Developing a digital mapping of soil organic carbon on a national scale using Sentinel-2 and hybrid models at varying spatial resolutions

Ji, Xiande ; Purushothaman, Balamuralidhar; Prasad, R. Venkatesha; Aravind, P. V.

DOI
10.1016/j.ecolind.2024.112654

Publication date
2024

Document Version
Final published version

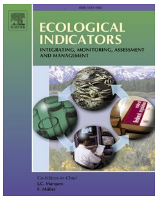
Published in
Ecological Indicators

Ecological Indicators 167 (2024) 112654

Contents lists available at ScienceDirect

Ecological Indicators

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



Original Articles

Developing a digital mapping of soil organic carbon on a national scale using Sentinel-2 and hybrid models at varying spatial resolutions

Xiande Ji[a,*], Balamuralidhar Purushothaman[b], R. Venkatesha Prasad[c], P.V. Aravind[a,d]

[a] *Energy Conversion Group, Energy and Sustainability Research Institute Groningen, Faculty of Science and Engineering, University of Groningen, Nijenborgh 6, 9747AG Groningen, the Netherlands*

[b] *TCS Research, Tata Consultancy Services, Bangalore, India*

[c] *Delft University of Technology, Leeghwaterstraat 39, 2628CB Delft, the Netherlands*

[d] *Department of Process and Energy, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Leeghwaterstraat 39, 2628CB, Delft, the Netherlands*

ARTICLE INFO

Keywords:
Soil organic carbon
Sentinel-2
Hybrid models
Digital soil mapping
Spatial autocorrelation
Germany

ABSTRACT

Mapping the spatial distribution of soil organic carbon (SOC) is crucial for monitoring soil health, understanding ecosystem functions, and contributing to global carbon cycling. However, few studies have directly compared the influence of hybrid models and individual models with varying spatial resolutions on SOC prediction at a national scale. In this study, by combining remote sensing data, we utilized the LUCAS 2018 soil dataset to evaluate the potential capacities of hybrid models for predicting SOC content at different spatial resolutions in Germany. The hybrid models PLSRK and RFK consisted of partial least square regression (PLSR) with residual original kriging (OK) models, and random forest (RF) models with residual OK models, respectively. Individual PLSR and RF models were used as reference models. All these models were applied to estimate SOC content at 10 m, 50 m, 100 m, and 200 m spatial resolutions. Sentinel-2 bands, band indices, and topography variables were as predictors. The results revealed that hybrid models had a more accurate prediction of SOC content with higher explanations and lower prediction errors compared with individual models. The RFK model at the spatial resolution of 100 m was the fittest model with $R^2 = 0.416$, RMSE $= 0.545$, and RPIQ $= 1.647$, which enhanced 3.74% of explanation compared with the performance of RF model. The results also showed that hybrid models at a relatively coarse resolution (100 m) had better accuracy instead of those at high spatial resolution (10 m, 50 m). Sentinel-2 remote sensing data showed significant predictive capabilities for estimating SOC content. The predicted spatial distribution of SOC content revealed that the high SOC concentrated in the northwest grassland, central and southwestern mountains, and the Alps in Germany. Our study provided a benchmark SOC map in Germany for monitoring the changes resulting from land use and climate impacts, and we illustrated the accuracy of hybrid models and the effects of spatial resolutions on SOC predictions at a national scale.

1. Introduction

Soil organic carbon (SOC) is an essential indicator for monitoring soil health and a fundamental component of the global carbon cycle (Stevens et al., 2008). The change of SOC has a significant impact on carbon storage, the ecological environment, and climate change (Crowther et al., 2016). Quantifying SOC content and mapping SOC distribution is a critical procedure for soil management and carbon monitoring. Traditional methods rely on abundant ground surveys and field measurement samples, which are time-consuming, labor-intensive, and destructive to the soil environment (Angelopoulou et al., 2019). Moreover, the traditional measurements are not suitable for effectively monitoring SOC on large scales, such as national scales and the global level. Digital soil mapping is a cost-effective approach for large-scale soil prediction. This approach constructs predictive models for analyzing the quantitative relationship between soil properties and environmental variables to achieve estimations of soil properties (Keskin et al., 2019).

To explore useful factors for estimating SOC content, many studies demonstrated that visible, near-infrared (NIR), and shortwave infrared (SWIR) spectral regions have a significant correlation with soil properties, including SOC. Multispectral bands and their indices derived from remote sensing technology can be used as promising variables to

[*] Corresponding author.
E-mail address: x.ji@rug.nl (X. Ji).

https://doi.org/10.1016/j.ecolind.2024.112654
Received 6 July 2024; Received in revised form 4 September 2024; Accepted 20 September 2024
Available online 28 September 2024

quantify SOC and map SOC distribution on a large scale without contact (Mulder et al., 2011; Ge et al., 2011). With the advanced development of remote sensors and satellite applications for land monitoring, therefore, combining remote sensing imagery with other environmental variables for predicting soil properties has become the primary method for digital soil mapping from regional scales to the global scale. For example, Wang et al. (2020) used Landsat data combined with topography and urban variables to predict SOC stocks in Dalian, China. The Landsat series satellites provide moderate spatial resolution remote sensing data that has been continuously applied to predict SOC and digital soil maps since the first generation satellite launched in the 1970s (Taghizadeh-Mehrjardi et al., 2016; Were et al., 2015; Mirzaee et al., 2016). Currently, the Landsat satellites provide comprehensive coverage of the Earth's surface at a 30-meter resolution approximately every two weeks. These satellite images include both multispectral and thermal data, allowing for a detailed analysis of various aspects of the Earth's surface. Some vegetation indices calculated from the ratio of spectral bands also reflect the growth situation of plants. However, slow revisit time and additional cloud cover affection reduce its application and accuracy for near real-time monitoring. Other public remote sensing data, like MODIS imagery, has been widely used for digital soil mapping on large scales and national scales within its advantage of a shorter repeat cycle (about 1–2 days) and wide-area coverage (Chen et al., 2019; Mishra et al., 2010; Gray et al., 2015), but its low spatial resolution with 250 m to 1 km is a limitation (Mulder et al., 2011).

The recently launched Sentinel-2 satellites by the European Space Agency (ESA) provide more options based on spatial and temporal resolutions for a wide range of monitoring and mapping of land use and land cover. The Sentinel-2 series includes two satellites (Sentinel-2A and Sentinel-2B) that work together in space to reduce the revisit time to 5 days. There are four bands (blue, green, red, NIR) of Sentinel-2 imagery with 10 m spatial resolution, and six bands including red edge 1–4, short-wave infrared (SWIR) 1–2 with 20 m spatial resolution, of which the spectral range matches well with SOC spectral features (Castaldi et al., 2019). With the benefits of high spatial resolution, effective spectral region, and short revisit time, Sentinel-2 data has been widely and successfully used to predict and map SOC distribution (Vaudour et al., 2019; Pham et al., 2021; Castaldi et al., 2019).

The widely used predictive models for SOC prediction and mapping involve geostatistics, traditional multivariate statistics, and machine learning methods (Keskin et al., 2019). Partial least squares regression (PLSR) is one of the important multivariate statistical models because it can reduce the impact on multicollinearity problems and effectively deal with models with plenty of variables (Ge et al., 2011). As many environmental variables are used for prediction, reducing the affection of multicollinearity is useful for getting a reasonable result based on the PLSR model (Vaudour et al., 2019; Dvorakova et al., 2020; Zhang et al., 2021). Recently, rapidly developed machine learning technologies have attracted more interest from researchers in exploring various algorithms for SOC prediction. For example, Zhou et al. (2021) used random forest (RF), boosted regression trees (BRT), and support vector machine (SVM) to analyze and compare the satellite sensors for estimating SOC content on the national scale (Switzerland). Lamichhane et al. (2021) compared the predictive prediction of stepwise multiple linear regression kriging (SMLRK) and random forest (RF) for SOC content prediction and mapping on the region of Nepal, and the results illustrated that the RF model performed better than SMLRK. In addition, Akpa et al. (2016) indicated that RF had better performance for estimating SOC content in Nigeria compared to Cubist and BRT models.

PLSR models and machine learning algorithms belong to feature-based models that can not account for spatial variability among soils (Guo et al., 2015). Soil properties are usually influenced by neighboring soil information. To improve the accuracy of prediction, geostatistics methods, such as ordinary kriging (OK), and simple kriging (SK) are added to model residuals of regression models or machine learning models (Keskin et al., 2019; Triantafilis et al., 2001). Because geostatistics methods are effective tools for explaining the spatial variability of soil properties. A combination with geostatistical models often has better predictions than individual models. For example, Mirzaee et al. (2016) demonstrated that the hybrid methods produced more reliable predictions than geostatistical methods alone based on Landsat 7 ETM + data on the regional scale of Iran. Guo et al. (2015) used a random forest plus residuals kriging (RFRK) approach to predict and map the spatial pattern of soil organic matter for rubber plantation with more accurate results than stepwise linear regression (SLR) based on MODIS 1 km data on a region of Hainan, China. It has been found that there are most hybrid methods being used on the regional scale, while few studies focus on national scales to compare the capacity and performance of hybrid models with individual models (Dai et al., 2014; Gasmi et al., 2022; Guo et al., 2017). Martin et al. (2014) tested the performance of hybrid models and individual models using the French national soil dataset. This study compared the impact of variable quantity on the model results and did not use remote sensing data as predictive variables nor analyze the influence of spatial resolution on the model results.

With the high demand and tendency of attention to national SOC information and distribution patterns, monitoring SOC content, and improving quantitative accuracy is crucial for every country to identify the capacity of soil and make detailed plans for carbon neutrality goals. In Germany, however, there is a lack of studies that exploring the national SOC spatial distribution, only some studies focus on SOC under single land use, such as forest, and cropland in Germany (Wellbrock et al., 2017; Jacobs et al., 2020; Sakhaee et al., 2022). Additionally, there are few references yet that directly compare the influence between hybrid models and individual models with different spatial resolutions for SOC prediction on a national scale. The object of this study aimed to analyze and compare the potential of hybrid models and individual models in predicting SOC content and mapping at different spatial resolutions based on Sentinel-2 data in Germany. The main objectives were (1) to use the PLSR model combined OK (PLSRK) model, RF model combined OK (RFK) model for predicting SOC mapping, and the PLSR and RF models as reference models in comparison of the prediction accuracy with hybrid models; (2) to evaluate the hybrid models' performance at 10 m, 50 m, 100 m, and 200 m different spatial resolution on the national scale; (3) to assess the effect of spatial correlation of soil features for SOC prediction on the national scale; (4) to analyze the relative importance of environmental variables in the different models; (5) to map the SOC pattern of Germany based on the optimal model.

2. Materials and methods

2.1. Study area

Germany is situated in central Europe and has a land area of 315,386 km^2, with a gradually rising altitude from flat plains in the north to medium-altitude mountains in the central region and the Alps in the South (Fig. 1) (Ginzky, 2021). The climate is diverse and mainly influenced by humid westerly winds and the oceanic climate in the western and coastal areas. The eastern regions are affected by the temperate continental climates (Lange et al., 2022; Zink et al., 2017). Due to the climate and diverse topography, at least 12 different soil types can be found in Germany (Ginzky, 2021). Land use in Germany is predominantly by croplands (53%) and forests (31%), with a small portion being covered by scattered grasslands. To discuss the SOC distribution on different land cover, the land cover data of Germany (Fig. 2) was used from the European Space Agency (ESA) WorldCover 10 m 2020 product.

2.2. Soil samples

The soil samples were obtained from the Land Use and Coverage Area Frame Survey (LUCAS). This survey has been conducting regular, consistent standard investigations and analyses for soil properties

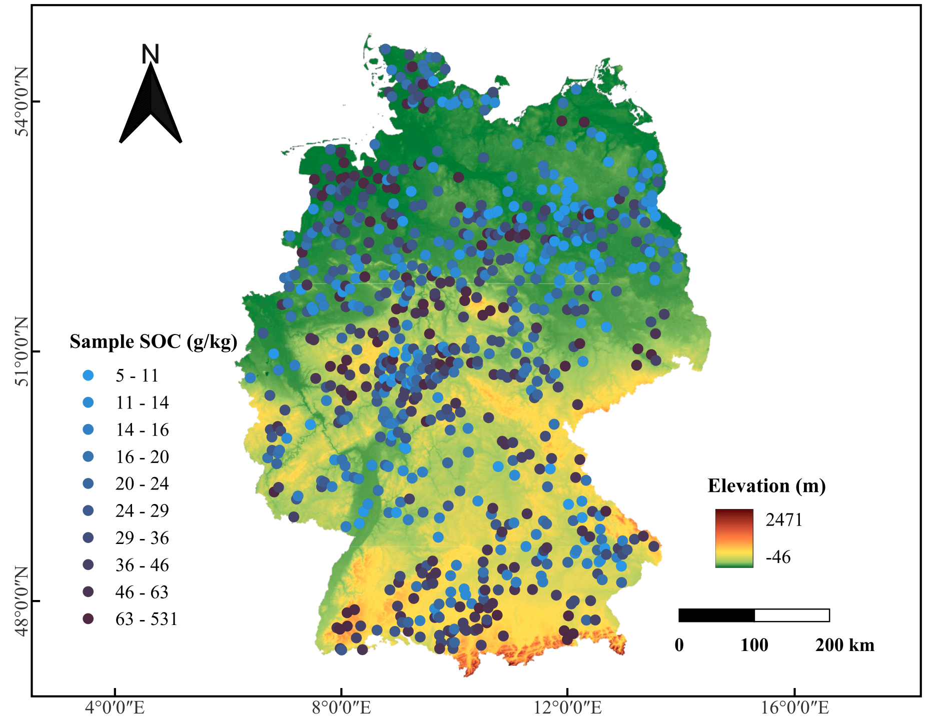


Fig. 1. Soil sampling points located on the digital elevation map in Germany.

