Estimation of Farmland Soil Organic Carbon Stocks by Using Single-Year and Multi-Year Landsat Imagery: A Comparison Across Different Time Periods and Indicator Combinations

Yumei Long and Yue Zhang *

Posted Date: 29 April 2026

doi: 10.20944/preprints202604.2005.v1

Keywords: soil organic carbon; digital soil mapping; indicator combinations; random forest; farm management; Mollisols

 doi:10.20944/preprints202604.2005.v1

Article

Estimation of Farmland Soil Organic Carbon Stocks by Using Single-Year and Multi-Year Landsat Imagery: A Comparison Across Different Time Periods and Indicator Combinations

Yumei Long [1] and Yue Zhang [1,2,3,*]

[1] College of Resources and Environment, Jilin Agricultural University, Changchun 130118, China
[2] Key Laboratory of Soil Resource Sustainable Utilization for Jilin Province Commodity Grain Bases Jilin Agricultural University, Changchun 130118, China
[3] Key Laboratory of Straw Comprehensive Utilization and Black Soil Conservation, Ministry of Education, Changchun 130118, China
* Correspondence: yuezhang@jlau.edu.cn

What are the main findings?

- Farmland SOC content can also be estimated indirectly at crop cover conditions.
- An optimized period of images for SOC prediction in agricultural soils is 3-5 years.

What are the implications of the main findings?

- This finding breaks the traditional limitation that SOC estimation is only feasible during bare-soil periods.
- This finding establishes a 3–5 year window as the optimal trade-off, providing a practical guideline for efficient and generalizable SOC prediction.

Abstract

Accurate soil organic carbon (SOC) estimation is vital for analyzing the global carbon cycle. Currently the bare soil compositing approaches for multi-temporal images are widely used, however the optimized length of compositing period and influence of different indicators on SOC estimaiton for both bare soil and crop cover conditions is unknown. In this study, a time series of Landsat 8 Operational Land Imager multitemporal images was obtained from 2013–2018, with the aim of generating datasets that represent SOC changes across single dates, single years, and multiple years. Soil properties (S), terrain attributes (T), vegetation conditions (V), and farm management practices (F) were employed to predict the spatial distribution of SOC by using the random forest model for both bare soil and crop cover conditions. The results revealed that multi-temporal images from three years and longer produced accurate SOC predictions, with coefficients of determination (R^2) and root mean squared errors (RMSEs) of 0.94-0.95 and 1.75-1.77 g kg^{-1}, respectively. The four types of indicator combinations (S+T+V+F) achieved the best model performance, followed by the T+V+F, S+V+F, and V+F combinations for the bare soil condition in 2016-2018 period. This study provides a possible way for obtaining farmland SOC sequestration under crop cover conditions.

Keywords: soil organic carbon; digital soil mapping; indicator combinations; random forest; farm management; Mollisols

1. Introduction

Soil organic carbon (SOC), especially in agricultural soils, plays an important role in the terrestrial carbon pool because it is closely related to soil quality, food production, and the global

 doi:10.20944/preprints202604.2005.v1

carbon cycle [1-2]. The amount of carbon that can be stored in agricultural soil is 3.3–6.8 billion tons globally each year, and small changes in the soil carbon stock can impact the carbon concentrations in the atmosphere [3-4]. However, temporal and spatial heterogeneity is an obstacle to SOC quantification in agricultural soils because SOC varies with climate, topography, vegetation, parent material, and farm management [5-6]. Accurate estimation of regional SOC stocks in farmlands is extremely important.

The traditional methods for SOC monitoring include extensive soil sampling and laboratory analysis, but this process demands extensive human resources and time, and is difficult to apply at regional scales [7]. In recent years, digital soil mapping (DSM) has been used to estimate cropland soil properties and provide SOC stock spatial patterns [8-9]. Current studies have demonstrated the potential of multispectral satellite images for regional SOC mapping either directly when bare soils are exposed [10] or indirectly when soils are covered by vegetation [11]. Soil reflectance composites (SRCs) based on remote sensing (RS) data have been applied for large-scale soil property prediction when most cropland soils are exposed [12-13]. Moreover, SRC techniques have also been widely applied to overcome the limitation of vegetation, residue, or snow covering the croplands in multitemporal images [14-15].

Various machine learning models are gaining popularity in DSM because they can work with the nonlinear and complex relationships among soil properties and environmental variables [12,16]. Furthermore, the relationships between SOC dynamics and various compositions of environmental covariates can also be obtained quantitatively [17-18]. Tree-based models such as the random forest (RF) model are feasible and reliable methods for more accurate predictions of soil properties, and are widely applied for time series SOC quantification purposes [19-21]. Nevertheless, multiple challenges arise for accurate SOC estimation by using time series images. First, the number of images that are sufficient for farmland SOC estimation is not clear. The SOC estimation accuracy from single-date versus multidate images in the same year, and multidate images in one year versus multidate images across several years, is still unknown [22-23]. Many studies have shown that the spectral indices derived from RS images are important covariates for SOC dynamics [24-25]. Multitemporal spectral indices derived from long-term satellite images can reflect vegetation variations and soil covering condition dynamics, and can be used for SOC monitoring [19,26]. A variety of indices, including the normalized difference vegetation index (NDVI), normalized burn ratio 2 (NBR2), and soil organic carbon index (SOC index), have been applied for soil compositing [11,27]. While there is some uncertainty around optimal time periods, research has provided insights into effective time frames for using spectral indices in SOC prediction [28-30].

Secondly, soil spectral reflectance in agricultural soils is also determined by soil properties, crop covering conditions, terrain attributes, and farm management practices. These factors contribute to the spatiotemporal heterogeneity of SOC in agricultural soils, and introduce uncertainties in the SOC estimation process [27,31]. Moreover, farm management practices, including fertility management, cover cropping, and tillage methods, also increase SOC storage in agricultural soils [32-33]. Anthropogenic disturbances introduce uncertainties in accurate SOC estimation and have received little attention for SOC mapping, especially in intensively cultivated farmlands [27,34]. Identifying how cropland SOC dynamics are linked to changes in agricultural practices, and identifying reliable practice combinations for SOC modeling, are essential. Owing to the spatial variability in SOC dynamics, environmental covariates are likely to vary in their relative influence at different spatial scales [35]. Topographical attributes, such as elevation, slope, and landscape position, may redistribute the effects of climatic factors on soils at local to regional scales [12]. Many studies have indicated the importance of various terrain parameters for predicting SOC at different scales [36-37]. Research has shown that terrain features at local to regional scales do indeed play a significant role in explaining the variation of cropland soil organic carbon (SOC) levels [38-39].

The Mollisol region of Northeast China is an important agricultural production base, and it faces serious problems with soil quality, fertility, and health. Nong'an County, which is characterized by a variety of typical soil types and main agricultural practices in the Mollisol region, was selected as an example case to estimate SOC stocks in the farmland. Although some studies have identified a relationship between the farmland SOC and environmental indicators, how to improve the prediction

 doi:10.20944/preprints202604.2005.v1

accuracy via investigating the synthesis peroid of multi-temporal images, selection of ground cover conditions, and combination of environmental indicators remains unclear. Therefore, the aim of this research is to (i) evaluate the potential of time series multitemporal images for SOC estimation under bare soil and crop cover conditions; (ii) assess the influence of different indicator combinations, especially agricultural management practices, on SOC estimation in agricultural soils; and (iii) obtain the spatial distribution of SOC stocks with associated uncertainty maps in Nong'an County, Northeast China.

2. Materials and Methods

2.1. Study Area

Nong'an County is located in the middle of Jilin Province (124°31′-125°45E, 43°55′-44°55′N) and belongs to the Mollisol region of Northeast China (Fig. 1). It covers a total area of 5400 km2, with the elevation varying from 147 m to 291 m above sea level. The climate of the study area has an average annual precipitation of 507.7 mm and an average annual temperature of 4.7 °C. As a center of commercial grain production, maize (Zea mays L.), rice (Oryza sativa L.) and peanut (Arachis hypogaea L.) are the main crop types in Nong'an County, with a traditional one-season one-year cropping practice and an annual cropping pattern. According to the FAO World Reference Base for Soil Resources [40], the soil types in the study area are Calcic Chernozems, Gleyic Phaeozems, Haplic Phaeozems, Calcaric Phaeozems, Gleyic Chernozems, Haplic Arenosols, Gleyic Solonetz, etc. The time period with bare soil is very short (April and late October), and generally the soil is covered by crops, snow, or straw. The flowchart of this study was shown in Fig. 2.

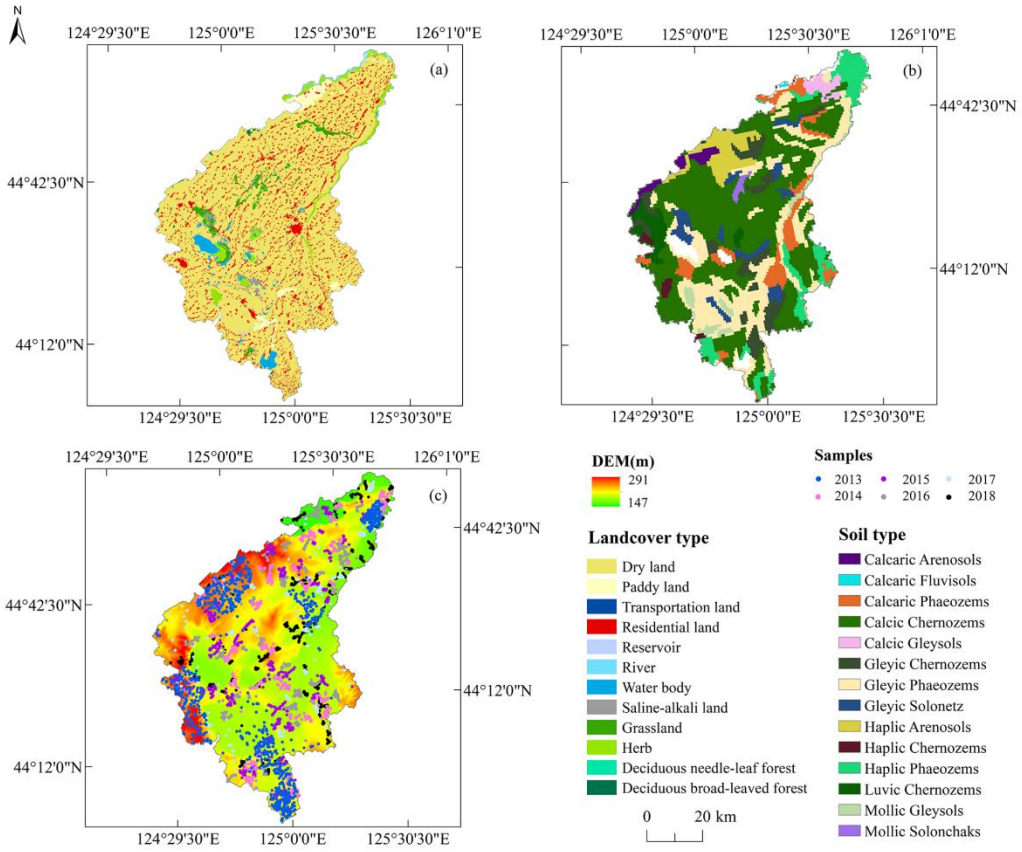

Figure 1. Land use types (a), soil types (b), and DEM and soil sampling sites (c) of Nong'an County.

 doi:10.20944/preprints202604.2005.v1

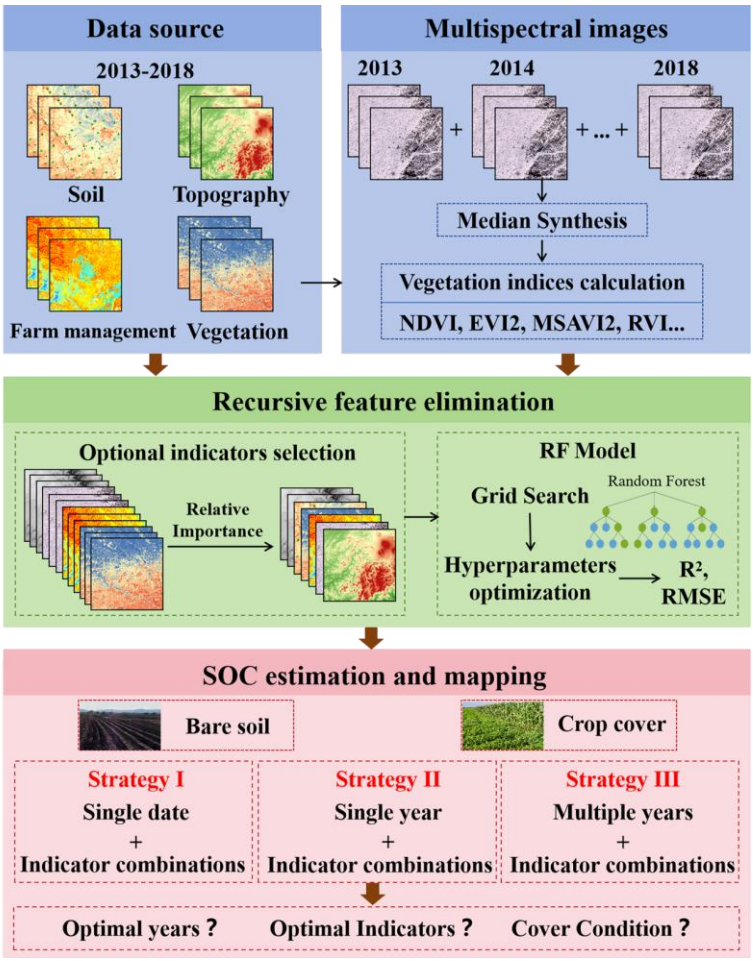

Figure 2. Flowchart of the study.

