/kg.

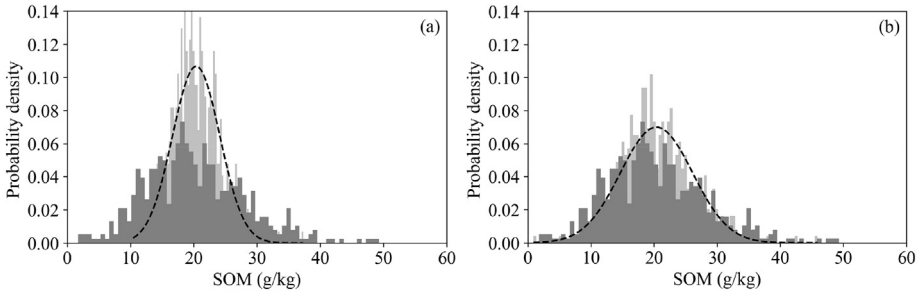


Fig. 4. Histograms of the original soil sample SOM and the predicted SOM based on the RF (a) and GBRT (b) models. Light gray bars indicate predicted SOM values, and dark gray bars indicate original SOM values from 562 soil validation samples over the Hubei province. Superimposed (black dotted curve) is a Gaussian probability distribution. The closer the black dotted curve is to the distribution of the original soil sample SOM, the better the prediction of the model. The mean and standard deviation (SD) of the original SOM are 20.45 and 8.35 g/kg, respectively. The means of the predicted SOM of RF and GBRT are 20.39 and 20.38 g/kg, respectively, and their standard deviations are 3.74 and 5.70 g/kg, respectively.