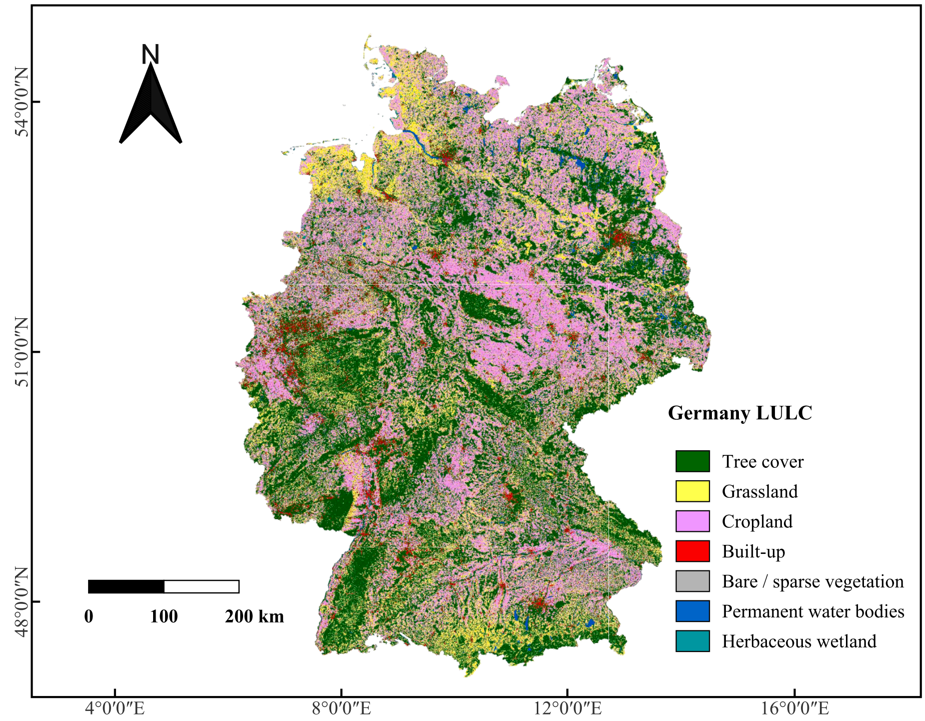


Fig. 2. Land cover types of Germany in 2020.

measurement across European members (28 countries) since 2009 (Ballabio et al., 2019). In 2018, the LUCAS survey selected 762 points in Germany and collected topsoil samples of 20 cm depth for properties analysis, including SOC in a single laboratory (Fig. 1). Each sample was collected at 0.5 kg and mixed from 5 subsamples. The centre subsample was recorded as the coordinate of the location, and the other 4 subsamples were collected in the north, south, east, and west direction from the centre subsample with a distance of 2 m. All LUCAS Samples were air-dried and SOC content was measured using ISO standard methods (Orgiazzi et al., 2018). Additionally, the main land cover class description of each sample was recorded in the LUCAS dataset. 328 samples, accounting for 54.67% of the total number, were mainly located in cropland, followed by grassland (208, 27.30%), and woodland (205, 26.90%). There are small accounts in the bare land, built-up land, shrubland, and wetland. The proportion of soil samples within each land cover type is generally consistent with the proportion of each land cover of areas in Germany. All these samples were randomly divided into two datasets: a training dataset (70%) for training prediction models and a testing dataset (30%) for validating the accuracy of models. Many studies have suggested that the ratio of 70:30 was a widely accepted and reasonable proportion for data splitting (Nguyen et al., 2021; Dobbin and Simon, 2011; Pham et al., 2018). Additionally, due to the moderate size of our dataset, the 30% of data for testing ensured it had sufficient numbers for a reliable assessment of model performance.

2.3. Environmental variables

In this study, Sentinel-2 multispectral bands, band indices, and topographic variables were selected for predicting SOC variation and comparing the ability of models. All these variables (Table 1) were collected and calculated from the Google Earth Engine (GEE). These variables were reprojected into WGS 84/ UTM zone 32 N coordinates and resampled spatially at 10 m, 50 m, 100 m, and 200 m spatial resolutions, separately.

2.3.1. Remote sensing imagery

High spatial resolution multispectral images were acquired from Sentinel-2A and Sentinel-2B, successively launched into space in 2015, and 2017 by the European Space Agency (ESA). The Sentinel-2 captures 13 multispectral bands for monitoring vegetation, soil, and water cover with a 5-day revisit cycle (Segarra et al., 2020). The most effective bands being relevant for SOC predictions are in the ranges of visible (400–700 nm), near-infrared (700–1400 nm), and shortwave infrared (1400–2500 nm) spectrum regions (Castaldi et al., 2019; Pham et al., 2021). The Sentinel-2 bands are located in these scopes and are consistent with the spectral bands required to predict SOC content. This study selected images from May to September, corresponding to the sampling time. Additionally, all chosen images had less than 10% cloud covers and were masked using the Sentinel-2 QA band. The average value of each band (B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12) derived from images as predictor variables for modeling.

Table 1
Environmental variables for prediction of SOC content.

Types	Variables	Definition	Reference
Sentinel-2 MSI	Band 2	Blue; central wavelength: 490 nm	ESA - Sentinel-2
	Band 3	Green; central wavelength: 560 nm	
	Band 4	Red; central wavelength: 665 nm	
	Band 5	Red edge 1; central wavelength: 705 nm	
	Band 6	Red edge 2; central wavelength: 740 nm	
	Band 7	Red edge 3; central wavelength: 783 nm	
	Band 8	NIR; central wavelength: 842 nm	
	Band 8A	Red edge 4; central wavelength: 865 nm	
	Band 11	SWIR 1; central wavelength: 1610 nm	
	Band 12	SWIR 2; central wavelength: 2190 nm	
Band index	NDVI	(Band 8 - Band 4)/(Band 8 + Band 4)	Pham et al. (2021)
	NBR2	(Band 11 - Band 12)/(Band 11 + Band 12)	Dvorakova et al. (2020)
	S2WI	(Band 8 - Band 11 - Band 12)/ (Band 8 + Band 11 + Band 12)	Vaudour et al. (2019)
	NSSI	(Band 8A - Band 7)/(Band 8A + Band 7)	Tian et al. (2021)
Topography	Elevation	Height above the Earth's sea level (m)	Huang et al. (2022)
	Slope	Average degree above flow path	Hu et al. (2021)
	Aspect	The compass direction of the maximum rate of change	Hu et al. (2021)

2.3.2. Band indices

In addition to a single band for retrieving SOC, some combined band indices effectively impact prediction models. The normalized difference vegetation index (NDVI) is traditionally used as a critical indicator to reflect vegetation cover for modeling in digital soil mapping (Xiao et al., 2019). The normalized burn ratio 2 (NBR2) was computed to indicate the difference between around 1600 nm and around 2100 nm, strongly correlated with soil moisture vulnerability to straw and crop residues (Castaldi et al., 2019). Several studies have proven that soil moisture has a negative effect on SOC performance in laboratory experiments, but there is not enough precise field measurement yet (Minasny et al., 2011; Rienzi et al., 2014). Vaudour et al. (2019) showed that using the soil surface moisture index (S2WI) could distinguish moist soil and dry soil in the time series prediction model. Thus, the S2WI was calculated to explore the explanation of soil moisture for modeling. Non-photosynthetic vegetation (NPV) is an essential component in sorts of vegetation-soil ecosystems, and Wang et al. (2018) found that it played a crucial role in predicting soil carbon content in dry graze in Australia. The NPV-soil separation index (NSSI) is a novel spectral index that has been proven effective in separating NPV and bare soil (Tian et al., 2021). However, few studies have used NSSI or other indices to focus on NPV for estimating SOC and digital soil mapping. Therefore, NSSI was applied as one of the indicators to assess the prediction models. The specific calculations of the band index based on Sentinel-2 are in Table 1.

2.3.3. Topographical variables

Topographical variables are crucial factors for influencing the spatial distribution of SOC by controlling precipitation, water accumulation, vegetation cover, and soil erosional process (Wiesmeier et al., 2019). Many studies have illustrated topographic parameters, such as elevation, slope, and aspect were significant variables to explain the variation of SOC (Wang et al., 2018; Zhou et al., 2020). In this study area, elevation data were obtained from the Shuttle Radar Topography Mission (SRTM) V3 product provided by NASA at a spatial resolution of 30 m. Slope and aspect were also important terrain attributes calculated as terrain variables based on elevation data throughout Germany.

2.4. Prediction models

2.4.1. PLSR

PLSR (partial least square regression) is the most frequently multivariate statistical model for predicting soil properties and digital soil mapping (Angelopoulou et al., 2019). This is because of its advantages for constructing prediction models that reduce multicollinearity among variables. There is usually a high correlation between environmental variables, especially among different VIS–NIR spectral bands (Wold et al., 2001). The PLSR algorithm generates orthogonal latent variables in variable space. These latent variables are linearly comprised of prediction variables or response variables to construct new regression relationships between them based on the maximum covariance. The numbers of latent variables have a critical impact on the accuracy and generalization of prediction models. In order to select optimal numbers of latent components, a 10-fold cross-validation method was used to train models based on the numbers of latent variables from 1 to 17 for selecting fitting one with minimum root mean square error (RMSE). The PLSR models were used by the ”pls” package in R.

2.4.2. RF

RF (random forest) method is a tree-based ensemble learning method and has been a popular approach for predicting SOC in recent years (Jia et al., 2021). The review of (Lamichhane et al., 2019) showed that RF had the best performance in 13 papers out of 17 comparative papers that used it as one of the predictive models. Additionally, previous studies have illustrated that RF performed better than other machine learning models, such as artificial neural network (ANN), and support vector regression (SVR) (Jeong et al., 2017; Siewert, 2018). The RF algorithm uses a series of independent regression trees to generate predictions by averaging the values of all trees' prediction results. To avoid overfitting problems, each regression tree is trained using a unique bootstrap sample of training data, making the RF algorithm more stable and accurate compared to decision trees. In the RF algorithm, two main parameters need to be defined in the model processing: the number of trees (n_{tree}) and the number of input variables used to split the nodes at each partitioning (m_{try}). Additionally, RF uses out-of-bag (OOB) samples as inner cross-validation to assess the accuracy of model prediction and estimate the importance of each predictor variable in the models. The OOB mean square error (MSE_{OOB}) is calculated by aggregating the prediction across all trees using Eq. (1):

$$MSE_{OBB} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \widehat{x}_i^{OOB}\right)^2 \tag{1}$$

where n is the number of observations, and $\widehat{x}_i^{OOB}$ is the OOB prediction for observation x_i.

In this study, to select optimal parameters of n_{tree} and m_{try}, the range number of n_{tree} was set from 100 to 1000 at 100 intervals, and the range of m_{try} was set from 1 to 17 as 1 interval. Combined every group of two parameters generated a series of prediction models based on training data. The final model was selected with the smallest prediction error. The training process was employed by the 'randomForest' package in R.

2.4.3. Hybrid models and spatial correlation evaluation

PLSRK (a combination of PLSR and residual ordinary kriging) and RFK (a combination of RF and residual ordinary kriging) are hybrid methods that combine PLSR, and RF with OK, respectively. The OK is regarded as an appropriate and effective geostatistical approach to be used for spatial distribution information about SOC content (Yao et al., 2019). It has been widely applied to interpolate residues from deterministic trend analyses (Gasmi et al., 2022; Lamichhane et al., 2019), and was reported to outperform other geostatistical models such as the SK for residual interpolation (Mirzaee et al., 2016). Thus, in this study, after prediction via PLSR or RF, the residual error of the predicted response variable (here is SOC content) was interpreted to explain the intrinsic spatial variability using OK methods. The residual error was defined as follows:

$$r(x_i) = Z(x_i) - \widehat{Z}_{PLSR}(x_i) \tag{2}$$

$$r(x_i) = Z(x_i) - \widehat{Z}_{RF}(x_i) \tag{3}$$

where $r(x_i)$ is the residual at location $x_i, Z(x_i)$ is the measured value, $\widehat{Z}_{PLSR}(x_i), \widehat{Z}_{RF}(x_i)$ is the predicted value by PLSR and RF, respectively.

Moran's I index and empirical semivariograms are the common methods to explore spatial autocorrelation and structure (Legendre and Fortin, 1989). The semivariograms examine the spatial dependence between data at a lag distance of h. The average variance between any pair of data is calculated as follows:

$$\gamma\left(h\right) = \frac{1}{2n}\sum_{i=1}^{n}\left[r(x_i) - r(x_i + h)\right]^2 \tag{4}$$

where $\gamma(h)$ is the average semivariograms of the SOC content, n is the number of pairs of sample points, and h is the lag distance between pairs of points.

In this study, Moran's I index was used to quantify the spatial autocorrelation levels of residue and measured SOC content based on training data, and empirical variogram models were used to access spatial variability of residue SOC. The parameters derived from empirical variogram models were conducted to build OK models for interpreting residues. The final results of $\widehat{Z}_{PLSR}(x_i)$ and $\widehat{Z}_{RF}(x_i)$ got the sum of the prediction by PLSR or RF and residual interpretation by OK:

$$\widehat{Z}(x_i) = \widehat{Z}_{PLSR}(x_i) + \widehat{Z}_{OK}(x_i) \tag{5}$$

$$\widehat{Z}(x_i) = \widehat{Z}_{RF}(x_i) + \widehat{Z}_{OK}(x_i) \tag{6}$$

where $\widehat{Z}(x_i)$ is the final predicted SOC content, and $\widehat{Z}_{OK}(x_i)$ is the result of interpretation by OK.