2.2. Data Source

On the basis of the scorpan model, we selected soil properties, terrain attributes, vegetation/organism conditions derived from satellite imagery, and farm management practices as predictors for SOC estimation in farmlands. This study was conducted at the county scale, so the atmospheric climate conditions were fairly consistent across the whole county, and were not considered. The topographic wetness index (TWI) was selected as a part of the soil climate predictor, and it also represents the effect of climate on SOC. Finally, we collected a total of 31 covariates, which respond to the scorpan model for SOC estimation in the farmlands of Nong'an County (Table 1).

Table 1. Covariates employed for farmland soil organic carbon (SOC) estimation.

Category	Covariates	Reference	Resolution	Source
Soil	Soil type	-	-	The Second National Soil Survey of China
	Soil pH			Soil sampling and measurement
Terrain	Elevation	-	30 m	DEM
	Slope	[41]		
	Aspect	[41]		
	Sin transformed aspect [sin (Aspect)]	[42]		
	Cos transformed aspect [cos (Aspect)]	[42]		
	Plan curvature	[42]		

 doi:10.20944/preprints202604.2005.v1

	Profile curvature	[42]		
	Average curvature	[42]		
	Terrain ruggedness index	[43]		
	Convergence index	[44]		
	TWI	[45]		
	Mass balance index	[46]		
	Vertical distance to channel network	[47]		
	Valley depth	[48]		
	Slope height	[41]		
	Normalized height	[41]		
	Standardized height	[41]		
	Mid-slope position	[41]		
	MRVBF	[49]		
	Length of the slope	[50]		
Vegetation/ Organism	NDVI	[51]	30 m	Landsat images
	EVI2	[52]		
	MSAVI2	[53]		
	RVI	[54]		
	SOC Index	[55]		
	NBR2	[56]	30 m	Landsat images
Farm management	Chemical fertilizer rates	-	-	Statistical yearbook
	Livestock production			
	Grain yield			

Abbreviations: DEM: digital elevation model; TWI: topographic wetness index; MRVBF: areas of deposited material in flat valley bottoms; NDVI: normalized difference vegetation index; EVI2: enhanced vegetation index 2; MSAVI2: modified soil-adjusted vegetation index 2; RVI: ratio vegetation index; SOC index: soil organic carbon index; NBR2: normalized burn ratio 2.

2.2.1. Soil Sampling and Measurement

A field survey of the farmland was conducted in May of each year from 2013–2018. With respect to the spatial distribution of crop types, five subsamples evenly distributed in a 5 m×5 m grid were gathered, and then a mixed sample was obtained for each sample. A total of 5,470 soil samples were collected from the 0–20 cm soil layer from 2013–2018. The number of soil samples across different soil types, and the soil type proportions of the study area, are listed in Table S2. These samples were air dried and sieved through a 2 mm mesh sieve, after which the SOC content was measured using the potassium dichromate volumetric method [57]. The soil pH was measured by using a glass electrode in a 1:5 suspension of soil in distilled water [58]. The soil type data were obtained from the Second National Soil Survey of China at a scale of 1:50,000.

2.2.2. Terrain Attributes

A multiscale digital terrain analysis approach was applied to achieve better SOC prediction performances, and analyzed the impacts of different terrain attributes across multiple scales on SOC prediction. We used the ASTER Global Digital Elevation Map with a spatial resolution of 30 m. The terrain attributes were divided into three groups: local scale, regional scale, and combined scale. As shown in Table 1, the terrain attributes at the local scale include elevation, slope, aspect, sin transformed aspect, cos transformed aspect, plan curvature, profile curvature, average curvature, the terrain ruggedness index, and the convergence index. The terrain attributes at the regional scale include TWI, mass balance index, vertical distance to the channel network, valley depth, slope height,

 doi:10.20944/preprints202604.2005.v1

normalized height, standardized height, mid-slope position, and MRVBF. The terrain attributes at the combined scale include only the length of the slope. The above terrain attributes were derived by using SAGA GIS software (https://saga-gis.org/) with the default parameter settings, and their algorithms and pre- or postprocessing can be found in [39], [37], and [59].

2.2.3. Satellite Imagery

This study selected all available Landsat 8 Operational Land Imager (OLI) images covering Nong'an from 2013–2018 via GEE, for a total of 42 images with less than 10% cloud coverage. The acquisition dates of Landsat images and the identification of ground cover conditions for each year are shown in Fig. 3. Landsat 8 OLI surface reflectance images, which were atmospherically corrected and archived in GEE, were used to estimate the SOC content. The imagery has 11 spectral bands ranging from the visible, near infrared (NIR), and shortwave infrared (SWIR) regions to the thermal infrared (TIR) region, at a spatial resolution of 30 m, which provides the main wavelengths for SOC estimation with high resolution. Finally, the median reflectance was calculated for each band to generate the composite reflectance of the collection images. Because the median synthesis method is rarely affected by possible extreme outliers, it has been a widely applied best pixel selection strategy in recent years [17,60].

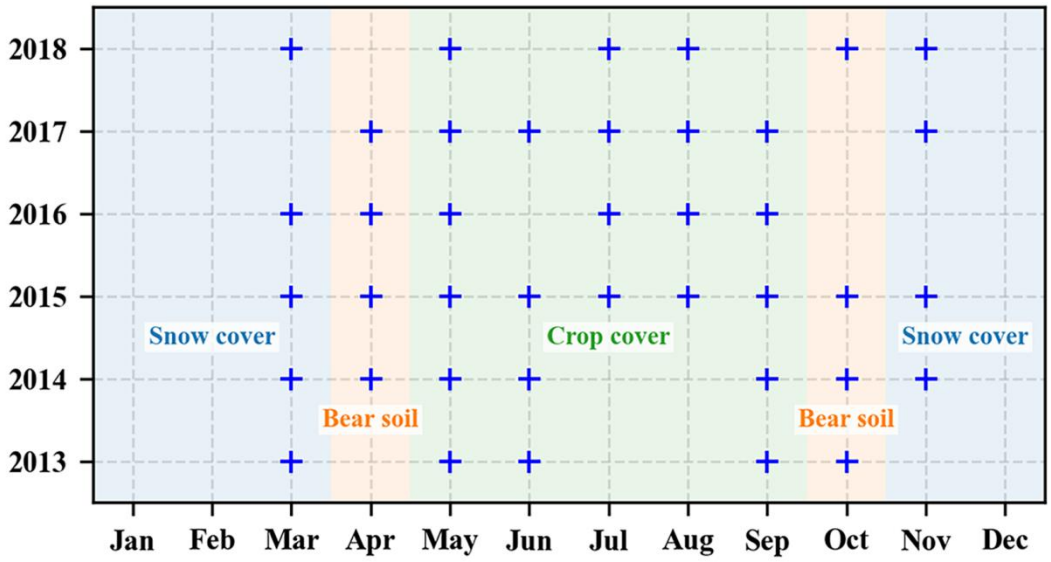

Figure 3. Acquisition dates of Landsat 8 Operational Land Imager (OLI) images and the identification of ground cover conditions from 2013-2018.

Subsequently, six spectral indices, which are commonly used in the mapping of SOC and other soil properties, were employed to reflect the vegetation/organism and farm covering conditions [25,61-62]. The selected spectral indices were the normalized difference vegetation index (NDVI)[51], enhanced vegetation index 2 (EVI2) [52], modified soil-adjusted vegetation index 2 (MSAVI2) [53], ratio vegetation index (RVI) [54], soil organic carbon index (SOC index) [55], and normalized burn ratio 2 (NBR2) [56]. Specifically, NBR2 was originally defined for Landsat imagery to detect crop residue cover in farmland under conservation tillage, so it was selected to derive management strategies such as straw cover. During the bare soil period, the pixels with NBR2 values greater than 0.09 represented crop residue management. These spectral indices were calculated by the following equations:

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \quad (1)$$

$$EVI2 = \frac{2.4 \times (NIR - RED)}{(NIR + RED + 1)} \quad (2)$$

 doi:10.20944/preprints202604.2005.v1

$$MSAVI2 = \frac{2\times NIR + 1 - \sqrt{(2\times NIR+1)^2 - 8\times(NIR - RED)}}{2} \quad (3)$$

$$RVI = \frac{NIR}{RED} \quad (4)$$

$$SOC\ Index = \frac{BLUE}{(RED\times GREEN)} \quad (5)$$

$$NBR2 = \frac{(SWIR1 - SWIR2)}{(SWIR1 + SWIR2)} \quad (6)$$

where BLUE, GREEN, RED, NIR, SWIR1, and SWIR2 correspond to the B2, B3, B4, B5, B6, and B7 bands of Landsat 8 images, respectively.

2.2.4. Farm Management

As mentioned above, NBR2 was selected as one indicator to represent farm management. Survey data, including chemical fertilizer rates, livestock production numbers, and grain yields, were also collected from farmlands across Nong'an County between 2013 and 2018, and were provided by the Statistical Yearbook of Jilin Province (http://tjj.jl.gov.cn/). The three parameters were spatially interpolated into a 30 m resolution raster map via the ordinary kriging method [63].

All the soil sampling points, twenty terrain attributes, six spectral indices, and farm management maps were transformed to the Universal Transverse Mercator (UTM) coordinate system with the World Geodetic System 1984 (WGS-84).

2.3. Semivariance Analysis

A crucial step of a successful DSM procedure is the sampling design [64-65]. A semivariance analysis was conducted to examine the effect of this sampling design on spatial autocorrelation between soil samples by using GS+9.0 software. The equation of semivariance is as follows [66]:

$$\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}[Z(x_i) - Z(x_i + h)]^2 \quad (7)$$

where $\gamma(h)$ represents the number of point pairs separated by h, and where Z(xi+h) and Z(xi) are the sample values at locations xi+h and xi, respectively. The optimal model (e.g., exponential, Gaussian, linear, or spherical) is selected on the basis of the lowest residual sum of squares (RSS) and the greatest coefficient of determination (R^2). Several important parameters were used to evaluate the spatial autocorrelation of the soil samples, including the nugget variance (C_0), sill variance ($C+C_0$), proportion $C_0/(C_0+C)$, and range. The spatial autocorrelation decreases with the increasing of $C_0/(C_0+C)$ values [67].

2.4. Recursive Feature Elimination

A combination of the recursive feature elimination (RFE) method [68-69] and the RF model was applied to reduce the redundant and irrelevant information from the input indicators. First, the importance of various indicators was calculated and sorted based on their impact on impurity reduction during the training process. Second, the least important 20% of indicators were eliminated, and a new prediction dataset was established by using the remaining indicators in each iteration. The iteration process was repeated until there were no or only a few indicators left. Then, the minimum indicator dataset corresponding to the optimal validation result can be obtained. Finally, the most relevant indicators for SOC estimation were identified. Here, the RFE was conducted by using the "rfe" function in the caret package [70].

2.5. RF Regression Algorithm

 doi:10.20944/preprints202604.2005.v1

The random forest (RF) regression model was selected as the estimation model for the SOC content. The RF model is an ensemble learning technique involving multiple decision trees and has been widely applied in classification and regression studies [71]. During the regression process, several regression trees were built on the basis of a unique bootstrap sample, and the averaged trees were used to estimate the soil properties. The RF model contains three user-defined parameters, including the number of trees (ntree), the number of variables used as predictors for each tree (mtry), and the minimum size in each terminal mode (node size) [72]. ntree was set from 500–2000 to train the model. The default value of mtry was generally set to one-third of the number of selected predictors. Node size was also tested with different values, such as 1, 2, 3, 4, and 5. Considering the above parameters and their ranges, hyperparameter tuning was applied to optimize the RF parameters, and the grid search method was implemented [73]. After grid search and validation procedure for parameter optimization, the optimal parameters were determined.

Moreover, the RF model can also calculate every predictor's relative importance on the basis of the regression prediction error of out-of-bag (OOB) predictions. The OOB data can compute the mean squared error (MSE) for each tree before and after permuting each predictor, and the mean difference between the two accuracies of all trees can measure the decrease in accuracy which result from the exclusion of the covariate [74]. The higher the decrease in accuracy, the higher the relative importance for the predictor. Particularly, the relative importance was calculated as follows [75]:

$$w_i = 0.5 \times A_i / \sum_{i=1}^{n} A_i + 0.5 \times B_i / \sum_{i=1}^{n} B_i \quad (8)$$

where w_i represents the relative importance of predictor i; A_i and B_i represent the values of %IncMSE and IncNodePurity for predictor i, respectively; and n represents the total number of predictors. %IncMSE is an increased percentage of the mean squared error, i.e., the decreased accuracy value of the prediction after removing the variable, and IncNodePurity is the increase in node purity, with greater node purity representing a more important variable [76].

