ISSN 1064-2293, Eurasian Soil Science, 2022, Vol. 55, No. 12, pp. 1689–1699. © Pleiades Publishing, Ltd., 2022..

Optimal Sample Size for SOC Content Prediction for Mapping Using the Random Forest in Cropland in Northern Jiangsu, China

Ting Wu[a], Qihang Wu[a], Qianlai Zhuang[b], Yifan Li[a], Yuan Yao[a], Liming Zhang[a, *], and Shihe Xing[a]

[a] *Fujian Provincial Key Laboratory of Soil Environmental Health and Regulation, College of Resources and Environment, Fujian Agriculture and Forestry University, Fuzhou, 350002 China*

[b] *Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, IN, 47907 USA*

**e-mail: lmzhang_1979@163.com*

Received April 27, 2022; revised September 26, 2022; accepted October 7, 2022

Abstract—A soil organic carbon (SOC) map of high accuracy is the basis for taking mitigation measures against crises of food security and global climate change. Predicting SOC based on a limited number of soil samples can reduce the cost and time for laboratory analysis. This study aimed to assess the influence of sample size on the prediction of SOC and to identify the optimal sample size of SOC prediction for cropland in northern Jiangsu, China. A total of 1182 soil samples were randomly split into calibration and validation sets. Ten calibration subsets of samples between 108 and 1064 were selected by using a parent material-based stratified random resampling strategy. The random forest algorithm was used to develop 10 calibration models validated based on the same validation sample set. These 10 models were evaluated through the explained variance (EV) and the root mean square error (RMSE). The results showed that the calibration model based on 960 soil samples had the best performance in SOC prediction. Significantly biased predictions were produced by the calibration models based on more or less than 960 soil samples due to underrepresentation or overrepresentation. Relief and climate were demonstrated to be the predominant factors influencing SOC prediction in this study area. These results may provide theoretical support for studies relevant to SOC mapping.

Keywords: organic matter, random forest, stratified random resampling, variable importance, Cambisols

DOI: 10.1134/S1064229322600816

INTRODUCTION

The soil organic carbon (SOC) pool is the largest carbon pool in terrestrial ecosystems. Soil organic carbon in cropland plays an important role in improving soil quality since it has a positive relationship with soil fertility [29]. Research has shown that cropland has a significant contribution to the overall global carbon emissions because of anthropogenic activities, while it has a carbon sequestration capacity of 20 Pg [25]. Thus, cropland offers significant opportunities for carbon sequestration to offset ongoing C losses, which will help to slow global warming. Optimizing agricultural management to reduce carbon emissions and increase carbon sequestration of cropland, therefore, is positive not only for food security [2, 23] but also for the achievement of the goal of “carbon centrality” in China.

Research corresponding to digital soil mapping of high precision provides decision support for the development of carbon sequestration and emission reduction measures on cropland. A high-quality soil dataset enables high-precision digital soil mapping, which is critical for understanding the spatiotemporal variations in SOC [18]. A high-quality soil dataset is positively related to soil sample size since more soil samples can capture more accurate SOC information [32]. Soil sample size influences the construction of prediction models for digital soil organic mapping, and the more complex the spatial distribution of SOC is, the more soil samples are needed to determine the SOC trend in detail. However, soil surveys and laboratory testing of soil samples are laborious, time-consuming, and costly. As a result, it is still challenging to make a tradeoff between the precision of digital soil mapping and the size of soil samples. Therefore, a suitable soil sample size must be identified for the high-precision and cost-effective soil organic carbon mapping of a specific region.

Selecting representative subsets of soil datasets from a large data pool is an effective method commonly used for identifying the optimal soil sample size. Although a few studies have conducted this selection by using different algorithms [17, 18], the influence of soil sample size on model prediction has received less attention [18]. Moreover, the optimal soil sample size is site-specific, varying with geographic scale and with the diversity of pedologic characteristics of the study area [7, 22]. In this respect, selecting the optimal soil subset from the existing legacy soil sample pool was run locally for the study area of this paper to improve the accuracy of SOC

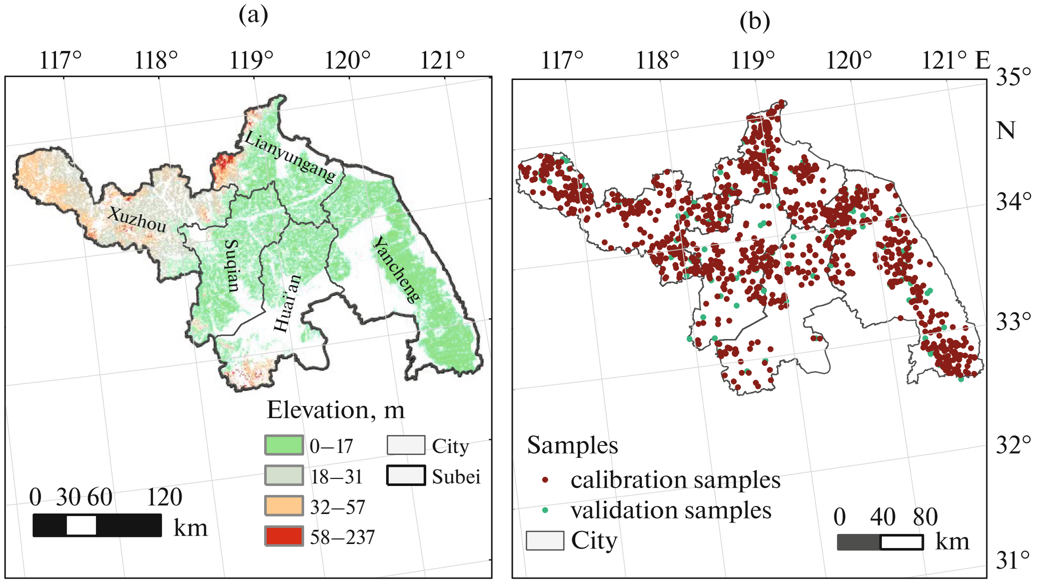

Fig. 1. The location of Subei and the spatial distributions of samples.

prediction, as well as to explore how sample size affects SOC prediction.

