Modeling Earth Systems and Environment
https://doi.org/10.1007/s40808-024-02158-1

ORIGINAL ARTICLE

Random forest regression kriging modeling for soil organic carbon density estimation using multi-source environmental data in central Vietnamese forests

Viet Hoang Ho[1,3] · **Hidenori Morita**[1] · **Felix Bachofer**[2] · **Thanh Ha Ho**[3]

Received: 24 June 2024 / Accepted: 11 September 2024

Abstract

Forest soil organic carbon plays a vital role in the terrestrial carbon cycle. Accurately analyzing the spatial distribution of soil organic carbon density (SOCD) is therefore necessary for sustainable forest management and climate change mitigation. Previous studies explored the potential of random forest (RF) in modeling forest SOCD using various environmental data sources. However, how forest SOCD prediction would be affected by using random forest regression kriging (RFRK), which integrates the predictive power of RF in generating deterministic trends and the capability of the ordinary kriging (OK) in handling spatial autocorrelation structure of residuals, based on the environmental data sources and their combinations remains elusive and deserves further exploration. For this purpose, 104 soil samples were collected at a depth of 30 cm in forest ecosystems of Central Vietnam, and 33 environmental covariates were derived from Sentinel-2 (S2) imagery, Advanced Land Observing Satellite-2 Phased Array L-band Synthetic Aperture Radar-2 (AL2) imagery, digital elevation model (DEM), and climatic data. Using a leave-one-out cross-validation procedure to evaluate and compare the model performances, four metrics, including coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), and relative improvement (RI), were calculated. The results showed that enhanced RFRK performance for forest SOCD estimation was found with the inclusion of additional environmental data sources, with RFRK based on all data sources achieving a high accuracy ($R^2 = 0.78$, MAE = 8.28 $t \cdot ha^{-1}$, and RMSE = 10.54 $t \cdot ha^{-1}$). The comparison of the RF and RFRK models exhibited that additionally interpolated residuals by OK were more accurate than only considering the influences of predictor covariates. The relative improvement of the RFRK models over the RF models in forest SOCD estimation was notable, with RI_{R^2} ranging from 8.20 to 65.00%, RI_{MAE} ranging from 8.18 to 21.07%, and RI_{RMSE} ranging from 6.76 to 18.18%. The result from our case study emphasizes the robustness of RFRK using S2, AL2, DEM, and climatic data in accurately predicting forest SOCD.

Keywords Digital soil mapping · Forest ecosystems · Multi-source environmental data · Random forest regression kriging · Soil organic carbon density

Introduction

The rise in carbon emissions has resulted in a significant focus on the worldwide terrestrial carbon cycle (Scharlemann et al. 2014; Zhao et al. 2021; Liu et al. 2022). Forest ecosystems contain two-thirds of terrestrial carbon, with the soil responsible for the majority (Kumar et al. 2018, 2022). The capacity of soils to retain organic carbon is a crucial characteristic that impacts both soil quality and functionality, as well as is decisive for climate regulation (Lal 2016; Jackson et al. 2017; Wiesmeier et al. 2019). Minor alterations in forest soil organic carbon (SOC) can substantially modify global atmospheric carbon levels (Don et al. 2011). Over several decades, there has been a significant decline in

✉ Viet Hoang Ho
paoe9ws1@s.okayama-u.ac.jp

[1] Graduate School of Environmental and Life Science, Okayama University, 1 Chome-1-1 Tsushimanaka, Kita Ward, Okayama 700-8530, Japan

[2] Earth Observation Center, German Aerospace Center (DLR), 82234 Wessling, Germany

[3] University of Agriculture and Forestry, Hue University, 102 Phung Hung Str, Hue City 53000, Thua Thien Hue, Vietnam

Published online: 27 September 2024

forest SOC pools under human disturbance, with land-use change emerging as a prominent driver (Don et al. 2011). Most of this decrease occurred in tropical nations, contributing 30% to the global carbon sink (Ameray et al. 2021; Satdichanh et al. 2023). Therefore, it is crucial to accurately analyze the soil organic carbon density (SOCD), particularly in tropical forests, to conduct spatially explicit assessments, which would aid in preserving soil quality and devising strategies for climate change mitigation (Vågen and Winowiecki 2013; Zhou et al. 2020a).

Digital soil mapping (DSM) approaches have growingly emphasized mapping SOCD in the last decade due to their efficiency and convenience for accurately predicting soil properties over unsampled locations (Keskin et al. 2019; Emadi et al. 2020). DSM characterizes the spatial variation of SOCD by establishing the quantitative relationships between soil observations and georeferenced information (McBratney et al. 2003). Numerous machine learning (ML) algorithms have been extensively applied to capture these relationships, including Random forest (RF), Boosted Regression Tree, Support Vector Regression, Bagged CART, Extreme Gradient Boosting, and Artificial Neural Networks (Guan et al. 2019; Wang et al. 2020a; Odebiri et al. 2020; Zhou et al. 2020b; Emadi et al. 2020; Shafizadeh-Moghadam et al. 2022). Among these algorithms, RF has consistently demonstrated exceptional performance in predicting SOCD in most studies due to its robust resistance to over-fitting and insensitivity to noise in data (Camera et al. 2017; Lamichhane et al. 2019; Mahmoudzadeh et al. 2020). However, a major drawback of this tree-based ML algorithm is that it only accounts for the relationship between SOCD and predictor covariates while ignoring the influences of nearby observed data, known as spatial autocorrelation, which can lead to suboptimal prediction of the spatial distribution of SOCD (Guo et al. 2015). Meanwhile, soil attributes, including organic carbon content, often exhibit strong spatial autocorrelation structure due to the combined effects of climate (e.g. temperature and precipitation), topography (e.g. elevation, aspect, and slope), biotic activities (e.g. organisms), and parent material (e.g. original minerals) interacting over time (Cambardella et al. 1994). To overcome this shortcoming, random forest regression kriging (RFRK), a hybrid approach of RF and kriging technique, is proposed for SOCD modeling in this study. This method involves interpolating the spatially correlated component of RF residuals by using ordinary kriging (OK) and then adding the kriged residuals to the deterministic trend component generated by RF, boosting the prediction accuracy with lower error (Pouladi et al. 2019). Veronesi and Schillaci (2019) pointed out that there is no 'one-size-fits-all algorithm' for estimating SOCD across all landscapes. Accordingly, although it demonstrated outperformance over RF in predicting the spatial distribution of soil properties in agro-ecosystems (Guo et al. 2015; Tziachris et al. 2019; Matinfar et al. 2021), RFRK still needs to be further explored in forest ecosystems.