2.6. Comparison, Validation, and Uncertainty Assessment

To predict the farmland SOC of Nong'an County in 2018, RF models were built using different indicator combinations across single-date, single-year, and multi-year imagery from 2013–2018. The indicator combinations included S+T, S+V, S+F, S+T+V, S+T+F, S+V+F, T+V, T+F, T+V+F, V+F, and S+T+V+F. The multi-year combinations included one year (2018), two years (2017–2018), three years (2016–2018), four years (2015–2008), five years (2014–2018), and six years (2013–2018). Moreover, we investigated the SOC prediction accuracy during the bare soil and crop coverage periods (Fig. 3). The soil surface in Northeast China is covered by snow from November of the previous year to March of the current year, followed by the crop coverage status from May to September, so the bare soil period is only in April and late October.

For all the SOC estimation models in the study area, the soil samples were split, with 70% for the model calibration set and 30% for the validation set. Two statistical parameters, the coefficient of determination (R^2) and root mean square error (RMSE), were subsequently used for RF model validation across different image numbers and indicator combinations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (z_i - z_i^*)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (z_i - z_i^*)^2} \quad (10)$$

 doi:10.20944/preprints202604.2005.v1

where and are the observed and predicted SOC contents at site , respectively, and where is the average of the observed values. n is the number of samples. A bootstrapping approach was applied to calculate the uncertainty of the SOC estimations [77].

Here, we constructed a quantile regression forest (QRF) model for the uncertainty calculation of SOC prediction. The QRF can estimate a target variable for any quantile that is needed to derive the lower and upper limits of a 95% prediction interval. The uncertainty was calculated on the basis of the prediction interval coverage probability (PICP) obtained from the QRF model, which is expressed as follows [78]:

$$PICP = \frac{count(PI_i^L < z_i < PI_i^U)}{n} \quad (11)$$

where PI_i^L and PI_i^U are the upper and lower 95% confidence limits for a pixel, respectively, and where is the number of data points in the validation dataset. On the basis of this method, the spatial patterns of SOC prediction uncertainty were obtained for different time periods and indicator combinations.

3. Results

3.1. Descriptive Statistics

The descriptive statistics for the measured SOC contents of the soil samples from 2013–2018 are presented in Table 2. The SOC content within Nong'an ranged from 7.2 g kg-1 to 44.7 g kg^{-1}, with a mean value of 25.9 g kg^{-1}. The coefficient of variation (CV) of the SOC was 16.8%, with skewness and kurtosis values of 0.1 and 1.7, respectively, indicating that it was basically normally distributed. The similarities between each year from 2013–2018, and the whole datasets, suggested that it was feasible to compare the performance of single-year and multi-year images for SOC estimation.

Table 2. Descriptive statistics of the measured soil organic carbon (SOC) content (g kg-1) from 2013–2018 in Nong'an County, the Mollisol region of Northeast China.

Year	n	Min	Max	Mean	SD	CV (%)	Median	Skewness	Kurtosis
2013	1016	7.3	33.2	26.3	3.6	13.5	26.8	-1.4	4.4
2014	954	12.5	34.1	25.1	3.6	14.4	25.3	-0.3	-0.3
2015	932	9.7	44.5	26.9	6.3	23.5	26.5	0.0	-0.2
2016	979	9.1	44.7	25.9	5.0	20.4	25.5	0.2	1.0
2017	598	14.6	36.6	26.9	2.9	10.8	27.0	-0.3	1.3
2018	991	7.2	36.2	24.7	2.8	11.3	24.7	-0.2	4.9
2013-2018	5470	7.2	44.7	25.9	4.4	16.8	25.8	0.1	1.7

n: number of soil samples, SD: standard deviation, CV: coefficient of variation.

3.2. Spatial Autocorrelation Analysis

The appropriate semivariance model was selected based on the maximum coefficient of determination R^2 (0.83) and the residual sums of squares RSS<0.001 (Table 3). The range of the semivariance was 900 m, and the mean sampling density was 1 point/km^2, which indicated that our averaged sampling interval was larger than the maximum distance of autocorrelation for the SOC content.

Table 3. Parameters of semivariograms of soil organic carbon (SOC) in Nong'an County.

Model	Nugget (C_0)	Sill (C_0+C)	Nugget/Sill (%)	Range (m)	R^2	RSS
Exponential	0.0117	0.0807	85.50	900	0.83	1.54E-04

 doi:10.20944/preprints202604.2005.v1

3.3. Indicator Selection and Model Performance

To conduct the feature elimination process, we used the RF model to calculate the feature importance across all the SOC prediction models with different time periods. Taking the prediction of 2018 under bare soil conditions as an example, the initial training with all 44 indicators yielded an R^2 of 0.79 and an RMSE of 2.03 g kg^{-1}. To achieve the highest R^2 and lowest RMSE, the 10 most influential indicators (grain yield, livestock production, chemical fertilizer, vertical depth, elevation, NBR2, soil type, SOC index, MSAVI2, sin (Aspect)) were selected for retraining. The model trained with these 10 indicators also achieved a consistent performance, with an R^2 of 0.79 and an RMSE of 2.01 g kg^{-1}. This result indicated that the RFE method could simplify the models without compromising their predictive ability.

After the indicator selection process, the performance of the SOC prediction models across different time periods and indicator combinations was tested by using the validation dataset, as shown in Table 4 for bare soil and Table 5 for crop cover conditions. For RF models of different indicator combinations, the range of optimal hyperparameters settings of ntree mainly ranged between 550-1000, and the optimal mtry mainly ranged between 5-10, respectively. Compared with the prediction model that uses multi-year indicators, the model that uses single-year indicators achieves a poorer performance for SOC estimation. During the 2016–2018 time period, the R^2 and RMSE of the S+T+V+F combination were 0.95 and 1.75 g kg^{-1}, respectively, and the prediction accuracy for farmland SOC was the highest for the bare soil condition (Table 4). The highest prediction accuracy for the best time period and indicator combination for crop cover conditions was consistent with that for bare soil conditions, with R2 (0.94) and lowest RMSE (1.77 g kg^{-1}) for the 2016–2018 time period and S+T+V+F combination (Table 5). The best time period for the selection of multi-temporal images was approximately three years, which achieved the best performance of the farmland SOC prediction model in 2018.

A variety of SOC prediction models showed significant differences in performance across a series of indicator combinations. A prediction model that uses two indicators generally performs poorly, and a model that uses three or four indicators performs well. Compared with other indicator combinations, the results revealed that adding management practices to the environmental covariates increased the prediction accuracy by 27.7-39.7% (in terms of R^2) and 13.0-43.6% (in terms of RMSE) relative to the S+T+V combination for the 2016-2018 period. As shown in Table 4, the farmland SOC prediction models also performed well with indicator combinations such as V+F, S+V+F, T+V+F, and S+T+V+F. These findings indicate that adding vegetation spectral features and management practices would also improve the prediction accuracy of farmland SOC for crop cover conditions.

Table 4. Model performance of different time periods and indicator combinations for soil organic carbon (SOC) estimation under bare soil conditions in Nong'an County, the Mollisol region of Northeast China.

	Single-date		**2018**		**2017-2018**		**2016-2018**		**2015-2018**		**2014-2018**		2013-2018	
Indicators	**R^2 (Mean ± SD)**	**RMSE (Mean ± SD)**	**R^2 (Mean ± SD)**	**RMSE (Mean ± SD)**	**R^2 (Mean ± SD)**	**RMSE (Mean ± SD)**	**R^2 (Mean ± SD)**	**RMSE (Mean ± SD)**	**R^2 (Mean ± SD)**	**RMSE (Mean ± SD)**	**R^2 (Mean ± SD)**	RMSE (Mean ± SD)	R^2 (Mean ± SD)	RMSE (Mean ± SD)
S+T	0.19 ±2.9 ×10^{-3}g	4.43 ±3.3 ×10^{-3}c	0.29 ±1.6 ×10^{-3}i	4.08 ±2.7 ×10^{-3}d	0.29 ±1.6 ×10^{-3}h	4.08 ±7.0 ×10^{-3}c	0.29 ±3.7 ×10^{-3}i	4.08 ±5.0 ×10^{-4}c	0.29 ±1.3 ×10^{-3}i	4.08 ±5.0 ×10^{-4}c	0.29 ±1.7 ×10^{-3}i	4.08 ±8.0 ×10^{-4}c	0.29 ±4.3 ×10^{-3}i	4.08 ±3.8 ×10^{-3}c
S+V	0.14 ±1.5 ×10^{-3}h	4.47 ±2.1 ×10^{-3}a	0.16 ±3.9 ×10^{-3}k	4.46 ±4.0 ×10^{-4}a	0.17 ±3.2 ×10^{-3}i	4.42 ±5.0 ×10^{-4}a	0.21 ±2.4 ×10^{-3}k	4.45 ±1.0 ×10^{-3}a	0.24 ±1.7 ×10^{-3}k	4.45 ±3.1 ×10^{-3}a	0.24 ±3.2 ×10^{-3}k	4.45 ±1.7 ×10^{-3}a	0.25 ±4.8 ×10^{-3}k	4.45 ±3.6 ×10^{-3}a

 doi:10.20944/preprints202604.2005.v1

S+F	0.25 ±2.3 ×10-3f	4.26 ±4.7 ×10-3e	0.39 ±3.1 ×10-3f	3.68 ±5.0 ×10-4e	0.41 ±8.0 ×10-4f	3.63 ±2.0 ×10-4e	0.42 ±1.1 ×10-3g	3.63 ±1.4 ×10-3f	0.42 ±2.9 ×10-3g	3.63 ±5.0 ×10-4e	0.42 ±1.1 ×10-3g	3.63 ±8.0 ×10-4a	0.42 ±2.7 ×10-3g	3.63 ±1.8 ×10-3e
T+V	0.15 ±2.9 ×10-3h	4.45 ±3.7 ×10-3b	0.23 ±2.8 ×10-3j	4.12 ±4.6 ×10-3b	0.26 ±3.0 ×10-3j	4.10 ±6.0 ×10-3b	0.25 ±3.0 ×10-3j	4.11 ±7.0 ×10-4b	0.26 ±2.1 ×10-3j	4.11 ±5.0 ×10-4b	0.26 ±3.3 ×10-3j	4.13 ±2.1 ×10-3b	0.26 ±2.8 ×10-3j	4.13 ±7.0 ×10-4b
T+F	0.20 ±3.0 ×10-3g	4.42 ±4.8 ×10-3d	0.31 ±4.2 ×10-3h	4.09 ±2.1 ×10-3c	0.35 ±2.6 ×10-3g	3.87 ±3.0 ×10-3d	0.37 ±4.3 ×10-3h	3.76 ±2.6 ×10-3e	0.37 ±4.3 ×10-3h	3.75 ±2.6 ×10-3d	0.37 ±1.1 ×10-3h	3.76 ±2.9 ×10-3d	0.37 ±1.1 ×10-3h	3.75 ±2.4 ×10-3d
V+F	0.39 ±2.1 ×10-3d	3.71 ±1.8 ×10-3i	0.47 ±1.9 ×10-3e	3.55 ±4.0 ×10-4g	0.62 ±4.0 ×10-4d	3.35 ±1.0 ×10-4f	0.74 ±3.5 ×10-3d	2.09 ±4.5 ×10-3i	0.74 ±4.0 ×10-3d	2.11 ±4.0 ×10-2h	0.74 ±9.0 ×10-4d	2.11 ±5.0 ×10-4g	0.74 ±1.2 ×10-3d	2.11 ±6.0 ×10-4h
S+T+V	0.35 ±1.3 ×10-3e	3.91 ±7.0 ×10-4g	0.46 ±2.8 ×10-3g	3.61 ±7.0 ×10-4f	0.55 ±3.2 ×10-3e	3.28 ±8.0 ×10-4h	0.65 ±2.8 ×10-4f	3.10 ±1.0 ×10-4g	0.68 ±1.5 ×10-3f	3.11 ±3.3 ×10-3f	0.68 ±3.3 ×10-3f	3.08 ±3.6 ×10-3f	0.68 ±3.5 ×10-3f	3.08 ±3.5 ×10-3f
S+T+F	0.36 ±2.0 ×10-3e	4.10 ±5.0 ×10-4f	0.54 ±4.0 ×10-3c	3.36 ±5.0 ×10-4i	0.55 ±2.9 ×10-3e	3.29 ±9.0 ×10-4g	0.70 ±1.3 ×10-3e	3.07 ±3.5 ×10-3h	0.70 ±3.3 ×10-3e	3.05 ±8.0 ×10-4g	0.70 ±4.1 ×10-3e	3.10 ±6.0 ×10-3e	0.70 ±4.2 ×10-3e	3.07 ±3.5 ×10-3g
S+V+F	0.42 ±3.2 ×10-3c	3.78 ±2.4 ×10-3h	0.51 ±3.6 ×10-3d	3.46 ±3.0 ×10-4h	0.66 ±3.7 ×10-4c	3.13 ±2.0 ×10-4i	0.77 ±2.7 ×10-3c	2.05 ±2.4 ×10-3i	0.77 ±7.0 ×10-4c	2.05 ±2.3 ×10-3i	0.77 ±1.8 ×10-3c	2.05 ±9.0 ×10-3h	0.76 ±1.0 ×10-3c	2.05 ±4.1 ×10-3i
T+V+F	0.51 ±3.9 ×10-2b	3.38 ±7.0 ×10-4k	0.65 ±3.7 ×10-4b	3.15 ±2.0 ×10-3j	0.75 ±3.2 ×10-3b	2.13 ±2.0 ×10-3j	0.84 ±3.4 ×10-3b	1.97 ±5.0 ×10-4j	0.84 ±6.0 ×10-4b	1.96 ±3.8 ×10-3j	0.84 ±2.0 ×10-4b	1.96 ±1.6 ×10-3i	0.84 ±2.0 ×10-4b	1.97 ±3.0 ×10-3j
S+T+V+F	0.58 ±4.0 ×10-4a	3.40 ±3.7 ×10-3j	0.79 ±3.2 ×10-3a	2.01 ±5.0 ×10-4k	0.87 ±9.0 ×10-4a	1.85 ±3.0 ×10-4k	0.95 ±3.4 ×10-3a	1.75 ±6.0 ×10-4k	0.95 ±9.0 ×10-4a	1.75 ±1.6 ×10-3k	0.95 ±1.0 ×10-3a	1.75 ±5.0 ×10-3j	0.94 ±6.0 ×10-4a	1.75 ±1.3 ×10-3k

Note: S-soil properties; T-terrain attributes; V-vegetation indices; F-farm management. Different lowercase letters indicate signiffcant statistical differences between groups of the same confidence intervals ($p<0.05$).