Several soil sampling designs have been developed for soil subset selection in recent decades, including random sampling, type-based stratified sampling, and equal interval grid sampling [10, 34]. Simple random sampling can ensure unbiased subset selection, and grid sampling is conveniently performed with geographic information system (GIS) technology. However, there are apparent differences in SOC and variability among different soil types, land use patterns, vegetation types or parent materials [10, 34]. Thus, simple random sampling may result in a sample subset that is unrepresentative of the SOC variability. Similarly, it is difficult to determine suitable grid sizes for grid sampling with little prior knowledge. As a result, a parent material-based stratified resampling method was utilized in this paper. In this sampling design, samples were randomly resampled in proportion by taking the parent material as the stratum, ensuring that each selected subset covered the variability of SOC in different parent materials.

The aim of this work was to (1) describe the spatial variations in SOC content in cropland in northern Jiangsu, China, by using a prediction model built on random forest algorithms; (2) investigate how sample size affects the performance of the SOC predictive model; and (3) identify the optimal sample size and uncover the dominant variables that have the greatest effect on SOC content prediction in the study area.

MATERIALS AND METHODS

Study area. Northern Jiangsu, China, labeled 'Subei', is located within 116°5′–120°3′ E and 22°5′–34°2′ N (Fig. 1). Subei is the lowest plain of Jiangsu Province, having a monsoon climate with four distinctive seasons [33]. Cropland here covers an area of 28 844 km^2, of which paddy fields occupy 36.68% with an area of 10 580 km^2, while rainfed land occupies the rest. The elevation of cropland ranges from 0 to 237 m with an average value of 17.62 m. The annual rainfall ranges from 790 to 1182 mm, gradually decreasing from the southeast to the northwest. The annual average temperature changes between 14.45 and 16.70°C and gradually decreases from south to north. Four soil groups, including Cambisols, Solonchaks, Luvisols and Leptosols, can be identified in the study area according to the International Classification IUSS Working Group WRB (2015) [24]. Among these four groups, Cambisols is the dominant, accounting for 71.87% of the area. Fourteen parent materials are found here, and Yellow River alluvial deposits and marine sediments are dominant, occupying 37.94 and 20.12% of the area, respectively.

Calculation of soil organic carbon content. A total of 1182 soil samples in 2008, providing information on soil organic matter in topsoil (0–20 cm), were sourced from the Ministry of Agriculture and Rural Affairs of Jiangsu Province in the project of Precision Fertilization. In the laboratory, the soil samples were air dried and passed through a 2-mm sieve at room temperature, and then soil organic matter was determined via the $K_2Cr_2O_7$ volumetric method. With this method, excess $K_2Cr_2O_7$ was used in acid medium to oxidize the soil organic carbon, and the organic matter was calculated according to volume of $FeSO_4$ that was used to titrate the remaining $K_2Cr_2O_7$. Formulation of sampling density and sampling layout was guided by expert knowledge corresponding to soil types, vegetation

Table 1. Fourteen environmental variables considered in this study

Categories	Variables included
Soil	Soil group, land use
Climate	Average annual precipitation, average annual mean temperature
Organism	Normalized difference vegetation index
Relief	Elevation, slope, aspect, plan curvature, profile curvature, topographic position index, topographic roughness index, convergence index
Parent Material	Parent material

types, parent material types, etc. Therefore, these samples covered the variability of SOC across landscapes here.

SOC content of topsoil at sampling points is the product of 0.58 and organic matter (OM), and 0.58 is the Bemmelen conversion coefficient.

$$\text{SOC content} = 0.58 \times \text{OM}. \quad (1)$$

Sampling and laboratory analysis of these samples occurred in the same campaign in 2008, which guaranteed contemporaneity of the SOC content data.

Environmental variables. Fourteen environmental variables (Table 1) were selected as the original model input according to a generalization of Jenny's five factors [20]:

$$S = f(s, c, o, r, p), \quad (2)$$

where S, the soil organic carbon to predict, is a function of soils (s), climate (c), organisms (o), relief (r), and parent materials (p).

Variables relevant to relief were computed in ArcGIS based on a digital elevation model (DEM) at a 30 m resolution. This DEM was extracted from the global DEM jointly released by *METI* in Japan and *NASA* in the USA in 2015. A DEM of 30 m is the most commonly used topographic data at the regional scale in this field.

Raster data of parent material and soil group, which were used to characterize parent materials and soils, respectively, were converted from vector soil data at a scale of 1 : 50000 established during the Second National Soil Survey of China. Land use data, which was used to characterize soils in 2008, was visually interpreted based on Landsat TM/ETM+ at a 1000 m resolution. Normalized difference vegetation index (NDVI) data, which was used to characterize organisms, were averaged from annual maximum MODIS-NDVI products with a resolution of 1000 m during 2006–2008. Average annual precipitation and average annual mean temperature data at a resolution of 1000 m, which were used to characterize climate during 2006–2008, were interpolated from daily observations of meteorological stations by using ANUSPLIN software.

All these data, including parent material, soil group, land use, NDVI, average annual precipitation, and average annual mean temperature, were resampled to 30 m to maintain consistency with relief data. Spatial distributions and descriptive statistics of the 14 environmental variables across the study area were shown in Fig. S1.

Collinearity diagnostics. The variance inflation factor (VIF) was applied to detect the possible existence of collinearity between these 14 variables. The results showed that the VIF for each variable was not higher than the traditionally accepted threshold of 10 (Table 2) [3], indicating that there was not a troubling degree of multicollinearity existing in the predictive model of SOC content. Therefore, all the 14 variables were used to construct the predictive models in this study.

Resampling. The whole dataset (n = 1182) was randomly split into a validation set (including 10% of the total samples, n = 118) and a full calibration set (remaining 90% of the total samples, n = 1064). And then, the calibration set was resampled by using a parent material-based stratified random resampling strategy. Specifically, in each stratum (parent material), a number of calibration samples were randomly selected in a decreasing proportion from 100 to 10% with an interval of 10%. As a result, ten calibration subsets in total were obtained and the sizes of samples in each subset and within different parent materials were shown in Table 3.

Random forest prediction of SOC. *Prediction of SOC.* The random forest algorithm, which has been widely used in the field of digital soil mapping, was applied to build the prediction models of SOC content. To predict the target variable, the random forest algorithm firstly generates a number of predictor trees and then gives the overall expression by aggregating these trees.