2.5. Model evaluation

To evaluate and compare the capability of prediction methods and effects of spatial resolution, PLSR, RF, PLSRK, and RFK models were trained to predict SOC content using all of the predictor variables in the 10 m, 50 m, 100 m, and 200 m spatial resolution based on training data. The testing data was used to assess model performances. Three evaluation indices (1) the coefficient of determination (R^2), the root mean squared error (RMSE), (3) the ratio of performance to interquartile range (RPIQ) were calculated to evaluate the accuracy of the predicted SOC content by testing data. R^2 and RMSE are widely used to measure the degree of linear correlation between predicted and measured values (Wang et al., 2018). Additionally, the RPID is a way to normalize the RMSE of prediction to compare the prediction accuracy between the different models and is preferred over the ratio of prediction to deviation (RPD) (Bellon-Maurel and McBratney, 2011). The definition of these three indices as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(z_i - \widehat{z}_i\right)^2}{\sum_{i=1}^{n}\left(z_i - \overline{z}\right)^2} \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z_i - \widehat{z}_i\right)^2} \tag{8}$$

$$RPIQ = \frac{IQ}{RMSE_P} \tag{9}$$

where n is the number of samples, z_i is the measured SOC value for the sample $i, \widehat{z}_i$ is the predicted SOC value, $\overline{z}$ is the mean value of the measured SOC, IQ is the interquartile range (IQ = Q3 - Q1) of the measured SOC from the testing data. The fittest model would be used to predict the SOC map. The uncertainty of the model was validated by the bootstrap approach. We used ten-time iterations to generate the SOC maps, thereby calculating the standard deviation of each raster cell to present the uncertainty map of SOC (Baltensweiler et al., 2021; Zhou et al., 2023).

3. Results

3.1. Descriptive statistics of SOC

Table 2 presents the basic statistics of SOC content for soil samples in Germany, including training and testing datasets. The SOC content of the training data ranged from 4.80 g/kg to 530.70 g/kg, with an average of 37.49 g/kg, and a median of 23.90 g/kg. The SOC content of the

Table 2
Descriptive statistics of SOC content (g/kg).

Data	Min	Max	Mean	Median	Standard Deviation	Skewness
SOC (Training)	4.80	530.70	37.49	23.90	49.84	5.72
LnSOC (Training)	1.57	6.27	3.27	3.17	0.76	0.78
SOC (Testing)	5.30	471.00	35.26	23.00	47.02	5.52
LnSOC (Testing)	1.67	6.15	3.23	3.13	0.71	0.95

Note: LnSOC: log-transformed SOC.

testing data ranged from 5.30 g/kg to 471.00 g/kg, with an average of 35.26 g/kg, and a median of 23.00 g/kg. The testing data had a similar data distribution as the training data as indicated by their results of mean, median, standard deviation (SD), and skewness. These statistical parameters reflected that both datasets were strongly skewed, with skewness values of 5.72 and 5.52 respectively. Therefore, both SOC content values were transformed by applying the natural logarithm function to remove skewness. After transformation, the median values became similar to the mean values, and the skewness of training and testing datasets was reduced to 0.78 and 0.95, respectively, which were both close to 0.

3.2. Performance of prediction models

The assessments of the performance of each model in predicting SOC content are shown in Table 3. The results showed that the type of models, and spatial resolutions had significant impacts on SOC prediction. In detail, the PLSR model at the 50 m spatial resolution achieved the best performance with $R^2 = 0.361$ on the testing data. The accuracy of the 100 m spatial resolution was slightly less than that of the 50 m resolution, while the worst precision quality was observed at 200 m resolution with $R^2 = 0.277$. RF models exhibited a similar trend to the PLSR models in terms of the spatial resolutions but they performed better than the PLSR models at each corresponding resolution. At the 50 m spatial resolution, the RF model achieved the best performance, with the lowest RMSE and the highest value of the R^2 and RPIQ which were 13.30%, 4.13% higher than those of the PLSR model, respectively. RF models had better accuracy for predicting SOC content on the national scale compared with PLSR models.

Compared to individual models, the interpretation of residues by combined-OK models improved the performance at all four spatial resolutions (Fig. 3). For instance, in the PLSRK model, the R^2 and RPIQ improved by 6.65%, and 1.65%, respectively, and the RMSE decreased by 1.58%, at the 50 m spatial resolution. However, PLSR models achieved the highest accuracy at the 50 m spatial resolution, while PLSRK performed best at the 100 m spatial resolution, with $R^2 = 0.386$, RMSE $= 0.559$, and RPIQ $= 1.606$. The same situation also occurred in the RFK models, where the performance results had the highest accuracy at the 100 m spatial resolution instead of at the 50 m spatial resolution as in the RF model. The comparative analysis of model performance indicated that hybrid models performed better at a relatively coarse spatial resolution (100 m), while the performance was unsatisfactory at high resolutions (10 m, 50 m) and a coarser resolution of 200 m.

Table 3
Accuracy of predicted SOC content produced by using PLSR, RF, PLSRK, and RFK models at 10 m, 50 m, 100 m, and 200 m spatial resolutions.

Model	Resolution	R^2	RMSE	RPIQ
PLSR	10 m	0.337	0.583	1.540
	50 m	**0.361**	**0.574**	**1.575**
	100 m	0.352	0.574	1.563
	200 m	0.277	0.608	1.477
PLSRK	10 m	0.356	0.581	1.545
	50 m	0.385	0.561	1.601
	100 m	**0.386**	**0.559**	**1.606**
	200 m	0.315	0.592	1.518
RF	10 m	0.405	0.551	1.629
	50 m	**0.409**	**0.547**	**1.640**
	100 m	0.401	0.552	1.627
	200 m	0.327	0.587	1.529
RFK	10 m	0.409	0.550	1.632
	50 m	0.412	0.547	1.642
	100 m	**0.416**	**0.545**	**1.647**
	200 m	0.350	0.577	1.557

Note: PLSR: partial least square regression; RF: random forest; PLSRK: partial least square regression plus original kriging; RFK: random forest plus original kriging; R^2: the coefficient of determination; RMSE: the root mean squared error; RPIQ: the ratio of performance to the interquartile range.

Compared to OK models interpreting the residue of RF models, the residual interpretation of PLSR models significantly enhanced SOC prediction at the same spatial resolution in hybrid models (Fig. 3). For instance, when the PLSR model was combined with the OK model to predict SOC, the R^2 of the prediction increased by 9.66%, and the RMSE dropped by 2.61% at the 100 m spatial resolution, while the R^2 of the RFK model only increased by 3.74% and the RMSE decreased by 1.27% compared to the RF model. Although PLSR models combined with OK models improved SOC prediction significantly, the RFK models consistently performed best at all spatial resolutions. When the resolution was set at 100 m, the RFK model was the fittest model to predict SOC content in this study with $R^2 = 0.416$, RMSE $= 0.545$, RPIQ $= 1.647$.

3.3. Spatial autocorrelation

The results of analyzing the spatial autocorrelation and variability of original and residual SOC are exhibited in Table 4, including Moran's I index and variogram function with optimal parameters. Moran's I index measures spatial autocorrelation on a scale from −1 (completely negative spatial autocorrelation) to 1 (completely positive spatial autocorrelation), where a value of 0 indicates complete spatial randomness. The value of Moran's I for original SOC samples was 0.141 which indicated they existed positive spatial autocorrelation, but it was weak. The values for the residues of PLSR and RF models were less than those for SOC, excluding the residue of PLSR on the 10 m spatial resolution. These results illustrated the residue retained spatial variability determined by intrinsic effects, while some complex relationships between residue and extrinsic effects had been explained by PLSR and RF models.

The experimental variogram functions fitted to the residual SOC were optimized using exponential, spherical, and Gaussian models. The ratios of nuggets and sill for these models were less than 0.25, indicating a high dependence on the spatial variability of residual SOC. The large range values were between 16000 m and 24000 m which indicated the scale of occurring spatial autocorrelation is relatively large.

3.4. Importance of environmental variables

The relative importance of environmental variables used in PLSR and RF models for SOC prediction at a 50 m resolution is displayed in Fig. 4. Based on the prediction accuracy of these models, the environmental variables showed the highest influence for accounting for SOC content at the 50 m pixels, but the related importance ranking of environmental variables was different. In the PLSR prediction model, the relative importance of all variables did not reach 10%, their performance is relatively flat concentrating on 4% to 8%. On the contrary, variables had an evident difference in importance in the RF model. Elevation ranked top in the RF model, while it is also ranked third in the PLSR model, which indicated elevation was an important explanatory variable in this study. Among the band indices, NSSI and NDVI were the top two variables in the PLSR model, while NBR2 had significant importance and ranked second in the RF model. However, S2WI did not have a high rank in both models. Regarding the Sentinel-2 bands, B4 and B2 were the most important band variables in the two models, another visible band B3 had a moderate ranking in both models. Other bands had different performances in these two models. Red edge B7 had over 5% relative importance located in the moderate ranking in the PLSR model, while it was the bottom variable in the importance ranking of the RF model. Whatever in the PLSR or RF model, the sum of Sentinel-2 bands was the main domination for explaining SOC content. Band-based indices were remarkable variables. It can be concluded that Sentinel-2 data can be effectively used for SOC estimation with good performance accuracy within the PLSR and RF models.

3.5. Spatial distribution of SOC

Fig. 5 and Fig. 6 show the spatial distribution of SOC content in

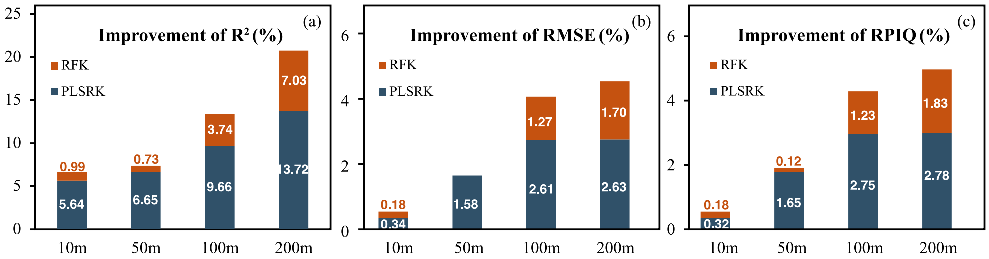


Fig. 3. Improvement for predicting SOC using PLSRK and RFK compared to PLSR and RF at four spatial resolutions.

Table 4
The analysis of Moran's I index and experimental variogram for original SOC and residuals predicted by PLSR and RF models.

Model	Resolution	Moran's I	Model type	Nugget	Sill	Nugget/Sill	Range	R^2
Original		0.141	Exponential	0.035	0.476	0.074	24600	0.774
PLSR	10	0.143	Spherical	0.093	0.302	0.308	23100	0.805
	50	0.137	Gaussian	0.039	0.321	0.121	17667	0.791
	100	0.111	Gaussian	0.043	0.327	0.131	17840	0.738
	200	0.109	Gaussian	0.042	0.331	0.127	17840	0.913
RF	10	0.114	Gaussian	0.007	0.044	0.159	16628	0.795
	50	0.097	Exponential	0.005	0.048	0.104	22200	0.803
	100	0.091	Spherical	0.001	0.052	0.019	19300	0.671
	200	0.110	Exponential	0.006	0.055	0.109	23400	0.906

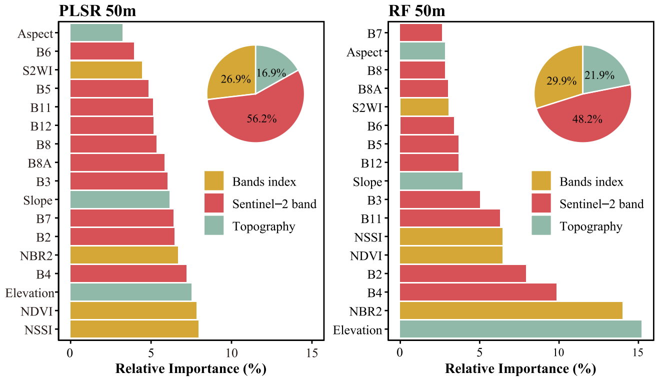


Fig. 4. Relative importance of the environmental variables used for SOC content prediction in the PLSR and RF models at 50 m spatial resolution. B2 to B12 correspond to band 2 to band 12 of Sentinel-2 data; NVDI, normalized difference vegetation index; NBR2, normalized burn ratio 2; S2WI, soil moisture index; NSSI, NPV-soil separation index.

Germany based on the RFK model and RF model at a 100 m spatial resolution. The SOC content of the spatial distribution was classified into ten levels, with the majority of areas exhibiting SOC content ranging from 8–47 g/kg. The SOC distribution predicted by the RFK model is generally similar to that predicted by the RF model. Some regions with abundant SOC content were located in the plains of northwestern, in the central and southwestern mountains, and in the Alps region of southern Germany. These two predictive maps both showed the general spatial distribution of SOC. However, some details were different. For instance, it was found that the SOC content produced by the RFK model is higher than that produced by the RF model in the central area expanded plots in Fig. 5, and 6. According to the prediction of the RFK model, approximately 9.49% of the pixels had SOC content over 50 g/kg, with 0.07% of the pixels exhibiting over 100 g/kg across Germany. This proportion was larger than that predicted by the RF model (8.83%, 0.05%). The Fig. 7 presented the fittest RFK model with a low uncertainty map. In most areas, this model showed a steady prediction of SOC content. Only the northwestern areas that had high SOC concentration presented the high uncertainty of this model.