Table 5. Model performance of different time periods and indicator combinations for soil organic carbon (SOC) estimation on the basis of crop cover conditions in Nong'an County, the Mollisol region of Northeast China.

Indicators	Single-date		2018		2017-2018		2016-2018		2015-2018		2014-2018		2013-2018	
	R^2 (Mean ± SD)	RMSE (Mean ± SD)	R^2 (Mean ± SD)	RMSE (Mean ± SD)	R^2 (Mean ± SD)	RMSE (Mean ± SD)	R^2 (Mean ± SD)	RMSE (Mean ± SD)	R^2 (Mean ± SD)	RMSE (Mean ± SD)	R^2 (Mean ± SD)	RMSE (Mean ± SD)	R2 (Mean ± SD)	RMSE (Mean ± SD)
S+T	0.15 ±1.7 ×10-3k	4.40 ±4.1 ×10-3a	0.29 ±1.3 ×10-3k	4.08 ±3.7 ×10-3c	0.29 ±2.7 ×10-3j	4.08 ±6.9 ×10-3a	0.29 ±1.7 ×10-3k	4.08 ±4.4 ×10-3b	0.29 ±3.6 ×10-3k	4.08 ±9.5 ×10-3a	0.29 ±1.9 ×10-3k	4.08 ±4.8 ×10-3b	0.29 ±1.3 ×10-3k	4.08 ±3.3 ×10-3a

 doi:10.20944/preprints202604.2005.v1

S+V	0.25 ±1.3 ×10-3i	4.32 ±3.1 ×10-3d	0.48 ±2.0 ×10-3g	3.24 ±5.5 ×10-3g	0.53 ±1.9 ×10-3f	3.41 ±5.2 ×10-3e	0.54 ±1.0 ×10-3g	3.42 ±2.3 ×10-3e	0.54 ±2.9 ×10-3g	3.41 ±8.1 ×10-3c	0.53 ±3.8 ×10-3g	3.41 ±1.0 ×10-2e	0.54 ±2.6 ×10-3g	3.41 ±5.8 ×10-3d
S+F	0.26 ±4.0 ×10-3h	4.35 ±8.9 ×10-3c	0.40 ±1.3 ×10-3h	3.65 ±4.5 ×10-3d	0.43 ±8.0 ×10-4h	3.62 ±2.3 ×10-3c	0.45 ±1.1 ×10-3i	3.53 ±2.9 ×10-3d	0.45 ±1.5 ×10-3i	3.53 ±5.0 ×10-3b	0.45 ±1.1 ×10-3i	3.53 ±3.9 ×10-3d	0.45 ±2.4 ×10-3i	3.53 ±7.9 ×10-3c
T+V	0.27 ±3.6 ×10-3g	4.13 ±8.9 ×10-3e	0.38 ±2.1 ×10-3i	4.12 ±5.4 ×10-3a	0.45 ±1.9 ×10-3g	3.67 ±5.2 ×10-3b	0.48 ±3.0 ×10-3h	3.54 ±7.9 ×10-3c	0.48 ±2.4 ×10-3h	3.54 ±5.7 ×10-3b	0.48 ±2.4 ×10-3h	3.84 ±6.0 ×10-3c	0.48 ±1.4 ×10-3h	3.54 ±3.9 ×10-3b
T+F	0.21 ±2.8 ×10-3j	4.41 ±4.9 ×10-3b	0.36 ±1.9 ×10-3j	4.11 ±5.0 ×10-3b	0.38 ±2.6 ×10-3i	4.08 ±7.7 ×10-3a	0.38 ±1.1 ×10-3j	4.09 ±3.9 ×10-3a	0.38 ±1.1 ×10-3j	4.09 ±3.9 ×10-3a	0.38 ±1.1 ×10-3j	4.09 ±3.7 ×10-3a	0.38 ±1.1 ×10-3j	4.09 ±.3. 6×10-3g
V+F	0.40 ±3.0 ×10-3d	3.86 ±7.7 ×10-3h	0.52 ±1.6 ×10-3f	3.46 ±4.0 ×10-3e	0.66 ±3.0 ×10-4c	3.15 ±1.1 ×10-3h	0.77 ±1.3 ×10-3d	2.12 ±4.3 ×10-3h	0.77 ±1.7 ×10-3d	2.12 ±5.4 ×10-3e	0.77 ±1.8 ×10-3d	2.12 ±6.6 ×10-3h	0.77 ±2.4 ×10-3d	2.12 ±8.2 ×10-3h
S+T+V	0.39 ±2.7 ×10-3e	3.88 ±7.8 ×10-3g	0.61 ±5.0 ×10-4d	3.22 ±1.3 ×10-3h	0.62 ±4.0 ×10-4d	3.17 ±1.9 ×10-3g	0.68 ±3.0 ×10-4e	3.11 ±1.4 ×10-3g	0.68 ±1.0 ×10-4e	3.12 ±8.0 ×10-4d	0.68 ±5.0 ×10-4e	3.12 ±1.9 ×10-3g	0.68 ±3.0 ×10-4e	3.12 ±1.4 ×10-3f
S+T+F	0.35 ±2.0 ×10-3f	4.02 ±5.4 ×10-3f	0.53 ±2.0 ×10-3e	3.41 ±5.0 ×10-3f	0.55 ±2.5 ×10-3e	3.43 ±5.9 ×10-3d	0.60 ±6.0 ×10-4f	3.41 ±1.9 ×10-3f	0.60 ±4.0 ×10-4f	3.41 ±1.6 ×10-3c	0.60 ±2.0 ×10-4f	3.41 ±8.0 ×10-4f	0.60 ±6.0 ×10-4f	3.41 ±2.4 ×10-3e
S+V+F	0.45 ±1.4 ×10-3c	3.60 ±3.5 ×10-3i	0.64 ±5.0 ×10-4c	3.16 ±1.2 ×10-3i	0.70 ±9.0 ×10-4b	2.25 ±2.5 ×10-3f	0.82 ±1.0 ×10-3c	2.04 ±5.8 ×10-3i	0.82 ±7.0 ×10-4c	2.04 ±3.5 ×10-3f	0.81 ±7.0 ×10-4c	2.04 ±6.2 ×10-3i	0.81 ±8.0 ×10-4c	2.04 ±5.4 ×10-3i
T+V+F	0.52 ±2.5 ×10-3b	3.45 ±6.4 ×10-3j	0.69 ±4.0 ×10-4b	3.15 ±1.6 ×10-3j	0.76 ±1.6 ×10-3a	1.96 ±5.5 ×10-3i	0.90 ±7.0 ×10-4b	1.85 ±4.8 ×10-3j	0.90 ±5.0 ×10-4b	1.85 ±5.2 ×10-3g	0.90 ±8.0 ×10-4b	1.85 ±4.9 ×10-3j	0.90 ±1.0 ×10-3b	1.85 ±4.9 ×10-3j
S+T+V+F	0.56 ±2.0 ×10-4a	3.20 ±5.5 ×10-3k	0.75 ±2.0 ×10-3a	2.01 ±7.1 ×10-3k	0.76 ±1.1 ×10-3a	2.05 ±4.5 ×10-3j	0.94 ±7.0 ×10-4a	1.77 ±6.0 ×10-3k	0.95 ±5.0 ×10-4a	1.76 ±5.1 ×10-3h	0.94 ±6.0 ×10-4a	1.76 ±4.1 ×10-3k	0.95 ±4.0 ×10-4a	1.76 ±4.0 ×10-3k

Note: S-soil properties; T-terrain attributes; V-vegetation indices; F-farm management. Different lowercase letters indicate signiffcant statistical differences between groups of the same confidence intervals ($p<0.05$).

Since the model performance of the S+T+V+F combination for the bare soil condition was almost consistent with that of the same indicator combination for the crop cover condition, the SOC prediction accuracies for the crop cover condition were selected to be listed here. Moreover, the bare soil period of the study area was short, and SOC prediction during crop cover conditions would be useful for most years. The prediction accuracies across single-date images from 2018, multi-temporal images from 2018, multi-temporal images from 2016–2018, and multi-temporal images from 2013–2018 for crop cover conditions were tested using the validation dataset (Fig. 4a-d). The SOC prediction results were satisfactory, as shown in Fig. 4c-d, with most points close to the 1:1 line; however, one problem with the validation results was somewhat overestimating the low SOC values. The validation results revealed that the use of multi-temporal images longer than three years can

 doi:10.20944/preprints202604.2005.v1

increase the prediction accuracy from only one-date image (R^2=0.56, RMSE=3.20 g kg^{-1}) or one-year images (R^2=0.75, RMSE=2.01 g kg^{-1}) to the multi-temporal images during 2013–2018 (R^2=0.95, RMSE=1.76 g kg^{-1}). Thus, the model performance can be improved by 21.1-41.1% (in terms of R^2) and 12.4-45.0% (in terms of RMSE) relative to the use of single-date or single-year for SOC prediction.

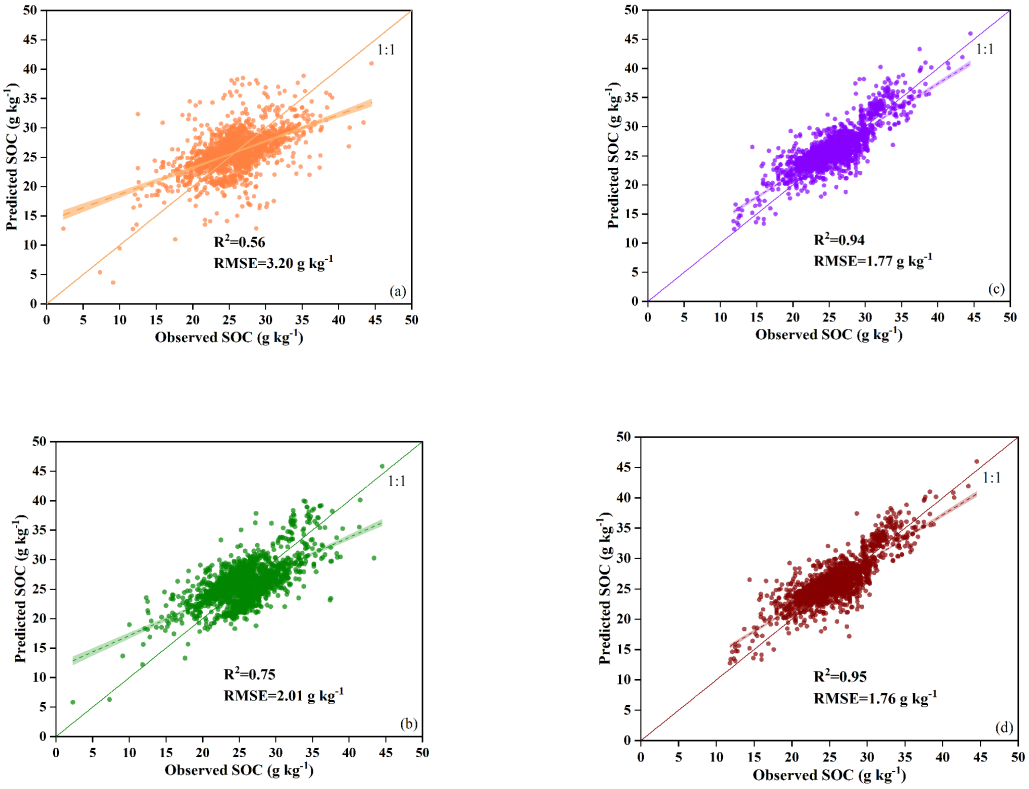

Figure 4. Scatter plots of observed versus predicted farmland soil organic carbon (SOC) (g kg^{-1}) derived from the RF model via the S+T+V+F indicator combination: (a) single-date image from 2018, (b) multitemporal images from 2018, (c) multitemporal images from 2016–2018, and (d) multitemporal images from 2013–2018 for crop cover conditions in Nong'an County, the Mollisol region of Northeast China. The dashed lines represent fitted curves, and the shaded area represent 95% confidence intervals.

To evaluate the robustness of our optimal modeling strategies, such as V+F, S+T+F, S+V+F, T+V+F, and S+T+V+F indicator combinations in 2016-2018, the bootstrap sampling method was also used to test the influence of reducing training samples at 90%, 70%, and 50% of the full calibration data set. The performance metrics were validated and presented in Table 6 (detailed in Table S3).

Table 6. Model robustness evaluation across different sample sizes, with S+T+V+F combination in 2016-2018 as an example.