Table 2. Collinearity statistics of the 14 environmental variables

	AAP	AP	CI	LU	AAMT	NDVI	PLC	PM	PRC	SG	TPI	TRI	ELE	SL
Tolerance	0.61	0.90	0.68	0.95	0.66	0.93	0.39	0.69	0.21	0.74	0.15	0.29	0.53	0.32
VIF	1.65	1.11	1.47	1.05	1.52	1.07	2.57	1.46	4.69	1.34	6.75	3.48	1.89	3.17

ELE—elevation, AP—aspect, AAP—average annual precipitation, CI—convergence index, LU—land use, AAMT—average annual mean temperature, NDVI—normalized difference vegetation index, PLC—plan curvature, PRC—profile curvature, SG—soil group, TRI—topographic roughness index, TPI—topographic position index, SL—slope, and PM—parent material.

Table 3. Statistical distributions of calibration samples within different types of parent material

PM	Size									
	1064	960	850	747	639	535	425	319	214	108
WMR	28	25	22	20	17	14	11	8	6	3
FMS	225	203	180	158	135	113	90	68	45	23
PAD	14	13	11	10	8	7	6	4	3	1
LS	71	64	57	50	43	36	28	21	14	7
RSWGG	34	31	27	24	20	17	14	10	7	3
YRAD	336	302	269	235	202	168	134	101	67	34
YRS	90	81	72	63	54	45	36	27	18	9
RTMBR	5	5	4	4	3	3	2	2	1	1
ADOR	190	171	152	133	114	95	76	57	38	19
SWDSS	3	3	2	2	2	2	1	1	1	1
SWDL	14	13	11	10	8	7	6	4	3	1
XL	51	46	41	36	31	26	20	15	10	5
SWDBR	3	3	2	2	2	2	1	1	1	1

The first column records the name of different parent material, and the first line records the size of different calibration subsets; WMR: Weathering of Metamorphic Rocks, FMS: Fluvial and Marine Sediments, PAD: Pluvial-Alluvial Deposit, LS: Lacustrine Sediments, RSWGG: Residual Slope Wash of Granite-Gneiss, YRAD: the Yellow River Alluvial Deposit, YRS: the Yellow River Sediments, RTMBR: Residual Talus Material of Basic Rocks, ADOR: Alluvial Deposit of Other Rivers, SWDSS: Slope Wash and Diluvium of Sandstone and Shale, SWDL: Slope Wash and Diluvium of Limestone, XL: Xiashu Loess, and SWDBR: Slope Wash and Diluvium of Basaltic Rocks.

The method referred to GridSearchCV from the sklearn library was used to tune the parameters of random forest modeling. In the tuning process, ten predictive models were trained on different sizes of calibration samples with parameters varying in a specified array respectively. The best parameters of each predictive model were selected according to Root Mean Squared Errors returned from 10-fold cross-validation. As a result, ten predictive models that performed the best on each calibration subset were obtained. With these ten optimal models, SOC content predictions were performed on the entire cropland in Subei, resulting in ten grid datasets of SOC content.

Variable Importance. The random forest algorithm learns and records variable importance by looking at how much prediction error increases when values of a specific variable are randomly permutated in the out of bag (OOB) data while values of all other variables remain unchanged [30]. Therefore, a variable with larger value of importance contributes more to SOC prediction. To examine the impact of individual variables and Jenny's factors on SOC prediction, the importance of each variable was calculated and added up by the groups referred to as Jenny's five factors for each calibration subset.

Model Validation. Predictive models were validated on the same validation set of 118 samples by using the root mean square error (RMSE) and explained variance (EV).

Root mean squared errors (RMSEs). The RMSE is a frequently used statistic in the literature to indicate the average deviation of predictions from observations [31].

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

where P_i is the ith predicted value, O_i is the ith observed value, and n is the number of observed values.

Explained variance (EV). Conceptually, EV is a percentage statistic (Eq. 4) measuring how much of the variation in the validation data is explained by the predictive models [16]. This metric is data- and scale-independent; thus, it is qualified for precision comparison among predictive models built on different datasets at various scales or with various variances.

$$\begin{aligned} \mathrm{EV} &= \left(1 - \frac{SSD}{SST}\right) \times 100\% \\ &= \left(1 - \frac{\sum_{1}^{n}(O_i - p_i)^2}{\sum_{1}^{n}(O_i - \bar{O})^2}\right) \times 100\%, \end{aligned} \tag{4}$$

where SSD refers to the sum of square departures, SST refers to the total sum of squares, SSD/SST is also termed relative square error (RSE), n is the number of observations, O_i is the ith observed value, P_i is the ith predicted value and $\bar{O}$ is the average of all observed values.

EV ranges from 0 to 100%. A value of 100% indicates that the predictive model is perfect, while 0% indicates that the predictions are as accurate as using the average of observed values as predictions.

RESULTS

Descriptive statistics of the observed SOC content. The SOC content of the calibration samples ranged from 4.41 to 23.78 g kg^{-1}, with an average value of 10.10 g kg^{-1} and a standard deviation of 2.56 g kg^{-1}. The SOC content of the validation samples ranged from 5.86 to 18.1 g kg^{-1}, with an average value of 10.33 g kg^{-1} and a standard deviation of 2.74 g kg^{-1}. The variation coefficients of the SOC content for the calibration and validation samples were 25.35 and 26.52%, respectively, indicating that the validation samples had higher variability in SOC content. The SOC content in both the calibration and validation sample sets showed no distributional outliers, and thus, both were in accordance with normal distributions.

Accuracy evaluation. Overall, EV increased while RMSE decreased as the size of the calibration samples grew progressively (Fig. 2), indicating that a decrease in the size of the calibration samples would diminish the accuracy of SOC prediction. This finding was also reported by a few previous studies [1, 14].

Moreover, the RF predictive model built on 960 calibration samples performed the best, which was seen from the lowest RMSE and the largest EV. Therefore, it can be concluded that 960 was the optimal sample size for this study. With smaller sizes of calibration samples, RF models made less accurate predictions for lack of SOC variability information; while with larger size of calibration samples (1064 here), the RF model made less accurate predictions as well for overrepresentation of SOC variability in a few parent materials. In the latter case, approximately 70.6% of the calibration samples were clustered into only three types of parent material (Fluvial and Marine Sediments, the Yellow River Alluvial Deposit, and the Alluvial Deposit of other Rivers), thereby exhibiting an imbalance in representing different landscapes characterized by parent materials. This kind of imbalance was deemed detrimental to model performance in SOC content prediction, which highlighted the necessity of sample size optimization.

