Theory and Practice of Science and Technology
2024, VOL. 5, NO. 4, 5-13
DOI: 10.47297/taposatWSP2633-456901.20240504

Spatial Prediction of Soil Organic Carbon in Cultivated Land Based on Multiple Variables: A Case Study of Guangdong Province

Shihao Zhang

The Chinese University of Hong Kong

1 Introduction

Traditionally, obtaining the spatial distribution pattern of soil properties through field sampling and analyses is a commonly used method, but the time-consuming and costly nature of this method and poor regional accessibility have limited its wide application at regional, national and even global scales [2]. With the rapid development of 3S technology (remote sensing, geographic information system and global positioning system), more and more scholars have realised the spatial mapping of regional soil attributes by establishing prediction models between auxiliary environmental variables and soil attributes [3].Currently, the prediction models used for digital soil mapping mainly include stepwise linear regression model [4], recursive partial least squares algorithm [5] and Random Forest (RF) algorithm [6]. Among them, RF algorithm is a machine learning algorithm consisting of Classification and Regression Tree, which is gradually widely used in the spatial prediction of soil attributes due to its significant advantage in handling multivariate nonlinear data [7].

In this paper, using Guangdong Province as the study area, a random forest model is used to select easily accessible remote sensing variables and climate factors as model inputs to train the RF model and conduct a comparative study on the spatial distribution pattern of organic carbon content. Through this study, it aims to provide a theoretical basis for the accurate estimation of organic carbon stocks in complex terrain areas, and to provide an effective method for the prediction of SOC stocks in regional topographically and geomorphologically complex areas.

2 Study area profiles and data

2.1 Study Area

Guangdong Province is located in the south of mainland China, bordering the South China Sea to the east, with geographical coordinates between 20°13′ and 25°31′N and 109°39′ and 117°19′E, and a total area of about 180,000 square kilometres. The topography of the region is complex, with high terrain in the north and low terrain in the south, with the highest elevation at 1,902 metres and the lowest at -52 metres, and the types of landforms include mountains, hills, plateaus and plains, of which mountains and hills account for more than half of the total area.

The climate type of Guangdong Province is a southern subtropical monsoon climate, with an average annual temperature of about 21.9°C, an average annual precipitation of 1,790 mm, and an average annual sunshine duration of 1, 755 hours. The climatic and hydrological conditions vary significantly within the study area due to the combined influence of topography and ocean. The difference in mean annual temperature between the northern mountainous area and the southern coastal plain is about 4.5°C, the difference in sunshine duration is 1108 hours, and the difference in annual precipitation is 2084 mm. The region has well-developed water systems and dense rivers, with a total surface water resource of 1.885 × 10^11 cubic metres in 2018, and 543 rivers with a rainfall catchment area of more than 100 km^2, with a total length of 28,600 km. According to the results of the Second Soil Census, the study area has a diversity of soil types, including 16 soil types such as rice soil, terracotta soil and tidal soil. The distribution of these soil types is jointly influenced by a variety of factors such as climate, topography, hydrology, vegetation and soil-forming parent material, showing significant spatial zonation and composite regional characteristics. From north to south, ferroaluminous soils such as red soil, reddish soil and brick red soil are widely distributed; while a variety of composite soil types are formed in the plains, red rock basins and limestone mountains. In 2018, the resident population of the study area was about 114 million, and the population density differed significantly in the core area of the Pearl River Delta, the coastal economic zone and the northern eco-development zone, which accounted for 55.53% of the total population, 29.60% and 14.87 per cent respectively. The region's agricultural spatial zones are divided into the northern mountainous agricultural zone, the central southern subtropical agricultural zone and the southwestern tropical agricultural zone, with a total cultivated area of about 225,900 hectares, of which paddy fields account for 61.89%, watered land for 28.24%, and dry land for 9.87%.

The complex geography, diverse climatic conditions and dense population distribution make Guangdong Province an ideal region for research on high-precision mapping of soil organic carbon in farmland.

2.2 Data sources

The measured soil organic carbon (SOC) data used in this study were obtained from the Soil Formulation and Fertilisation Survey (SFFS) sample point data of the Ministry of Agriculture and Rural Affairs of China at the end of 2017. These data include the geographic coordinates of the sampling sites and the values of soil organic carbon content. For spatial prediction of SOC, soil combined with remotely sensed variables and climatic variables as environmental variables. The remotely sensed variables include vegetation index and terrain factor, which were extracted based on Landsat 8 OLI imagery and ASTER GDEM elevation imagery, respectively. The above data were obtained from the United States Geological Survey (USGS) and Geospatial Data Cloud (GDC) website. The spatial resolution of the images is 30 m, and the projected coordinate system is WGS_ 1984_UTM_Zone_50 N. To ensure that the vegetation index can reflect the real condition of the ground surface at the time of sampling, the shooting time of the selected images is consistent with the acquisition time of the actual survey sample sites, and the weather is sunny and basically free of cloud cover.