Cover condition	Sample size	R^2 (Mean ± SD)	RMSE (Mean ± SD)
Bare soil	100%	0.95 ± 3.4×10^{-3}	1.75 ± 6.0×10^{-4}
	90%	0.93 ± 9.5×10^{-4}	1.60 ± 6.0×10^{-4}
	70%	0.96 ± 4.8×10^{-3}	1.50 ± 4.5×10^{-4}
	50%	0.94 ± 4.3×10^{-3}	1.55 ± 4.0×10^{-4}
Crop cover	100%	0.94 ± 7.0×10^{-4}	1.77 ± 6.0×10^{-3}
	90%	0.93 ± 8.5×10^{-4}	1.85 ± 6.7×10^{-3}
	70%	0.94 ± 7.1×10^{-4}	1.50 ± 7.9×10^{-3}
	50%	0.94 ± 6.9×10^{-4}	1.56 ± 6.5×10^{-3}

Note: The model accuracies across different sample sizes were obtained through 100 bootstrap sampling iterations. Full data in detail is in Supplementary Table S3.

3.4. Covariate Importance for SOC Predictions

 doi:10.20944/preprints202604.2005.v1

The indicator importance in the S+T+V+F combination for crop cover conditions in the four time periods of satellite images (single date, single year, three years, and more than three years) is shown in Fig. 5. The relative importance of the four types of indicators varied across different time periods of the selected satellite images. The soil properties accounted for 22.6-27.5% of the relative importance for SOC predictions in the single-date and single-year periods, and their importance decreased as the time periods changed, with importance scores ranging from 8.4-10.6% in the 2016–2018 and 2013–2018 periods. The relative importance of terrain attributes decreased from 37.5% to 15.7% as the time period increased. The importance scores of both vegetation and farm management increased over time, with values ranging from 23.5-48.7% (in terms of vegetation) and 11.5-27.3% (in terms of farm management), respectively. Moreover, the contribution of each type of indicator to the S+T+V+F combination changed with increasing time. The relative importance of each indicator was consistent in the 2016–2018 and 2013–2018 time periods, with soil properties accounting for 8.4-10.6%, terrain attributes accounting for 15.7-17.2%, vegetation accounting for 45.6-48.7%, and farm management accounting for 27.3-27.8% of the S+T+V+F combination for SOC predictions.

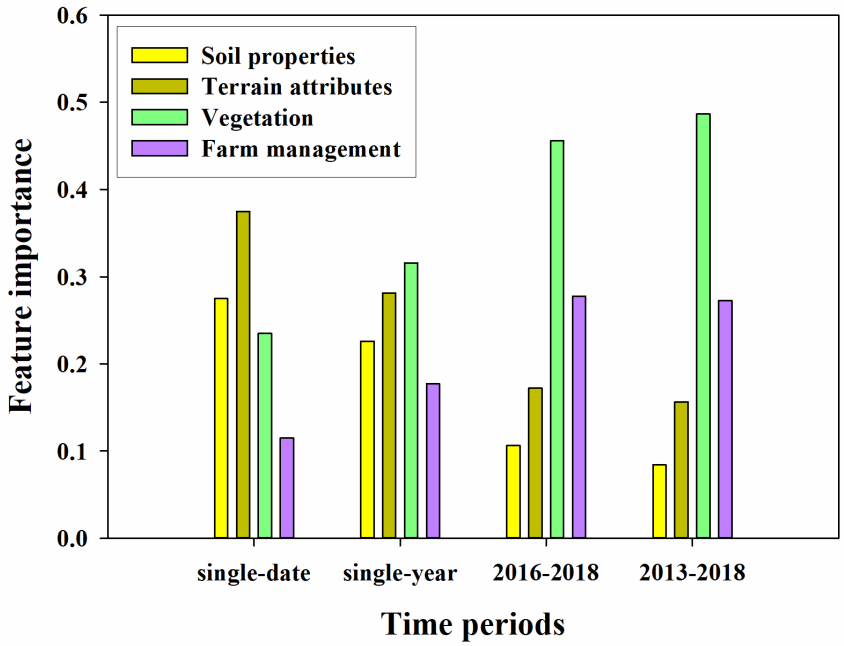

Figure 5. Covariate importance of soil organic carbon (SOC) prediction for crop cover conditions via the RF model for the S+T+V+F indicator combination across four time periods: single-date images from 2018, multitemporal images from 2018, multitemporal images from 2016–2018, and multitemporal images from 2013–2018 in Nong'an County, the Mollisol region of Northeast China.

3.5. Spatial Distribution of SOC and Uncertainty Mapping

Fig. 6 shows the spatial distribution maps of farmland SOC in Nong'an County during four different time periods (single-date, single-year, three-year, and more than three-year periods) under crop cover conditions. The spatial patterns of the four maps were mostly consistent, with the lower SOC values distributed in the western part and the higher SOC values distributed in the northeastern and southwestern regions. However, the spatial details differed across the four predicted SOC maps. The SOC spatial maps from single-date and single-year images presented more heterogeneity than did the maps from the multi-year images, especially in the central part of the study area. To obtain more detailed differences and errors among the four predicted SOC maps, the PICP was calculated on the basis of the 95% prediction interval as the uncertainty map (Fig. 7a-d). The PICP values from multiple years were generally higher than those from single dates and single year, which indicated the reliability and stability of the prediction model derived from three or more years.

 doi:10.20944/preprints202604.2005.v1

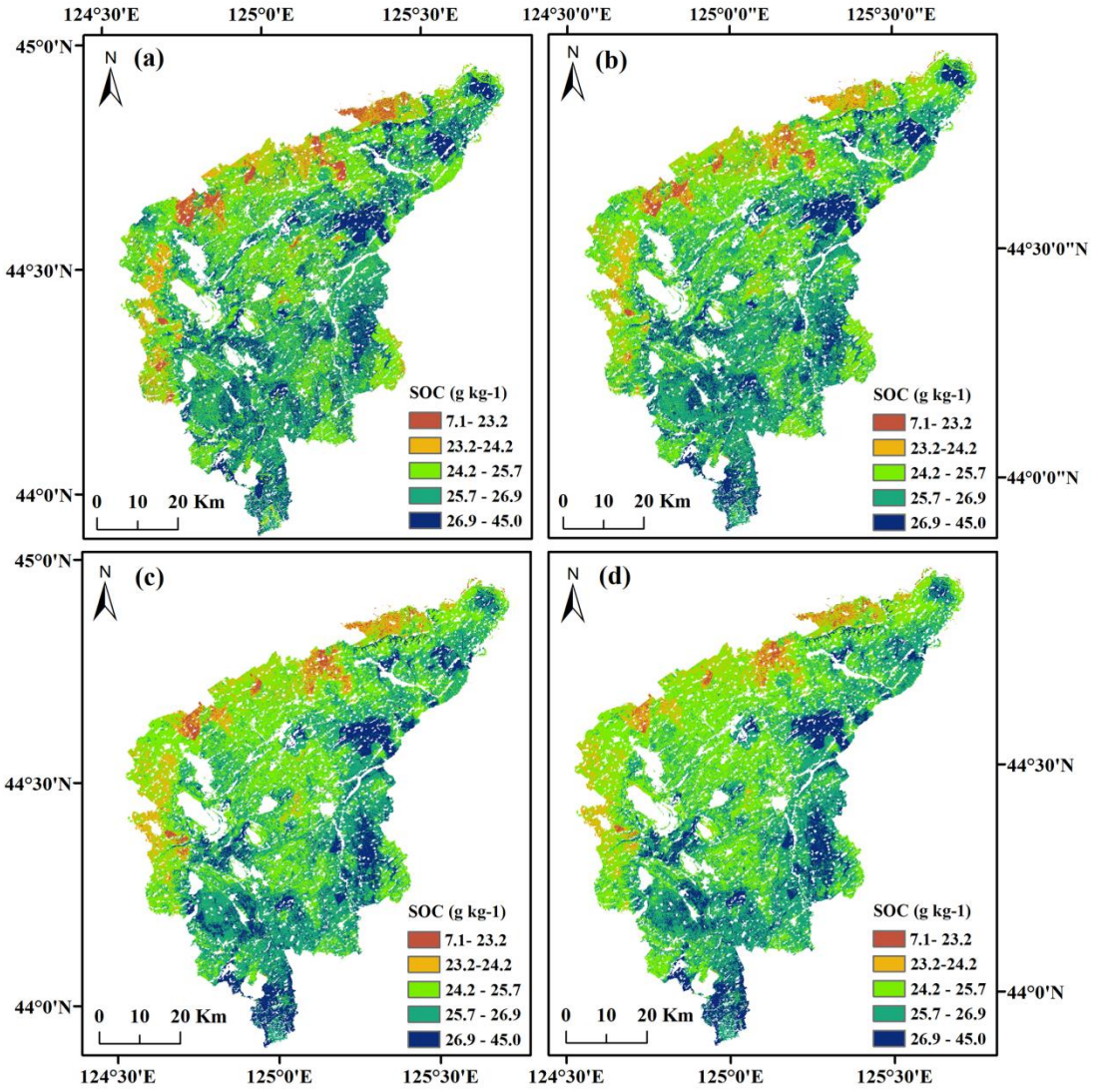

Figure 6. Spatial distribution map of soil organic carbon for the S+T+V+F combination on crop cover conditions over four time periods of Landsat images. (a) Single-date image from 2018, (b) multitemporal images from 2018, (c) multitemporal images from 2016–2018, and (d) multitemporal images from 2013–2018 in Nongan County, the Mollisol region of Northeast China.

 doi:10.20944/preprints202604.2005.v1

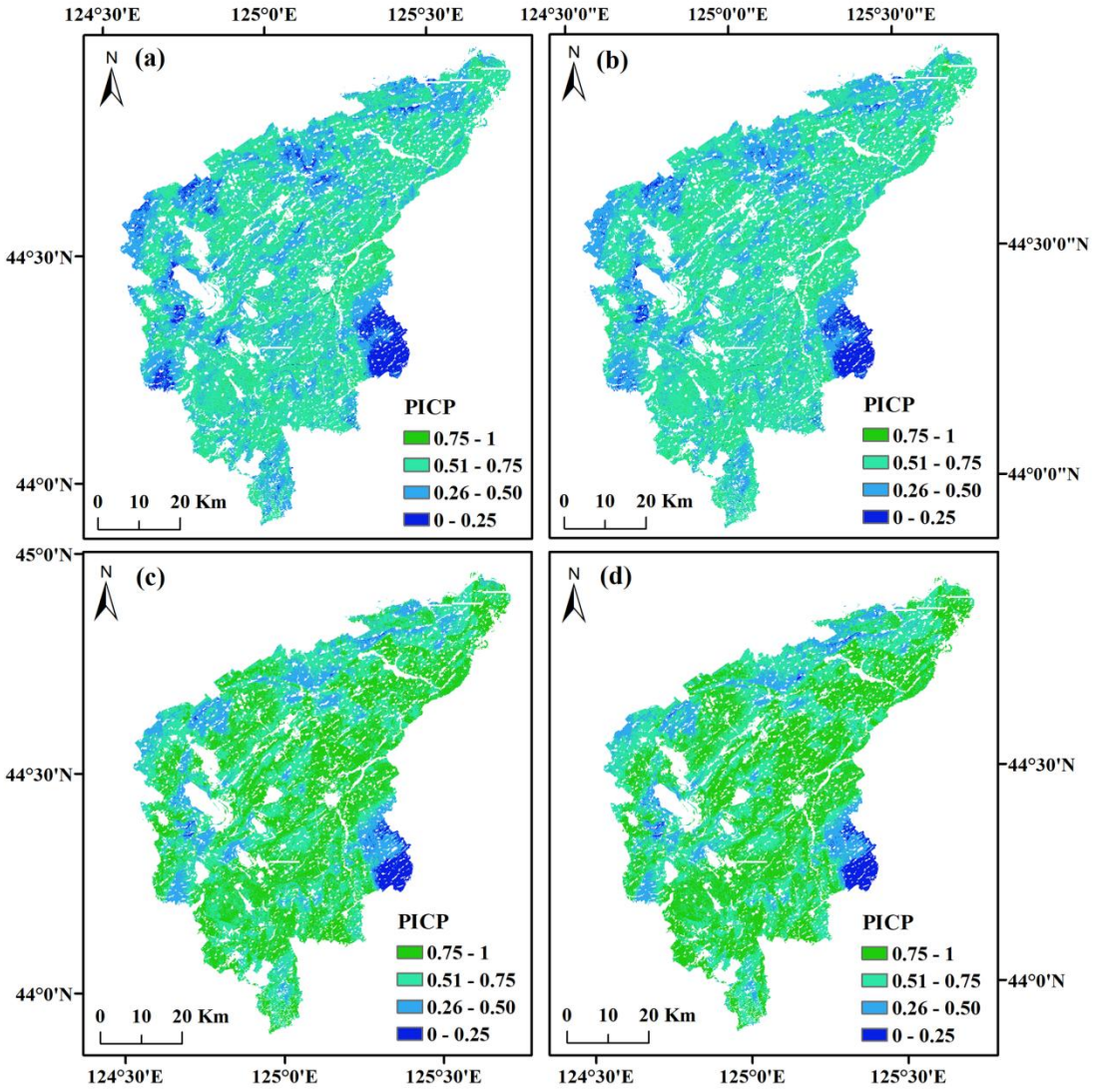

Figure 6. The prediction uncertainty of soil organic carbon expressed by the PICP in the farmland of Nong'an. (a) Single-date image from 2018, (b) multitemporal images from 2018, (c) multitemporal images from 2016–2018, and (d) multitemporal images from 2013–2018 for crop cover conditions in Nong'an County, the Mollisol region of Northeast China.