Variable importance. Land use was definitely the least important variable since it kept holding the least values in the importance permutations (Fig. 3). Both paddy fields and rainfed land, which were the only two types of land use in Subei, were continuous and concentrated in space, showing a lower level of spatial heterogeneity. Therefore, land use contributed the least to the spatial variance in SOC content.

Average annual precipitation was the most important variable in that it captured the largest values in the importance permutations. Rainfed land, which occupied approximately 64% of this study area, was largely dependent on precipitation in terms of crop production. Consequentially, SOC content, which is positively relevant to crop production, tended to be more sensitive to the changes in precipitation. This finding is consistent with studies conducted by Mahmoudzadeh et al. [19] and Davy and Koen [5], which have highlighted the importance of precipitation to SOC prediction.

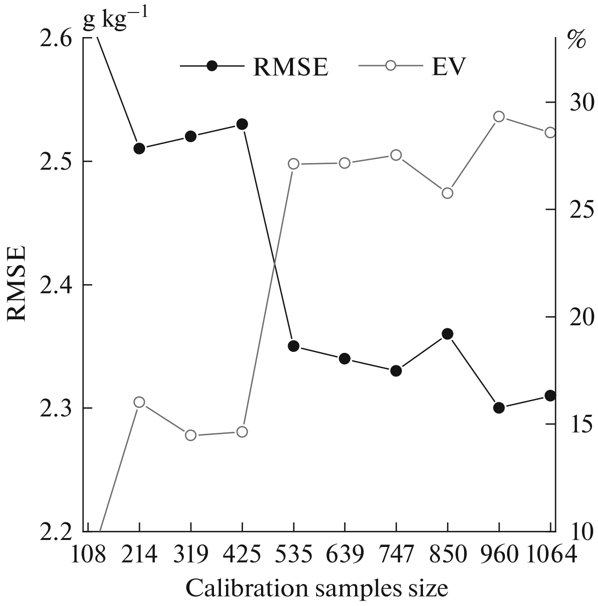

Fig. 2. Validation of SOC content prediction models based on different sizes of calibration samples.

The importance of individual variables was added up by five groups referred to relief, climate, organisms, soils, and parent materials, respectively (Fig. 4). Clearly, importance decreased in the sequence of relief, climate, organisms, soils, and parent materials in all predictions, indicating that the size of calibration samples has little effect on the relative importance of Jenny's five factors. The importance of the relief factor exhibited the highest with values exceeding 0.44, implying that the relief factor played the dominant role in the SOC prediction of this area. This finding has been confirmed by many previous studies. For example, Sherpa et al. [23] and Mahmoudzadeh et al. [19] reported that relief contributed the most to the spatial variability of SOC, and Minasny et al. [21] found that relief was the most important factor influencing SOC distribution through a systematic review of SOC-mapping study around the globe.

The climate factor contributed the second most to the spatial variability of SOC content in Subei. Of the two variables in the climate group, average annual precipitation during 2006–2008 played a more important role in SOC prediction than did the average annual mean temperature. This conclusion had

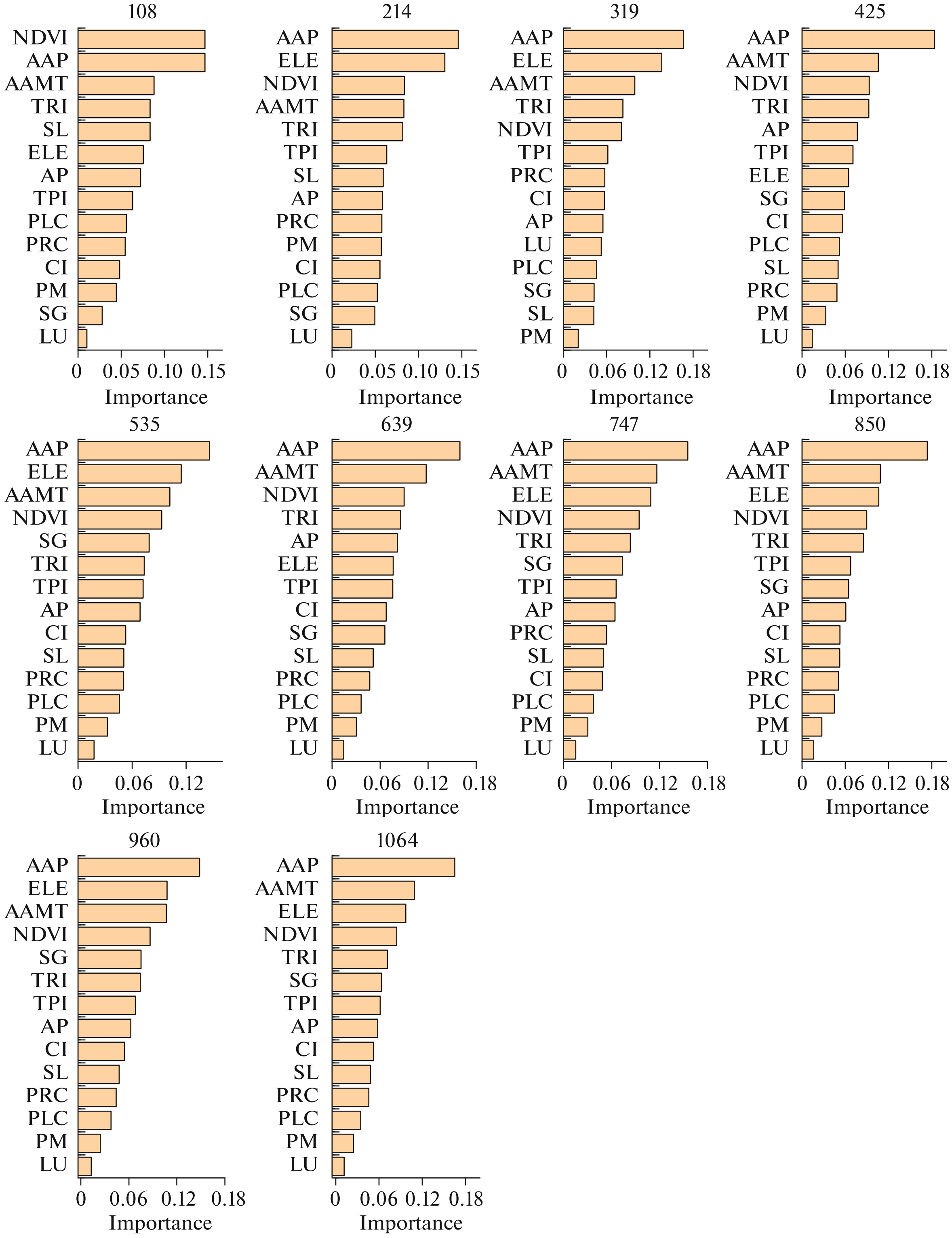

Fig. 3. Variable importance for SOC predictions based on different sizes of calibration samples.