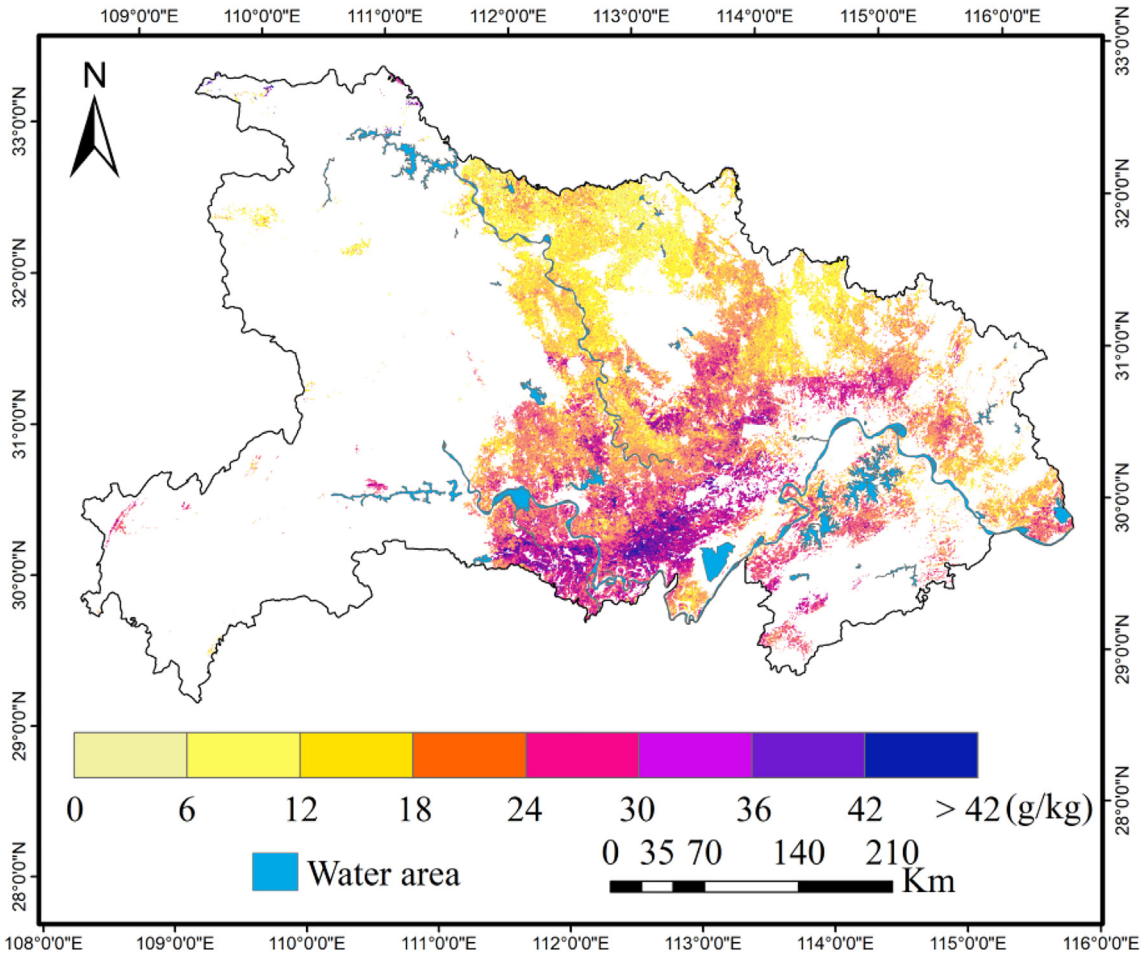


Fig. 5. Magnitude and spatial distribution of the predicted cropland SOM based on GBRT in Hubei province in 2017.

We compared the distribution of the predicted SOM between dry land and rice paddy fields in Hubei province in 2017 (Fig. 6). Dry land and rice paddy fields were identified using the cropland classification map obtained from the National land use/cover database of China in 2015 (http://www.resdc.cn/). Although the mean SOM of rice paddy fields (20.99 g/kg) was slightly higher than that of dry lands (20.20 g/kg), there was no significant difference in the predicted SOM content between these two cropland types ($p > 0.05$).

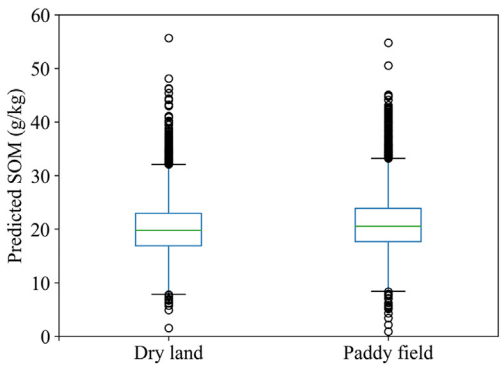


Fig. 6. The box plots of the predicted SOM in 2017 for the two dominant cropland types: dry land and rice paddy fields.

4.3. Spatio-temporal changes in predicted SOM

Using the SOM maps predicted by the trained GBRT model, we examined the trend of SOM between 2000 and 2017 for each cropland pixel in Hubei province. Our results showed that the SOM content slightly increased in most cropland areas during the past 18 years. The slope of the SOM trend in the newly added cropland ranged from −1.51 to 1.25 g/kg/yr and had a mean value of 0.074 g/kg/yr. The slope of the SOM change in the lost cropland ranged from −1.61 to 1.19 g/kg/yr and was averaged at 0.037 g/kg/yr. The slope of the SOM change in the unchanged cropland ranged from −1.55 to 1.66 g/kg/yr and had a mean value of 0.064 g/kg/yr. The mean increase rate of the predicted SOM content in the newly added cropland was higher than that in the lost cropland and unchanged cropland. A large portion of the new croplands in northern Hubei exhibited decreasing trends in the predicted SOM content (Fig. 7a). In the eastern part of the province, a large fraction of the cropland was converted to other land uses and the predicted SOM content had an increasing trend (Fig. 7b). In general, the predicted SOM content of cropland increased in southern Hubei and decreased in the central and northern parts of the province (Fig. 7c).

We calculated the spatially-averaged predicted SOM for Hubei's cropland for each year, and then examined the variability of predicted SOM during the period from 2000 to 2017 (Fig. 8). The mean predicted cropland SOM of the province showed relatively large inter-annual variability. The mean predicted SOM increased from 20.26 g/kg to 20.52 g/kg from 2000 to 2017, with a relative increase of 1.28%. Overall,