4. Discussion

4.1. Comparisons of model performance in SOC prediction

In this study, it was found that model types and spatial resolutions significantly influence the prediction accuracy of SOC content at the national scale. Although the PLSR model is a popular regression method and widely used for soil properties on the regional scale, it did not perform as well as the RF models at each spatial resolution in Germany. In previous studies, PLSR models were rarely used to predict SOC at the national scale. On the contrary, advanced machine learning methods were widely used for soil prediction on the national scale (Zhou et al., 2021; Wang et al., 2020; Odebiri et al., 2021). In previous comparison studies, machine learning algorithms had better performance than PLSR models (Laamrani et al., 2019; Heil et al., 2022). These reasons might be that PLSR models cannot effectively deal with the complicated and nonlinear relationships between SOC and environmental variables. Therefore, machine learning methods are the preferred selection on the national scale for retrieving SOC compared to the PLSR model. In this study, the RF model had satisfactory prediction results. This is supported by the result of Wang et al. (2020), which also used the RF model to predict SOC content in Mainland Spain with a reliable prediction based on the LUCAS database. However, for SOC prediction in Switzerland, Zhou et al. (2021) found that boosted regression trees (BRT) models performed better than RF models, while Paul et al. (2020) showed that the RF model overperformed the BRT model in all accuracy measures for digital SOC mapping in the agricultural land on the regional scale. These differences were likely due to the influence of various environmental situations, selected remote sensors, and study scales.

Compared with the individual RF model, the hybrid model RFK added the residual interpretation improved the prediction performance whatever the resolutions. Similar results were observed for PLSRK models. To further analyze the spatial autocorrelation at the national scale, the calculation results of Moran's I index revealed that residual errors persisted after predicting with RF and PLSR models and that these errors still exhibited spatial characteristics similar to those of the original SOC content. Thus, neither PLSR nor RF models explained the inner spatial variability or spatial dependence. Even if RF performed well in accounting for SOC, hybrid combination RFK had better performance. Although the spatial autocorrelation was not strong among soil content

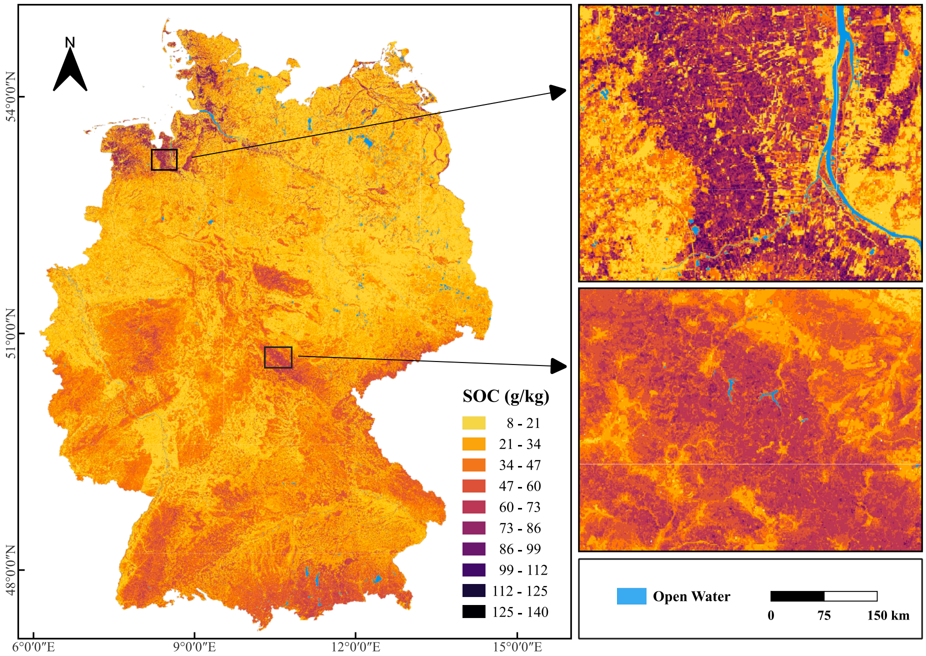


Fig. 5. The spatial distribution of SOC content in Germany predicted by RFK at 100 m resolution.

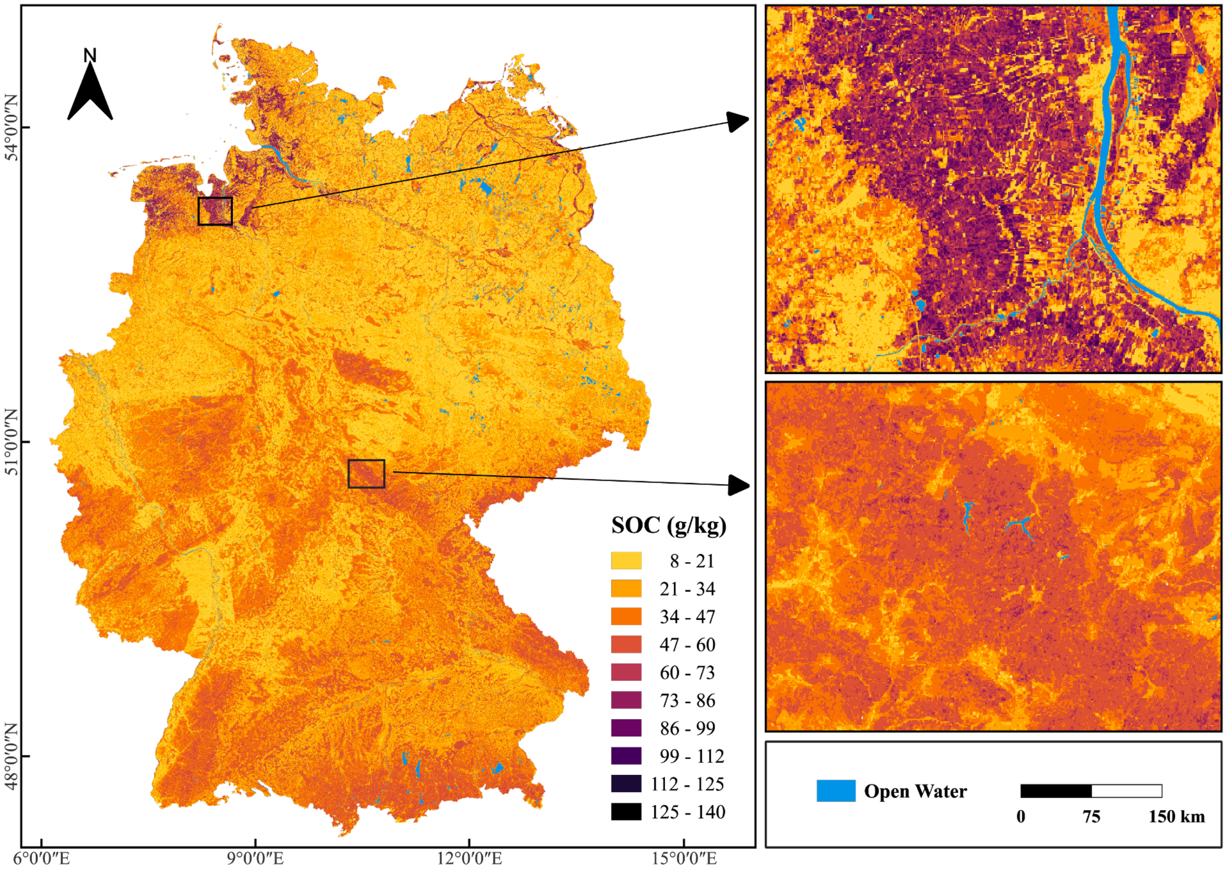


Fig. 6. The spatial distribution of SOC content in Germany predicted by RF at 100 m resolution.

at the national scale, it still improved the performance of the prediction when combined with geostatistical methods. A similar combination model was used by Gasmi et al. (2022), who found when comparing the RF model to the interpolation of RF residuals by the OK method, the predictive results were higher than RF models. Similarly, Guo et al. (2015) illustrated that the RFK model could map the SOM spatial distribution for a rubber plantation region and account for unexplained spatial information in the RF model residuals. The difference is these studies concentrated on regional areas and occurred spatial autocorrelation at a mediate level while the global spatial autocorrelation was not

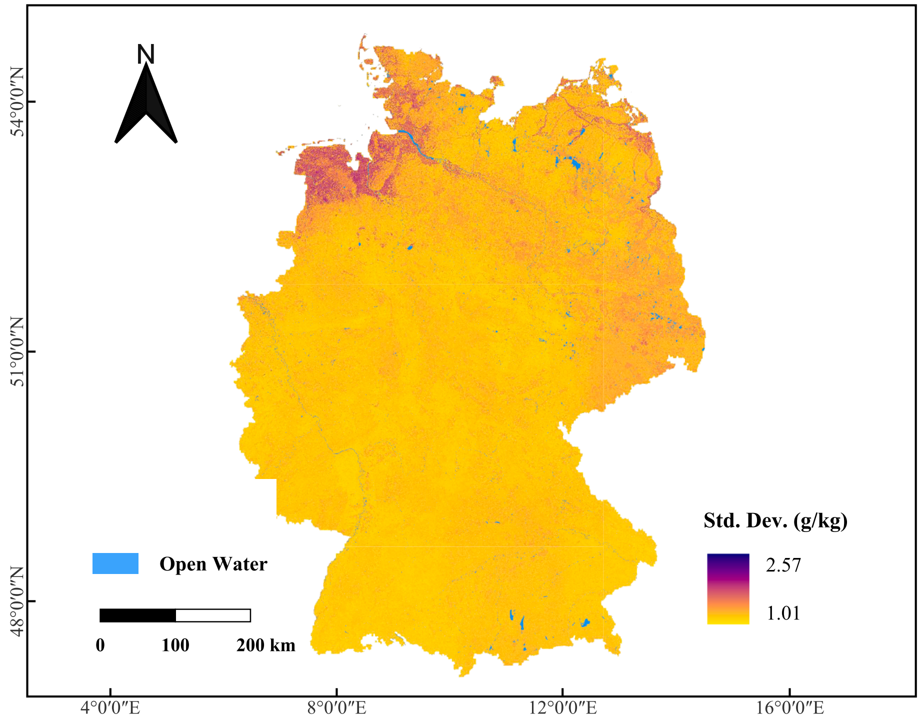


Fig. 7. The standard deviation map of SOC content in Germany by the RFK model at 100 m resolution.

strong among soil content at the national scale in our study (Moran's I = 0.141). At a large scale, SOC spatial autocorrelation is influenced by a broader range of factors, and some of these factors are important for SOC content and interactively affect its distribution. However, these factors may exhibit different spatial variability patterns on the local area of a large scale, and their variability weakens the overall autocorrelation of SOC on large scales (Liu et al., 2012; Gorai et al., 2013). For example, elevation was the most critical factor for predicting SOC in RF models, and its heterogeneity from northern low plain to southern rugged mountains combined with vegetation indices can disrupt the continuity of SOC. On the local scale, these environmental factors exhibit continuous patterns. As a result, partial strong autocorrelation may exist, demonstrated according to the value of the nugget/still in variogram functions. In addition, the spatial autocorrelation may also be influenced by the sampling density and distribution (Oliver and Webster, 2014). In our study, the sampling points are spaced too far apart and unevenly distributed, capturing the fine-scale spatial autocorrelation may be challenging. Thus, most spatial analyses of soil properties focus on local or regional scales, and the studies on large scales exit the challenge of the impact of environmental heterogeneity and sampling distributions.

On the national scale, Martin et al. (2014) used the BRT model and the hybrid model to predict SOC content in France and compared the impact of variables on the performance of models. Their study showed that the performance of hybrid models was influenced by the number of environmental variables. The predictive performance of the BRT model was not significantly improved by geostatistical interpolation when numerous environmental variables were applied to the models. However, when the environmental predictors were limited in the prediction models, the hybrid model was a significantly effective method to improve prediction performance. This difference might be various covariates sufficiently explain deterministic trends of featured space for SOC compared to limited covariates. So the errors of residues were not likely to show significant spatial variability and hybrid models did not improve the accuracy of SOC prediction. In this study, the selected covariates are limited to only involving Sentinel-2 bands, bands-based indices, and topographical factors. On the one hand, variables with high spatial resolution were limited. McBratney et al. (2003) proposed a fundamental framework of environmental variables for soil properties indicating that climate and parent materials were influential to SOC estimations. However, it was difficult to acquire high spatial resolution resources of these factors. For example, climate factors like temperature and precipitation are common with 1 km resolution. Resampling these data to high resolutions (10 m) might produce much noise and errors decreasing the performance of models. On the other hand, although it is possible to obtain numerous variables from remote sensing images, it would generate redundancy, model overfitting, and covariates collinearity issues (Rahmani et al., 2022). For example, Suleymanov et al. (2021) calculated 17 relief variables from the digital elevation model to map SOC, and there are only three variables with the most importance for modeling. In most cases, it is a challenge to include all significant variables or to easily acquire them. Therefore, with a limited number of variables, our study demonstrated that, at the national scale, employing a hybrid model is a promising method to enhance the accuracy of SOC prediction. Kühn and and Dormann (2012) suggested considering the missed residual spatial process to improve the mapping accuracy. Thus, the hybrid model, which integrated feature-based models with geostatistical methods, effectively captured the complex relationships of SOC and spatial variability across diverse land covers. It could be precious for countries with limited legacy inventories, as it allowed for more accurate national-scale SOC mapping by residual analysis, which offered a reliable reference for them seeking soil strategies on a national scale.