By integrating remote sensing and climate variables, the accuracy of soil organic carbon spatial distribution prediction can be improved. This approach not only improves the accuracy of the prediction model, but also provides an important theoretical basis and technical means for the study of regional soil organic carbon.

3 Methodology

3.1 Acquisition of Environment Variables

In this study, we combined remotely sensed variables and meteorological factors as input variables. The main objective of this investigation is to analyze the usefulness of multi-source data in

Table 1 The composition of environmental variables

Categories of variables		Indexes
Remote Sensing (RS)	Vegetation Index (VI)	Normalized Difference Vegetation Index(NDWI) Transformed Vegetation Index(Tv1)
	Topographical Factor(TF)	Digital Evaluation Model(DEM) Slope Aspect
Climale Factor(CF)		Max Annual Temperature(Maxt) Min Annual Temperature(Mint) Mean Annual Temperature(Meant) Mean Annual Rainfall(Rainfall)

explaining the spatial distribution of soil organic carbon (SOC) by constructing a high-precision SOC spatial prediction model. To begin, the study utilizes easily attainable remotely sensed variables (RS) and climate factors (CF) for initial modelling to further understand their effectiveness in explaining the spatial variability of SOC.

The remotely sensed variables will include vegetation indices and topographic factors derived from Landsat 8 OLI imagery and ASTER GDEM elevation data, as well as climate factors consisting of precipitation and air temperature, etc., to capture the climate properties in the study area. All these data work together to provide a comprehensive picture of the spatial distribution and variability of SOC. This is expected to elucidate the spatial distribution characteristics of SOC.

3.2 RF model construction

Random Forest (RF) model is a fusion learning method of decision trees. The core idea of Random Forest model is to use multiple bootstrap sampling to generate and create arbitrary multiple random sample subsets, and then separately build decision trees with these subsets, and then combine these decision trees together to form a random forest. When the model is used for regression prediction, the final prediction result is the averaging results of all decision trees prediction[8]. The number of decision trees is represented by n_estimators. The maximum depth of decision trees is using a depth-first search, which is represented by max_depth. The other important parameters of the RF model are and so on. These parameters also have some influence on the performance of the model. In this study, we also use the ordinary kriging interpolation model (OK) to compare the prediction results with the RF model. We use the different error assessment methods including MAE, RMSE, and so on to compare the prediction performance of the RF model[9]. This assessment method not only measures the prediction performance of the models comprehensively, but also effectively compares the performance of different models in SOC spatial prediction, providing a scientific basis for selecting the best prediction model.

3.3 Data processing methods

The data processing in this study mainly covers three aspects: environmental variable extraction, RF model construction and validation, and generation of spatial distribution maps of soil organic carbon (SOC) in arable land. Firstly, for environmental variable extraction, the soil quantitative attribute data were interpolated by Ordinary Kriging (OK) in the Geostatistical Interpolation Module of ArcGIS 10.2 to generate 30 m $\times$ 30 m raster data. The Normalised Difference Vegetation Index (NDVI) and Triangular Vegetation Index (TVI) of the remotely sensed variables were obtained by Landsat 8 OLI image banding. Specifically, after pre-processing the Landsat 8 OLI image with atmospheric correction, mosaicing and cropping using ENVI 5.3 software, the surface reflectance of the study area was calculated with the following equations:

$$NDVI = \frac{\text{NIR}-\text{red}}{\text{NIR}+\text{red}} \tag{1}$$

$$SAVI = \left(\frac{\text{NIR}-\text{red}}{\text{NIR}+\text{red}} + 0.5\right)^{\frac{1}{2}} \times 100 \tag{2}$$

Near Infrared (NIR) and Red (Red) band reflectance were extracted based on the preprocessed Landsat 8 OLI image data respectively. The Digital Elevation Model (DEM) of the topographic factors was extracted from 19 ASTER GDEM image data of Guangdong Province through coordinate transformation, mosaicing and study area extent extraction. In addition, the other topographic factors were calculated based on the DEM using the Spatial Analyst module of ArcGIS 10.2 software. The raster data of meteorological factors were obtained from the province's climate information dataset by interpolating through the Inverse Distance Weighted (IDW) method in the Geostatistical Interpolation Module of ArcGIS 10.2, and then extracted by the extent of the study area, and the spatial resolution was consistent with that of the remotely sensed variables, which was 30 m.