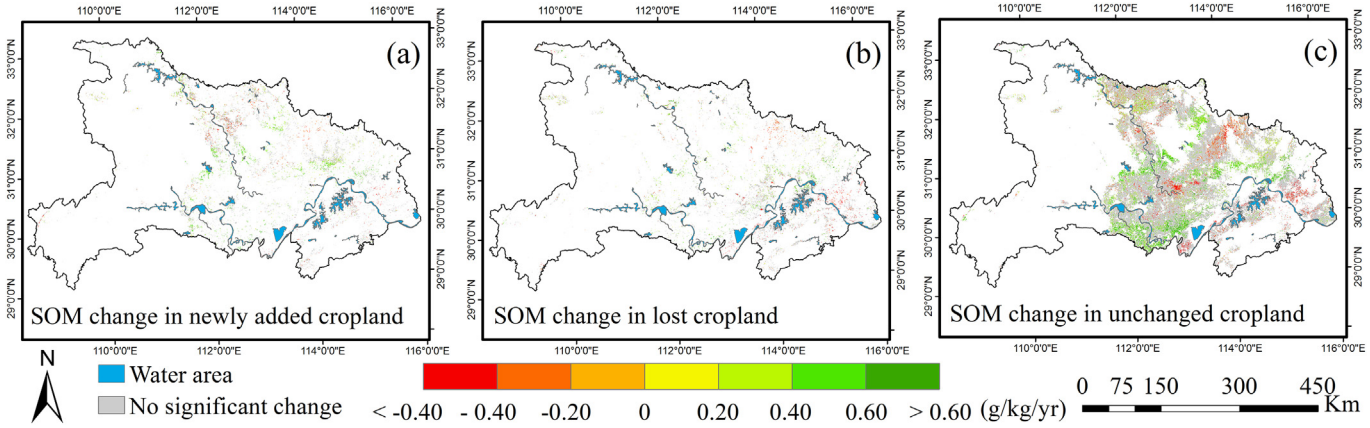


Fig. 7. Linear trend of predicted SOM in the newly added cropland (a), lost cropland (b), and unchanged cropland (c) on a per-pixel basis between 2000 and 2017.

the mean predicted SOM exhibited a statistically significant increasing trend ($p < 0.05$) during the period 2000–2017.

5. Discussion

5.1. Relative importance of explanatory variables

Previous studies showed that the important explanatory variables were different for different machine learning approaches and/or different study areas (Kumar and Lal, 2011; Yang et al., 2016). Our results showed that TWI was the first important factor for all four models, while the order of factor importance was different starting with the second important factor. TWI accounted for 6.27%–10.30% of the 25 explanatory variables for DT, RF and GBRT (Fig. 2). Previous studies revealed that soil with high moisture content could promote vegetation growth and slow down organic matter decomposition, thus increasing SOM content (Starr et al., 2000). Soil moisture could be quantitatively modeled by terrain factors (Mueller and Pierce, 2003), and therefore, there could be significant relationships between terrain factors and SOM content (Luca et al., 2007). TWI is a terrain factor that can quantitatively describe the balance between water accumulation and drainage conditions (Beven and Kirkby, 1979). TWI can capture soil moisture distribution at different landscape positions where overland flow dominates water transport processes and is correlated with SOM (Zhang et al., 2012). Hubei is dominated by a subtropical monsoon climate with abundant light energy, heat and rainfall. The temperature and precipitation were interpolated from daily observations of meteorological stations. MAT and MAP were the mean values of climate factors over the 20 years, and the change was relatively stable. The province is characterized by a high proportion of mountains and hills and relatively large variations in the topography. TWI was derived from the 90-m DEM, and is more sensitive to soil moisture distribution at different landscape positions. Therefore, TWI is of higher importance than MAP and MAT.

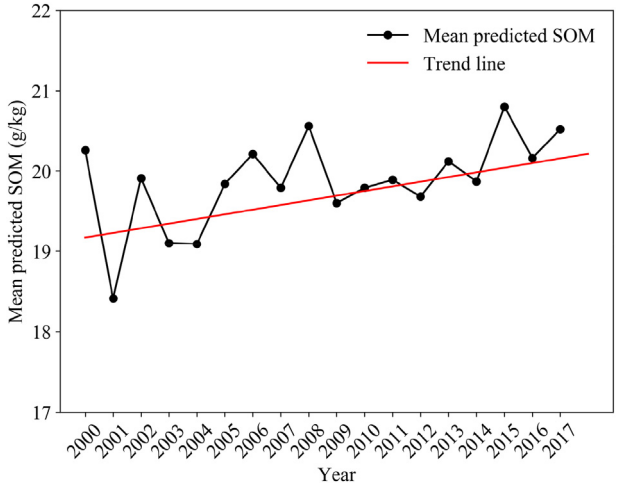


Fig. 8. Trend of the mean predicted SOM across Hubei's cropland over the period of 2000–2017.