In the DSM system, a wide range of environmental covariates rooted in the SCORPAN framework (including soil, climate, organisms, relief, parent material, age, and space) were proposed to enhance the performance of predictive models for estimating SOCD (Zhou et al. 2022; Xia et al. 2022). These environmental variables can be retrieved from various available data sources, such as remote sensing (RS) data, digital elevation model (DEM), and climatic data (Lamichhane et al. 2019; Zhou et al. 2020a). Numerous RS sources have been applied for forest SOCD prediction, each with certain advantages over the others (Radočaj et al. 2024). Optical multispectral (MSI) sensors are the most commonly used (Zhou et al. 2022). Nevertheless, they often experience information loss due to sensor failures and unfavorable weather conditions, especially in tropical regions with frequent cloud cover and rainfall (Wang et al. 2020c; Li et al. 2022). Among MSI imageries, Sentinel-2, Landsat, and MODIS are the most popular in forest SOCD prediction (Kumar et al. 2018, 2022; Zhou et al. 2019; Wang et al. 2020a, b; Odebiri et al. 2020; Suleymanov et al. 2023). Compared to MSI sensors, Synthetic Aperture Radar (SAR) sensors offer the advantages of all-day and all-weather monitoring, as well as the ability to penetrate vegetation canopies (Balzter 2001; Lausch et al. 2019; Wu et al. 2020; Zhang et al. 2020; Zhou et al. 2020a). However, its full potential for application in DSM has not been extensively studied due to its complexity, diversity, and limited accessibility (Zribi et al. 2019). Several studies have investigated the capability of SAR sensors to map SOCD in forest regions, among which the most commonly used are C-band from Sentinel-1 and L-band from Advanced Land Observing Satellite Phased Array L-band SAR (ALOS PALSAR) (Ceddia et al. 2017; Sothe et al. 2022; Zhou et al. 2022; Shafizadeh-Moghadam et al. 2022; Vatandaşlar and Abdikan 2022). Combining MSI and SAR data is a feasible approach to increase the accuracy of SOCD prediction in the forests (Zhou et al. 2022). Previous studies have explored the MSI-SAR fusion of different satellite datasets, such as Sentinel-1 and Sentinel-2 data, Landsat and ALOS PALSAR data, Landsat and Sentinel-1 data, and the combination of Landsat, MODIS, Sentinel-1 and ALOS PALSAR data, and suggested superior performances for estimating forest SOCD than a standalone source (Ceddia et al. 2017; Zhou et al. 2020a, b; Sothe et al. 2022; Shafizadeh-Moghadam et al. 2022). The fusion of Sentinel-2 and ALOS PALSAR data offers new opportunities for predicting the spatial distribution of SOC content, as these sensors possess significant advantages. Specifically, Sentinel-2 imagery features three red-edge bands and a higher spatial resolution compared to Landsat imagery, while ALOS PALSAR imagery has the capability to penetrate deeper into vegetation canopies and soil compared to Sentinel-1 imagery (Ceddia et al. 2017; Zribi et al. 2019; Jha et al. 2021; Sothe et al. 2022). However, the fusion of Sentinel-2 and ALOS PALSAR data for forest SOCD estimation is still limited and rarely reported in the literature. Apart from RS data, topographic and climatic indices are also helpful in determining the variation in SOCD at regional scales (Zhou et al. 2019, 2020b; Shafizadeh-Moghadam et al. 2022). According to Lamichhane et al. (2019), due to the local variations in SOC dynamics, the relationships of environmental covariates to SOC levels depend upon the environmental conditions of the area under concern. Additionally, the selection of environmental covariates significantly impacts the performance of predictive models in estimating SOCD (Zhou et al. 2020b; Shafizadeh-Moghadam et al. 2022). Thus, how environmental covariates derived from Sentinel-2 imagery, ALOS PALSAR imagery, topographic data, and climatic data, as well as their combinations, affect the prediction accuracy of RFRK for predicting SOCD in forested areas deserves further assessment.

The main purposes of this study are to: (1) investigate the impact of environmental variables derived from different data sources, including Sentinel-2 (S2), Advanced Land Observing Satellite-2 Phased Array L-band SAR-2 (AL2), topographic data (T), and climatic data (C), as well as their combination on the performance of RFRK; (2) examine to what extent RFRK can improve prediction accuracy compared to RF for estimating SOCD in Central Vietnamese forests.

Materials and methods

Study area

The study area is located in Danang city, Central Vietnam (107.81°–108.34° E, 15.92°–16.22° N), covering approximately 960 km^2 (Fig. 1). The city consists of two primary topographic settings: mountains and plains. More than 50% of the city's territory is occupied by mountainous regions along the northern and western boundaries, and the coastal lowlands are characterized mainly by flat terrain in the city's southern and eastern areas, with altitudes varying between 0 and 1661 m.

Danang has a tropical monsoon climate, with an annual average temperature of 25 °C and annual average precipitation of 2134 mm, characterized by distinct rainy and dry seasons. The rainy season, with heavy rains and occasional typhoons, typically occurs from August to December, while the remainder of the year is the dry season, with hot and humid weather (Hoang Khanh Linh and Van Chuong 2015). Since Danang lies in a tropical rainforest zone, the forest ecosystems are dominated by evergreen broadleaf vegetation (Huy et al. 2016). The primary soil types found in the forested areas of the study site are Ferralic Acrisols (54.86%), Arenic Acrisols (34.11%), and Humic

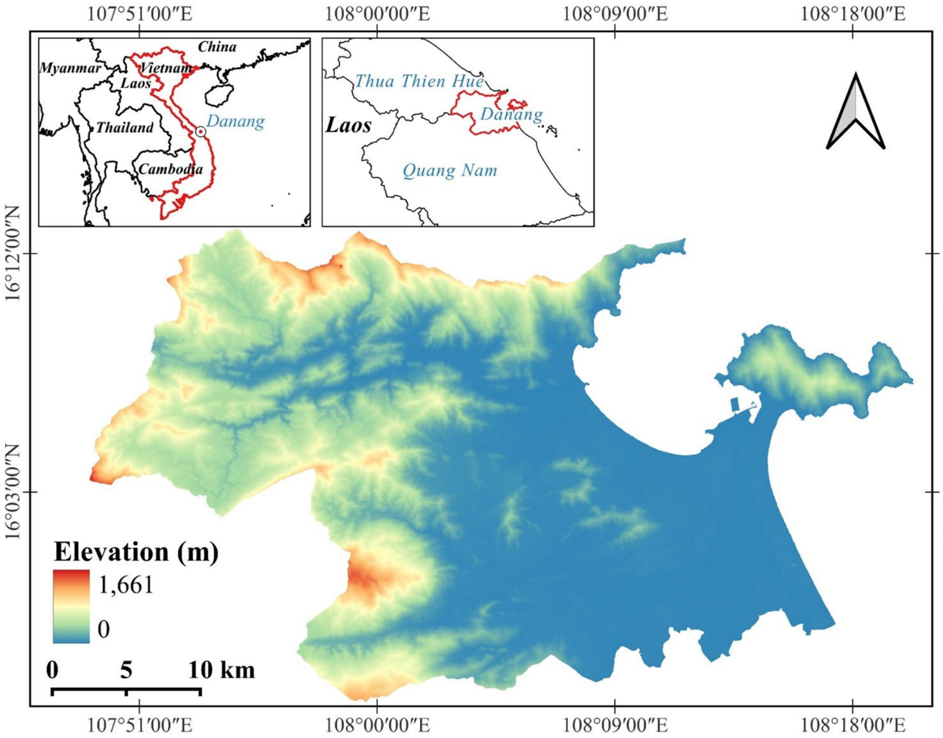


Fig. 1 The location of the study site

Acrisols (6.13%) (National Institute of Agricultural Planning and Projection of Vietnam 2005).