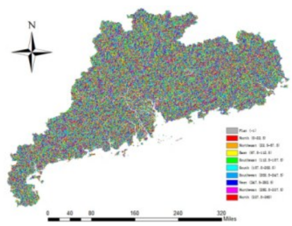

Figure 1 Aspect map of Guangdong Province

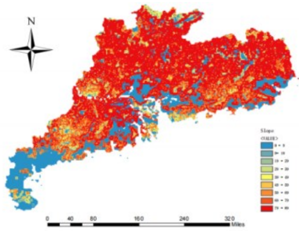

Figure 2 Slope Map of Guangdong Province

Python's scikit-learn library was utilized to build and predict RF models, and generated spatial distribution maps of soil organic carbon (SOC) in arable land through RF and Ordinary Kriging (OK) models. A total of 10,763 SOC samples were randomly aliquoted, with 80% of the samples used for RF calibration and 20% for RF verification[8]. The RF model was implemented by the Python scikit-learn library. RandomForestRegressor() in the RandomForestRegressor() module was used to implement the RF model. The climate factor rasters and soil attribute rasters were used to build the RF model using software.

The RF model was implemented by the Python library, and the climate factor rasters and soil attribute rasters, which had a spatial resolution of 30 m, were input into the RF model based on the corresponding variable combinations to obtain the spatial distribution map of SOC.

4 Results

4.1 RF model prediction accuracy

The training and prediction accuracies of the RF models were examined under different gradient sampling percentages, and the results are presented in Table 2. In the training dataset, the correlation coefficients r between the SOC values predicted by the RF model and the measured values are in the range of 0.75-0.85 and 0.89-0.92, respectively, under different sampling percentages. These results demonstrate that the overall accuracy of the RF model is reliable.

Table 2 Comparison table of the accuracy of RF model in the training and the validation dataset

Prediction models	Training dataset				Validation dataset			
	Gradient Sampling/%	MAE	RMSE	CV	Gradient Sampling/%	MAE	RMSE	CV
RF	80	2.41	3.08	22.26	20	3.7	4.62	15.83
	70	2.81	3.57	19.27	30	3.9	4.9	16.26
	60	2.73	3.47	20.28	40	3.89	4.92	17.85
	50	2.42	3.03	21.99	50	3.83	4.9	18.02

The range of the mean predicted values is much narrower compared to the actual range, therefore reducing the variability and standard deviation of the predictions. This suggests the RF model is good at predicting the mean SOC values, but not at predicting the extremes or capturing the full range of SOC values across the different sites. The RF model needs to be improved to be good at predicting extreme SOC values, to capture the full range of SOC values across the different sites, and to ensure the model shows the true spatial variability of SOC, so that we can make large-scale predictions accurately.

Table 3 Comparison of RF prediction results with measured values

Models	Min/(g*kg-1)	Mean/(g*kg-1)	Max/(g*kg-1)	SD/(g*kg-1)	CV/%
Data	1.51	14.04	31.09	5.22	37.18
RF	3.82	14.05	31.09	3.24	23.06

According to the data in Table 3 , the key environmental variables used for soil organic carbon prediction in the RS model showed a clear importance structure. Specifically, DEM (Digital Elevation Model) occupied the highest importance in the model at 23.22%. In addition, climate factors occupied the second to fifth positions in the overall importance ranking, and although individual factors may have lower importance, their cumulative contribution reached 58.20%. This indicates that climate factors as a whole play a key role in the model and together with the DEM form the main basis of the model's predictive ability.

In the process of soil organic carbon prediction, DEM, as a representative of topographic features, has an important influence on the spatial distribution of soil carbon stocks. Climate factors (e.g., temperature, precipitation, etc.), on the other hand, directly affect the formation, decomposition, and transformation processes of soil organic matter, and thus their role in the model cannot be ignored. This comprehensive model construction based on environmental variables effectively captures the changing law of soil organic carbon under different environmental conditions, providing a scientific basis and predictive capability for accurate soil management and carbon stock assessment.