Besides meteorological and terrain conditions, remote sensing reflectance (red band, near infrared band and blue band) and vegetation indices (RVI_max, DVI_mean, and EVI_mean) are also key factors for predicting topsoil SOM in these predictive models in Hubei province. Previous research demonstrated that remote sensing reflectance and vegetation indices can reflect vegetation productivity and biomass (Xiao et al., 2015). Vegetation, the main source of SOM, is highly correlated with the SOM content in topsoil (Bui et al., 2009). Remote sensing data have been widely used to predict SOM (Yang et al., 2016). The 30 m Landsat TM/ETM+ images are perhaps the most frequently used satellite data for SOM mapping, and both reflectance (e.g., red, near infrared) and vegetation indices have been used in previous studies (Taghizadeh-Mehrjardi et al., 2014; Were et al., 2015; Wang et al., 2017). In general, medium and high resolution remote sensing images are suitable for studies with limited soil sampling areas. MODIS data may be a better choice to predict soil properties (e.g., SOM, pH) at large spatial scales (Vågen et al., 2016). Unlike most previous studies, we used the 8-day MODIS data to calculate the mean value of surface reflectance and the mean and maximum values of various vegetation indices in this study. The EVI is strongly correlated with vegetation productivity (Lourenço et al., 2018). RVI is a sensitive indicator of vegetation biomass, and can also reflect soil degradation conditions (Dunagan et al., 2007). DVI can reflect the land surface moisture content (Dupigny-Giroux, 2007). Our results showed that the strength of the correlation between the mean and maximum values of vegetation indices and the SOM content varied with vegetation index. EVI_mean, RVI_max and DVI_mean were more suitable for predicting SOM than NDVI.

5.2. Applicability of machine learning algorithms

We found that the four tree-based machine learning algorithms: RF, BDT, DT and GBRT were suitable for SOM prediction in Hubei. These approaches have been used in previous studies (Carslaw and Taylor, 2009; Yang et al., 2016). BDT can avoid the problem of high variance of DT, while RF is an extension of BDT (Were et al., 2015). RF makes all the trees decorrelated by random perturbation and is less sensitive to over-fitting than DT and BDT (Heung et al., 2016). In addition, RF can incorporate a large number of predictors because it can reduce bias and

local feature predictors are important decision makers in tree structures, resulting in a significant increase in the accuracy of the prediction. Therefore, the performance of RF was better than that of DT and BDT in Hubei. Previous research showed that GBRT had a high predictive accuracy because it relies on stochastic gradient boosting, which allows for more accurate and faster computations through numerical optimization and regularization (Müller et al., 2013). Yang et al. (2016) compared the RF and GBRT for mapping SOC, and demonstrated that both of them were effective and powerful modeling approaches and accurately predicted the topsoil SOC in the northeastern Tibetan Plateau. However, previous research indicated that the different spatial extent of the study areas, sampling densities, or quantity and quality of the environmental factors could result in different model selection and accuracy (Were et al., 2015). Were et al. (2015) demonstrated that SVR was the best algorithm for predicting and mapping the SOC stocks in the Eastern Mau Forest Reserve of Eastern Africa, while Vågen and Winowiecki (2013) reported that RF had a slightly higher R^2 values in this region of Eastern Africa. Our results showed that among these tree-based models, RF and GBRT had the best performance with similar R^2 values, but the SOM spatial map based on GBRT better captured the SOM distribution of the original soil sample data than that based on RF.

5.3. Spatial and temporal changes of the predicted SOM

The predicted SOM content of cropland in northern Hubei was relatively low. From 2000 to 2017, the predicted SOM content of cropland showed a downward trend in northern Hubei, but a large portion of the cropland in this region was newly added. In the eastern parts of Hubei, the predicted SOM content of cropland was relatively high and also showed an upward trend, while a significant portion of the cropland was gradually lost during the 18 years. It has been confirmed that with the acceleration of urbanization, some high-quality croplands with convenient transportation, concentrated fields and perfect water supply and drainage system are now occupied by the construction land (~0.45 × 10^4 ha). However, the newly reclaimed cropland is almost the low-quality cropland with inconvenient transportation and fragmented fields and is susceptible to drought and flood (Liu et al., 2016).

A few previous studies examined the magnitude and distribution of SOM in Hubei province (Liu et al., 2016; Wang et al., 2012). Liu et al. (2016) collected soil fertility sample survey data from 26 counties (cities, districts) in Hubei province in 2015. The mean SOM content of cropland in the 26 counties (cities, districts) was 21.46 g/kg, and the range of the SOM content was from 10 to 30 g/kg (Liu et al., 2016). In our study, the mean SOM content based on our predicted SOM map in 2015 (20.80 g/kg) was almost identical with the estimate of Liu et al. (2016), and the land area with SOM content in the range of 10–30 g/kg accounted for 95% of the cropland area. Wang et al. (2012) conducted a sampling study of soil nutrient in rice paddy fields in Hubei province in 2008, and showed that the SOM content of rice soil in the province was averaged as 26.1 g/kg and ranged from 10 to 40 g/kg. The SOM content was relatively high in the south and east and relatively low in the north and west (Wang et al., 2012). In our research, the distribution of the predicted SOM in rice paddy fields in Hubei province also showed similar spatial patterns (i.e., high values in the south and east and low values in the north and west) (Fig. 9). Our estimate of the mean predicted SOM content value in Hubei's rice paddy fields (20.79 g/kg) was lower than that of Wang et al. (2012), and the area of the predicted SOM content in the range of 10–40 g/kg accounted for 99% of the rice paddy fields based on our predicted SOM map. The distribution of the predicted SOM content of this study was close to that of the soil samples.