In situ data and laboratory analysis

Soil samples were collected from 104 sampling locations (Fig. 2a) between 13th July and 21st September 2023. The sampling locations in the forest landscapes were determined using a systematic unaligned sampling design, where a sampling point was randomly chosen within each 2.5 × 2.5 km grid. However, due to accessibility limitations, some locations could not be reached, so the farthest accessible ones to these locations were replaced. At each location, soil samples were collected at three different depths (0–10 cm, 10–20 cm, and 20–30 cm) using soil probes and cores, labeled, and brought to the laboratory for SOCD measurements, following a standard protocol of Pearson et al. (2007). In the laboratory, the samples from soil probes were air-dried and sieved through a 2-mm sieve before being put in a CN-corder machine to measure their carbon concentration. Meanwhile, the samples in soil cores, which were used to determine the bulk density, were subjected to oven-drying at a temperature of 105 °C for 48 h and sieved using a 2-mm sieve to separate into coarse fragments (> 2 mm) and fine fractions (< 2 mm). The bulk density was estimated as follows:

$$BD = \frac{ODW}{CV - (Rf/Pd)} \tag{1}$$

where BD is bulk density of fine fractions (g.cm^{-3}), ODW is oven-dry mass of fine fractions (g), CV is core volume (cm^{3}), Rf is mass of coarse fragments (g), and Pd is density of rock fragments (given as 2.65 g.cm^{-3}).

The SOCD content of soil samples was measured by:

$$SOCD(t/ha) = BD \times D \times \%C \tag{2}$$

where D is the soil depth at which the sample was taken (cm), and $\%C$ is carbon concentration (%).

The SOCD data at a soil depth of 0–30 cm were calculated by summing the SOCD values from three sampled soil depths and were then used for SOCD modeling.

Environmental data

Remote sensing data

S2 MSI instrument, launched by the European Space Agency (ESA), has 13 spectral bands (spanning from the visible and the near-infrared to the short-wave infrared) with high spatial resolution and a frequent revisit of 05 days (Drusch et al. 2012). In this study, we downloaded S2 images with less than 10% cloud cover from the https://dataspace.copernicus.euwebsite. AL2 instrument, developed by the Japan Aerospace Exploration Agency (JAXA), offers four possible polarizations (HH, HV, VH, VV) with a 14-day revisit time and various resolutions depending on observation modes (Truong et al. 2019; Yang et al. 2021). AL2 fine beam dual–polarized (HV, HH) mode level 2.1

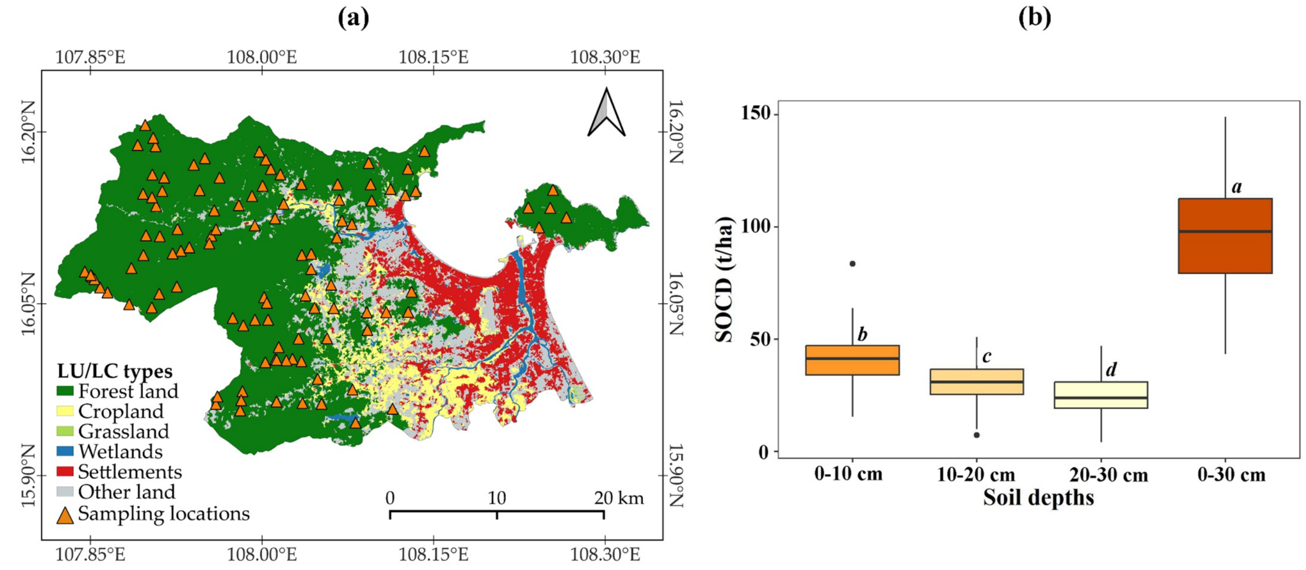


Fig. 2 LULC map and sampling locations (**a**), and boxplots of sample SOCD at different soil depths (**b**) with colors ranging from light brown to dark brown illustrating increasing mean values of SOCD and different letters denoting significant differences among soil depths (Turkey's test, p-value < 0.05)

Table 1 Detailed information on the acquired RS data

No	RS data	Number of images	Acquisition date (yy/mm/dd)	Processing level	Spatial resolution/pixel spacing (m)
1	AL2	02	23/06/04	Level 2.1	6.25
2	S2	09	23/05/07, 23/05/22, 23/06/01, 23/06/21, 23/07/06, 23/07/11, 23/07/21, 23/08/08, and 23/08/15	Level 1 C	10–20

