Spatial Prediction of Forest Soil Organic Carbon Based on Random Forest Algorithm and Remote Sensing

Xiong Yao (1), Jian Liu (1)

1 College of Forestry，Fujian Agriculture and Forestry University, Fuzhou 350002, China
Email: fjyaoxiong@163.com; fjliujian@fafu.edu.cn;

KEY WORDS: Soil organic carbon; Random forest model; Ordinary kriging; Spatial prediction

ABSTRACT: Forest soil organic carbon (SOC) as an important part of soil carbon pool, is not only closely related to soil quality and soil fertility, but also great influence on global climate change with the working capital cycle of carbon in the atmosphere. Accurate prediction of forest SOC is of great significance for carbon cycle research and precision forestry. In this study, a total of 59 topsoil (0-20cm) samples were collected from Hetian Town located at Fujian Province, China. Ten environmental variables, elevation (ELE), slope (S), normalized difference vegetation index (NDVI), green normalized difference vegetation index (GNDVI), soil adjusted vegetation index (SAVI), modified soil adjusted vegetation index (MSAVI), structure insensitive pigment index (SIPI), triangle vegetation index (TVI), mean annual precipitation (MAP) and mean annual temperature (MAT), were collected using digital elevation model and Landsat-8 OLI remote sensing data. We used out of bag error of random forest (RF) model to select model variables. RF and it integrate with ordinary kriging method (RF-OK) were applied to predict SOC. The coefficient of determination (R^2), root mean square error (RMSE) and average relative error (MAE) were used to evaluate and compare model prediction accuracy. The results showed that SOC ranged from 2.9 to 47.5 g kg^{-1} , with an average of 19.7 g kg^{-1}. as participated factors of the RF model and RF-OK model. In addition to SAVI, SIPI and MAT, other environmental factors, ELE, S, NDVI, GNDVI, MSAVI, TVI and MAP, could be used as predictor factors of the RF model and RF-OK model. Although the SOC spatial distribution patterns generated by using RF model and RF-OK model were almost the same, the R^2 of RF-OK model was 0.58, increasing by 34.4%, and there were different degrees of decrease in RMSE and MAE of the RF-OK model. This study demonstrated that the RF-OK model was superior to the RF model in mapping the spatial patterns of forest soil organic carbon, and it could be used to predict forest SOC.

1 Introduction

Soil carbon pool is the largest carbon pool in the terrestrial ecosystem, which is twice larger than the atmospheric carbon pool and three times larger than the terrestrial vegetation carbon pool (Yao et al., 2019). Even a small change in soil carbon pool could cause a large fluctuation in atmospheric CO_2 concentration, and subsequently influence both global carbon cycle and climate. As the main component of soil carbon pool, forest soil organic carbon (SOC) is not only closely related to soil quality and soil fertility, but also has a significant impact on global climate change (Yang et al., 2016). Therefore, a better predicting performance of SOC and its spatial distribution are of importance in estimating the status of carbon emissions in the terrestrial ecosystems.

It is difficult to obtain the spatial distribution characteristics and continuity of SOC using traditional sampling and mapping methods due to the high spatial and temporal variability of soil properties in the regional scale. In recent years, digital soil mapping (DSM) (Grunwald et al., 2011), using the mathematical models to simulate the statistical relationship between georeferenced soil information (i.e., dependent variables) and environmental covariates (i.e., independent variables) that represent soil-forming factors (e.g., vegetation, topographic factors), has become an effective method to accurately obtain the spatial distribution of SOC.

Although numerous models have been applied to map the spatial distribution of SOC, most of these models are based on traditional statistical or geostatistical methods, including multiple linear regression (Olaya-Abril et al., 2017), partial least squares regression (Nocita et al., 2014), and geographically weighted regression (Kumar et al., 2012; Song et al., 2016), ordinary kriging (OK) (Mishra et al., 2009). With the deep development of artificial intelligence technology, more and more scholars use machine learning techniques (e.g., artificial neural network, support vector machine) to explore the potential relationship between soil properties and environmental factors. Random forest (RF) method, in particular, has became one of the effective methods to predict the spatial distribution of SOC, especially in DSM (Castro-Franco et al., 2015; Guo et al., 2015). Nevertheless, RF ignores the spatial autocorrelation of variables, and thus affecting the prediction accuracy of SOC. In order to overcome the disadvantages, a hybrid model of random forest and ordinary kriging (RF-OK) is proposed to predict the SOC based on Landsat-8 OLI data and other environmental covariates, as well as spatial autocorrelations. The objectives of this study were to digitally map the spatial distribution of SOC using RF-OK and compare performances of RF-OK with RF.