Almost all previous studies assessing SOM changes in Hubei compared sample data in a given year with the second national soil census

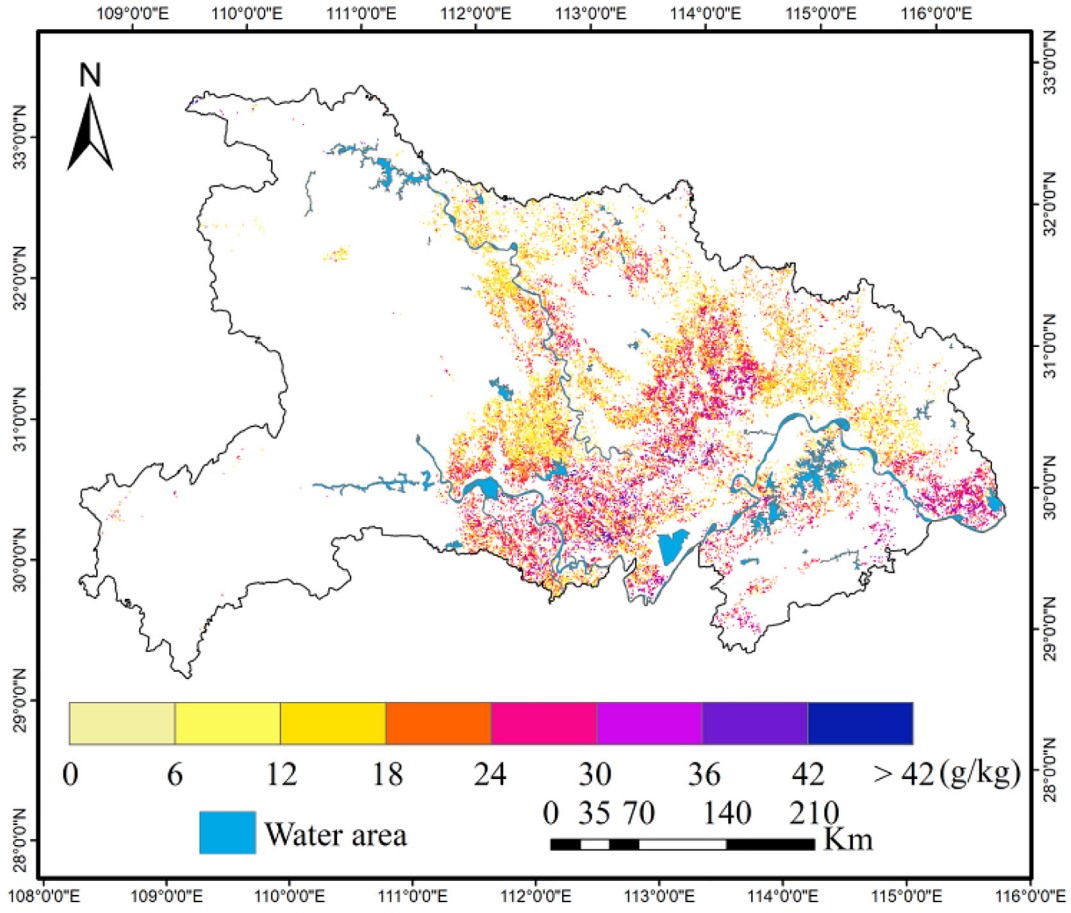


Fig. 9. The spatial distribution of the predicted SOM in rice paddy fields in Hubei province in 2008.