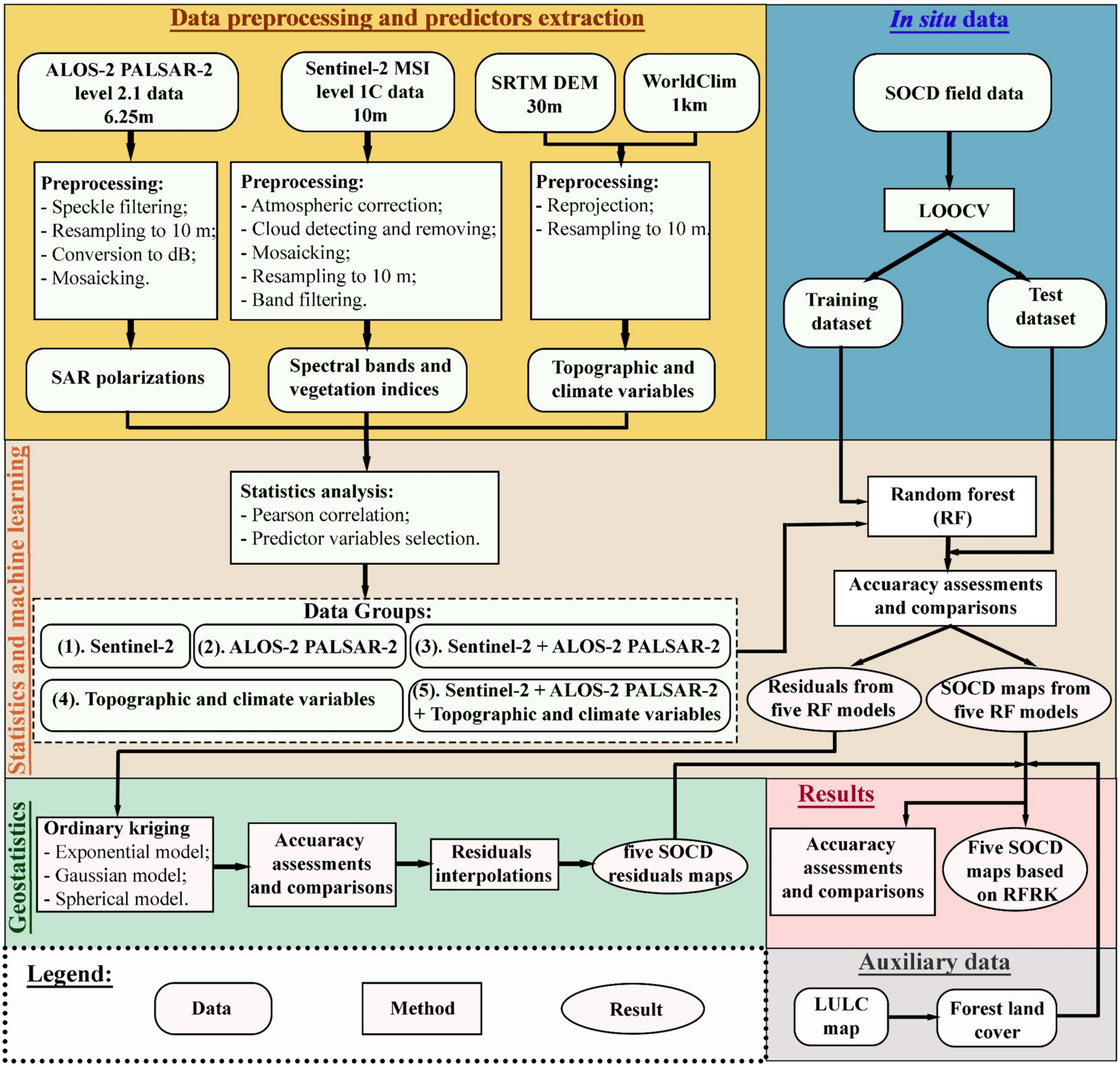

Data preprocessing and predictors extraction
In situ data
ALOS-2 PALSAR-2 level 2.1 data 6.25m
Sentinel-2 MSI level 1C data 10m
SRTM DEM 30m
WorldClim 1km
SOCD field data
Preprocessing:
- Speckle filtering;
- Resampling to 10 m;
- Conversion to dB;
- Mosaicking.
Preprocessing:
- Atmospheric correction;
- Cloud detecting and removing;
- Mosaicking;
- Resampling to 10 m;
- Band filtering.
Preprocessing:
- Reprojection;
- Resampling to 10 m.
LOOCV
Training dataset
Test dataset
SAR polarizations
Spectral bands and vegetation indices
Topographic and climate variables
Statistics and machine learning
Statistics analysis:
- Pearson correlation;
- Predictor variables selection.
Random forest (RF)
Accuaracy assessments and comparisons
Data Groups:
(1). Sentinel-2
(2). ALOS-2 PALSAR-2
(3). Sentinel-2 + ALOS-2 PALSAR-2
(4). Topographic and climate variables
(5). Sentinel-2 + ALOS-2 PALSAR-2 + Topographic and climate variables
Residuals from five RF models
SOCD maps from five RF models
Geostatistics
Ordinary kriging
- Exponential model;
- Gaussian model;
- Spherical model.
Accuaracy assessments and comparisons
Residuals interpolations
five SOCD residuals maps
Results
Accuaracy assessments and comparisons
Five SOCD maps based on RFRK
Legend:
Data
Method
Result
Auxiliary data
LULC map
Forest land cover

Fig. 3 Flow chart of predicting forest SOCD using RFRK and multi-source environmental data

CEOS format data was obtained from the https://www.eorc.jaxa.jpwebsite. The details of RS data are shown in Table 1.

The research procedures for preprocessing and using the RS data are displayed in Fig. 3. In our study, RS data was preprocessed using the ESA SNAP toolbox. The AL2

Table 2 Variable information

Types	Variables	Descriptions	References
SAR polarizations	HH	Horizontal/horizontal intensity data	
	HV	Horizontal/vertical intensity data	
	HH-HV	Difference between HH and HV	
	HH/HV	Ratio between HH and HV	
Spectral bands	B2	Blue, 490 nm	
	B3	Green, 560 nm	
	B4	Red, 665 nm	
	B5	Red edge, 705 nm	
	B6	Red edge, 749 nm	
	B7	Red edge, 783 nm	
	B8	Near-infrared, 842 nm	
	B8A	Near-infrared, 865 nm	
	B11	Short wave infrared, 1610 nm	
	B12	Short wave infrared, 2190 nm	
Vegetation indices	NDVI	Normalized difference vegetation index, $(B8 - B4)/(B8 + B4)$	Tucker (1979)
	GNDVI	Green Normalized Difference Vegetation Index, $B7 - B3/B7 + B3$	Gitelson and Merzlyak (1998)
	DVI	Difference Vegetation Index, $B8 - B4$	Richardsons and Wiegand (1977)
	SAVI	Soil Adjusted Vegetation Index,$(1 + 0.725) \times B8 - B4/B8 + B4 + 0.725$	Huete (1988)
	MSAVI	Modified Soil Adjusted Vegetation Index, $(1 + L)(B8 - B4)/B8 + B4 + L$ Where: L = 1–2 × s × NDVI × WDVI and s is the soil line slope	Qi et al. (1994)
	EVI	Enhance Vegetation Index, $2.5 \times B8 - B4/(B8 + 6 \times B4 - 7.5 \times B2 + 1)$	Matsushita et al. (2007)
	TVI	Transformed Vegetation Index, $\sqrt{NDVI} + 0.5$	Nellis and Briggs (1992)
Topography	ELEV	Elevation	
	ASPECT	Aspect	
	SLOPE	Slope	
	LSF	Slope steepness factor	
	MBI	Mass Balance Index	
	CNBL	Channel Network Base Level	
	PC	Plan curvature	
	TRI	Terrain Ruggedness Index	
	TWI	Topographic Wetness Index	
	VD	Valley depth	
Climate	MAP	Mean annual precipitation	
	MAT	Mean annual temperature	

scenes were preprocessed in several steps, including a Lee speckle filter with a 3 × 3-pixel kernel, resampling to a pixel spacing of 10 m, and conversion to normalized radar backscattering coefficients (γ^0) as shown in Eqs. 3 and 4 before mosaicked together.

$$\sigma^0 = 10.log_{10}(DN)^2 + CF \tag{3}$$

$$\gamma^0 = \frac{\sigma^0}{cos\phi} \tag{4}$$

where σ^0 is sigma-nought backscattering coefficients in decibels (dB), DN is digital numbers, ϕ is the incidence angle, and CF is the calibration factor. For AL2 imageries, the CF is -83.0 dB (Shimada et al. 2009).