2 Material and methods

2.1 Study area

The study area is located in the Hetian Town, Changting County, Fujian Province, China (25°33′-25°48′N,116°18′-116°31′E), and the total land area reaches approximately 296 km^2 (Fig.1). This town belongs to a typical subtropical monsoon climate, with a mean annual temperature of 18.3℃. There are about 265 frost-free days with a mean annual precipitation of 1700mm, which ranging from 1074 mm to 2522 mm. Based on soil parent material, the area is primarily covered by acid red soil, with smaller areas of purple soil and yellow soil. The dominant land use of the town is woodland which mainly consist of *Pinus massoniana* forests.

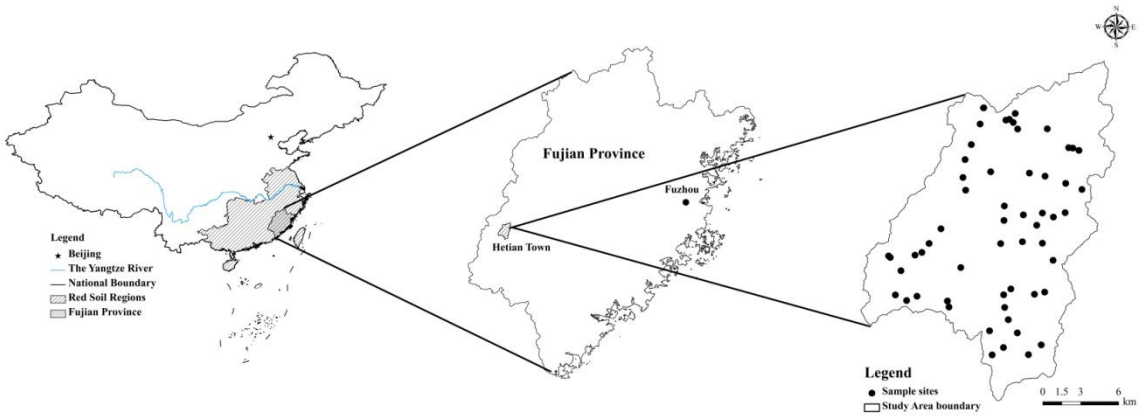

Figure 1 Geographic locations of the study area.

2.2 Soil sampling and analysis

In January 2015, 59 representative sample plots (32 Masson pine plantations, 17 Chinese fir plantations and 10 broad-leaved natural forestland) with a size of 666.7 m^2 were established in Hetian Town. GPS was used to determine the coordinate positions of the plots. For each plot, a typical profile was established and soil samples were collected from topsoil (0-20 cm). Soil samples from each plot, approximately 1 kg were shipped to the laboratory for the further SOC analysis. In the laboratory, rocks and plant residues were removed from each soil sample, and then all soil samples were air-dried and passed through 0.15mm sieves prior to measuring the SOC.

After that, SOC were measured using the $K_2Cr_2O_7$–H_2SO_4 wet oxidation approach (Tang et al., 2017).

2.3 Environmental variables

The formation and change of SOC are not isolated, but closely related to soil parent material, topography, vegetation, climate and other factors. The topographic factors were mainly obtained from digital elevation model (DEM) with a spatial resolution of 30m. Vegetation index is an important parameter to characterize SOC content. In this study, six vegetation index indicators closely related to SOC were extracted from Landsat-8 OLI remote sensing data. Climate variables are important factors causing spatial variability of SOC content in the region. The mean annual precipitation and mean annual temperature were generated using inverse distance weighted interpolation with a spatial resolution of 30 m. The basic climate dataset were provided by the Changting County Meteorological Bureau. Soil parent material was not used as the environmental variable due to the soil parent material is mainly granite in our study. More detailed information of the environmental variables are shown in Table 1.

Table1 Description and source of environmental variables.

Variable types	Name of variables	Description and calculation	Source
Topographic variables	Elevation (ELE)	Extracted from DEM	DEM
	Slope (S)	Using spatial analysis module in ArcGIS	DEM
Vegetation variables	Normalized difference vegetation index (NDVI)	$NDVI = (B_4 - B_3)/(B_4 + B_3)$	OLI
	Green normalized difference vegetation index (GNDVI)	$GNDVI = (B_4 - B_2)/(B_4 + B_2)$	OLI
	Soil adjusted vegetation index（SAVI）	$SAVI = ((B_4 - B_3) \times 1.5)/(B_4 - B_3 + 0.5)$	OLI
	Modified soil adjusted vegetation index（MSAVI）	$MSAVI = \left[(2B_4 + 1) - \sqrt{(2B_4 + 1)^2 - 8(B_4 - B_3)}\right]/2$	OLI
	Structure insensitive pigment index (SIPI)	$SIPI = (B_4 - B_1)/(B_4 + B_1)$	OLI
	Triangle vegetation index (TVI)	$TVI = 0.5[120(B_4 - B_2) - 200(B_3 - B_2)]$	OLI
Climate variables	Mean annual precipitation（MAP）	Inverse distance weighted interpolation	Changting County Meteorological Bureau
	Mean annual temperature（MAT）	Inverse distance weighted interpolation	Changting County Meteorological Bureau