theoretical support because it had been demonstrated that a twofold increase in the rate of organism decomposition would be triggered by a 10°C increase in temperature [13] or by only a 1 mm increase in annual precipitation [4]. Across the study area, the average annual mean temperature during 2006–2008 ranged from 14.6 to 16.7°C with a standard deviation of 3.2°C; the average annual precipitation during the same period changed between 791 and 1182 mm with a standard deviation of 727 mm. Accordingly, the spatial variance in the rate of soil organic matter decomposition caused by average annual precipitation was much more significant than that caused by average annual mean temperature. Consequently, average annual precipitation was recorded to be more important than average annual mean temperature in this study.

Spatial and statistical characteristics of predicted SOC content. Ten raster datasets of SOC content prediction, ranging from 6.68 to 19.2 g kg^{-1}, were classified into five classes by using the natural break classification technique and rendered in a gray-blue to brown gradient, as shown in Fig. 5. Descriptive statistics of these ten raster datasets are shown in Table 4.

SOC content prediction based on 960 calibration samples was used as the baseline due to its highest accuracy. The SOC content in this prediction ranged from 7.60 to 18.93 g kg^{-1} with an average value of 10.56 g kg^{-1}, which was smaller than the study of Zhang et al. [33] and comparable to the study of Liao et al. [15]. In Zhang's study, the average topsoil predicted SOC content in Subei ranged from 13.66 g kg^{-1} in 1999 to 13.13 g kg^{-1} in 2007. In Liao's study, the average topsoil predicted SOC content in Jiangsu Province increased from 9.45 g kg^{-1} in 1982 to 10.9 g kg^{-1} in 2004. In the soil type of Cambisols, which is the dominant in this area, the minimum, maximum and average values of the predicted SOC was 7.60, 18.93, and 10.44 g kg^{-1}, respectively.

In geographic space, very high values (>13.33 g kg^{-1} in this study) in the baseline prediction were primarily distributed in the southern part of Suqian, as well as in the middle and southeastern parts of Huai'an. All these regions have elevations below 7 m and, therefore, are the lowest areas in Subei. In contrast, very low values (<9.06 g kg^{-1} in this study) were mainly distributed in southwestern, northwestern, and northeastern Xuzhou and western Lianyungang. These regions, generally having elevations above 30 m, were the highest places in Subei and accounted for only 15.4% of the whole study area. In conclusion, the spatial distribution of the predicted SOC content in Subei generally followed the topographic trend, which further confirmed the predominant role of relief in SOC content prediction.

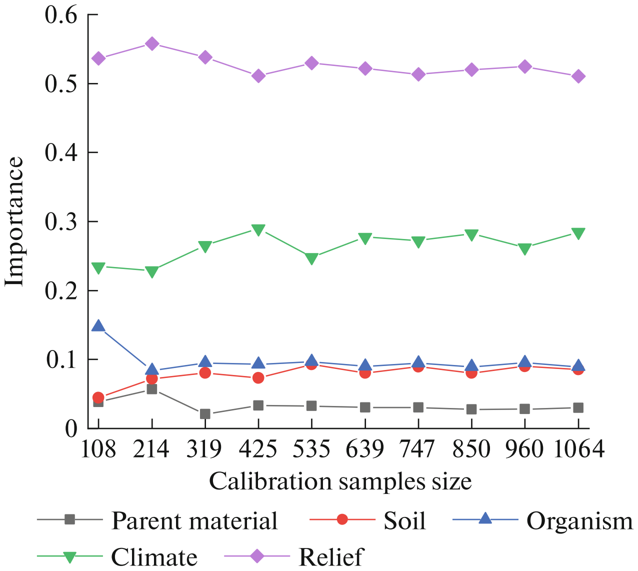

Fig. 4. Importance of the Jenny's five factors for SOC predictions based on different sizes of calibration samples.

In addition, the variation coefficients of these ten predictions based on different sizes of calibration samples were lower than 11%, indicating a lower level of SOC content variability in this study area. This was reasonable since relief, a factor that contributed more than 40% of the variation in the spatial heterogeneity of the predicted SOC content, was extremely flat with a lower level of variability.

Table 4. Descriptive statistics of SOC content predictions based on different sizes of calibration samples

Calibration sample size	MIN	MAX	Mean	STD	Variation coefficient, %
	g kg^{-1}				
108	7.69	13.77	10.22	0.97	8.86
214	7.01	15.38	10.38	0.92	10.31
319	7.00	19.20	10.38	1.07	9.00
425	6.68	14.18	10.44	0.94	8.53
535	7.32	16.92	10.67	0.91	6.56
639	7.66	14.45	10.52	0.69	8.49
747	7.41	18.71	10.60	0.90	8.83
850	6.74	18.92	10.53	0.93	8.05
960	7.60	18.93	10.56	0.85	8.02
1064	7.64	19.08	10.60	0.85	8.86