Considering cloud cover, a cloud-free mosaic was created using the S2 time series. The S2 level 1 C orthorectified images were processed into orthoimage bottom-of-atmosphere corrected reflectance level 2 A products and clouds removed using the Sen2Cor plug-in and the Idepix-assembly plug-in in the SNAP toolbox. These images were then mosaicked, resampled to a 10 m spatial resolution, and processed using a 3 × 3-pixel mean filter to reduce band noise.

Twenty-one environmental variables were derived from RS data, including four from AL2 and seventeen from the S2 mosaic, as shown in Table 2. These variables were reported to be strongly correlated with forest SOCD in previous studies (Wang et al. 2020a, 2020c; Odebiri et al. 2020; Zhou et al. 2020a; Abbaszad et al. 2024).

Topographic and climatic variables

Topography and climate are among the most commonly used predictor variables for SOCD prediction (Luo et al. 2017; Ayele et al. 2019; Odebiri et al. 2020). Shuttle Radar Topography Mission (SRTM) DEM at a resolution of 30 m, downloaded from Google Earth Engine (https://code.earthengine.google.com), used to calculate topographic indices in the SAGA-GIS software platform. In addition, mean annual precipitation (MAP) and mean annual temperature (MAT) data with a spatial resolution of 1 km, downloaded from the Worldclim website (https://www.worldclim.org), were used as climatic variables in this study. WorldClim provides interpolated climate data for global land areas, developed using thin-plate splines to interpolate weather station data, with a detailed methodology described by Fick and Hijmans (2017). All topographic and climatic variables, as shown in Table 2, were resampled to 10 m resolution.

Auxiliary data

In addition to the above data, a 2023 land use/land cover (LULC) map of Danang city was used as auxiliary data. This LULC map, created based on simple non-iterative clustering segmentation, the interpretation of time series Sentinel-1 and S2 images, and kernel-principal component analysis supporting the RF classifier, classified the study area into six LULC types: forest land, cropland, grassland, wetlands, settlements, and other land (bare soil and all land areas that do not fall into any of the other five categories) as the guideline of Intergovernmental Panel on Climate Change (IPCC 2006). The LULC classification was considered acceptable if it met the following criteria: a Kappa statistic (K) greater than 0.80 and an overall accuracy (OA) exceeding 85% (Nyamekye et al. 2021).

Statistical analysis

A Pearson's product-moment correlation analysis was performed to ascertain the relationship between field-based SOCD and environmental indices. The role of the Pearson correlation coefficient (r) has been demonstrated in the context of noise reduction (Benesty et al. 2008). The strength of the correlation is identified based on r values, where r values of 0.00–0.29 indicate little to no correlation, 0.30–0.49 indicate low correlation, 0.50–0.69 indicate moderate correlation, 0.70–0.89 indicate high correlation and 0.90–1.00 indicate very high correlation (Asuero et al. 2006). Environmental variables that did not show a significant correlation with observed SOCD (p-value $\geq$ 0.05) were excluded from the modeling process. Correlation analysis results were also used to determine the collinearity among each data source. Predictor covariates from the same source with r values greater than 0.8 were considered collinearity and eliminated from predictor covariates of the modeling (Xu et al. 2018). Besides, the differences in SOCD between soil depths (0–10 cm, 10–20 cm, 20–30 cm, and 0–30 cm) were examined by ANOVA using Tukey's test for multiple pairwise comparisons.

Modeling techniques

Random forest

RF is a popular ensemble learning technique using tree-based models for classification and regression tasks (Breiman 2001). The model training process involves generating numerous decision trees from the training dataset using bootstrap samples (Zhou et al. 2020b). For RF implementation, the training parameters that should be taken into consideration include (1) the number of trees to grow in the forest (n_estimator), (2) the maximal number of randomly selected predictor variables at each node (max_feature), and (3) the minimal number of samples at leaf nodes (min_sample_leaft) (Forkuor et al. 2017). These were set to 100, 1 (indicating "max_features" is the number of training features), and 1, respectively, as defaults in the Scikit-learn package. The training data not included in the bootstrap samples, also known as out-of-bag (OOB) samples, can be used for validation by comparing it to the model outputs and calculating the corresponding relative errors (Camera et al. 2017). An additional feature of RF is the capacity to assess the relative importance of the variables. In this study, Gini importance, known as the impurity importance and computed by the mean impurity decrease measures of all nodes in the forest (Nembrini et al. 2018), was used to identify the variable importance.

Random forest regression kriging

RFRK is a hybrid technique that integrates the prediction of RF for the deterministic trend of the response variable and the interpolation of the OK technique for residuals. The execution included three sequential stages. Firstly, based on the relationships between the observed SOCD and the environmental variables, RF was used to estimate the predicted SOCD values ($SOCD_{RF}$). Secondly, the residuals (RES) obtained by RF were interpolated using OK. Thirdly, SOCD prediction by RFRK ($SOCD_{RFRK}$) was calculated by adding the values of $SOCD_{RF}$ and RES.

$$SOCD_{RFRK} = SOCD_{RF} + RES \tag{5}$$

In OK, the weights were estimated using the semivariogram model, and the values of the response variable at unsampled locations were determined using Eq. 6 (Kravchenko and Bullock 1999). The assumption regarding the distribution of data, namely normality and stationarity, needs to be satisfied to utilize the OK method (Meul and Van Meirvenne 2003). In this study, we confirmed the normality by using the Kolmogorov-Smirnov (K-S) test (p-value ≥ 0.05) and the stationarity by utilizing the intrinsic hypothesis (Webster and Oliver 2007).

$$Z_{u,OK}\left(x_0\right) = \sum_{i=1}^{n} \lambda_{ui} Z_u\left(x_i\right) \tag{6}$$

where $Z_{u,OK}\left(x_0\right)$ is the interpolated residual value of SOCD, λ_{ui} is the weight at location x_i, $Z_u\left(x_i\right)$ is the SOCD residual at location x_i, and n is the number of sample points used for interpolation.