data in the 1980s (Liu et al., 2016; Yang et al., 2017). Despite the different distribution of soil samples and prediction methods, we compared our predicted SOM maps with the results of these previous studies quantitatively. According to the evaluation of cropland productivity and fertility in 26 counties in Hubei province by Liu et al. (2016), the SOM content of cropland in Hubei decreased by 0.93% from the 1980s to 2015. Yang et al. (2017) compared the data of soil formula fertilization project from 2005 to 2014 against the data of the second national soil census, and found that the mean SOM content in Hubei province had a slight increase of 0. 71 g/kg and an average growth rate of 0.11% per year. Our research showed that the predicted SOM content of cropland in Hubei province exhibited a slight decreasing trend, with a decrease of 0. 86 g/kg and an average reduction rate of 0.13% per year from the 1980s to 2015. However, the average growth rate of the predicted SOM content from 2000 to 2017 was 0.07% per year from 2000 to 2017. This suggests that, from 1980s to 2017, predicted SOM content showed a tendency to decrease first and then increase. Minasny et al. (2011) also showed that soil carbon showed different trends in different periods. However, it should be noted that this slight increase could offset global anthropogenic greenhouse gas emissions to some extent, and there is still a gap to achieve '4 per mille Soils for Food Security and Climate' (Minasny et al., 2017). Previous research has shown that the SOM of cropland topsoil in China has increased (Yang et al., 2017; Minasny et al., 2017). However, the SOM in Hubei was relatively stable. Therefore, future management practices such as straw returning, organic matter input and rotation system (Meng et al., 2005) are essential for improving SOM in croplands in Hubei province.

5.4. Limitations and future research

There are several limitations with our study, and future research is needed for evaluating other machine learning methods and reducing the uncertainty in predicted SOM at regional and decadal scales. First, we chose only four algorithms to predict SOM, and they are all based on tree models. Some other machine learning algorithms (e.g., neural networks (NN)) have also been applied in digital soil mapping (Li et al., 2013; Taghizadeh-Mehrjardi et al., 2016). These studies indicated that NN methods require more soil samples or explanatory variables to achieve a good training performance, and tree models are more suitable for studies with a relatively small number of soil samples and explanatory variables. Nevertheless, future research should compare all these approaches along with emerging methods (e.g., deep learning) for varying sizes of soil samples and explanatory variables. Second, we did not explicitly take agricultural practices into consideration in our prediction. Agricultural practices (e.g., C input, crop rotation, irrigation and fertilization) play an important role in determining the magnitude, pattern and trend of SOM, while the explanatory variables that we used cannot explicitly account for these practices. The incorporation of spatially explicit information on management practices is anticipated to improve the accuracy of the predicted SOM. Third, we only validated the predicted SOM in 2017 and were not able to evaluate the inter-annual variability and trends of the predicted SOM using field-based SOM data due to the lack of a soil sample time series. Changes in management practices and environmental factors could cause significant changes in SOM which were only evaluated by comparing our results with those from previous studies. Having field-based SOM data for multiple years will allow us to quantitatively evaluate the inter-annual variability and trends of predicted SOM.

6. Conclusions

We evaluated the performance of four different machine learning methods: RF, BDT, DT and GBRT for predicting cropland SOM in Hubei province, China, and also mapped SOM for each 500 m × 500 m cropland pixel for each year over the period 2000–2017. Compared with the other three algorithms, GBRT had a better performance for mapping SOM content. In addition to climate and terrain factors, surface reflectance (red band, near infrared band and blue band) and vegetation indices (RVI_max, DVI_mean, EVI_mean) derived from MODIS were also key factors for the prediction of SOM content. Our results showed that the predicted SOM content of cropland was relatively high in the south of Hubei and relatively low in the north. Overall, the mean predicted SOM content of cropland in Hubei province exhibited a slight increasing trend from 2000 to 2017. Spatially, the predicted SOM content of cropland increased in southern Hubei and decreased in central and northern parts of the province. A significant portion of the low-quality cropland in northern Hubei was recently reclaimed, while a part of the high-quality cropland in the east was lost due to land use change (e.g., urbanization).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (41871356). J. Xiao was supported by the National Aeronautics and Space Administration (NASA) (the Carbon Cycle Science Program: Grant No. NNX14AJ18G). We thank the anonymous reviewers for their constructive and insightful comments on our manuscript.