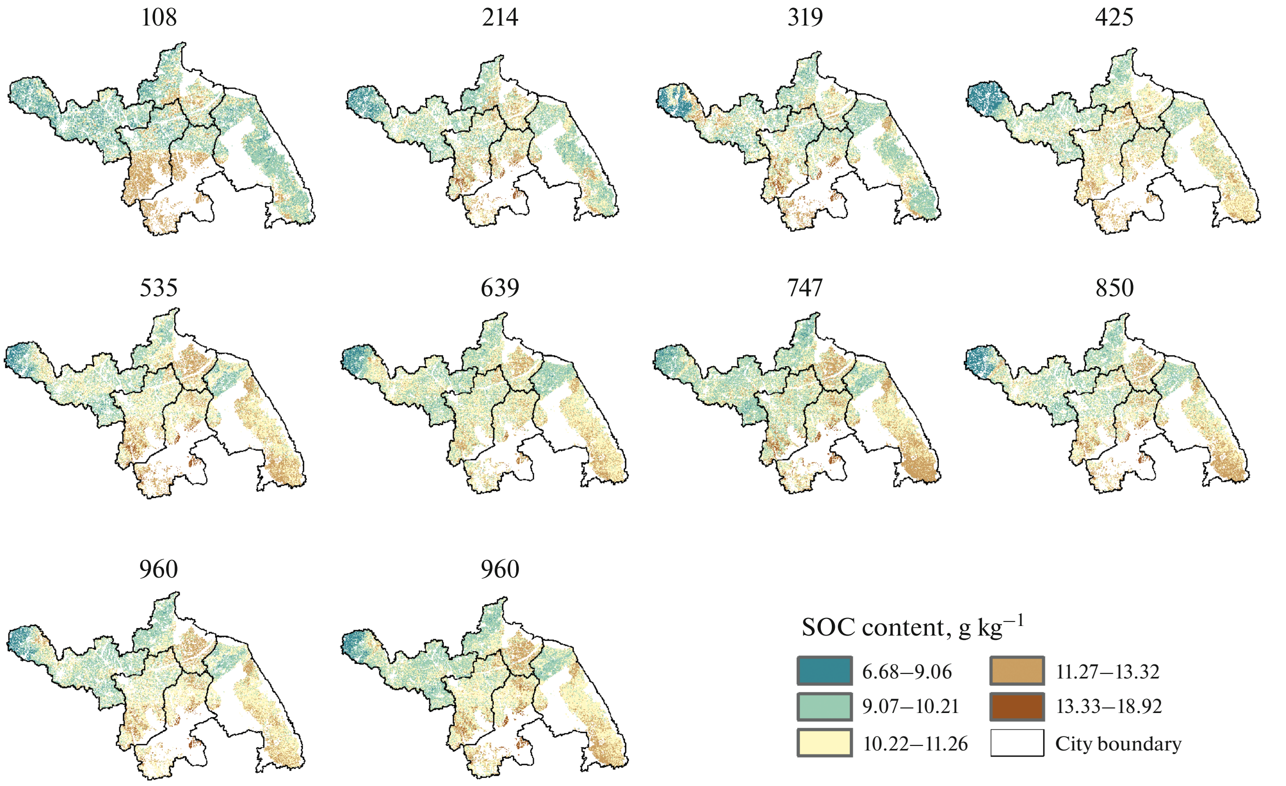

Fig. 5. Spatial distributions of SOC content predictions based on different sizes of calibration samples.

Analysis of difference. Predicted values of SOC content in the ten gridded datasets were extracted at 1182 soil samples, resulting in point data with ten fields recording predicted SOC content based on different sizes of calibration samples.

Since these ten groups of predictive values did not follow the normal distribution, nonparametric statistical tests [7] were used to judge the existence of significant differences among these groups. For multiple comparison, Friedman test was applied and the result, $p = 0.000 < 0.01$, indicated that SOC content predictions based on different sizes of calibration samples were significantly different at the significance level of 1%. For pairwise comparison, Wilcoxon signed rank test was performed by making the prediction based on 960 calibration samples as the control group (Table 5). The results showed that the prediction based on 960 calibration samples was significantly different from the predictions based on 108, 214, 319, 425, 535, 747, and 1064 calibration samples. In addition, the baseline prediction was not significantly different from the predictions based on 639 and 850 calibration samples.

Table 5. Results of Wilcoxon signed rank test

	Z	DF	p
108/960	11.605	1181	0.000***
214/960	8.208	1181	0.000***
319/960	6.95	1181	0.000***
425/960	5.202	1181	0.000***
535/960	7.702	1181	0.000***
639/960	0.417	1181	0.676
747/960	3.02	1181	0.002***
850/960	0.166	1181	0.868
1064/960	4.493	1181	0.000***

*** Represent the significance levels of 1%.

DISCUSSION

This study highlights two potential issues corresponding to large-scale SOC prediction. One of the issues is associated with the effect of sample size on SOC prediction. Here, the results have shown sample size does affect representations of samples as well as relationships between covariates and SOC prediction. Predictions based on more or less than 960 calibration samples were less accurate than baseline prediction, due to overrepresentation of the predictive model or unsatisfactory coverage of SOC spatial variance. Additionally, the importance permutation and importance of individual variables differed across calibration subsets, indicating that relationships between covariates and SOC prediction varied with sample size. Accordingly, both inadequate and oversized samples may lead to biased SOC prediction and misinterpretation of SOC variation, which will mislead the land management strategies corresponding to carbon emission reduction and carbon sequestration.

Another issue this study dealt with is the optimization of sample size. The number of 960 (33/1000 km^2) was identified as the optimal size of calibration samples here. However, the optimal sample size is site-specific, and it varies with resampling strategy, predictive model, or even with environmental variables that participate in modeling. Therefore, we suggest the number of 960 simply as a reference for studies corresponding to SOC prediction in Subei, and one should undertake the optimization procedure to obtain the optimal size for his own study. In addition, the optimal sample size, which accounts for over 80% of the total calibration set, does not save much cost here. Improvement in accuracy of SOC prediction and reliability of relationships between covariates and SOC, however, has highlighted the importance and necessity of sample optimization. Therefore, to accurately inform management strategies related to SOC accounting and sequestration, studies relevant to SOC prediction, especially at large scales with plenty of samples, should start with the optimization of sample size.

The stratified random resampling method has been proven efficient in avoiding sample clustering and obtaining more even coverage within domains and in environmental variables [1]. Various environmental variables could be used for stratification, such as DEM [11], land use [28], parent material (lithology) [26], and soil class [12]. The parent material-based stratified resampling strategy used in this study ensured that the resulting calibration subsets were representative of the variability in SOC content across different parent materials. This resampling strategy, however, is limited in use at small scales where parent material is homogeneous or at places with few types of parent material. Besides, at places where samples are distributed evenly or randomly, grid or random resampling is suggested to guarantee the representation of the variability in observed SOC.

Relief and climate are the predominant factors influencing SOC prediction in Subei. Relief plays a crucial role in soil formation as they control the redistribution of water, solar radiation, sediments, and solutes, which in turn affects soil development and the spatial distribution of soil properties. Many previous studies have highlighted its importance in SOC prediction. For example, Mahmoudzadeh et al. [19] found that 71% of SOC variability was described by terrain attributes, and Taghizadeh-Mehrjardi et al. [27] found that topography had the potential to explain large parts of the variation in SOC. In this study, the association between the relief and SOC is largely due to the effect of topographic variability on rainfall and temperature. Globally, areas with higher rainfall and low temperature are conducive to SOC, since soils in these areas have higher biomass and lower decomposition. In conclusion, climate is the predominant control at large scales, which is further highlighted by the largest value of importance of average annual precipitation and temperature here.