B1, B2, B3, B4 are represent the reflectivity of the blue, green, red and near infrared bands respectively, which corresponding to the 2nd, 3rd, 4th and 5th bands of OLI image data.

2.4 Modeling techniques

The RF model, proposed by Breiman (2001), is a machine learning method based on multiple decision trees. The basic idea of RF is: 1) extracting samples from the original training set through bootstrap method, and the size of each sample is consistent with the size of the original training set; 2) constructing a decision tree to obtain a modeling result for each sample; 3) according to the modeling results of all decision trees, the final prediction result is obtained by voting. Two parameters, including number of trees to be grown (n_{tree}), number of predictor variables

used to split the nodes at each partitioning (m_{try}) were set to 1000 and 3 respectively.

The RF-OK is an extension of RF. We used RF model to obtain the predicted SOC value, then the residuals from RF were interpolated to prediction grids using the OK method, and finally the interpolated residuals were added to the RF prediction results as the RF-OK prediction results. The predicted SOC value of RF-OK can be calculated as follows:

$$\hat{C}_{RF}(x_i) = f(a_1, a_2, \cdots, a_n) \tag{1}$$

$$\hat{r}_{OK}(x_i) = \sum_{i=1}^{n} w_i r(x_i) \tag{2}$$

$$\hat{C}_{RF-OK}(x_i) = \hat{C}_{RF}(x_i) + \hat{r}_{OK}(x_i) \tag{3}$$

where $\hat{C}_{RF}(x_i)$ is the predicted SOC value at location x_i using RF; f is the mathematical relationship between SOC and environmental variables; $a_1, a_2, \cdots, a_n$ is environmental variables; $\hat{r}_{OK}(x_i)$ is estimated residual value at location x_i using OK; $\hat{C}_{RF-OK}(x_i)$ is the predicted SOC value at location x_i using RF-OK.

2.5 Model validation

The 59 samples were randomly divided into a training set (number of samples = 44) and testing set (number of samples = 15). Model evaluation indices, including determination coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE), were used to compare the accuracy of RF and RF-OK model. The formula of these evaluation indices were listed as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{6}$$

where n is the number of measured points, y_i and $\hat{y}_i$ are the predicted and measured values, respectively, at location i, and $\bar{y}$ is the mean value of the measured values.

3 Results

3.1 Descriptive statistics of SOC

The descriptive statistics of SOC is presented in Table 2. The SOC contents ranged from 2.9-47.5 g kg^{-1} and the mean SOC was 19.7 g kg^{-1}. The coefficient of variation of SOC was 52.28%, indicating a moderate variability. One-Sample Kolmogov-Smirnov Test manifested that all of the values of SOC were normality distributed.

Table 2 Descriptive statistics of SOC (g kg^{-1}).

Minimum	Maximum	Mean	Standard deviation	Coefficient of Variation /%	Distribution type
2.9	47.5	19.7	10.3	52.28	Normal distribution

3.2 Environmental variables selection

There are many environmental factors affecting SOC, and different environmental factors have different

degrees of importance to their impact. In order to reduce the impact of the out of bag error (OOB error) on the prediction accuracy of SOC, we removed the variables with less importance. It is mainly judged whether the factor remains by increasing or decreasing the OOB error after removing the factor one by one. If the OOB error increases, the factor is remained, and vice versa. As shown in Table 3, the OOB error decreased when SAVI, SIPI and MAT were removed. Therefore, seven environmental variables (i.e., ELE, S, NDVI, GNDVI, MSAVI, TVI and MAP) were selected as the variables of RF and RF-OK model.

Table 3 Screening of environmental variables.

Environmental variables	OOB errors	Variables remained or not	Environmental variables	OOB errors	Variables remained or not
ALL	0.8733				
ELE	0.8837	Y	MASVI	0.8764	Y
S	0.8776	Y	SIPI	0.8635	N
NDVI	0.8812	Y	TVI	0.8751	Y
GNDVI	0.8881	Y	MAP	0.8696	Y
SAVI	0.8653	N	MAT	0.8572	N

ALL is all environmental variables are involved in the RF model.