References

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. https://doi.org/10.1016/j.chemolab.2011.12.002.
Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. J. 24, 43–69. https://doi.org/10.1080/02626667909491834.
Breiman, L., 1996. Out-of-bag estimation, ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Bui, E., Henderson, B., Viergever K., 2009. Using knowledge discovery with data mining from the Australian soil resource information system database to inform soil carbon mapping in Australia. Glob. Biogeochem. Cycles 23, GB4033. doi: https://doi.org/10.1029/2009GB003506.
Camera, C., Zomeni, Z., Noller, J.S., Zissimos, A.M., Christoforou, I.C., Bruggeman, A., 2017. A high resolution map of soil types and physical properties for Cyprus: a digital soil mapping optimization. Geoderma 285, 35–49.
Carslaw, D.C., Taylor, P.J., 2009. Analysis of air pollution data at a mixed source location using boosted regression trees. Atmos. Environ. 43 (22−23), 3563–3570.
Cooperative Research Group on Chinese Soil Taxonomy, 2001. Keys to Chinese Soil Taxonomy. Press of University of Science and Technology of China. Hefei, China (in Chinese).
Doetterl, S., Stevens, A., van Oost, K., Quine, T.A., van Wesemael, B., 2013. Spatially-explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. Geoderma. 204–205, 31–42.
Dunagan, S.C., Gilmore, M.S., Varekamp, J.C., 2007. Effects of mercury on visible/near - infrared reflectance spectra of mustard spinach plants. Environ. Pollut. 148, 301–311.
Dupigny-Giroux, L.L., 2007. Using AirMISR data to explore moisture- driven land use- land cover variations at the Howland Forest, Maine: a case study. Remote Sens. Environ. 107, 376–384.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.
Friedl, M., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X., 2010. MODIS collection 5 global land cover: algorithm refinements and characterization of new datasets. Remote Sens. Environ. 114, 168–182.
Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.
Gao, B., 1996. NDWI-A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266.
Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island-digital soil mapping using random forests analysis. Geoderma 146, 102–113.
Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma. 265, 62–77.
Huang, Y., Sun, W., 2006. Changes in topsoil organic carbon of croplands in mainland China over the last two decades. Chin. Sci. Bull. 51, 1785–1803.
Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25, 53–70.
Huete, A.R., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213.
Jordan, C.F., 1969. Derivation of leaf-area index from quality of light on the forest floor. Ecology 50, 663–666.
Kumar, S., Lal, R., 2011. Mapping the organic carbon stocks of surface soils using local spatial interpolator. J. Environ. Monit. 13, 3128–3135.
Li, C., Zhuang, Y., Frolking, S., Galloway, J., Harriss, R., Moore, B., Schimel, D., Wang, X., 2003. Modeling soil organic carbon change in croplands of China. Ecol. Appl. 13, 327–336.