Additionally, in our study, we found that the performance of hybrid models was affected by spatial resolutions on the national scale. On the one hand, the PLSRK and RFK models performed well at 100 m instead of 10 m spatial resolution. The high spatial resolution (10 m) did not have the best performance in all of the models. Similarly, Zhou et al. (2021) predicted SOC content and C: N ratio with multi-spatial scales (20 m, 100 m, 400 m, and 800 m) in Switzerland, and the best performance was also found at the 100 m spatial resolution. In other soil properties mapping studies, Chi et al. (2019) predicted total nitrogen using the PLSR model achieved the best accuracy at 100 m spatial resolution compared with 200 m, 400 m, and 800 m scales. Correspondingly, Xu et al. (2017) suggested that high spatial resolutions(<10 m)-

based soil prediction models did not have better performance than coarse spatial resolution (30 m) models in the exchange potassium prediction. The spatial resolutions have similar impacts on hybrid models.

On the other hand, hybrid models with the best performance were not consistent with individual models in terms of spatial resolution. The interpolation of OK in the residue was affected by spatial resolutions of environmental indices as well. Thus, the performance of hybrid models was affected by the integrated effection of featured space models plus residual interpretation on spatial scales. In this study, we found individual models whether PLSR or RF, had the highest accuracy at 50 m spatial resolution. At the fine resolution, more detailed information can be captured, which provides higher accuracy and lower uncertainty in deterministic trend analyses (Chi et al., 2019). In this situation, the interpretation of OK models might be limited, in particular RF individual models. The residual OK model only improved R^2 by 0.73%, while the RMSE did not improve at 50 m resolution (Fig. 3). However, at the 100 m resolution, despite RF or PLSR models performing worse than those at the 50 m resolution, the residual OK model improved the integrated performance of hybrid models at a coarser resolution. In the coarse scale, certain details are smoothed out and averaged, resulting in reduced spatial variability within the local area. Thus, at the 200 m resolution, OK models were more influential to residues than those at the 100 m resolution. However, due to the bad performance of RF and PLSR models at 200 m resolution, the integrated performance did not exceed models at 100 m resolution. It was found that when the spatial resolution was 200 m, all models performed the worst compared to other scales in our studies. Hybrid models are sensitive to the scales of spatial resolutions of environmental variables. It plays an essential role in soil prediction and mapping, the best accurate prediction of hybrid models might happen at a related coarser resolution than that of individual models. These results might be interpreted cautiously because only OK models were considered for interpolating the spatial distribution of residual SOC in our study. Although many studies showed OK combining with various machine learning models had better performance than individual models, there is a lack of comparisons in using different geostatistical models for hybrid model combinations. A further study focusing on selecting various geostatistical models, such as Inverse distance weighting (IDW), should be done to investigate the accuracy of hybrid models..

4.2. Remote sensing variables controlling SOC prediction

Previous studies have demonstrated the crucial role of remote sensing data in digital soil mapping (Xiao et al., 2019; Wang et al., 2018; Vaudour et al., 2022). In this study, we investigated the contribution of Sentinel-2 variables as environmental predictors for predicting SOC content and creating a digital mapping of SOC in Germany. Our results showed that Sentinel-2 bands accounted for 56.2%, and 48.2% of the relative importance of all variables in the PLSR and RF models, respectively, at a spatial resolution of 50 m. Similar studies have also reported the importance of Sentinel-2 variables in explaining soil properties (Castaldi et al., 2019; Vaudour et al., 2019; Gholizadeh et al., 2018). The relative importance of bands varied across different models, however, the visible bands (B2, B3, B4) at 490, 560, and 659 nm consistently performed well in explaining SOC features, which is consistent with previous studies (Nocita et al., 2015; Wang et al., 2022). This is because high levels of organic carbon generally lead to lower reflectance in these spectral bands, as organic materials tend to absorb more light (Nocita et al., 2014; Stuart, 2004). Soils with higher organic carbon content often appear darker, particularly in the red band, which is sensitive to color changes induced by organic matter (Ladoni et al., 2010). Castaldi et al. (2019), in the northern region of Germany, found that Sentinel-2 performed similarly to airborne hyperspectral data in predicting SOC content. The spatial resolution and spectral characteristics of Sentinel-2 were adequate for mapping SOC variability at the regional and even large-scale distribution. Compared to Landsat or MODIS data, Sentinel-2 has four extra red-edge bands (B5, B6, B7, B8A), which are sensitive to SOC. However, these bands were ranked lower in the importance lists, particularly in terms of the RF models. This was different from previous studies (Xie et al., 2022; Guo et al., 2021). The performance of the red-edge band and red-edge information should be explored for future studies. Xie et al. (2022) showed their designed new spectral indices, where B5, B6, and B7 replaced B4, and B8A replaced B8, involved in the models had better prediction for SOC content.

In addition to the multispectral data, the band indices performed significantly in the prediction models. The combinations of multis-bands represented the impact of different environmental factors on SOC prediction. Among all band indices, NSSI had the highest relative importance in the PLSR model and a moderately important role in the RF model. Unlike most indices that rely on measuring green vegetation, such as NDVI, NSSI effectively separates non-photosynthetic vegetation and bare soil from the 750 nm-900 nm spectral range (Jia et al., 2021). It reflects the surface cover situation of non-photosynthetic vegetation, which influenced carbon cycling and soil erosion and had still not been appropriately learned for predicting SOC content. However, this study used NSSI as a prediction variable and verified it to be an essential predictor for estimating SOC variability. Non-photosynthetic vegetation should be considered a necessary element for predicting soil properties, particularly in regions where it is prevalent, such as drylands or post-harvest agricultural fields. By integrating NSSI, the prediction models can capture a more complete picture of the surface cover, improving SOC estimates' accuracy. NDVI also made similar contributions to both prediction models. It is often used to assess vegetation health and density, which is strongly correlated with SOC. Areas with high-density and healthy plant cover can contribute to high organic matter in the soil through leaf litter, root decay, and other processes (Zhang et al., 2019). As NDVI captures variations in vegetation, it indirectly reflects variations in SOC content. This result was consistent with previous studies (Taghizadeh-Mehrjardi et al., 2021; Jeong et al., 2017). Particularly, in tropical regions or countries where vegetation dynamics significantly influence SOC, using vegetation indices like NDVI could provide critical insights into its predictions. Among other band indices, NBR2 had a very high ranking in the importance of variables in the RF model with moderate importance in the PLSR model. It reflected the influence of crop residues and soil moisture. Crop residues could prevent weed growing to improve soil aeration and enhance SOC (Berger et al., 2021). Sentinel-2-derived data effectively represent biophysical properties related to vegetation cover, and residues reducing the time-consuming of field measurements. However, S2WI as an indicator calculated from Sentinel-2 data for analyzing the influence of soil moisture did not play a key role in exploiting soil moisture with features of SOC variability. Regarding the limitation of optical images, acquiring accurate information on soil moisture relies on land surface spectral data, which is a challenge. To fill these gaps, active SAR sensors, such as Sentinel-1, were proven to be sensitive to the variation in soil moisture (Bauer-Marschallinger et al., 2019; Paloscia et al., 2013). In addition, previous studies have illustrated that combining SAR data with multispectral data improved the accuracy of SOC prediction (Hamzehpour et al., 2019; Shafizadeh-Moghadam et al., 2022), particularly in dense vegetation areas (Yang and Guo, 2019). It is a promising approach to use multi-remote sensing data fusion for future studies' digital mapping of SOC.

4.3. Spatial distribution of SOC content in Germany

In this study, the fittest digital SOC map of Germany was predicted by the RFK model at 100 m spatial resolution. The map generally revealed the spatial distribution and variability of SOC content in the topsoil. The range of SOC content predicted was from 7.57 g/kg to 137.68 g/kg. The maximum value is far less than that in LUCAS soil samples. The evaluation of prediction results did not perform well at the scale of high SOC value, and many of the samples that have high soil organic carbon are underestimated (Figure S1). Furthermore, in the northwest region where soil organic carbon was very high, the uncertainty map (Fig. 7) revealed that the uncertainty in the predictions was also stronger. The first limitation is related to imbalanced data distribution. The variability of SOC content of Germany in inventory was significant, ranging from 4.80 g/kg to 530.70 g/kg, while only 2.62% of all soil samples contained more than 150 g/kg SOC. Due to the limited number of soil samples, only 762, there are not enough samples with high SOC values to train predictive models. The models may be inclined to learn patterns from low SOC samples, leading to underestimating samples and high uncertainty with high SOC values. Supplying extra databases should be considered to increase and balance the number of samples in further studies. Sakhaee et al. (2022) used soil datasets from German agricultural soil inventory to supply LUCAS 2013. Then, they divided German soil samples into organic soils and mineral soils in agricultural land for SOC prediction due to the highly variable range of SOC content, which improved the performance compared to the original data. The capacity of predictive models probably causes the second limitation. Because these models might not be sufficient to capture highly complex or nonlinear interactions effectively, leading to less accurate predictions. Although the RF has successfully demonstrated its excellent performance for SOC prediction in many studies, its algorithms usually exist in incommensurable results across complicated and various ecological systems, limiting its general effection and application in different situations (Simon et al., 2023). Recently, some deep learning methods, such as deep neural networks (DNN) have been used to estimate SOC and found with competitive abilities to estimate SOC (Odebiri et al., 2022; Zhong et al., 2021). Nonetheless, the performance of each model should be discussed and compared according to specific environmental situations.

To further discuss the SOC distribution in different land types, the SOC content on the main land cover was described, separately (Table 5, Fig. 8). The high concentration of SOC content was found in the tree cover lands, mainly located in the central and southwestern areas with mediate to high altitude and Alps mountain zones. High-altitude mountain areas in Germany, usually with high averaged SOC content, showed that topography was a crucial environmental variable for SOC estimation. Moreover, the high tree cover SOC reserves were attributed to the vegetation biomass, forest transformation, parent materials, and climate (Akpa et al., 2016; Wiesmeier et al., 2013). The German National Forest Inventory (NFI) 2012 reported an increase in the growth of broadleaved trees in German forests compared to NFI 2002. This increase is partly due to the regeneration of coniferous forests with broadleaved species, resulting in multilayer stands. These transformations may influence carbon pools by potentially translocating soil organic carbon (SOC) from the organic layer into the mineral soil (Grüneberg et al., 2014). Additionally, Wellbrock et al. (2017) showed that broadleaved forests and limed plots store more carbon. This is attributed to the lower C:N ratio in broadleaved forests, which enhances the humus layer, while liming or elevated pH levels may lead to carbon leaching from the humus layer into the upper mineral soil. Similarly, Grüneberg et al. (2014) mentioned soil carbon sequestration was affected by the interactions of tree species in conjunction with specific parent material. Their study indicated that calcareous soils stored higher amounts of carbon than noncalcareous soils. The evident result was found in the Alps, where calcareous substrates were predominant and had high SOC concentrations. Similar studies were reported by (Vesterdal et al., 2008). Besides that, human activities and forest management, such as thinning, timber harvest, and drainage, also play important roles in carbon dynamics in forest ecosystems (Nave et al., 2010). For example, different harvest treatments will likely lead to significant long-term differences in ecosystem carbon during later stages of stand development (Johnson et al., 2002). Additionally, forest lands are under threat because of the increasing interest in wood products, resulting in the carbon sink rate of forest stands with the potential reduction in the coming years (Wellbrock et al., 2017). To avoid the reduction of SOC and increase the economic benefits of Germany's forests, the government should consider specific management and ecological conditions to create more site-adapted orientated forests. German grassland stored plenty of organic carbon underground in soils. The northwestern low plains of Germany, where grassland was the predominant land use, had the highest concentration of SOC. The consistent result was reported in the first comprehensive inventory of Germany (Poeplau et al., 2020). One of the reasons for this high SOC distribution was the concentration of organic soils in the north and peat soils in the northwest areas (Roßkopf et al., 2015). There were also some small areas with organic soils in the moraine landscape and the foothills of the Alps (Sakhaee et al., 2022). Similarly, results of high SOC distribution in the Alps were reported by Zhou et al. (2021), who explained the high contribution of SOC in the Alps in Switzerland due to the abundant plant litter under the forest cover and cold environment. Additionally, grasslands differ from other ecosystems by storing the bulk of their sequestered carbon underground in the root zone, which provides greater resilience against natural disturbances like wildfires (Odebiri et al., 2022). Nonetheless, grassland soil carbon has been shown to be vulnerable to management intensity (van Wesemael et al., 2010; Smith et al., 2007). Different grassland management practices have varying impacts on soil carbon accumulation. On the one hand, in extensive grassland management, such as management with minimal human intervention, the accumulation of SOC tended to decrease over time as plant productivity declined due to nutrient limitations (Allard et al., 2007). On the other hand, intensive management practices, such as the heavy application of fertilizers, could harm SOC content, leading to its decline (Soussana et al., 2004). Conversely, applying fertilizers in modest amounts under intermediate grazing management has been shown to enhance the accumulation of SOC (Leifeld et al., 2011). At present, Lange et al. (2022) reported Alpine foothills had a high grazing intensity and a high share of fertilized grasslands, and northwestern areas had a high grazing intensity in Germany. To avoid overgrazing resulting in soil carbon loss, grassland management should balance the intensity of land use with soil health for sustainable development. Most croplands had low SOC contents under 20 g/kg than tree covers and grasslands. Because native ecosystems such as unmanaged forests and grasslands usually conserve much higher SOC stocks than croplands (Paustian et al., 2016). In Germany, agricultural soils are threatened by contamination by fertilizers, erosion, and loss of SOC (Ginzky, 2021). Carbon loss and input rely greatly on land use and management practices (Poeplau et al., 2011; Lal, 2004). Traditional deep tillage disrupts soil structure, increasing the exposure of SOC and accelerating its decomposition (Hussain et al., 2021), while long-term monocropping

Table 5
Summary statistics of the spatial distribution of SOC (g/kg) throughout main land types in Germany.