The semivariograms were modeled in the Gstat package in the R-Studio software environment, employing exponential, Gaussian, and spherical functions. These models are characterized by three primary parameters: (1) range, which presents the distance beyond which little or no spatial autocorrelation occurs; (2) sill, which is the semivariance value where the spatial distance between two locations reaches the range; and (3) nugget, which reflects the small-scale variability of the data (Hohn 1991). Positive sill values indicate measurable spatial variability influenced by substrate effects and sampling errors (Yao et al. 2019). Nugget values reflect the degree of spatial randomness caused by sampling error, and undetectable and inherent variability (Brownstein et al. 2012; Yao et al. 2019). The nugget-to-sill value (N/S), which defines the proportion of short-range variability that geostatistics cannot describe based on a variogram, has been used to quantify the strength of spatial autocorrelation structure (Zhu and Lin 2010). N/S values of < 0.25, 0.25–0.75, and > 0.75 represent a strong, moderate, and weak spatial autocorrelation, respectively (Cambardella et al. 1994).

Model performance evaluation

The performances of the RF, OK, and RFRK models were evaluated using the leave-one-out cross-validation (LOOCV) procedure. This method entails partitioning the dataset into training and validation sets, where during each iteration, one data point is eliminated and used as the validation set, while the remaining data points are employed as the training set (Yue et al. 2018). Three distinct metrics were selected to compare the performance of the models: the mean absolute error (MAE), the root-mean-square error (RMSE), and the coefficient of determination (R^2). A model is considered superior performance if the values of MAE and RMSE are lower while the value of R^2 is greater (John et al. 2020).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{9}$$

where n is the number of observed values, y_i is the observed SOCD value for observation i, $\hat{y}_i$ is the predicted SOCD value for observation i, and $\bar{y}$ is the mean of the observed SOCD values.

Additionally, the relative improvement (RI) indices were used to evaluate the performance gain of RFRK compared to RF (Mishra et al. 2010).

$$RI_{MAE} = \frac{MAE_{RF} - MAE_{RFRK}}{MAE_{RF}} \times 100 \tag{10}$$

$$RI_{RMSE} = \frac{RMSE_{RF} - RMSE_{RFRK}}{RMSE_{RF}} \times 100 \tag{11}$$

$$RI_{R^2} = \frac{R^2{}_{RFRK} - R^2{}_{RF}}{R^2{}_{RF}} \times 100 \tag{12}$$

where RI_{MAE}, RI_{RMSE}, and $RI_{R^2}$are the relative improvement values of MAE, RMSE, and R^2, respectively. A positive RI indicates that RFRK has shown improvement over RF, whereas a negative RI indicates the opposite. In addition, higher values of the RI suggest more substantial enhancements.

Results

Land Cover Mapping and descriptive statistics of SOCD

A total of 427 global positioning system (GPS) points, representing the ground truth for LULC classification, were collected from July to September 2023. The dataset was divided

Table 3 Summary of classification accuracies

LULC type	Producer accuracy (%)	User accuracy (%)	Kappa statistics	Overall accuracy (%)
Forest land	97.50	95.12	0.88	91.86
Cropland	70.00	77.78		
Grassland	100.00	100.00		
Wetlands	100.00	100.00		
Settlements	92.31	100.00		
Other land	86.67	81.25		

into 80% (341 observations) for calibrating the RF classifier and 20% (86 observations) for validating the model.

Table 3 shows that the RF classifier achieved a large K value of 0.88 and a high OA value of 91.86% for LULC classification. K value was above 0.80, showing a strong level of agreement with classification accuracy, and the OA value was higher than the accepted OA threshold of 85.00% for LULC classification. Furthermore, all classes exhibited a high level of accuracy in terms of both the producer's and user's accuracy, surpassing 70.00% and 77.78%, respectively. Therefore, the LULC map of the study site in Fig. 2a was deemed suitable for further processing.

The boxplots of SOCD at different soil depths are also shown in Fig. 2b. The ANOVA (p-value < 0.05) and Turkey test results revealed a difference in SOCD among soil depths. The soil layer of 0–10 cm had the greatest figure of SOCD, followed sequentially by the soil layer of 10–20 cm and 20–30 cm. Moreover, the observed SOCD of topsoil (0–30 cm) ranged from 43.40 to 149.07 t·ha^{-1} and had a mean value of 96.97 t·ha^{-1}.

Correlation analysis

According to Pearson's product-moment correlation analysis (Table 4), 24 out of 33 predictor variables were significantly correlated to SOCD (p-value < 0.05), and no collinearity was detected within each data source ($r_{variables}$ < 0.8). MAT from climatic factors exhibited the strongest but negative correlation with forest SOCD (r = –0.67), indicating that the increasing MAT values related to lower SOCD values. It was followed by elevation and CNBL with the same r value (0.65). In SAR-derived variables, HH-HV showed the closest relationship with forest SOCD. It revealed that variations of HH-HV conveyed more useful information than HH and HV on forest SOCD. Meanwhile, vegetation indices showed the most critical influence among S2 variables. Several MSI-derived variables recorded higher values of r than SAR-derived ones, revealing that variables from S2 provided more valuable information regarding forest SOCD in the study area. From the results, r values of all environmental variables ranged from 0.23 to 0.67, indicating little to moderate linear relationships with forest SOCD. This suggests that non-linear modeling was necessary to capture the complexities of the data.

Table 4 Correlation coefficients of predictor variables derived from MSI, L-band SAR, topography, and climate for forest SOCD estimation. '***', '**', '*' means that p values were below 0.001, 0.01, and 0.05, respectively

Sources	Types	Predictor variables	r
S2	Reflectance	B2	-0.55^{***}
		B3	-0.41^{***}
		B4	-0.55^{***}
		B5	-0.30^{**}
		B11	-0.32^{**}
		B12	-0.49^{***}
	Vegetation indices	NDVI	0.58^{***}
		GNDVI	0.50^{***}
		SAVI	0.48^{***}
		MSAVI	0.45^{***}
		DVI	0.36^{***}
		EVI	0.44^{***}
		TVI	0.58^{***}
AL2	Backscatter	HH	0.23^{*}
		HV	0.44^{***}
		HH-HV	-0.50^{***}
Topography and climate	Topographic indices	CNBL	0.65^{***}
		Elevation	0.65^{***}
		Slope	0.34^{***}
		TRI	0.27^{**}
		TWI	-0.33^{***}
		VD	-0.23^{*}
	Climatic indices	MAP	0.60^{***}
		MAT	-0.67^{***}

Random forest models

Evaluation of random forest models

To assess the influence of environmental data from different sources and their combinations on predictive models for topsoil SOCD (0–30 cm) estimation in the Central Vietnamese forests, we created the following data groups: Group One and Group Two included only S2 and AL2 data, respectively, while Group Three used both S2 and AL2 data. Group Four (T + C) and Group Five (S2 + AL2 + T + C) were a combination of topographic and climatic data without and with RS data, respectively. Table 5 shows the evaluation metrics of RF using different data groups. Overall, an increase in the number of environmental data sources can enhance the