3.3 Spatial prediction of SOC

Spatial predictions of SOC were performed using RF and RF-OK, respectively. Prediction maps of SOC were shown in Fig. 2. The general spatial patterns of SOC were similar for RF and RF-OK. The high SOC contents mainly distributed in the north of the study area, but the low SOC contents mainly distributed in the central and southern region. The main reason for this phenomenon is that the forest structure in the north is rich, which is conducive to the accumulation of organic carbon in the surface soil, while the vegetation in the central and southern areas is sparse. Compared to RF, the ranges of SOC in the prediction maps by RF-OK were much closer to the observed values.

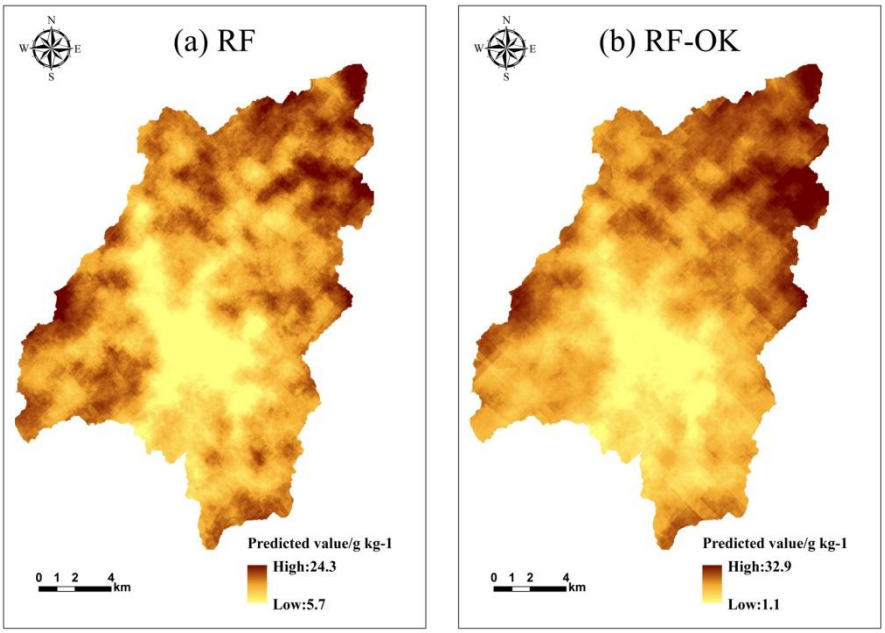

Figure 2 Prediction maps of forest SOC.

3.4 Validation

The independent validation dataset (n = 15) was used to test the model performances. And the performance indicators (R^2, RMSE, and MAE) were presented in Fig. 3. Compared to RF, RF-OK significantly improved the prediction by reducing the RMSE and MAE by 14.3% and 10.9%, respectively, while increasing the R^2 about 34.4%.