4. Discussion

4.1. Multi-Temporal Satellite Images for SOC Prediction

Currently, the challenge is to find the shortest time length of the multi-temporal images to capture the features of SOC changes, especially in agricultural soils. Here, the best model performance obtained in the three-year period might be linked to the cycle of input and transformation of carbon, which was just the carbon sequestration period for agricultural soils. The model performance did not improve when the period was longer than three years, possibly because adding more multi-temporal images away from the SOC prediction year cannot provide more related features for SOC sequestration. The results of this study were also consistent with those of [79], who compared one and three years of compositing Sentinel-2 images for SOC predictions in croplands, and the best model performances were obtained for three years or longer. Zepp et al. [30] noted that the model performance over three years (R^2: 0.66, RMSE: 1.33%) when the NBR2 index was used was similar to that over fifteen years (R^2: 0.72, RMSE: 1.16%) when the PV+BLUE combination was used for SOC prediction in topsoil croplands. Moreover, comparable prediction performances were also obtained here from the models from the V+F (R^2:0.77, RMSE:2.12 g kg^{-1}), S+V+F (R^2:0.82, RMSE:2.04 g kg^{-1}), and T+V+F (R^2:0.90, RMSE:1.85 g kg^{-1}) combinations, even for the vegetation cover conditions. However, the study region was covered by snow from early November to late March of the following

 doi:10.20944/preprints202604.2005.v1

year, and the number of cloudless scenes may be limited in some years. Although three years is the shortest period for obtaining the best model performance, a longer period (3–5 years) is advisable for farmland SOC prediction to reduce the uncertainty from multi-temporal images across multiple years.

SOC estimation and mapping is challenging, due to the temporal or permanent crop and crop residue cover. Compared with previous studies that focused on the SOC prediction during bare soil periods only [16,28,60,], this paper proposed a new method for SOC prediction under crop cover conditions by integrating multi-temporal images from crop cover periods across different time lengths. The similar performance was obtained on crop cover conditions compared with that from the bare soil condition by optimising environmental indicator combinations and time periods. The results from our study are reliable, at least at this study site, and are expected to be used to overcome the limitations of short bare soil periods, especially in the farmlands of Northeast China. As Zepp et al. [30] suggested, investigating how long the soil reflectance composited length to capture SOC changes, and the shortest time period for optimizing SOC prediction performance, was essential. It is challenging to quantify and monitor SOC changes over short bare soil periods every year. Time series of satellite images have the potential to capture SOC dynamics through long-term varying soil reflectance composites.

4.2. Contributions of the Indicators

Vegetation/organism conditions, including spectral indices (e.g., MSAVI2 and the SOC index), are the most influential indicators for farmland SOC prediction in the best time periods (2016–2018) of the satellite images (Fig. 5). Here, the results, which presented the close relationships between the SOC and spectral indices, were inconsistent with those of previous studies which also indicated loose relationship between them. Some studies have shown that spectral indices can represent carbon input and are the dominant factor in SOC sequestration, especially in agricultural soils [25,27]. Some studies have shown very inconsistent results for SOC relationships with spectral indices, especially for indices derived from a single-date image [24]. Moreover, spectral indices sometimes cannot accurately represent the temporal variations in vegetation growth, and may work ineffectively in SOC estimation. This can be explained by the relationships between SOC and spectral indices being influenced by various random spatial components, such as soil moisture, surface roughness, and the presence of crops on the soil surface [80]. To best describe the spatial variability of the SOC, the random component needs to be minimized. Similar to previous studies, only long-term changes in carbon input can be detected by multi-temporal spectral indices. This occurred because the presence of confounding random errors in such long-term relationships would be weakened. Accordingly, the correlation of SOC and spectral indices can be captured by multi-temporal RS images via a more reliable approach.

In agricultural soils, SOC release occurs when farmland is subjected to continuous cropping or intensive cultivation practices. Improvements in soil conservation and fertility are considered suitable agricultural practices. We introduced the straw return practice (NBR2 index) to represent soil conservation and three other covariates (chemical fertilizer, livestock production, and grain yield) as soil fertility indicators. Here, farm management was the secondary influential indicator for SOC prediction from 2016–2018 to 2013–2018, accounting for approximately 27.7% of the SOC prediction. We also found that livestock production was employed to represent mature fertilizer. As mature fertilizer is a direct carbon input to farmland, it is the most influential indicator of agricultural practices. In intensively cultivated Mollisols in Northeast China, chemical fertilizer has been widely used for many years, and it has made a significant contribution to crop production[33]. Accordingly, greater crop production increased the amount of straw returned to the farmland and promoted SOC storage. It is necessary to reconsider SOC estimation and mapping in agricultural soils, and the use of agricultural practices has the ability to improve the performance of SOC prediction models.

The quality of SOC prediction is also affected by the explanatory power of variables with scale-specific representativeness [39,59]. Twenty terrain attributes across three spatial scales, including the local, regional, and combined scales, were identified for SOC prediction via the RF data-mining model. In addition to the terrain features at the local scale, our results indicated that introducing

 doi:10.20944/preprints202604.2005.v1

topographic indicators at other scales affected model performance for all the models and at different time periods. The most influential terrain indicators included valley depth, elevation, MRVBF, TWI, and slope height, with contributions of 4.1%-4.8%, 4.0%-4.2%, 2.4%-2.9%, 1.9%-2.2%, and 1.7%-2.1%, respectively, for the prediction models from 2016–2018 to 2013–2018. Because Nong'an, which is located in the Mollisol region of Northeast China, is a rolling and hilly area, the above indicators affect the vertical distribution of precipitation, influencing the decomposition and transformation of soil organic carbon. As Guo et al. [37] reported, elevation, valley depth, and the TWI are the key attributes influencing SOC distribution and should not be ignored in SOC prediction. Except for the elevation indicator, the other terrain variables were all regional-scale indicators. This occurred because coarsening regional sizes by the aggregation of terrain features could loose the overall spatial information and thereby improve soil variable predictions. Notably, most areas of cropland in the study region are located in flat terrains, which somewhat weakens the impact of terrain variables on SOC variation.

The soil properties accounted for 8.4-10.6% of the SOC prediction, with the best model performance occurring in the 2016–2018 and 2013–2018 time periods for crop cover conditions. First, the bare soil period of the Mollisol region of Northeast China is very short every year, and it is covered mainly by snow or crops for most years; second, the collection of soil samples is labor intensive, especially for crop cover conditions. The results indicated that soil organic carbon could be predicted not only during the bare soil period but also during the crop coverage period. Moreover, the explanatory power for the bare soil and crop cover periods was similar.

4.3. Spatial Distribution of Farmland SOC in Nong'an

The predicted SOC maps, including single-date, single-year, 2016–2018, and 2013–2018 maps, presented highly heterogeneous spatial patterns. This occurred because of the complex interactions among various variables, such as the terrain attributes, vegetation, soil, and human activities. The SOC maps from 2016–2018 and 2013–2018 provide more reliable spatial patterns than those from single date and single year, as shown in the four uncertainty maps (Fig. 7a-d). The detailed trends in the digital soil maps from 2016–2018 and 2013–2018 were consistent with those in the other soil maps from single date and single year. Moreover, SOC showed significant spatial variation across different elevations and farm management practices. Generally, high SOC contents occur mainly in areas at low elevations, where precipitation and water combine to improve the transformation of carbon inputs. Similar to the relative importance of the terrain indicators shown in Section 3.4, a variety of indicators (e.g., valley depth, elevation, MRVBF, TWI, and slope height) at different scales contribute to SOC variation. The spatial distribution of SOC maps was also affected by farm management practices. The lower SOC concentrations were mainly distributed in the western regions of the study area, where few residue cover practices were applied. The results of the SOC maps confirmed that terrain features and farm management had an impact on the spatial distribution of the SOC content, a phenomena that was also reported in other soil mapping studies.

4.4. Limitations and Prospects

This study has several limitations, and future research should consider these problems and improve the SOC prediction performance. First, only four indicators of farm management were employed in the SOC prediction models because of the limited data availability. Other agricultural practices that may induce SOC dynamics, such as carbon inputs, tillage practices, and cropping practices [27], could be introduced into the models to improve the input datasets. Specifically, the conservation practices applied widely in the Mollisols of Northeast China, including no tillage, mulching, incorporation of crop residue, and organic farming, play important roles in carbon storage and SOC dynamics. Secondly, the RF algorithm was selected as the prediction model here due to its high predictive power and the flexibility in DSM applications. Other machine learning and deep learning algorithms, such as convolutional neural networks (CNNs) and long short term memory (LSTM), which can capture the dependencies in spatial features and time-series RS images, could also be applied for more accurate SOC prediction in further research. Finally, coupling the knowledge-driven process model with the data-driven machine learning model may advance the long-term SOC

 doi:10.20944/preprints202604.2005.v1

simulation and regional expansion, and obtain more robust soil carbon storage dynamics. It may expand this work in both space and time, and offer a robust solution for comprehensive SOC modelling and its interpretability.

5. Conclusions

Estimation of SOC in farmlands by using multi-temporal satellite images and environmental variables has been widely applied in current studies. However, the consideration of the minimum time periods for establishing SOC estimation models has been largely overlooked in previous research. In this study, a novel approach for estimating SOC in agricultural soils was proposed. The SOC predictions in the farmland of Nong'an by using single-date, single-year, and multi-year Landsat 8 OLI images were investigated to obtain the optimized time periods for the best model performance. Four types of indicators, including soil properties, terrain attributes, vegetation conditions, and farm management practices, were selected to represent SOC changes, and their different combinations for both bare soil and crop cover conditions were compared for SOC sequestration. The overall purpose of this research was to determine the optimized time length and best indicator combination for farmland SOC prediction, especially for crop cover periods. The results indicated that the use of multiple years of multi-temporal images achieved greater accuracy in the estimation of farmland SOC than did the use of single-date and single-year images. We propose that the time period for SOC prediction in agricultural soils should be 3-5 years. Across the four types of indicators for SOC estimation, spectral indices and farm management accounted for significant importance in estimating SOC, and their contributions ranged from 45.6-48.7% and 27.3-27.8% for the best model performance under crop cover conditions, respectively.

Author Contributions: Conceptualization, Y.Z.; methodology, Y.Z.; software, Y.L.; validation, Y.L.; formal analysis, Y.L.; investigation, Y.L.; resources, Y.Z.; data curation, Y.L.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z.; visualization, Y.L.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China [grant number: 42301074]; the Science Research Project of the Education Department of Jilin Province [grant number: JJKH20261426KJ].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, X.; Zhang, Y.; Atkinson, P.M.; Yao, H. Predicting soil organic carbon content in Spain by combining Landsat TM and ALOS PALSAR images. *International Journal of Applied Earth Observation Geoinformation* **2020**, 92, 102182. https://doi.org/10.1016/j.jag.2020.102182.
2. Oldfield, E.E.; Eagle, A.J.; Rubin, R.L.; Rudek, J.; Sanderman, J.; Gordon, D.R. Crediting agricultural soil carbon sequestration. *Science* **2022**, 375(6586), 1222-1225. https://doi.org/10.1126/science.abl799.
3. Zhou, Y.; Hartemink, A.E.; Shi, Z.; Liang, Z.; Lu, Y. Land use and climate change effects on soil organic carbon in North and Northeast China. *Science of the Total Environment* **2019**, 647, 1230-1238. https://doi.org/10.1016/j.scitotenv.2018.08.016.
4. Xia, L.; Cao, L.; Yang, Y.; Ti, C.; Liu, Y.; Smith, P.; van Groenigen, K.J.; Lehmann, J.; Lal, R.; Butterbach-Bahl, K.; Kiese, R.; Zhuang, M.; Lu, X.; Yan, X. Integrated biochar solutions can achieve carbon-neutral staple crop production. *Nature Food* **2023**, 1-11. https://doi.org/10.1016/10.1038/s43016-023-00694-0.
5. Wiesmeier, M.; Urbanski, L.; Hobley, E.; Lang, B.; von Lutzow, M.; Marin-Spiotta, E.; van Wesemael, B.; Rabot, E.; Liess, M.; Garcia-Franco, N.; Wollschlager, U.; Vogel, H.J.; Kogel-Knabner, I. Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. *Geoderma* **2019**, 333, 149-162. https://doi.org/10.1016/j.geoderma.2018.07.026.
6. Wang, J.; Feng, L.; Palmer, P.I.; Liu, Y.; Fang, S.X.; Bösch, H.; O'dell, C.W.; Tang, X.P.; Yang, D.X.; Liu, L.X.; Xia, C.Z. Large Chinese land carbon sink estimated from atmospheric carbon dioxide data. *Nature* **2020**, 586(7831), 720-723. https://doi.org/10.1038/s41586-020-2849-9.