Land type	Tree cover	Grassland	Cropland	Built-up	Bare/ Sparse vegetation	Herbaceous Wetland
Mean SOC	35.75	34.58	19.99	19.45	19.96	40.00
Min SOC	7.75	7.56	7.60	7.86	7.63	8.14
25% SOC	23.11	20.89	14.40	15.42	13.81	26.94
Median SOC	36.93	32.66	17.82	17.93	17.17	36.64
75% SOC	45.25	43.02	23.21	21.55	22.89	48.72
Max SOC	124.70	137.68	129.92	108.61	111.86	137.08

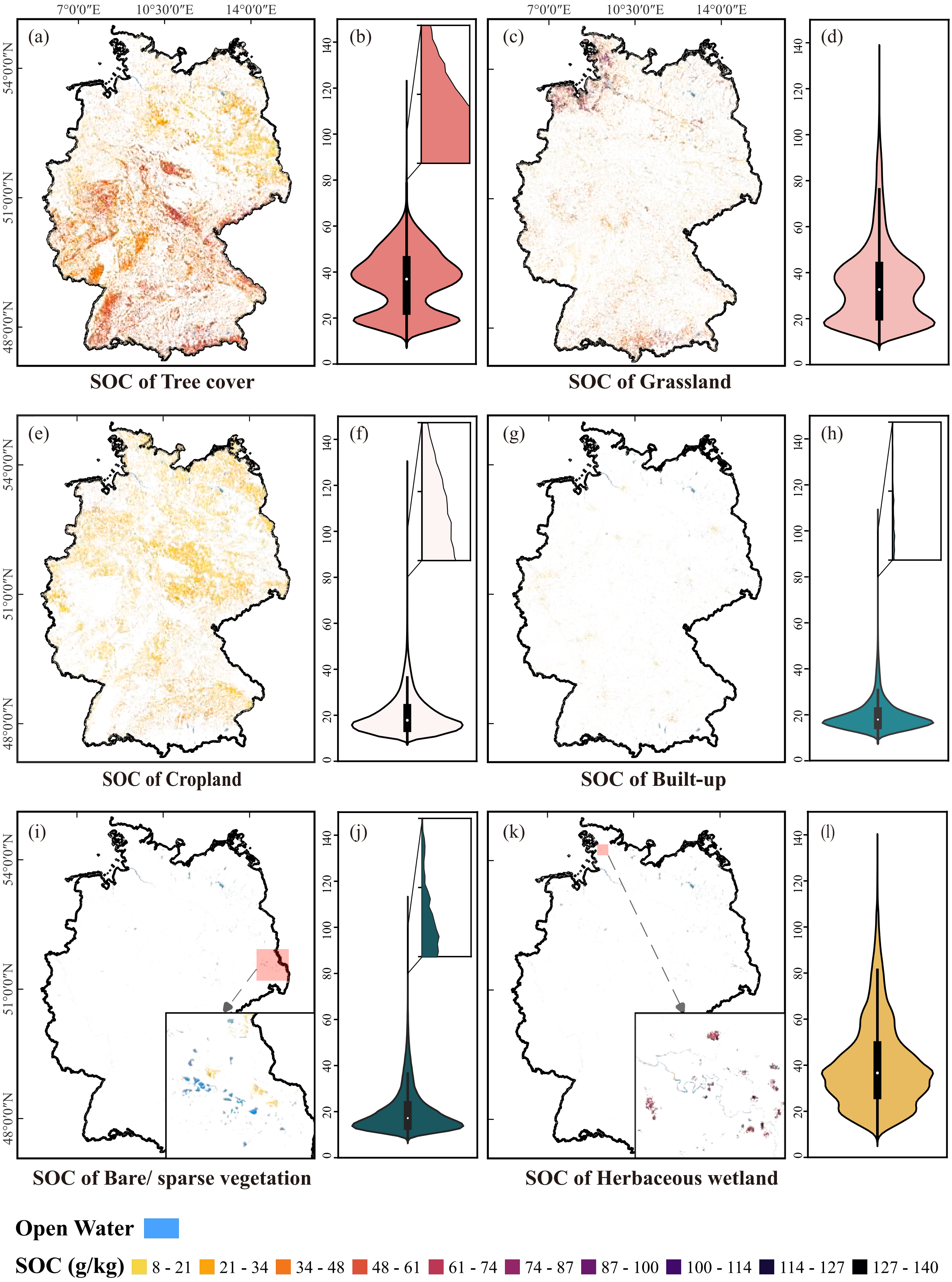


Fig. 8. SOC content spatial distribution and Violin plot of SOC content for six main land cover types including (a) and (b): Tree cover; (c) and (d): Grassland; (e) and (f): Cropland; (g) and (h): Built-up; (i) and (j): Bare/ sparse vegetation; (k) and (l): Herbaceous wetland.

depletes soil nutrients and reduces soil biodiversity, thereby leading to the loss of SOC (Zhang et al., 2016). At the same time, improved agricultural management can reduce greenhouse gas emissions and store carbon from the atmosphere. Increasing the capacities of croplands for storing organic carbon has been regarded as a necessary approach to reducing greenhouse gas effects on climate change (Amelung et al., 2020). As more than half of Germany's land is used for agricultural purposes, improving organic matter in agricultural soils is essential. Within the study of Seitz et al. (2023), cover crops were a promising method for improving the storage of SOC in croplands in Germany. In addition, organic fertilizer and cropping system shifts were suggested to enhance the large regional SOC stock (Deng et al., 2018). Regarding bare or sparse vegetation areas, utilizing naked soils for growing crops might effectively reduce SOC losses. Herbaceous wetlands were less distributed in Germany. However, soils under wetlands were predicted with high SOC content. Some regions with SOC values over 100 g/kg were concentrated on the coast side in the north. Whereas herbaceous wetlands are not widely distributed in Germany, protecting the wetland environment and reducing wetland degradation is crucial for carbon stock in wetlands.

5. Conclusion

This work compared the predictive capabilities of the hybrid models with individual PLSR models and RF models in mapping SOC content using Sentinel-2 data at four spatial resolutions (10 m, 50 m, 100 m, 200 m) in Germany. The comparative analysis identified the best-performing models, the optimal spatial resolutions, and the important variables for predicting SOC content. At the national scale, the RF model outperformed the statistical model PLSR at four resolutions, additionally, the hybrid models that combined with OK models showed better predictive performance compared to individual models. The best prediction ($R^2 = 0.416$) for SOC content was constructed by the RFK model at a spatial resolution of 100 m, and the hybrid model enhanced the explanation by 3.7% of prediction compared to the individual RF model. Although spatial autocorrelation was low on the national scale, geostatistical models should be considered as a residual prediction to analyze the inner spatial variability of SOC content based on machine learning methods within limited covariates. Additionally, selecting an appropriate spatial resolution is crucial for prediction accuracy at the national scale. Our results demonstrated that the prediction results of hybrid models were affected by spatial resolutions, and the best performance of the hybrid model occurred at a related coarser spatial resolution (100 m) compared to individual models. The spatial resolution at which the hybrid model achieves the best prediction effect is inconsistent with that of a single model with the best performance at 50 m resolution. Multiscale spatial resolution comparison could be explored to select optimal prediction models based on the study areas. In this study, the selected Sentinel-2 bands and band indices accounted for 63.0% and 78.1% of all environmental variables in the PLSR and RF models, respectively, indicating Sentinel-2 remote sensing data had predictive capabilities for SOC content prediction at the national scale. Visible bands, vegetation indices, and elevation were important factors for predicting SOC content. A digital SOC content map of Germany was displayed, revealing the spatial distribution of SOC in different land types. The significant high SOC distribution was concentrated in the northwestern grassland, in the forest covers of central and southwestern altitudes, and in the Alps in the south. This map offered a national SOC distribution in Germany, which can guide researchers to monitor the changes of SOC content under the impacts on climate and land use change. It worked toward being a useful reference to fill up the application of hybrid methods on national SOC monitoring and management. It aimed to offer new feasibility for using remote sensing data and hybrid models to achieve digital soil mapping and analyze the important and potential factors for SOC prediction on the national scale.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors are particularly grateful to the European Soil Data Center for its support of soil data. This research was supported by the China Scholarship Council PhD program (Xiande Ji).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ecolind.2024.112654.

References

Akpa, S.I., Odeh, I.O., Bishop, T.F., Hartemink, A.E., Amapu, I.Y., 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. Geoderma 271, 202–215. https://doi.org/10.1016/j.geoderma.2016.02.021.

Allard, V., Soussana, J.F., Falcimagne, R., Berbigier, P., Bonnefond, J.M., Ceschia, E., D'hour, P., Hénault, C., Laville, P., Martin, C., Pinarès-Patino, C., 2007. The role of grazing management for the net biome productivity and greenhouse gas budget (co2, n2o and ch4) of semi-natural grassland. Agric., Ecosyst. Environ. 121, 47–58. https://doi.org/10.1016/j.agee.2006.12.004.

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., 2020. Towards a global-scale soil climate mitigation strategy. Nat. Commun. 11, 5427. https://doi.org/10.1038/s41467-020-18887-7.

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D., 2019. Remote sensing techniques for soil organic carbon estimation: A review. Remote Sens. 11, 1–18. https://doi.org/10.3390/rs11060676.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., Panagos, P., 2019. Mapping lucas topsoil chemical properties at european scale using gaussian process regression. Geoderma 355, 113912. https://doi.org/10.1016/j.geoderma.2019.113912.

Baltensweiler, A., Walthert, L., Hanewinkel, M., Zimmermann, S., Nussbaum, M., 2021. Machine learning based soil maps for a wide range of soil properties for the forested area of Switzerland. Geoderma Regional 27, e00437. https://doi.org/10.1016/j.geodrs.2021.e00437.

Bauer-Marschallinger, B., Freeman, V., Cao, S., Paulik, C., Schaufler, S., Stachl, T., Modanesi, S., Massari, C., Ciabatta, L., Brocca, L., Wagner, W., 2019. Toward global soil moisture monitoring with sentinel-1: Harnessing assets and overcoming obstacles. IEEE Trans. Geosci. Remote Sens. 57, 520–539. https://doi.org/10.1109/TGRS.2018.2858004.

Bellon-Maurel, V., McBratney, A., 2011. Near-infrared (nir) and mid-infrared (mir) spectroscopic techniques for assessing the amount of carbon stock in soils – critical review and research perspectives. Soil Biol. Biochem. 43, 1398–1410. https://doi.org/10.1016/j.soilbio.2011.02.019.

Berger, K., Hank, T., Halabuk, A., Rivera-Caicedo, J.P., Wocher, M., Mojses, M., Gerhátová, K., Tagliabue, G., Dolz, M.M., Venteo, A.B.P., Verrelst, J., 2021. Assessing non-photosynthetic cropland biomass from spaceborne hyperspectral imagery. Remote Sens. 13, 4711. https://doi.org/10.3390/rs13224711.

Castaldi, F., Chabrillat, S., Don, A., van Wesemael, B., 2019. Soil organic carbon mapping using lucas topsoil database and sentinel-2 data: An approach to reduce soil moisture and crop residue effects. Remote Sens. 11. https://doi.org/10.3390/rs11182121.

Castaldi, F., Chabrillat, S., van Wesemael, B., 2019. Sampling strategies for soil property mapping using multispectral sentinel-2 and hyperspectral enmap satellite data. Remote Sens. 11. https://doi.org/10.3390/rs11030309.

Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M., van Wesemael, B., 2019. Evaluating the capability of the sentinel 2 data for soil organic carbon prediction in croplands. ISPRS J. Photogram. Remote Sens. 147, 267–282. https://doi.org/10.1016/j.isprsjprs.2018.11.026.

Chen, D., Chang, N., Xiao, J., Zhou, Q., Wu, W., 2019. Mapping dynamics of soil organic matter in croplands with modis data and machine learning algorithms. Sci. Total Environ. 669, 844–855. https://doi.org/10.1016/j.scitotenv.2019.03.151.

Chi, Y., Zhao, M., Sun, J., Xie, Z., Wang, E., 2019. Mapping soil total nitrogen in an estuarine area with high landscape fragmentation using a multiple-scale approach. Geoderma 339, 70–84. https://doi.org/10.1016/j.geoderma.2018.12.040.

Crowther, T.W., Todd-Brown, K.E.O., Rowe, C.W., Wieder, W.R., Carey, J.C., Machmuller, M.B., Snoek, B.L., Fang, S., Zhou, G., Allison, S.D., Blair, J.M., Bridgham, S.D., Burton, A.J., Carrillo, Y., Reich, P.B., Clark, J.S., Classen, A.T., Dijkstra, F.A., Elberling, B., Emmett, B.A., Estiarte, M., Frey, S.D., Guo, J., Harte, J., Jiang, L., Johnson, B.R., Kröel-Dulay, G., Larsen, K.S., Laudon, H., Lavallee, J.M., Luo, Y., Lupascu, M., Ma, L.N., Marhan, S., Michelsen, A., Mohan, J., Niu, S., Pendall, E., Peñnuelas, J., Pfeifer-Meister, L., Poll, C., Reinsch, S., Reynolds, L.L., Schmidt, I. K., Sistla, S., Sokol, N.W., Templer, P.H., Treseder, K.K., Welker, J.M., Bradford, M. A., 2016. Quantifying global soil carbon losses in response to warming. Nature 540, 104–108. doi:10.1038/nature20150.