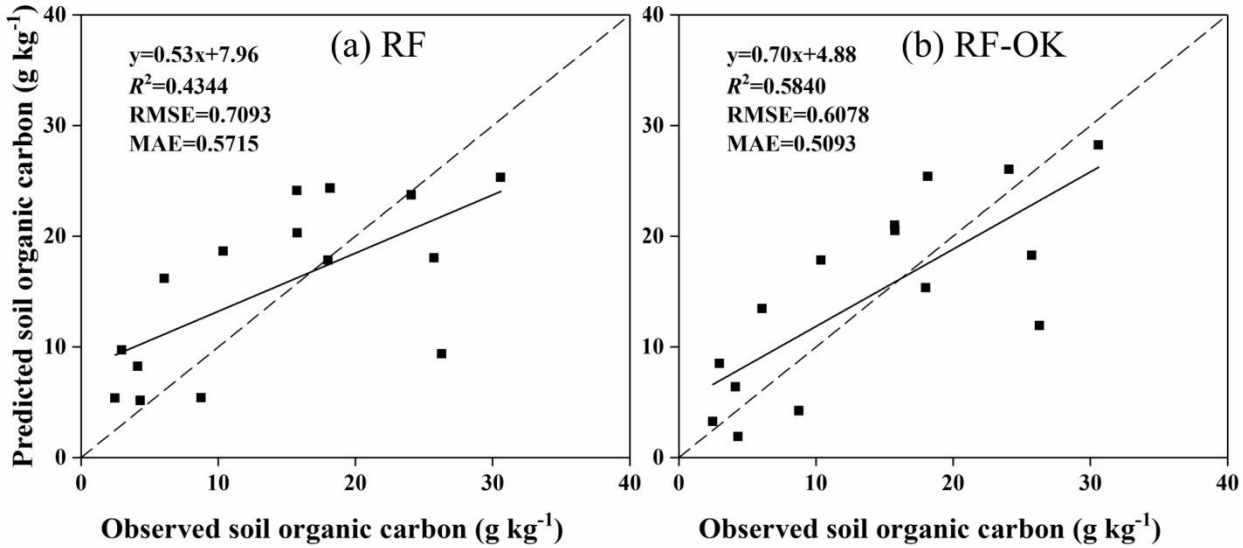

Figure 3 Performances of the RF and RF-OK methods in predicting spatial distribution of SOC.

4 Discussion

The formation and change of soil are not isolated, but closely related to the surrounding natural geographical environment. Due to the complexity of the soil-forming environment, soil properties are simultaneously affected by various environmental factors, such as soil parent material, climate, vegetation and topography. Therefore, the selection of environmental factors has an important impact on the prediction accuracy of soil properties. Our study selected 10 environmental variables. In general, temperature is an important factor affecting SOC content, but in this study, the mean annual temperature is of low importance to SOC content (Table 3), mainly due to the low variability of mean temperature between different regions of the study area.

Because soil properties are affected by many environmental factors, the relationship between SOC and environmental variables is very complex and difficult to describe using simple linear relationships. Therefore, it is necessary to use suitable nonlinear models to fit the relationship between SOC and environmental variables. In the present study, the RF model was selected to predict the spatial distribution of SOC. Meanwhile, an extension method of the RF, i.e., hybrid the RF and OK model (RF-OK), was constructed. The prediction accuracy showed that the RF-OK model is more effective in spatial prediction of SOC content than the RF model.

In this study, the environmental variables are selected by the OOB error in the RF model. The selected vegetation indices (i.e., NDVI, GNDVI, MSAVI, and TVI) still have collinearity. Nevertheless, the RF and RF-OK model is insensitive to multivariate collinearity. In addition, the RF-OK model in this study showed better prediction accuracy than the RF model. The reason is that the RF model only considers the relationship between SOC content and environmental variables, and ignoring the spatial autocorrelation of SOC content. The RF-OK model overcomes the shortcomings of the RF model due to the addition of OK model. Therefore, the prediction accuracy of the RF-OK model is higher than that of the RF model.

5 Conclusions

In the present study, the RF-OK approach produced quite good results for predicting and mapping spatial distribution of SOC for forestland at regional scale. Compared to RF, RF-OK performed much better in predicting and mapping spatial pattern of SOC because RF-OK well accounted for the spatial structure of model residuals, and it could be used to predict forest SOC.

References

Breiman, L., 2011. Random forest. Machine Learning, 45(1): 5-32.

Castro-Franco, M., Costa, J. L., Peralta, N., Aparicio, V., 2015. Prediction of soil properties at farm scale using a model-based soil sampling scheme and random forest. Soil Science, 180, 74-85.

Grunwald, S., Thompson, J. A., Boettinger, J. L., 2011. Digital soil mapping and modeling at continental scales: finding solutions for global issues. Soil Science Society of America Journal, 75(4), 1201-1213.

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

Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma, 189-190, 627-634.

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D. S., Meirvenne, M. V., 2009. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. Soil Science Society of America Journal, 73, 614-621.

Nocita, M., Stevens, A., Toth, G., Panagos, P., Wesemael, B., Montanarella. L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biology & Biochemistry, 68, 337-347.

Olaya-Abril, A., Parras-Alcántara, L., Lozano-García, B., Obregón-Romero, R. 2017. Soil organic carbon distribution in mediterranean areas under a climate change scenario via multiple linear regression analysis. Science of the Total Environment, 592, 134-143.

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. Geoderma, 261, 11-22.

Tang, X., Xia, M., Pérez-Cruzado, C., Guan, F., Fan, S., 2017. Spatial distribution of soil organic carbon stock in Moso bamboo forests in subtropical China. Scientific Reports. 7, 42640.

Yang, R., Zhang, G., Liu, F., Lu, Y., Yang, F., Yang, F., Yang, M., Zhao, Y., Li, D., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators, 60, 870-878.

Yao, X., Yu, K., Wang, G., Deng, Y., Lai, Z., Chen, Y., Jiang, Y., Liu, J., 2019. Effects of soil erosion and reforestation on soil respiration, organic carbon and nitrogen stocks in an eroded area of Southern China. Science of the Total Environment, 683, 98-108.