The RMSE here was more accurate than a few previous studies. For example, Zhang et al. [34] obtained an RMSE between 5.10 and 6.12 g kg^{-1} in a study on SOC content prediction for Yujiang County in southeastern China, and Guo et al. [8] obtained an RMSE between 58.76 and 65.19 g kg^{-1} in a study across central Jutland in Denmark. EV values, however, were not high in this study, indicating a poor performance of these predictive models. This may be attributed to (1) the intrinsic large spatial variability of SOC with the interplay of a series of variables [9] and (2) the existence of other variables affecting SOC content variability that were not investigated in this study, such as crop growth conditions and agricultural management. Further research, therefore, shall be undertaken to explore whether parameters that characterize crop growth or agricultural management could improve the performance of SOC predictive models.

Inconsistency in data resolutions may be another error source for the poor performance of predictive models. Resolutions of relief, climate, organisms, soils and parent materials data were different mainly due to (1) the availability of data and (2) disparate spatial autocorrelations and variability of these factors across a given landscape. As 30 m resembled more closely the inherent spatial variability of soil properties [6], all the data of other factors were resampled to the resolution of 30 m, which may introduce imponderable errors that can be propagated to the predictive model. Therefore, collecting more detail data on climate, organisms, soils and parent materials shall be an effective way to improve the performance of the SOC predictive model in future studies.

The average annual mean temperature and precipitation over the 3 years prior to the date of sampling for all 1182 soil sites were calculated to characterize the climate. These time-aggregated variables were spatially and temporally explicit and thus enhanced Jenny's modeling framework since they accounted for the three-year impacts of climate. A 3-year period, however, may be another weakness that brings about errors of SOC prediction, since as proposed by Grunwald et al. in 2011 [6], 30 years is an accepted time scale to characterize the short-term climate. Therefore, climate variables over a 30-year period are suggested for future studies to improve the accuracy of SOC prediction.

CONCLUSIONS

Ten RF models were built on different sizes of calibration samples to predict the SOC content in Subei, aiming to explore the impact of the calibration sample size on SOC content prediction. The number 960 ($33/1000$ km^2) was identified as the optimal calibration sample size in this study, and less or more than this size would lead to significant differences in predicted SOC content due to underrepresentation or overrepresentation of the predictive model, respectively. The results highlighted the necessity of sample size optimization before SOC prediction, which may provide theoretical support for studies relevant to SOC mapping. Moreover, this study can provide useful information and technical support for selecting the optimal sample size for SOC prediction, which can avoid pursuing unnecessarily high density of samples.

FUNDING

The work was supported by the National Natural Science Foundation of China (no. 41971050), the Natural Science Foundation of Fujian Province in China (no. 2020J05027), and the Collaborative Innovation for Juncao Ecology Industry (no. K80ND8002).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest,

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at https://doi.org/10.1134/S1064229322600816.

Fig. S1. Spatial distribution of the 14 environmental variables.

REFERENCES

1. A. Biswas and Y. K. Zhang, "Sampling designs for validating digital soil maps: a review," Pedosphere **28** (1), 1–15 (2018).
https://doi.org/10.1016/S1002-0160(18)60001-3
2. W. Burghardt, D. Heintz, and N. Hocke, "Soil fertility characteristics and organic carbon stock in soils of vegetable gardens compared with surrounding arable land at the center of the urban and industrial area of Ruhr, Germany," Eurasian Soil Sci. **51** (9), 1067–1079 (2018).
https://doi.org/10.1134/S106422931809003X
3. J. H. Cheng, J. Sun, K. S. Yao, M. Xu, and Y. Cao, "A variable selection method based on mutual information and variance inflation factor," Spectrochim. Acta, Part A **268**, 120652 (2022).
https://doi.org/10.1016/j.saa.2021.120652
4. R. C. Dalal and R. J. Mayer, "Long term trends in fertility of soils under continuous cultivation and cereal cropping in southern Queensland. II. Total organic carbon and its rate of loss from the soil profile," Aust. J. Soil Res. **24** (2), 281–292 (1986).
https://doi.org/10.1071/sr9860281
5. M. C. Davy and T. B. Koen, "Variations in soil organic carbon for two soil types and six land uses in the Murray Catchment, New South Wales, Australia," Soil Res. **51** (8), 631 (2013).
https://doi.org/10.1071/sr12353
6. S. Grunwald, J. A. Thompson, and J. L. Boettinger, "Digital soil mapping and modeling at continental scales: finding solutions for global issues," Soil Sci. Soc. Am. J. **75** (4), 1201–1213 (2011).
https://doi.org/10.2136/sssaj2011.0025
7. C. Guerrero, R. Zornoza, I. Gómez, and J. Mataix-Beneyto, "Spiking of NIR regional models using samples from target sites: effect of model size on prediction accuracy," Geoderma **158** (1), 66–77 (2010).
https://doi.org/10.1016/j.geoderma.2009.12.021
8. Z. X. Guo, K. Adhikari, M. Chellasamy, M. B. Greve, P. R. Owens, and M. H. Greve, "Selection of terrain attributes and its scale dependency on soil organic carbon prediction," Geoderma **340**, 303–312 (2019).
https://doi.org/10.1016/j.geoderma.2019.01.023
9. O. K. L. Hounkpatin, F. Op De Hipt, A. Y. Bossa, G. Welp, and W. Amelung, "Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso)," Catena **166**, 298–309 (2018).
https://doi.org/10.1016/j.catena.2018.04.013
10. B. Huang, W. X. Sun, Y. C. Zhao, J. Zhu, R. Q. Yang, Z. Zou, F. Ding, and J. P. Su, "Temporal and spatial variability of soil organic matter and total nitrogen in an agricultural ecosystem as affected by farming practices," Geoderma **139** (3), 336–345 (2007).
https://doi.org/10.1016/j.geoderma.2007.02.012
11. A. Jafari, H. Khademi, P. A. Finke, J. Van De Wauw, and S. Ayoubi, "Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran," Geoderma **232–234**, 148–163 (2014).
https://doi.org/10.1016/j.geoderma.2014.04.029
12. S. B. Karunaratne, T. F. A. Bishop, J. A. Baldock, and I. O. A. Odeh, "Catchment scale mapping of measurable soil organic carbon fractions," Geoderma **219–220**, 14–23 (2014).
https://doi.org/10.1016/j.geoderma.2013.12.005
13. J. N. Ladd, J. M. Oades, and M. Amato, "Microbial biomass formed from 14C, 15N-labelled plant material decomposing in soils in the field," Soil Biol. Biochem. **13** (2), 119–126 (1981).
https://doi.org/10.1016/0038-0717(81)90007-9
14. X. M. Lai, Q. Zhu, Z. W. Zhou, and K. H. Liao, "Influences of sampling size and pattern on the uncertainty of correlation estimation between soil water content and its influencing factors," J. Hydrol. **555**, 41–50 (2017).
https://doi.org/10.1016/j.jhydrol.2017.10.010
15. Q. L. Liao, X. H. Zhang, Z. P. Li, G. X. Pan, P. Smith, Y. Jin, and X. M. Wu, "Increase in soil organic carbon stock over the last two decades in China's Jiangsu Province," Global Change Biol. **15**, 861–875 (2009).
https://doi.org/10.1111/j.1365-2486.2008.01792.x
16. J. Li, "Assessing spatial predictive models in the environmental sciences: Accuracy measures, data variation