Li, Q., Yue, T., Wang, C., Zhang, W., Yu, Y., Li, B., Yang, J., Bai, G., 2013. Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach. Catena. 104, 210–218.
Liu, Y., Zhang, Y., Guo, L., 2010. Towards realistic assessment of cultivated land quality in an ecologically fragile environment: a satellite imagery-based approach. Appl. Geogr. 30, 271–281.
Liu, F., Liang, H., Liu, T., Zhang, S., He, X., He, L., Xu, N., 2016. Soil fertility changes of farmland in Hubei Province for the last three decades. J. Huazhong Agric. Univ. 35, 79–85 (in Chinese).
Louppe, G., 2014. Understanding Random Forests from Theory to Practice. University of Liège.
Lourenço, P., Alcaraz-Segura, D., Reyes-Díez, A., Requena-Mullor, J.M., Cabello, J., 2018. Trends in vegetation greenness dynamics in protected areas across borders: what are the environmental controls? Int. J. Remote Sens. 39, 4699–4713.
Luca, C., Si, B.C., Farrell, R.E., 2007. Upslope length improves spatial estimation of soil organic carbon content. Can. J. Soil Sci. 87, 291–300.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
McLauchlan, K., 2006. The nature and longevity of agricultural impacts on soil carbon and nutrients: a review. Ecosystems 9, 1364–1382.
Meersmans, J., De Ridder, F., Canters, F., De Baets, S., Van Molle, M., 2008. A multiple regression approach to assess the spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 143, 1–13.
Meng, L., Ding, W., Cai, Z., 2005. Long-term application of organic manure and nitrogen fertilizer on N_2O emissions, soil quality and crop production in a sandy loam soil. Soil Biol. Biochem. 37, 2037–2045.
Minasny, B., Sulaeman, Y., Mcbratney, A.B., 2011. Is soil carbon disappearing? The dynamics of soil organic carbon in Java. Glob. Chang. Biol. 17, 1917–1924.
Minasny, B., Sulaeman, Y., Mcbratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C., Vågen, T., Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86.
Mueller, T.G., Pierce, F.J., 2003. Soil carbon maps: enhancing spatial estimates with simple terrain attributes at multiple scales. Soil Sci. Soc. Am. J. 67, 258–267.
Mulder, V.L., de Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping -a review. Geoderma 162, 1–19.
Müller, D., Leitão, P.J., Sikor, T., 2013. Comparing the determinants of cropland abandonment in Albania and Romania using boosted regression trees. Agric. Syst. 117, 66–77.
Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67, 215–226.
Popeplau, C., Don, A., Vesterdal, L., Leifeld, J., Van Wesemael, B., Schumacher, J., Gensior, A., 2011. Temporal dynamics of soil organic carbon after land-use change in the temperate zone - carbon response functions as a model approach. Glob. Chang. Biol. 17, 2415–2427. https://doi.org/10.1111/j.1365-2486.2011.02408.x.
Richardson, A.J., Weigand, C., 1977. Distinguishing vegetation from soil background information. Photogramm. Eng. Remote. Sens. 43, 1541–1552.
Rouse Jr., J.W., Haas, R., Schell, J., Deering, D., 1974. Monitoring vegetation systems in the great plains with ERTS. NASA Special Publication. 1, 309–317.
Somarathna, P.D.S.N., Malone, B.P., Minasny, B., 2016. Mapping soil organic carbon content over New South Wales, Australia using local regression kriging. Geoderma. Regional 7, 38–48.
Song, X., Brus, D.J., Liu, F., Li, D., Zhao, Y., Yang, J., Zhang, G., 2016. Mapping soil organic carbon content by geographically weighted regression: a case study in the Heihe River Basin, China. Geoderma 261, 11–22.
Starr, G.C., Lal, R., Malone, R., Hothem, D., Owens, L., Kimble, J., 2000. Modeling soil carbon transported by water erosion processes. Land Degrad. Dev. 11, 83–91.
Stockmann, U., Padarian, J., McBratney, A., Minasny, B., Brogniez, D., Montanarella, L., Hong, S.Y., Rawlins, B.G., Field, D.J., 2015. Global soil organic carbon assessment. Glob. Food Sec. 6, 9–16.
Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., Malone, B.P., 2014. Digital mapping of soil salinity in Ardakan region, central Iran. Geoderma 213, 15–28.
Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma. 266, 98–110.
Tien Bui, D., Pradhan, B., Lofman, O., Revhaug, I., 2012. Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naïve Bayes models. Math. Probl. Eng. 2012, 1–26.
Vågen, T., Winowiecki, L.A., 2013. Mapping of soil organic carbon stocks for spatially explicit assessments of climate change mitigation potential. Environ. Res. Lett. 8, 15011.
Vågen, T., Winowiecki, L.A., Tondoh, J.E., Desta, L.T., Gumbricht, T., 2016. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. Geoderma 263, 216–225.
Wang, L., Qiu, J., Tang, H., Li, H., Li, C., Van Ranst, E., 2008. Modelling soil organic carbon dynamics in the major agricultural regions of China. Geoderma 147, 47–55.
Wang, W., Lu, J., Lu, M., Dai, Z., Li, X., 2012. Status quo and variation of soil fertility in paddy field - a case study of Hubei province. Acta Pedol. Sin. 2, 319–330 (in Chinese).
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.
Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52, 394–403.
Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7–24.
Xiao, J., Zhou, Y., Zhang, L., 2015. Contributions of natural and human factors to increases in vegetation productivity in China. Ecosphere 6 (233). https://doi.org/10.1890/ES14-00394.1.
Yang, R., Zhang, G., Liu, F., Lu, Y., Yang, F., Yang, F., Yang, M., Zhao, Y., Li, D., 2016b. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecol. Indic. 60, 870–878.
Yang, F., Xu, Y., Cui, Y., Meng, Y., Dong, Y., Li, R., Ma, Y., 2017. Variation of soil organic matter content in croplands of China over the last three decades. Acta Pedol. Sin. 54, 1047–1056 (in Chinese).
Yeomans, J.C., Bremner, J.M., 1988. A rapid and precise method for routine determination of organic carbon in soil. Commun. Soil Sci. Plant Anal. 19, 1467–1476. https://doi.org/10.1080/00103628809368027.
Zhang, F., Li, C., Wang, Z., Wu, H., 2006. Modeling impacts of management alternatives on soil carbon storage of farmland in Northwest China. Biogeosciences 3, 451–466.
Zhang, S., Huang, Y., Shen, C., Ye, H., Du, Y., 2012. Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information. Geoderma. 171–172, 35–43.
Zhang, F., Wang, Z., Glidden, S., Wu, Y., Tang, L., Liu, Q., Li, C., Frolking, S., 2017. Changes in the soil organic carbon balance on China's cropland during the last two decades of the 20th century. Sci. Rep. 7, 7144. https://doi.org/10.1038/s41598-017-07237-1.