Dai, F., Zhou, Q., Lv, Z., Wang, X., Liu, G., 2014. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in tibetan plateau. Ecol. Ind. 45, 184–194. https://doi.org/10.1016/J.ECOLIND.2014.04.003.

Deng, X., Chen, X., Ma, W., Ren, Z., Zhang, M., Grieneisen, M.L., Long, W., Ni, Z., Zhan, Y., Lv, X., 2018. Baseline map of organic carbon stock in farmland topsoil in east china. Agric., Ecosyst. Environ. 254, 213–223. https://doi.org/10.1016/j.agee.2017.11.022.

Dobbin, K.K., Simon, R.M., 2011. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med. Genomics 4, 31. https://doi.org/10.1186/1755-8794-4-31.

Dvorakova, K., Shi, P., Limbourg, Q., van Wesemael, B., 2020. Soil organic carbon mapping from remote sensing: The effect of crop residues. Remote Sens. 12. https://doi.org/10.3390/rs12121913.

Gasmi, A., Gomez, C., Chehbouni, A., Dhiba, D., El Gharous, M., 2022. Using prisma hyperspectral satellite imagery and gis approaches for soil fertility mapping (fertimap) in northern morocco. Remote Sens. 14. https://doi.org/10.3390/rs14164080.

Ge, Y., Thomasson, J.A., Sui, R., 2011. Remote sensing of soil properties in precision agriculture: A review. Front. Earth Sci. 5, 229–238. https://doi.org/10.1007/s11707-011-0175-0.

Gholizadeh, A., Žižala, D., Saberioon, M., Borůvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and sentinel-2 spectral imaging. Remote Sens. Environ. 218, 89–103. https://doi.org/10.1016/j.rse.2018.09.015.

Ginzky, H., 2021. Soil protection governance in Germany. In: Ginzky, H., Dooley, E., Heuser, I.L., Kasimbazi, E., Kibugi, R., Markus, T., Qin, T., Ruppel, O. (Eds.), International Yearbook of Soil Law and Policy 2019. Springer International Publishing, Cham. International Yearbook of Soil Law and Policy, pp. 295–333. doi:10.1007/978-3-030-52317-615.

Gorai, T., Bhushan, M., Kumar, S.B., 2013. Application of geostatistical techniques in spatial variability mapping of soil fertility–a review. Int. J. Adv. Agric. Sci. Technol. 1, 100–111.

Gray, J.M., Bishop, T.F., Wilson, B.R., 2015. Factors controlling soil organic carbon stocks with depth in eastern australia. Soil Sci. Soc. Am. J. 79, 1741–1751. https://doi.org/10.2136/sssaj2015.06.0224.

Grüneberg, E., Ziche, D., Wellbrock, N., 2014. Organic carbon stocks and sequestration rates of forest soils in germany. Glob. Change Biol. 20, 2644–2662. https://doi.org/10.1111/gcb.12558.

Guo, L., Fu, P., Shi, T., Chen, Y., Zeng, C., Zhang, H., Wang, S., 2021. Exploring influence factors in mapping soil organic carbon on low-relief agricultural lands using time series of remote sensing data. Soil Till. Res. 210, 104982. https://doi.org/10.1016/j.still.2021.104982.

Guo, L., Zhao, C., Zhang, H., Chen, Y., Linderman, M., Zhang, Q., Liu, Y., 2017. Comparisons of spatial and non-spatial models for predicting soil carbon content based on visible and near-infrared spectral technology. Geoderma 285, 280–292. https://doi.org/10.1016/j.geoderma.2016.10.010.

Guo, P.T., Li, M.F., Luo, W., Tang, Q.F., Liu, Z.W., Lin, Z.M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma 237–238, 49–59. https://doi.org/10.1016/j.geoderma.2014.08.009.

Hamzehpour, N., Shafizadeh-Moghadam, H., Valavi, R., 2019. Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture. CATENA 182, 104141. https://doi.org/10.1016/j.catena.2019.104141.

Heil, J., Jörges, C., Stumpe, B., 2022. Fine-scale mapping of soil organic matter in agricultural soils using uavs and machine learning. Remote Sens. 14, 3349. https://doi.org/10.3390/rs14143349.

Hu, W., Shen, Q., Zhai, X., Du, S., Zhang, X., 2021. Impact of environmental factors on the spatiotemporal variability of soil organic matter: A case study in a typical small mollisol watershed of northeast china. J. Soils Sediments 21, 736–747. https://doi.org/10.1007/s11368-020-02863-1.

Huang, H., Yang, L., Zhang, L., Pu, Y., Yang, C., Wu, Q., Cai, Y., Shen, F., Zhou, C., 2022. A review on digital mapping of soil carbon in cropland: Progress, challenge, and prospect. Environ. Res. Lett. 17, 123004. https://doi.org/10.1088/1748-9326/aca41e.

Hussain, S., Hussain, S., Guo, R., Sarwar, M., Ren, X., Krstic, D., Aslam, Z., Zulifqar, U., Rauf, A., Hano, C., El-Esawi, M.A., 2021. Carbon sequestration to avoid soil degradation: A review on the role of conservation tillage. Plants 10, 2001. https://doi.org/10.3390/plants10102001.

Jacobs, A., Poeplau, C., Weiser, C., Fahrion-Nitschke, A., Don, A., 2020. Exports and inputs of organic carbon on agricultural soils in germany. Nutr. Cycl. Agroecosyst. 118, 249–271. https://doi.org/10.1007/s10705-020-10087-5.

Jeong, G., Oeverdieck, H., Park, S.J., Huwe, B., Ließ, M., 2017. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. CATENA 154, 73–84. https://doi.org/10.1016/j.catena.2017.02.006.

Jia, X., Cao, Y., O'Connor, D., Zhu, J., Tsang, D.C., Zou, B., Hou, D., 2021. Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. Environ. Pollut. 270, 116281. https://doi.org/10.1016/j.envpol.2020.116281.

Johnson, D.W., Knoepp, J.D., Swank, W.T., Shan, J., Morris, L.A., Van Lear, D.H., Kapeluck, P.R., 2002. Effects of forest management on soil carbon: Results of some long-term resampling studies. Environ. Pollut. 116, S201–S208. https://doi.org/10.1016/S0269-7491(01)00252-4.

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58. https://doi.org/10.1016/j.geoderma.2018.12.037.

Kühn, I., Dormann, C.F., 2012. Less than eight (and a half) misconceptions of spatial analysis. J. Biogeogr. 39, 995–998. https://doi.org/10.1111/j.1365-2699.2012.02707.x.

Laamrani, A., Berg, A.A., Voroney, P., Feilhauer, H., Blackburn, L., March, M., Dao, P.D., He, Y., Martin, R.C., 2019. Ensemble identification of spectral bands related to soil organic carbon levels over an agricultural field in southern ontario, canada. Remote Sens. 11, 1298. https://doi.org/10.3390/rs11111298.

Ladoni, M., Bahrami, H.A., Alavipanah, S.K., Norouzi, A.A., 2010. Estimating soil organic carbon from soil reflectance: A review. Precision Agric. 11, 82–99. https://doi.org/10.1007/s11119-009-9123-3.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304, 1623–1627. https://doi.org/10.1126/science.1097396.

Lamichhane, S., Kumar, L., Adhikari, K., 2021. Digital mapping of topsoil organic carbon content in an alluvial plain area of the terai region of nepal. Catena 202, 105299. https://doi.org/10.1016/j.catena.2021.105299.

Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma 352, 395–413. https://doi.org/10.1016/j.geoderma.2019.05.031.

Lange, M., Feilhauer, H., Kühn, I., Doktor, D., 2022. Mapping land-use intensity of grasslands in germany with machine learning and sentinel-2 time series. Remote Sens. Environ. 277, 112888. https://doi.org/10.1016/j.rse.2022.112888.

Legendre, P., Fortin, M.J., 1989. Spatial pattern and ecological analysis. Vegetatio 80, 107–138. https://doi.org/10.1007/BF00048036.

Leifeld, J., Ammann, C., Neftel, A., Fuhrer, J., 2011. A comparison of repeated soil inventory and carbon flux budget to detect soil carbon stock changes after conversion from cropland to grasslands. Glob. Change Biol. 17, 3366–3375. https://doi.org/10.1111/j.1365-2486.2011.02471.x.

Liu, Z.P., Shao, M.A., Wang, Y.Q., 2012. Large-scale spatial variability and distribution of soil organic carbon across the entire loess plateau, china. Soil Research 50, 114. https://doi.org/10.1071/SR11183.

Martin, M.P., Orton, T.G., Lacarce, E., Meersmans, J., Saby, N.P., Paroissien, J.B., Jolivet, C., Boulonne, L., Arrouays, D., 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. Geoderma 223–225, 97–107. https://doi.org/10.1016/j.geoderma.2014.01.005.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Minasny, B., McBratney, A.B., Bellon-Maurel, V., Roger, J.M., Gobrecht, A., Ferrand, L., Joalland, S., 2011. Removing the effect of soil moisture from nir diffuse reflectance spectra for the prediction of soil organic carbon. Geoderma 167–168, 118–124. https://doi.org/10.1016/j.geoderma.2011.09.008.

Mirzaee, S., Ghorbani-Dashtaki, S., Mohammadi, J., Asadi, H., Asadzadeh, F., 2016. Spatial variability of soil organic matter using remote sensing data. CATENA 145, 118–127. https://doi.org/10.1016/J.CATENA.2016.05.023.

Mirzaee, S., Ghorbani-Dashtaki, S., Mohammadi, J., Asadi, H., Asadzadeh, F., 2016. Spatial variability of soil organic matter using remote sensing data. Catena 145, 118–127. https://doi.org/10.1016/j.catena.2016.05.023.

Mishra, U., Lal, R., Liu, D., Van Meirvenne, M., 2010. Predicting the spatial variation of the soil organic carbon pool at a regional scale. Soil Sci. Soc. Am. J. 74, 906–914. https://doi.org/10.2136/sssaj2009.0158.

Mulder, V.L., de Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping - a review. Geoderma 162, 1–19. https://doi.org/10.1016/j.geoderma.2010.12.018.

Nave, L.E., Vance, E.D., Swanston, C.W., Curtis, P.S., 2010. Harvest impacts on soil carbon storage in temperate forests. For. Ecol. Manage. 259, 857–866. https://doi.org/10.1016/j.foreco.2009.12.009.

Nguyen, Q.H., Ly, H.B., Ho, L.S., Al-Ansari, N., Le, H.V., Tran, V.Q., Prakash, I., Pham, B. T., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering 2021, 4832864. https://doi.org/10.1155/2021/4832864.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347. https://doi.org/10.1016/J.SOILBIO.2013.10.022.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E. K., Vargas, R., Wetterlind, J., 2015. Chapter four - soil spectroscopy: An alternative to wet chemistry for soil monitoring, in: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press. volume 132, pp. 139–159. doi:10.1016/bs.agron.2015.02.002.

Odebiri, O., Mutanga, O., Odindi, J., 2022. Deep learning-based national scale soil organic carbon mapping with sentinel-3 data. Geoderma 411, 115695. https://doi.org/10.1016/j.geoderma.2022.115695.

Odebiri, O., Mutanga, O., Odindi, J., Naicker, R., 2022. Modelling soil organic carbon stock distribution across different land-uses in south africa: A remote sensing and deep learning approach. ISPRS J. Photogram. Remote Sens. 188, 351–362. https://doi.org/10.1016/j.isprsjprs.2022.04.026.

Odebiri, O., Odindi, J., Mutanga, O., 2021. Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review. Int. J. Appl. Earth Obs. Geoinf. 102, 102389. https://doi.org/10.1016/J.JAG.2021.102389.

Oliver, M.A., Webster, R., 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. CATENA 113, 56–69. https://doi.org/10.1016/j.catena.2013.09.006.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. Lucas soil, the largest expandable soil dataset for europe: A review. Eur. J. Soil Sci. 69, 140–153. https://doi.org/10.1111/ejss.12499.

Paloscia, S., Pettinato, S., Santi, E., Notarnicola, C., Pasolli, L., Reppucci, A., 2013. Soil moisture mapping using sentinel-1 images: Algorithm and preliminary validation. Remote Sens. Environ. 134, 234–248. https://doi.org/10.1016/j.rse.2013.02.027.

Paul, S.S., Coops, N.C., Johnson, M.S., Krzic, M., Chandna, A., Smukler, S.M., 2020. Mapping soil organic carbon and clay using remote sensing to predict soil workability for enhanced climate change adaptation. Geoderma 363, 114177. https://doi.org/10.1016/j.geoderma.2020.114177.

Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G.P., Smith, P., 2016. Climate-smart soils. Nature 532, 49–57. https://doi.org/10.1038/nature17174.

Pham, B.T., Prakash, I., Jaafari, A., Bui, D.T., 2018. Spatial prediction of rainfall-induced landslides using aggregating one-dependence estimators classifier. J. Indian Soc. Remote Sens. 46, 1457–1470. https://doi.org/10.1007/s12524-018-0791-1.