and variance explained," Environ. Modell. Software **80**, 1–8 (2016).
https://doi.org/10.1016/j.envsoft.2016.02.004

17. J. Liu and Y. B. Xu, "T-friedman test: a new statistical test for multiple comparison with an adjustable conservativeness measure," Int. J. Comput. Intell. Syst. **15**, 29 (2022).
https://doi.org/10.1007/s44196-022-00083-8

18. F. Lucà, M. Conforti, A Castrignanò, G. Matteucci, and G. Buttafuoco, "Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy," Geoderma **288**, 175–183 (2017).
https://doi.org/10.1016/j.geoderma.2016.11.015

19. H. Mahmoudzadeh, H. R. Matinfar, R. Taghizadeh-Mehrjardi, and R. Kerry, "Spatial prediction of soil organic carbon using machine learning techniques in western Iran," Geoderma Reg. **21**, e00260 (2020).
https://doi.org/10.1016/j.geodrs.2020.e00260

20. A. B. Mcbratney, M. L. Mendonça Santos, and B. Minasny, "On digital soil mapping," Geoderma **117** (1–2), 3–52 (2003).
https://doi.org/10.1016/s0016-7061(03)00223-4

21. B. Minasny, A. B. Mcbratney, B. P. Malone, and I. Wheeler, "Digital mapping of soil carbon," Adv. Agron. **118**, 1–47 (2013).
https://doi.org/10.1016/B978-0-12-405942-9.00001-3

22. S. Nawar and A. M. Mouazen, "Optimal sample selection for measurement of soil organic carbon using online vis-NIR spectroscopy," Comput. Electron. Agric. **151**, 469–477 (2018).
https://doi.org/10.1016/j.compag.2018.06.042

23. S. R. Sherpa, D. W. Wolfe, and H. M. Van Es, "Sampling and data analysis optimization for estimating soil organic carbon stocks in agroecosystems," Soil Sci. Soc. Am. J. **80**, 1377–1392 (2016).
https://doi.org/10.2136/sssaj2016.04.0113

24. X. Z. Shi, D. S. Yu, S. X. Xu, E. D. Warner, H. J. Wang, W. X. Sun, Y. C. Zhao, and Z. T. Gong, "Cross-reference for relating Genetic Soil Classification of China with WRB at different scales," Geoderma **155** (3–4), 344–350 (2010).
https://doi.org/10.1016/j.geoderma.2009.12.017

25. P. Smith, "Carbon sequestration in croplands: the potential in Europe and the global context," Eur. J. Agron. **20** (3), 229–236 (2004).
https://doi.org/10.1016/j.eja.2003.08.002

26. X. L. Sun, S. C. Wu, H. L. Wang, Y. G. Zhao, Y. C. Zhao, G. L. Zhang, Y. B. Man, and M. H. Wong, "Uncertainty analysis for the evaluation of agricultural soil quality based on digital soil maps," Soil Sci. Soc. Am. J. **76** (4), 1379–1389 (2012).
https://doi.org/10.2136/sssaj2011.0426

27. R. Taghizadeh-Mehrjardi, K. Nabiollahi, and R. Kerry, "Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran," Geoderma **266**, 98–110 (2016).
https://doi.org/10.1016/j.geoderma.2015.12.003

28. G. M. Vasques, S. Grunwald, N. B. Comerford, and J. O. Sickman, "Regional modeling of soil carbon at multiple depths within a subtropical watershed," Geoderma **156**, 326–336 (2010).
https://doi.org/10.1016/j.geoderma.2010.03.002

29. H. Wang, J. Wang, Z. Teng, W. Fan, P. Deng, Z. Wen, K. Zhou, and X. Xu, "Nitrogen and phosphorus additions impact statility of soil organic carbon and nitrogen in subtropical evergreen broad-leaved forest," Eurasian Soil Sci. **55** (4), 425–436 (2022).
https://doi.org/10.1134/S1064229322040159

30. K. Were, D. T. Bui, Ø. B. Dick, and B. R. Singh, "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 (2015).
https://doi.org/10.1016/j.ecolind.2014.12.028

31. C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Clim. Res. **30** (1), 79–82 (2005).
https://doi.org/10.3354/cr030079

32. D. S. Yu, Z. Q. Zhang, H. Yang, X. Z. Shi, M. Z. Tan, W. X. Sun, and H. J. Wang, "Effect of soil sampling density on detected spatial variability of soil organic carbon in a red soil region of China," Pedosphere **21** (2), 207–213 (2011).
https://doi.org/10.1016/S1002-0160(11)60119-7

33. L. M. Zhang, Q. L. Zhuang, X. D. Li, Q. Y. Zhao, D. S. Yu, Y. L. Liu, X. Z. Shi, S. H. Xing, and G. X. Wang, "Carbon sequestration in the uplands of Eastern China: an analysis with high-resolution model simulations," Soil Tillage Res. **158**, 165–176 (2016).
https://doi.org/10.1016/j.still.2016.01.001

34. Z. Q. Zhang, Y. Q. Sun, D. S. Yu, P. Mao, and L. Xu, "Influence of sampling point discretization on the regional variability of soil organic carbon in the red soil region, China," Sustainability **10** (10), 3603 (2018).
https://doi.org/10.3390/su10103603