Table 4 RF model variable importance ranking

Models	Category	Relative importance
RF	RS-TF	DEM(23.22)
	CF	Mint(16.64)
	CF	Maxt(16.24)
	CF	Meant(13.45)
	RS-VI	NDVI(5.73)

4.2 Spatial distribution of organic carbon content

The coastal areas of Guangdong Province in southern China have low to medium levels of SOC content (1.56-3.97 kg·m^-2 shown in blue and light green). The coastal SOC contents also indicate the inherited marine environment influence on SOC distribution. That is why the land use change in coastal areas, including urban and agriculture land use impact the coastal SOC content greatly. The SOC distribution may also be influenced by other marine factors such as saltwater intrusion and coastal management practices. In this figure, the spatial patterns of SOC in coastal areas are not uniform, likely due to the influence of vegetation cover, land use, climate regimes and human activities. For example, the SOC levels in northern regions are higher, where the vegetation cover is denser and the climate is cooler. The spatial patterns of SOC in parts of the central and eastern regions are highly variable, likely due to the dominance of mixed land uses in central and eastern areas. And the coastal and western regions have low SOC content, which could have been influenced by intensive land use and climate regimes. .

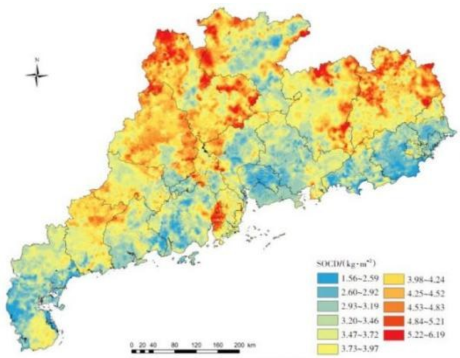

Figure 3 Distribution map of soil organic carbon in Guangdong Province

Figure 4 shows that the soil organic carbon (SOC) content increases with increasing digital elevation model (DEM) values. The detailed trend is that the SOC content increases as the elevation increases. However, a combination of all factors is considered that once the altitude reaches a certain level, the temperature decreases, which has a significant impact on the rate of decomposition of soil organic matter. In colder climates, microbial activity slows down, which means more organic carbon remains in the soil. As a result, organic carbon accumulates in the soil.

This is in line with the findings of Yang Shunhua [9], who also observed that in the transition zone between plains and hills, soil organic carbon increases as the altitude rises. Their study attributed the look-alike phenomenon to slower decomposition rates and lower temperatures. In more vegetated areas, organic matter increases soil carbon through leaf litter and root decay, thus contributing to the increase in soil SOC. At higher elevations where vegetation is abundant, a steady stream of organic matter enters the soil. Organic matter accumulates in these areas due to the slower rate of decomposition. Therefore, lower temperatures, slower decomposition rates and more vegetation at higher altitudes are conducive to the accumulation of SOC.

Figure 5 displays the correlation between Soil Organic Carbon (SOC) and Normalized Difference Vegetation Index (NDVI). In Figure 5, SOC and NDVI have a positive correlation and SOC increases as NDVI increases. The positive correlation emphasizes the important role that vegetation plays in the dynamics of SOC. Higher NDVI values show denser and healthier vegetation, and denser vegetation correlates with higher levels of SOC. Vegetation is a key source of biomass production. As biomass decomposes, it forms soil organic matter. Higher NDVI values represent healthier plants that produce more biomass, resulting in higher SOC. Additionally, vegetation with high NDVI indicates photosynthesis is occurring efficiently. Photosynthesis is the process that plants convert

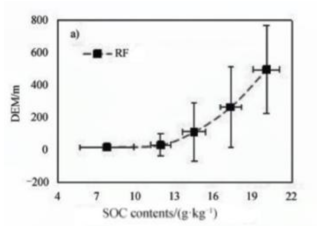

Figure 4 DEM and SOC Relationship

atmospheric carbon dioxide into organic compounds through photosynthesis, and they are stored in the plant tissue. The plant tissue decomposes and contributes a substantial portion of organic carbon to the soil. With dense vegetation, the soil surface is protected from potential erosion, protects existing organic matter, and promotes more organic matter being added to the soil.

The patterns presented in Figure 5 are important to the health, productivity, and overall sustainability of the ecosystem. By understanding the relationship among net difference vegetation index (NDVI) and soil organic carbon (SOC), land managers and environmental scientists can predict and manage soil carbon resources more accurately. Manageable soil conservation and carbon sequestration programs will be more successful.

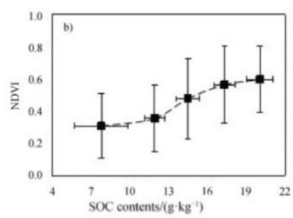

Figure 5 NDVI and SOC Relationship

Figure 6 shows how the minimum temperature, or MinT, affects soil organic carbon (SOC). The embedded figure shows that there's a negative relationship between soil organic carbon (SOC) and minimum annual temperature. This means that there's a strong negative relationship between SOC and temperature. As temperatures drop, the amount of SOC in the soil goes up. The effect of altitude on carbon accumulation is similar to that mentioned above: as temperature decreases, microbial activity slows down corresponding to slower decomposition and more organic carbon is retained in the soil.