Pham, T.D., Yokoya, N., Nguyen, T.T.T., Le, N.N., Ha, N.T., Xia, J., Takeuchi, W., Pham, T.D., 2021. Improvement of mangrove soil carbon stocks estimation in north vietnam using sentinel-2 data and machine learning approach. GIScience and Remote Sens. 58, 68–87. https://doi.org/10.1080/15481603.2020.1857623.

Poeplau, C., Don, A., Vesterdal, L., Leifeld, J., Van Wesemael, B., Schumacher, J., Gensior, A., 2011. Temporal dynamics of soil organic carbon after land-use change in the temperate zone – carbon response functions as a model approach. Glob. Change Biol. 17, 2415–2427. https://doi.org/10.1111/j.1365-2486.2011.02408.x.

Poeplau, C., Jacobs, A., Don, A., Vos, C., Schneider, F., Wittnebel, M., Tiemeyer, B., Heidkamp, A., Prietz, R., Flessa, H., 2020. Stocks of organic carbon in german agricultural soils—key results of the first comprehensive inventory. J. Plant Nutr. Soil Sci. 183, 665–681. https://doi.org/10.1002/jpln.202000113.

Rahmani, S.R., Ackerson, J.P., Schulze, D., Adhikari, K., Libohova, Z., 2022. Digital mapping of soil organic matter and cation exchange capacity in a low relief landscape using lidar data. Agronomy 12, 1338. https://doi.org/10.3390/agronomy12061338.

Rienzi, E.A., Mijatovic, B., Mueller, T.G., Matocha, C.J., Sikora, F.J., Castrignanò, A., 2014. Prediction of soil organic carbon under varying moisture levels using reflectance spectroscopy. Soil Sci. Soc. Am. J. 78, 958–967. https://doi.org/10.2136/sssaj2013.09.0408.

Roßkopf, N., Fell, H., Zeitz, J., 2015. Organic soils in germany, their distribution and carbon stocks. CATENA 133, 157–170. https://doi.org/10.1016/j.catena.2015.05.004.

Sakhaee, A., Gebauer, A., Ließ, M., Don, A., 2022. Spatial prediction of organic carbon in german agricultural topsoil using machine learning algorithms. SOIL 8, 587–604. https://doi.org/10.5194/soil-8-587-2022.

Segarra, J., Buchaillot, M.L., Araus, J.L., Kefauver, S.C., 2020. Remote sensing for precision agriculture: Sentinel-2 improved features and applications. Agronomy 10. https://doi.org/10.3390/agronomy10050641.

Seitz, D., Fischer, L.M., Dechow, R., Wiesmeier, M., Don, A., 2023. The potential of cover crops to increase soil organic carbon storage in german croplands. Plant Soil 488, 157–173. https://doi.org/10.1007/s11104-022-05438-w.

Shafizadeh-Moghadam, H., Minaei, F., Talebi-khiyavi, H., Xu, T., Homaee, M., 2022. Synergetic use of multi-temporal sentinel-1, sentinel-2, ndvi, and topographic factors for estimating soil organic carbon. CATENA 212, 106077. https://doi.org/10.1016/j.catena.2022.106077.

Siewert, M.B., 2018. High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: A case study in a sub-arctic peatland environment. Biogeosciences 15, 1663–1682. https://doi.org/10.5194/bg-15-1663-2018.

Simon, S.M., Glaum, P., Valdovinos, F.S., 2023. Interpreting random forest analysis of ecological models to move from prediction to explanation. Scientific Reports 13, 3881. https://doi.org/10.1038/s41598-023-30313-8.

Smith, P., Martino, D., Cai, Z., Gwary, D., Janzen, H., Kumar, P., McCarl, B., Ogle, S., O'Mara, F., Rice, C., Scholes, B., Sirotenko, O., Howden, M., McAllister, T., Pan, G., Romanenkov, V., Schneider, U., Towprayoon, S., Wattenbach, M., Smith, J., 2007. Greenhouse gas mitigation in agriculture. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 789–813. https://doi.org/10.1098/rstb.2007.2184.

Soussana, J.F., Loiseau, P., Vuichard, N., Ceschia, E., Balesdent, J., Chevallier, T., Arrouays, D., 2004. Carbon cycling and sequestration opportunities in temperate grasslands. Soil Use Manag. 20, 219–230. https://doi.org/10.1111/j.1475-2743.2004.tb00362.x.

Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., Ben-Dor, E., 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma 144, 395–404. https://doi.org/10.1016/J.GEODERMA.2007.12.009.

Stuart, B.H., 2004. Infrared Spectroscopy: Fundamentals and Applications. John Wiley & Sons.

Suleymanov, A., Abakumov, E., Suleymanov, R., Gabbasova, I., Komissarov, M., 2021. The soil nutrient digital mapping for precision agriculture cases in the trans-ural steppe zone of russia using topographic attributes. ISPRS International Journal of Geo-Information 10, 243. https://doi.org/10.3390/ijgi10040243.

Taghizadeh-Mehrjardi, R., Hamzehpour, N., Hassanzadeh, M., Heung, B., Ghebleh Goydaragh, M., Schmidt, K., Scholten, T., 2021. Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. Geoderma 399, 115108. https://doi.org/10.1016/j.geoderma.2021.115108.

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

Tian, J., Su, S., Tian, Q., Zhan, W., Xi, Y., Wang, N., 2021. A novel spectral index for estimating fractional cover of non-photosynthetic vegetation using near-infrared bands of sentinel satellite. Int. J. Appl. Earth Obs. Geoinf. 101, 102361. https://doi.org/10.1016/j.jag.2021.102361.

Triantafilis, J., Odeh, I., McBratney, A., 2001. Five geostatistical models to predict soil salinity from electromagnetic induction data across irrigated cotton. Soil Sci. Soc. Am. J. 65, 869–878. https://doi.org/10.2136/sssaj2001.653869x.

van Wesemael, B., Paustian, K., Meersmans, J., Goidts, E., Barancikova, G., Easter, M., 2010. Agricultural management explains historic changes in regional soil carbon stocks. Proc. Nat. Acad. Sci. 107, 14926–14930. https://doi.org/10.1073/pnas.1002592107.

Vaudour, E., Gholizadeh, A., Castaldi, F., Saberioon, M., Borůvka, L., Urbina-Salazar, D., Fouad, Y., Arrouays, D., Richer-De-forges, A.C., Biney, J., Wetterlind, J., Van Wesemael, B., 2022. Satellite imagery to map topsoil organic carbon content over cultivated areas: An overview. Remote Sens. 14. https://doi.org/10.3390/rs14122917.

Vaudour, E., Gomez, C., Fouad, Y., Lagacherie, P., 2019. Sentinel-2 image capacities to predict common topsoil properties of temperate and mediterranean agroecosystems. Remote Sens. Environ. 223, 21–33. https://doi.org/10.1016/j.rse.2019.01.006.

Vaudour, E., Gomez, C., Loiseau, T., Baghdadi, N., Loubet, B., Arrouays, D., Ali, L., Lagacherie, P., 2019. The impact of acquisition date on the prediction performance of topsoil organic carbon from sentinel-2 for croplands. Remote Sens. 11. https://doi.org/10.3390/rs11182143.

Vesterdal, L., Schmidt, I.K., Callesen, I., Nilsson, L.O., Gundersen, P., 2008. Carbon and nitrogen in forest floor and mineral soil under six common european tree species. For. Ecol. Manage. 255, 35–48. https://doi.org/10.1016/j.foreco.2007.08.015.

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., Liu, D.L., 2018. High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern australia. Sci. Total Environ. 630, 367–378. https://doi.org/10.1016/j.scitotenv.2018.02.204.

Wang, S., Adhikari, K., Wang, Q., Jin, X., Li, H., 2018. Role of environmental variables in the spatial distribution of soil carbon (c), nitrogen (n), and c:n ratio from the northeastern coastal agroecosystems in china. Ecol. Ind. 84, 263–272. https://doi.org/10.1016/j.ecolind.2017.08.046.

Wang, S., Adhikari, K., Zhuang, Q., Gu, H., Jin, X., 2020. Impacts of urbanization on soil organic carbon stocks in the northeast coastal agricultural areas of china. Science of The Total Environment 721, 137814. https://doi.org/10.1016/j.scitotenv.2020.137814.

Wang, S., Guan, K., Zhang, C., Lee, D., Margenot, A.J., Ge, Y., Peng, J., Zhou, W., Zhou, Q., Huang, Y., 2022. Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing. Remote Sens. Environ. 271, 112914. https://doi.org/10.1016/j.rse.2022.112914.

Wang, X., Zhang, Y., Atkinson, P.M., Yao, H., 2020. Predicting soil organic carbon content in spain by combining landsat tm and alos palsar images. Int. J. Appl. Earth Obs. Geoinf. 92, 102182. https://doi.org/10.1016/j.jag.2020.102182.

Wellbrock, N., Grüneberg, E., Riedel, T., Polley, H., 2017. Carbon stocks in tree biomass and soils of german forests. Central European Forestry Journal 63, 105–112. https://doi.org/10.1515/forj-2017-0013.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an afromontane landscape. Ecol. Ind. 52, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028.

Wiesmeier, M., Prietzel, J., Barthold, F., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., Schilling, B., von Lützow, M., Kögel-Knabner, I., 2013. Storage and drivers of organic carbon in forest soils of southeast germany (bavaria) – implications for carbon sequestration. For. Ecol. Manage. 295, 162–172. https://doi.org/10.1016/j.foreco.2013.01.025.

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.J., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - a review of drivers and indicators at various scales. Geoderma 333, 149–162. https://doi.org/10.1016/j.geoderma.2018.07.026.

Wold, S., Sjöström, M., Eriksson, L., 2001. Pls-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Xiao, J., Chevallier, F., Gomez, C., Guanter, L., Hicke, J.A., Huete, A.R., Ichii, K., Ni, W., Pang, Y., Rahman, A.F., Sun, G., Yuan, W., Zhang, L., Zhang, X., 2019. Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. Remote Sens. Environ. 233, 111383. https://doi.org/10.1016/j.rse.2019.111383.

Xie, B., Ding, J., Ge, X., Li, X., Han, L., Wang, Z., 2022. Estimation of soil organic carbon content in the ebinur lake wetland, xinjiang, china, based on multisource remote sensing data and ensemble learning algorithms. Sensors 22, 2685. https://doi.org/10.3390/s22072685.

Xu, Y., Smith, S.E., Grunwald, S., Abd-Elrahman, A., Wani, S.P., 2017. Evaluating the effect of remote sensing image spatial resolution on soil exchangeable potassium prediction models in smallholder farm settings. J. Environ. Manage. 200, 423–433. https://doi.org/10.1016/j.jenvman.2017.06.017.

Yang, R.M., Guo, W.W., 2019. Modelling of soil organic carbon and bulk density in invaded coastal wetlands using sentinel-1 imagery. Int. J. Appl. Earth Obs. Geoinf. 82, 101906. https://doi.org/10.1016/j.jag.2019.101906.

Yao, X., Yu, K., Deng, Y., Zeng, Q., Lai, Z., Liu, J., 2019. Spatial distribution of soil organic carbon stocks in masson pine (pinus massoniana) forests in subtropical china. CATENA 178, 189–198. https://doi.org/10.1016/j.catena.2019.03.004.

Zhang, H., Wang, L., Tian, T., Yin, J., 2021. A review of unmanned aerial vehicle low-altitude remote sensing (uav-lars) use in agricultural monitoring in china. Remote Sens. 13. https://doi.org/10.3390/rs13061221.

Zhang, Y., Guo, L., Chen, Y., Shi, T., Luo, M., Ju, Q.L., Zhang, H., Wang, S., 2019. Prediction of soil organic carbon based on landsat 8 monthly ndvi data for the jianghan plain in hubei province, China. Remote Sens. 11. https://doi.org/10.3390/rs11141683.

Zhang, Z., Qiang, H., McHugh, A.D., He, J., Li, H., Wang, Q., Lu, Z., 2016. Effect of conservation farming practices on soil organic matter and stratification in a mono-cropping system of northern china. Soil and Tillage Research 156, 173–181. https://doi.org/10.1016/j.still.2015.10.008.

Zhong, L., Guo, X., Xu, Z., Ding, M., 2021. Soil properties: Their prediction and feature extraction from the lucas spectral library using deep convolutional neural networks. Geoderma 402, 115366. https://doi.org/10.1016/j.geoderma.2021.115366.

Zhou, T., Geng, Y., Chen, J., Liu, M., Haase, D., Lausch, A., 2020. Mapping soil organic carbon content using multi-source remote sensing variables in the heihe river basin in china. Ecol. Ind. 114, 106288. https://doi.org/10.1016/j.ecolind.2020.106288.

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., Bumberger, J., Haase, D., Lausch, A., 2021. Prediction of soil organic carbon and the c:n ratio on a national scale using machine learning and satellite data: A comparison between sentinel-2, sentinel-3 and landsat-8 images. Sci. Total Environ. 755, 142661. https://doi.org/10.1016/j.scitotenv.2020.142661.

Zhou, T., Geng, Y., Lv, W., Xiao, S., Zhang, P., Xu, X., Chen, J., Wu, Z., Pan, J., Si, B., Lausch, A., 2023. Effects of optical and radar satellite observations within google earth engine on soil organic carbon prediction models in spain. J. Environ. Manage. 338, 117810. https://doi.org/10.1016/j.jenvman.2023.117810.

Zink, M., Kumar, R., Cuntz, M., Samaniego, L., 2017. A high-resolution dataset of water fluxes and states for germany accounting for parametric uncertainty. Hydrol. Earth Syst. Sci. 21, 1769–1790. https://doi.org/10.5194/hess-21-1769-2017.