 doi:10.20944/preprints202604.2005.v1

7. Mirchooli, F.; Kiani-Harchegani, M.; Darvishan, A.K.; Falahatkar, S.; Sadeghi, S.H. Spatial distribution dependency of soil organic carbon content to important environmental variables. *Ecological Indicators* **2020**, 116, 106473. https://doi.org/10.1016/j.ecolind.2020.106473.
8. Minasny, B.; and McBratney, A.B. Digital soil mapping: a brief history and some lessons. *Geoderma* **2016**, 264, 301-311. https://doi.org/10.1016/j.geoderma.2015.07.017.
9. Amelung, W.; Bossio, D.; de Vries, W.; Kögel-Knabner, I.; Lehmann, J.; Amundson, R.; Bol, R.; Collins, C.; Lal, R.; Leifeld, J.; Minasny, B.; Pan, G.; Paustian, K.; Rumpel, C.; Sanderman, J.; van Groenigen, J.W.; Mooney, S.; van Wesemael, B.; Wander, M.; Chabbi, A. Towards a global-scale soil climate mitigation strategy. *Nature Communications* **2020**, 11, 5427. https://doi.org/10.1038/s41467-020-18887-7.
10. Biney, J.K.M.; Houška, J.; Volánek, J.; Abebrese, D.K.; Cervenka, J. Examining the influence of bare soil UAV imagery combined with auxiliary datasets to estimate and map soil organic carbon distribution in an erosion-prone agricultural field. *Science of the Total Environment* **2023**, 870, 161973. https://doi.org/10.1016/j.scitotenv.2023.161973.
11. Dvorakova, K.; Heiden, U.; Pepers, K.; Staats, G.; van Os, G.; van Wesemael, B. Improving soil organic carbon predictions from a Sentinel-2 soil composite by assessing surface conditions and uncertainties. *Geoderma* **2023**, 429, 116128. https://doi.org/10.1016/j.geoderma.2022.116128.
12. Lamichhane, S.; Kumar, L.; Wilson, B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma* **2019**, 352, 395-413. https://doi.org/10.1016/j.geoderma.2019.05.031.
13. Wadoux, A. M. J.-C.; and McBratney, A. B. Digital soil science and beyond. *Soil Science Society of America Journal* **2021**, 85(5), 1313-1331. https://doi.org/10.1002/saj2.20296.
14. Adams, B.; Iverson, L.; Matthews, S.; Peters, M.; Prasad, A.; Hix, D.M. Mapping forest composition with Landsat time series: An evaluation of seasonal composites and harmonic regression. *Remote Sensing* **2020**, 12 (4), 610. https://doi.org/10.3390/rs12040610.
15. Vaudour, E.; Gomez, C.; Lagacherie, P.; Loiseau, T.; Baghdadi, N.; Urbina-Salazar, D.; Loubet, B.; Arrouays, D. Temporal mosaicking approaches of Sentinel-2 images for extending topsoil organic carbon content mapping in croplands. International *Journal of Applied Earth Observation and Geoinformation* **2021**, 96, 102277. https://doi.org/10.1016/j.jag.2020.102277.
16. Broeg, T.; Don, A.; Gocht, A.; Scholten, T.; Taghizadeh-Mehrjardi, R.; Erasmi, S. Using local ensemble models and Landsat bare soil composites for large-scale soil organic carbon maps in cropland. *Geoderma* **2024**, 444, 116850. https://doi.org/10.1016/j.geoderma.2024.116850.
17. Luo, C.; Zhang, X.; Meng, X.; Zhu, H.; Ni, C.; Chen, M.; Liu, H. Regional mapping of soil organic matter content using multitemporal synthetic Landsat 8 images in Google Earth Engine. *Catena* **2022**, 209, 105842. https://doi.org/10.1016/j.catena.2021.105842.
18. Wang, N.; Peng, J.; Xue, J.; Zhang, X.; Huan, J.; Biswas, A.; He, Y.; Shi, Z. A framework for determining the total salt content of soil profiles using time-series Sentinel-2 images and a random forest-temporal convolution network. *Geoderma* **2022**, 409, 115656. https://doi.org/10.1016/j.geoderma.2021.115656.
19. Pouladi, N.; Møller, A.B.; Tabatabai, S.; Greve, M.H. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. *Geoderma* **2019**, 342, 85-92. https://doi.org/10.1016/j.geoderma.2019.02.019.
20. Dai, L.; Ge, J.; Wang, L.; Zhang, Q.; Liang, T.; Bolan, N.; Lischeid, G.; Rinklebe, J. Influence of soil properties, topography, and land cover on soil organic carbon and total nitrogen concentration: A case study in Qinghai-Tibet plateau based on random forest regression and structural equation modeling. *Science of the Total Environment* **2022**, 821, 153440. https://doi.org/10.1016/j.scitotenv.2022.153440.
21. Gaber, K.S.; Singla, M.K. Predictive analysis of groundwater resources using random forest regression. *Journal of Artificial Intelligence and Metaheuristics* **2025**, 9(1), 11-19. https://doi.org/10.54216/JAIM.090102.
22. Dou, X.; Wang, X.; Liu, H.; Zhang, X.; Meng, L.; Pan, Y.; Yu, Z.; Cui, Y. Prediction of soil organic matter using multi-temporal satellite images in the Songnen Plain, China. *Geoderma* **2019**, 356, 113896. https://doi.org/10.1016/j.geoderma.2019.113896.

 doi:10.20944/preprints202604.2005.v1

23. Shafizadeh-Moghadam, H.; Minaei, F.; Talebi-khiyavi, H.; Xu, T.; Homaee, M. Synergetic use of multi-temporal Sentinel-1, Sentinel-2, NDVI, and topographic factors for estimating soil organic carbon. *Catena* **2022**, 212, 106077. https://doi.org/10.1016/j.catena.2022.106077.
24. Kumar, N.; Velmurugan, A.; Hamm, N.A.S.; Dadhwal, V.K. Geospatial mapping of soil organic carbon using regression kriging and remote sensing. *Journal of the Indian Society of Remote Sensing* **2018**, 46, 705-716. https://doi.org/10.1007/s12524-017-0738-y.
25. De Almeida Minhoni, R.T.; Scudiero, E.; Zaccaria, D.; Saad, J.C.C. Multitemporal satellite imagery analysis for soil organic carbon assessment in an agricultural farm in southeastern Brazil. *Science of the Total Environment* **2021**, 784, 147216. https://doi.org/10.1016/j.scitotenv.2021.147216.
26. Zhang, Y.; Jiang, Y.; Jia, Z.; Qiang, R.; Gao, Q. Identifying the scale-controlling factors of soil organic carbon in the cropland of Jilin Province, China. *Ecological Indicators* **2022**, 139, 108921. https://doi.org/10.1016/j.ecolind.2022.108921.
27. Wang, Q., No, J.L., Li, Q., Lan, T., Gao, X., Deng, O., Li, Y. Incorporating agricultural practices in digital mapping improves prediction of cropland soil organic carbon content: The case of the Tuojiang River Basin. *Journal of Environmental Management* **2023**, 330, 117203. https://doi.org/10.1016/j.jenvman.2022.117203.
28. Dvorakova, K.; Heiden, U.; van Wesemael, B. Sentinel-2 exposed soil composite for soil organic carbon prediction. *Remote Sensing* **2021**, 13(9), 1791. https://doi.org/10.3390/rs13091791.
29. Tian, X.; Consoli, D.; Witjes, M.; Schneider, F.; Parente, L.; Şahin, M.; Ho, Y.-F.; Minařík, R.; Hengl, T. Time-series of Landsat-based bi-monthly and annual spectral indices for continental Europe for 2000-2022. *Earth System Science Data* **2024**, 17(2), 741-772. https://doi.org/10.5194/essd-2024-266.
30. Zepp, S.; Heiden, U.; Bachmann, M.; Möller, M.; Wiesmeier, M.; van Wesemael, B. Optimized bare soil compositing for soil organic carbon prediction of topsoil croplands in Bavaria using Landsat. *ISPRS Journal of Photogrammetry and Remote Sensing* **2023**, 202, 287-302. https://doi.org/10.1016/j.isprsjprs.2023.06.003.
31. Guo, L.; Sun, X.; Fu, P.; Shi, T.; Dang, L.; Chen, Y.; Linderman, M.; Zhang, G.; Zhang, Y.; Jiang, Q.; Zhang, H.; Zeng, C. Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas. *Geoderma* **2021**, 398, 115118. https://doi.org/10.1016/j.geoderma.2021.115118.
32. Lal, R. Digging deeper: a holistic perspective of factors affecting soil organic carbon sequestration in agroecosystems. *Global Change Biology* **2018**, 24, 3285-3301. https://doi.org/10.1111/gcb.14054.
33. Liang, A.; Zhang, Y.; Chen, X.; Zhang, S.; Huang, D.; Yang, X.; Zhang, X.; Li, X.; Tian, C.; McLaughlin, N.B.; Xiang, Y. Development and effects of conservation tillage in the black soil region of Northeast China. *Scientia Geographica Sinica* **2022**, 42(8), 1325-1335. https://doi.org/ 10.13249/j.cnki.sgs.2022.08.001.
34. Sanderman, J.; Hengl, T.; Fiske, G.J. Soil carbon debt of 12,000 years of human land use. *Proceeding of the National Academy of Sciences of USA* **2017**, 114 (36), 9575-9580. https://doi.org/10.1073/pnas.1706103114.
35. Hobley, E.; Wilson, B.; Wilkie, A.; Gray, J.; Koen, T. Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. *Plant Soil* **2015**, 390 (1-2), 111-127. https://doi.org/10.1007/s11104-015-2380-1.
36. Grimm, R.; and Behrens, T. Uncertainty analysis of sample locations within digital soil mapping approaches. *Geoderma* **2020**, 155(3), 154-163. https://doi.org/10.1016/j.geoderma.2009.05.006.
37. Guo, Z.; Adhikari, K.; Chellasamy, M.; Greve, M. B.; Owens, P. R.; Greve, M. H. Selection of terrain attributes and its scale dependency on soil organic carbon prediction. *Geoderma* **2019**, 340, 303-312. https://doi.org/10.1016/j.geoderma.2019.01.023.
38. Jendoubi, D.; Liniger, H.; Ifejika Speranza, C. Impacts of land use and topography on soil organic carbon in a Mediterranean landscape (north-western Tunisia). *Soil* **2019**, 5(2), 239-251. https://doi.org/10.5194/soil-5-239-2019.
39. Möller, M.; and Volk, M. Effective map scales for soil transport processes and related process domains-Statistical and spatial characterization of their scale-specific inaccuracies. *Geoderma* **2015**, 247-248, 151-160. https://doi.org/10.1016/j.geoderma.2015.02.003.
40. WRB, I.W.G. World reference base for soil resources 2006. A framework for international classification, correlation and communication. *World Soil Resources Reports* **2006**, 103.

 doi:10.20944/preprints202604.2005.v1

41. Conrad, O.; Bechtel, B.; Bock, M.; Dietrich, H.; Fischer, E.; Gerlitz, L.; Wehberg, J.; Wichmann, V.; Böhner, J. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development* **2015**, 8, 1991-2007. https://doi.org/10.5194/gmd-8-1991-2015.
42. Zevenbergen, L.W.; and Thorne, C.R. Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms* **1987**, 12, 47-56. https://doi.org/10.1002/esp.3290120107.
43. Riley, S.j.; DeGloria, S.D.; Elliot, R. A terrain ruggedness that quantifies topo-graphic heterogeneity. *Intermountain Journal of Science* **1999**, 5 (1-4), 23-27.
44. Kiss, R. Determination of drainage network in digital elevation models, utilities and limitations. *Journal of Hungarian Geomathematics* **2004**, 2, 16-29.
45. Gallant, J.C.; and Dowling, T.I. A multiresolution index of valley bottom ffatness formapping depositional areas. *Water Resources Research* **2003**, 39 (12), 1347-1360. https://doi.org/10.1029/2002WR001426.
46. Möller, M.; Volk, M.; Friedrich, K.; Lymburner, L. Placing soil-genesis and transport processes into a landscape context: A multiscale terrain-analysis approach. *Journal of Plant Nutrition and Soil Science* **2008**, 171, 419-430. https://doi.org/10.1002/jpln.200625039.
47. Bock, M.; and Köthe, R. Predicting the Depth of Hydromorphic Soil HaracteristicsInffuenced by Ground Water **2009**.
48. Rodriguez, F.; Maire, E.; Courjault-Radé, P.; Darrozes, J. The Black Top Hat function applied to a DEM: a tool to estimate recent incision in a mountainous wa-tershed (Estibére Watershed, Central Pyrenees). *Geophysical Research Letters* **2002**, 29 (6) (9-1-9-4). https://doi.org/10.1029/2001GL014412.
49. Gallant, J.C.; and Dowling, T.I. A multiresolution index of valley bottom ffatness formapping depositional areas. *Water Resources Research* **2003**, 39 (12), 1347-1360. https://doi.org/10.1029/2002WR001426.
50. Böhner, J.; and Selige, T. Spatial prediction of soil attributes using terrain analysis andclimate regionalisation. In: Böhner, J., McCloy, K.R., Strobl, J. (Eds.), SAGA-Analysesand Modelling Applications. *Göttinger Geographische Abhandlungen* **2006**, 115, 13-28.
51. Rouse, J.W.; Haas, J.R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. Proceedings of the 3rd ERTS Symposium. Washington **1974**, USA.
52. Jiang, Z.; Huete, A.R.; Didan, K., Miura, T. Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment* **2008**, 112 (10), 3833-3845. https://doi.org/10.1016/j.rse.2008.06.006.
53. Qi, J.; Kerr, Y.; Chehbouni, A. External Factor Consideration in Vegetation Index Development. Proceeding of International Symposium on Physical Measurements and Signatures in Remote Sensing **1994**. Val D'Isere, France.
54. Rondeaux, G.; Steven, M.; Baret, F. Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment* **1996**, 55(2), 95-107. https://doi.org/10.1016/0034-4257 (95)00186-7.
55. Thaler, E.A.; Larsen, I.J.; Yu, Q. A new index for remote sensing of soil organic carbon based solely on visible wavelengths. *Soil Science Society of America Journal* **2019**, 83, 1443-1450. https://doi.org/10.2136/sssaj2018.09.0318.
56. Van Deventer, A.P.; Ward, A.D.; Gowda, P.H.; Lyon, J.G. Using Thematic Mapper Data to Identify Contrasting Soil Plains and Tillage Practices. *American Society for Photogrammetry and Remote Sensing* **1997**, 63, 87-93. https://doi.org/10.1117/12.277087.
57. Wang, K.; Zhang, C.; Li, W. Comparison of geographically weighted regression and regression Kriging for estimating the spatial distribution of soil organic matter. *GIScience & Remote Sensing* **2012**, 49(6), 915-932. https://doi.org/10.2747/1548-1603.49.6.915.
58. British Standard Institute (BSI). Determination of particle size distribution in mineral soil material-method by sieving and sedimentation **1998**, BSI 389. Chiswick High Road, London W4 4AL, UK.
59. Behrens, T.; Schmidt, K.; MacMillan, R. A.; Rossel, R. A. V. Multiscale contextual spatial modelling with the Gaussian scale space. *Geoderma* **2018**, 310, 128-137. https://doi.org/https://doi.org/10.1016/j.geoderma.2017.09.015.
60. Sorenson, P.; Shirtliffe, S.; Bedard-Haughn, A. Predictive soil mapping using historic bare soil composite imagery and legacy soil survey data. *Geoderma* **2021**, 401, 115316. https://doi.org/10.1016/j.geoderma.2021.115316.