Table 5 Evaluation metrics of RF and RFRK based on different data groups

Models	Group	Variable sources	MAE (T ha^{-1})	RMSE (T ha^{-1})	R^2	RI$_{MAE}$ (%)	RI$_{RMSE}$ (%)	RI$_{R2}$ (%)
RF	One	S2	14.31	17.83	0.38	–	–	–
	Two	AL2	16.37	20.16	0.20	–	–	–
	Three	S2 + AL2	12.62	15.84	0.51	–	–	–
	Four	T + C	11.50	14.05	0.61	–	–	–
	Five	S2 + AL2 + T + C	10.49	12.88	0.67	–	–	–
RFRK	One	S2	13.14	16.11	0.49	8.18	9.65	28.95
	Two	AL2	14.87	18.50	0.33	9.16	8.23	65.00
	Three	S2 + AL2	11.45	14.14	0.61	9.27	10.73	19.61
	Four	T + C	10.30	13.10	0.66	10.43	6.76	8.20
	Five	S2 + AL2 + T + C	8.28	10.54	0.78	21.07	18.18	16.42

performance of RF. Among the RF models, RF using all environmental variables performed best, whereas the AL2-based RF model showed the lowest accuracy.

Regarding RS data, RF using Group One outperformed RF using Group Two, suggesting that the predictive power of S2-derived variables was superior to that of AL2-derived ones for SOCD estimation. The prediction accuracy of RF increased when SAR data were integrated with MSI data. Specifically, the addition of AL2 to S2 using RF resulted in enhanced MAE (decreasing from 14.31 to 12.62 t·ha^{-1}), RMSE (decreasing from 17.83 to 15.84 t·ha^{-1}), and R^2 (increasing from 0.38 to 0.51). Besides, the combination of topographic and climatic data (Group Four) achieved a high accuracy for RF prediction (MAE = 11.50 t·ha^{-1}, RMSE = 14.05 t·ha^{-1}, and R^2 = 0.61). It even outperformed RF for SOCD estimation based on dual-source RS data. Similar to the observed improvement in RS data fusion over standalone RS data, the prediction accuracy of RF was raised when topographic and climatic variables were added to dual-source RS data. This enhancement was confirmed by the Group Five-based RF model, which achieved the highest R^2 of 0.67 and the lowest MAE and RMSE of 10.49 and 12.88 t·ha^{-1}, respectively.

Attribute importance

The attribute importance of the five RF models is depicted in Fig. 4. For single-source RF models, NDVI and HH-HV were shown to be the most significant variables for S2 data and AL2 data, respectively. In dual-source RF models, SAR-derived variables had a lower impact than MSI predictors in estimating SOCD. The reflectance of MSI bands and vegetation indices were less important after being combined with L-band SAR backscatters. When combined with topographic and climatic factors, the influence of MSI bands, vegetation indices, and L-band SAR backscatters was marginal, with only HV, TVI, and NDVI showing some importance. This suggests that topographic and climatic indices had a greater influence than both SAR and MSI-derived variables in predicting forest SOCD. Among correlated predictor variables, elevation was identified as the primary determinant for SOCD estimation in forested regions.

Random forest regression kriging models

Residuals of random forest-derived soil organic carbon density and semivariogram analysis

Spatial autocorrelation of the RF residuals was detected by visual inspection of the semivariograms (Fig. 5). Prior to semivariogram analysis, residual data was tested to determine whether it met the normality and stationarity assumption. Figure 5a, e, i, m, q summarize the descriptive statistics of the residuals from RF across the five data groups. The absolute skewness values (ranging from 0.05 to 0.26) and absolute kurtosis values (ranging from 0.04 to 0.58) deviated from the expected values of 0 and 3, respectively, which are indicative of a normal distribution. Thus, we implemented the Box-Cox transformation to obtain data distributions more closely approximate normality (Fig. 5b, f, j, n, r). The results from the K–S test showed that the transformed residuals from Group One (R$_{S2}$, skewness = 0.00, kurtosis = –0.18), Group Two (R$_{AL2}$, skewness = 0.02, kurtosis = 0.01), Group Three (R$_{S2+AL2}$, skewness = 0.00, kurtosis = 0.05), Group Four (R$_{T+C}$, skewness = –0.04, kurtosis = –0.58), and Group Five (R$_{S2+AL2+T+C}$, skewness = –0.02, kurtosis = –0.51) all had p-values above 0.05 (Fig. 5b, f, j, n, r), suggesting that they possessed a normal distribution. Moreover, the semivariogram clouds of Box-Cox transformed residuals (Fig. 5c, g, k, o, s) exhibited that there was no trend in the semivariograms and semivariances were nearly constant across the entire study area, indicating that the transformed data satisfied the intrinsic stationarity assumption. After confirming the data normality and stationarity condition, those five groups of the transformed residuals were used