Across climates, the colder the climate, the more organic carbon is retained. It can therefore be inferred that there may be better ways to store carbon in colder places or during colder, more seasonal periods. This contributes to our understanding and management of soil carbon storage in relation to global climate change..

The results of this study demonstrate the importance of considering different types of ecosystems and the environment as a whole when it comes to soil conservation. By looking at

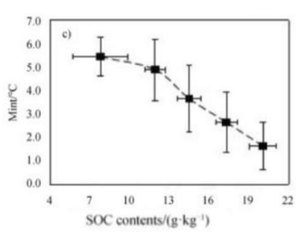

Fig. 6 Mint and SOC Relationship

factors such as altitude, vegetation cover, and temperature, we can find soil management practices that help store SOC. Thus, sound advice on maintaining and increasing natural vegetation cover, managing soil nutrients, and taking climate into account, and incorporating these insights into land management policies and practices, can ensure the long-term health and productivity of soils, thereby contributing to the achievement of sustainable development goals.

5 Discussions

5.1 Lack of Soil Physicochemical Property Variables

The RF model presented in this paper is primarily based on remotely sensed variables and climatic factors in order to predict the distribution of SOC. Upon consideration of the aforementioned variables, it became evident that key soil physicochemical variables, such as effective phosphorus content, were not adequately addressed. Nevertheless, these variables exert a pivotal influence on the spatial heterogeneity of SOC. The exclusion of these key soil properties from this experiment may limit the ability of the model to accurately capture the complexity of SOC distribution, especially in areas with diverse geomorphological features. The absence of variables may result in a poor model fit, thereby reducing the overall reliability and accuracy of predictions.

5.2 Limitations on the Predictive Power of the Model

The random forest model is useful in situations where data is limited or there are economic constraints, but it doesn't perform well overall. Not including important soil physical properties in the model makes it less useful, especially in regions with complex topography. This limits the model's effectiveness in practical land management and soil health assessment applications.

6 Conclusion

In this paper, the Random Forest Model (RFM) was used to select easily accessible remotely sensed variables and climatic factors as model inputs, and the result was that the RF model was jointly constructed with the remotely sensed variables and climatic factors to digitally map the soil organic carbon (SOC) content in Guangdong Province. The results of variable contribution showed that the most important environmental variable affecting the spatial distribution of the SOC prediction results of arable land in the study area was DEM, which was directly related to SOC.

References

[1] ZHOU Y, LIU L Y, LU Y L, et al. Regional scale mapping of soil organic matter using remote sensing and visible-near infrared spectroscopy[J]. Journal of Remote Sensing, 2015, 19(6): 998-1006.

[2] Forkuor G, Hounkpatin O K L, Welp G, et al. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models [J]. PLoS One, 2017, 12(1): e0170478. DOI:10.1371/journal.pone.0170478

[3] Zhang C T, Yang Y, He L Y, et al. Prediction of spatial distribution of soil organic matter based on environmental factors and a joint probability method (In Chinese)[J]. Acta Pedologica Sinica, 2014, 51(3): 666-673.

[4] Sun X L, Zhao Y G, Zhao L, et al. Prediction and mapping of spatial distribution of soil attributes by using soil-landscape models (In Chinese)[J]. Soils, 2008, 40(5): 837-842.

[5] Jia S Y, Yang X L, Li G, et al. Quantitatively determination of available phosphorus and available potassium in soil by near infrared spectroscopy combining with recursive partial least squares (In Chinese)[J]. Spectroscopy and Spectral Analysis, 2015, 35(9): 2516-2520.

[6] Zhou Z Y, Huang W, Xu W, et al. Updating traditional soil maps based on random forest algorithm (In Chinese)[J]. Journal of Huazhong Agricultural University, 2019, 38(3): 53-59.

[7] Fang K N, Wu J B, Zhu J P, et al. A review of technologies on random forests (In Chinese)[J]. Statistics & Information Forum, 2011, 26(3): 32-38. DOI:10.3969/j.issn.1007-3116.2011.03.006

[8] Qi Y B, Wang Y Y, Chen Y, et al. Soil organic matter prediction based on remote sensing data and random forest model in Shaanxi Province (In Chinese)[J]. Journal of Natural Resources, 2017, 32(6): 1074-1086.

[9] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32. DOI:10.1023/A:1010933404324