 doi:10.20944/preprints202604.2005.v1

61. Gholizadeh, A.; Žižala, D.; Saberioon, M.; Borůvka, L. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sensing of Environment* **2018**, 218, 89-103. https://doi.org/10.1016/j.rse.2018.09.015.
62. Zeraatpisheh, M.; Garosi, Y.; Owliaie, H.R.; Ayoubi, S.; Taghizadeh-Mehrjardi, R.; Scholten, T.; Xu, M. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. *Catena* **2022**, 208, 105723. https://doi.org/10.1016/j.catena.2021.105723.
63. Oliver, M.A.; and Webster, R. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems* **1990**, 4(3), 313-332. https://doi.org/10.1080/02693799008941549.
64. Brus, D. J. Sampling for digital soil mapping: A tutorial supported by R scripts. *Geoderma* **2019**, 338, 464-480. https://doi.org/https://doi.org/10.1016/j.geoderma.2018.07.036.
65. Ließ, M. At the interface between domain knowledge and statistical sampling theory: Conditional distribution based sampling for environmental survey (CODIBAS). *Catena* **2020**, 187, 104423. https://doi.org/10.1016/j.catena.2019.104423.
66. Goovaerts, P. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma* **1999**, 89 (1-2), 1-45. https://doi.org/10.1016/S0016-7061(98)00078-0.
67. Cambardella, C.A.; Moorman, T.B.; Novak, J.; Parkin, T.; Karlen, D.; Turco, R.; Konopka, A. Field-scale variability of soil properties in central Iowa soils. *Soil Science Society of America Journal* **1994**, 58 (5), 1501-1511. https://doi.org/10.2136/sssaj1994.03615995005800050033x.
68. Guyon, I.; Weston, J.; Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46(1/3), 389-422. https://doi.org/10.1023/A:1012487302797.
69. Svetnik, V., Liaw, A., Tong, C., Wang, T. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. Springer Berlin Heidelberg **2004**. https://doi.org/10.1007/978-3-540-25966-4_33.
70. Kuhn, M., 2021. Caret: Classification and Regression Training [R package version 6.0-88,]. https://CRAN.R-project.org/package=caret
71. Breiman, L. Random forests. *Machine Learning* **2001**, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324.
72. Friedman, J.H.; and Meulman, J.J. Multiple additive regression trees with application in epidemiology. *Statistics in Medicine* **2003**, 22 (9), 1365-1381. https://doi.org/10.1016/10.1002/sim.1501.
73. Bergstra, J.; and Bengio, Y. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research* **2012**, 13, 281-305.
74. Siqueira, R.G.; Moquedace, C.M.; Fernandes-Filho, E.I.; Schaefer, C.; Francelino, M.R.; Sacramento, I.F.; Michel, R. Modelling and prediction of major soil chemical properties with Random Forest: Machine learning as tool to understand soil-environment relationships in Antarctica. *Catena* **2024**, 235, 107677. https://doi.org/10.1016/j.catena.2023.107677.
75. Niu, L.; Guo, Y.; Li, Y.; Wang, C.; Hu, Q.; Fan, L.; Wang, L.; Yang, N. Degradation of river ecological quality in Tibet plateau with overgrazing: A quantitative assessment using biotic integrity index improved by random forest. *Ecological Indicators* **2021**, 120, 106948. https://doi.org/10.1016/j.ecolind.2020.106948.
76. Angermueller, C.; Pärnamaa, T., Parts, L.; Stegle, O. Deep learning for computational biology. *Molecular Systems Biology* **2016**, 12 (7), 878. https://doi.org/10.15252/msb.20156651.
77. Fathololoumi, S.; Vaezi, A.R.; Alavipanah, S.K.; Ghorbani, A.; Saurette, D.; Biswas, A. Improved digital soil mapping with multitemporal remotely sensed satellite data fusion: a case study in Iran. *Science of the Total Environment* **2020**, 721, 137703. https://doi.org/10.1016/j.scitotenv.2020.137703.
78. Shrestha, D.L.; and Solomatine, D.P. Machine learning approaches for estimation of prediction interval for the model output. *Neural Network* **2006**, 19, 225-235. https://doi.org/10.1016/j.neunet.2006.01.012.
79. Castaldi, F.; Hueni, A.; Chabrillat, S.; Ward, K.; Buttafuoco, G.; Bomans, B.; Vreys, K.; Brell, M.; van Wesemael, B. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS Journal of Photogrammetry and Remote Sensing* 2019, 147, 267-282. https://doi.org/10.1016/j.isprsjprs.2018.11.026.

 doi:10.20944/preprints202604.2005.v1

80. Xu, Y.; Smith, S.E.; Grunwald, S.; Abd-Elrahman, A.; Wani, S.P.; Nair, V.D. Estimating soil total nitrogen in smallholder farm settings using remote sensing spectral indices and regression kriging. *Catena* **2018**, 163, 111-122. https://doi.org/10.1016/j.catena.2017.12.011.

81.

Table S1. Abbreviations in this study.

Full name	Abbreviation
Digital soil mapping	DSM
Soil organic carbon	SOC
Soil reflectance composite	SRC
Remote sensing	RS
Random forest	RF
Quantile regression forest	QRF
Normalized difference vegetation index	NDVI
Normalized burn ratio 2	NBR2
Soil organic carbon index	SOC index
Enhanced vegetation index 2	EVI2
Ratio vegetation index	RVI
Modified soil-adjusted vegetation index 2	MSAVI2
Digital elevation model	DEM
Topographic wetness index	TWI
Recursive feature elimination	RFE
Prediction interval coverage probability	PICP

Table S2. Distribution proportions and sample numbers of different soil types.

Soil type	Proportion (%)	Sample number
Calcic Chernozems	41.6	2,442
Gleyic Phaeozems	19.7	976
Haplic Phaeozems	7.3	548
Calcaric Phaeozems	6.5	393
Gleyic Chernozems	6.4	286
Haplic Arenosols	4.7	243
Gleyic Solonetz	4.1	204
Luvic Chernozems	3	93
Mollic Gleysols	1.9	54
Calcaric Gleysols	1.7	115
Calcic Gleysols	1.5	88
Haplic Chernozems	0.8	17
Mollic Solonchaks	0.5	6
Calcaric Fluvisols	0.1	5

Table S3. Model robustness evaluation across different sample sizes.

 doi:10.20944/preprints202604.2005.v1

Cover condition	Indicator combination	Sample size	R^2 (Mean ± SD)	RMSE (Mean ± SD)
Bare soil	V+F	100%	$0.74 \pm 3.5\times10^{-3}$	$2.09 \pm 4.5\times10^{-3}$
		90%	$0.74 \pm 1.1\times10^{-3}$	$2.63 \pm 6.0\times10^{-3}$
		70%	$0.73 \pm 6.3\times10^{-4}$	$3.56 \pm 2.8\times10^{-3}$
		50%	$0.73 \pm 4.1\times10^{-4}$	$2.62 \pm 5.4\times10^{-3}$
	S+T+F	100%	$0.70 \pm 1.3\times10^{-3}$	$3.07 \pm 3.5\times10^{-3}$
		90%	$0.71 \pm 1.0\times10^{-3}$	$2.14 \pm 4.5\times10^{-3}$
		70%	$0.72 \pm 1.1\times10^{-3}$	$1.97 \pm 5.2\times10^{-3}$
		50%	$0.69 \pm 9.3\times10^{-4}$	$1.96 \pm 4.7\times10^{-3}$
	S+V+F	100%	$0.77 \pm 2.7\times10^{-3}$	$2.05 \pm 2.4\times10^{-3}$
		90%	$0.75 \pm 7.9\times10^{-4}$	$3.59 \pm 4.4\times10^{-3}$
		70%	$0.76 \pm 8.9\times10^{-4}$	$3.51 \pm 5.4\times10^{-3}$
		50%	$0.77 \pm 5.3\times10^{-4}$	$1.55 \pm 4.3\times10^{-3}$
	T+V+F	100%	$0.84 \pm 3.4\times10^{-3}$	$1.97 \pm 5.0\times10^{-4}$
		90%	$0.83 \pm 6.5\times10^{-4}$	$1.62 \pm 4.1\times10^{-4}$
		70%	$0.84 \pm 3.4\times10^{-3}$	$1.55 \pm 3.9\times10^{-4}$
		50%	$0.85 \pm 3.5\times10^{-3}$	$1.59 \pm 4.6\times10^{-4}$
	S+T+V+F	100%	$0.95 \pm 3.4\times10^{-3}$	$1.75 \pm 6.0\times10^{-4}$
		90%	$0.93 \pm 9.5\times10^{-4}$	$1.60 \pm 6.0\times10^{-4}$
		70%	$0.96 \pm 4.8\times10^{-3}$	$1.50 \pm 4.5\times10^{-4}$
		50%	$0.94 \pm 4.3\times10^{-3}$	$1.55 \pm 4.0\times10^{-4}$
Crop cover	V+F	100%	$0.77 \pm 1.3\times10^{-3}$	$2.12 \pm 4.3\times10^{-3}$
		90%	$0.77 \pm 1.2\times10^{-3}$	$1.65 \pm 6.0\times10^{-3}$
		70%	$0.76 \pm 2.43\times10^{-4}$	$1.57 \pm 2.1\times10^{-3}$
		50%	$0.77 \pm 1.2\times10^{-3}$	$1.63 \pm 3.1\times10^{-3}$
	S+T+F	100%	$0.60 \pm 6.0\times10^{-4}$	$3.41 \pm 1.9\times10^{-3}$
		90%	$0.61 \pm 1.0\times10^{-3}$	$4.14 \pm 4.5\times10^{-3}$
		70%	$0.59 \pm 5.4\times10^{-3}$	$5.96 \pm 2.12\times10^{-3}$
		50%	$0.60 \pm 9.0\times10^{-4}$	$3.96 \pm 4.7\times10^{-3}$
	S+V+F	100%	$0.82 \pm 1.0\times10^{-3}$	$2.04 \pm 5.8\times10^{-3}$

 doi:10.20944/preprints202604.2005.v1

	90%	$0.81 \pm 1.0\times10^{-3}$	$1.60 \pm 6.8\times10^{-3}$
	70%	$0.82 \pm 6.5\times10^{-3}$	$2.52 \pm 2.2\times10^{-3}$
	50%	$0.82 \pm 6.5\times10^{-3}$	$1.57 \pm 2.7\times10^{-3}$
T+V+F	100%	$0.90 \pm 7.0\times10^{-4}$	$1.85 \pm 4.8\times10^{-3}$
	90%	$0.92 \pm 8.8\times10^{-4}$	$1.63 \pm 4.2\times10^{-3}$
	70%	$0.90 \pm 6.1\times10^{-4}$	$1.55 \pm 4.7\times10^{-3}$
	50%	$0.89 \pm 7.2\times10^{-4}$	$1.61 \pm 4.1\times10^{-3}$
S+T+V+F	100%	$0.94 \pm 7.0\times10^{-4}$	$1.77 \pm 6.0\times10^{-3}$
	90%	$0.93 \pm 8.5\times10^{-4}$	$1.85 \pm 6.7\times10^{-3}$
	70%	$0.94 \pm 7.1\times10^{-4}$	$1.50 \pm 7.9\times10^{-3}$
	50%	$0.94 \pm 6.9\times10^{-4}$	$1.56 \pm 6.5\times10^{-3}$

Note: The accuracy of the models at 90%, 70%, and 50% sample sizes was obtained through 100 bootstrap resampling iterations (mean ± standard deviation).