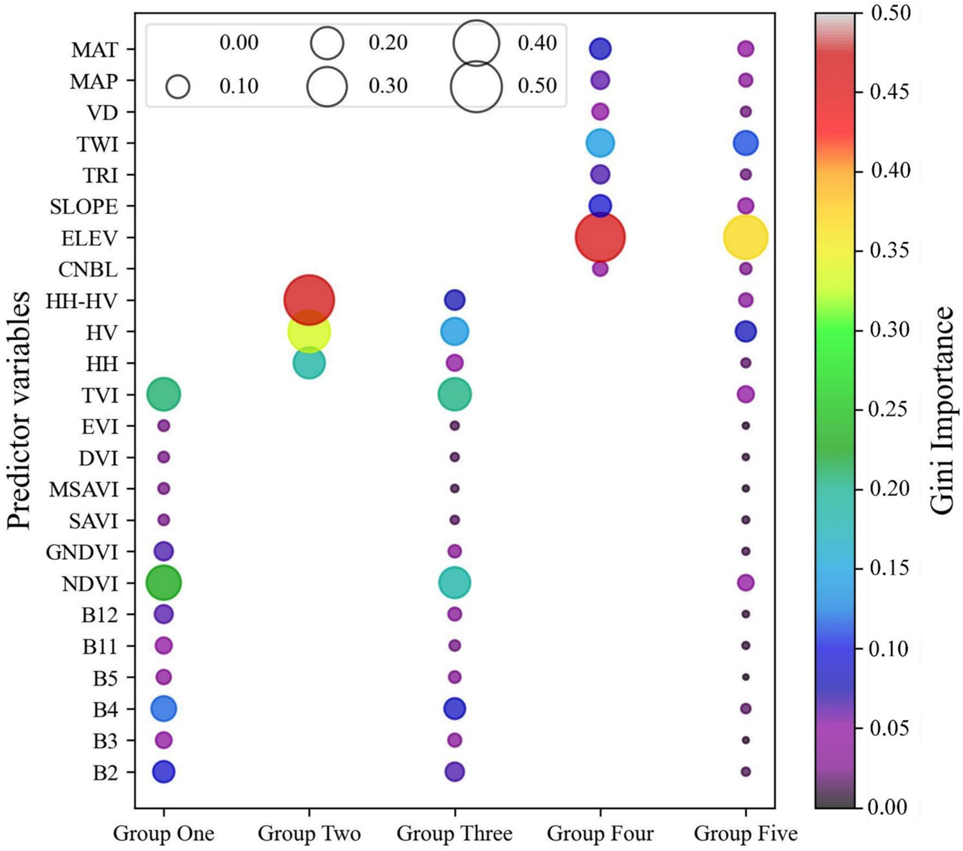


Fig. 4 The attribute importance of predictor variables across five data groups

Table 6 Parameter estimations for semivariogram analysis

Model parameters	Functions	Nugget	Sill	Nugget/Sill	Range (m)	MAE	RMSE	R^2
R_{S2}	Exponential	8.65	27.15	0.32	2879.29	3.62	4.43	0.18
	Gaussian	13.36	26.49	0.50	3714.76	3.64	4.45	0.18
	Spherical	9.99	25.89	0.39	6043.34	3.65	4.44	0.18
R_{AL2}	Exponential	0.00	188.60	0.00	1524.19	9.93	12.40	0.15
	Gaussian	74.20	187.25	0.40	2610.74	9.99	12.40	0.15
	Spherical	35.05	185.03	0.19	4176.95	10.04	12.44	0.15
R_{S2+AL2}	Exponential	7.33	21.79	0.34	2411.10	3.21	4.01	0.18
	Gaussian	12.33	21.83	0.56	4043.92	3.22	3.99	0.19
	Spherical	8.76	21.24	0.41	5633.43	3.22	4.00	0.19
R_{T+C}	Exponential	0.00	249.12	0.00	1590.34	10.99	13.97	0.13
	Gaussian	67.53	244.04	0.28	1947.13	11.07	14.09	0.12
	Spherical	30.30	242.48	0.12	3936.97	11.05	14.01	0.12
$R_{S2+AL2+T+C}$	Exponential	0.00	4.94	0.00	2203.12	1.36	1.72	0.33
	Gaussian	0.76	4.73	0.16	1710.50	1.39	1.75	0.31
	Spherical	0.00	4.69	0.00	4438.32	1.37	1.72	0.33

to calculate experimental semivariogram models for OK interpolation.

Table 6 shows the basic information on experimental semivariogram models of transformed residuals across five data groups. In general, the transformed residuals of all data groups displayed spatial autocorrelation structures. The functions that achieved the highest R^2 and lowest MAE and RMSE were chosen to fit the experimental semivariograms. As a result, the exponential function was selected to best fit the experimental semivariograms in most cases, except for

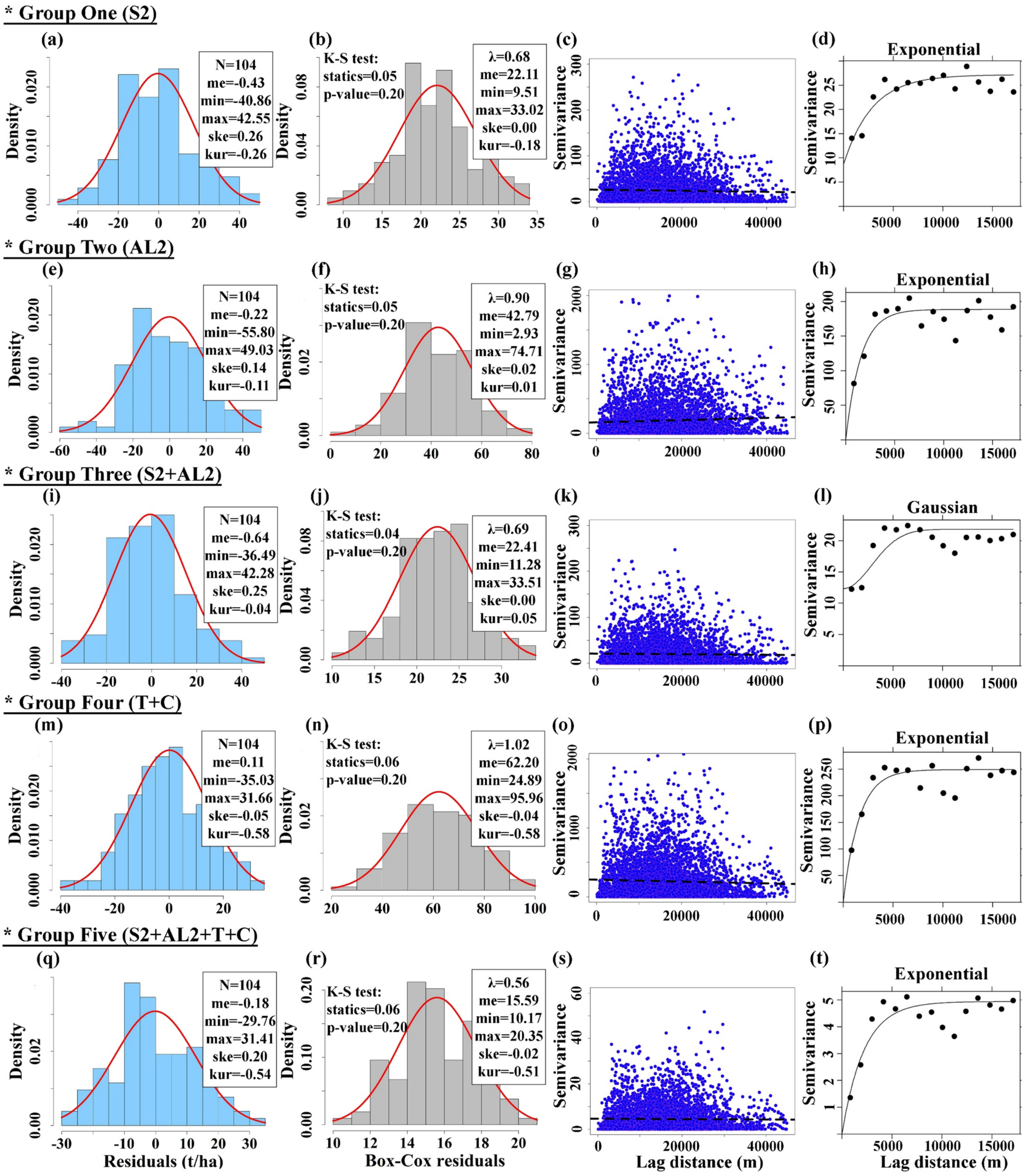


Fig. 5 Histograms/PDF curves of residuals (**a**, **e**, **i**, **m**, **q**) and transformed residuals (**b**, **f**, **j**, **n**, **r**), semivariogram clouds of transformed residuals (**c**, **g**, **k**, **o**, **s**), and experimental semivariogram models of transformed residuals (**d**, **h**, **l**, **p**, **t**) based on five data groups. N, me, min, max, ske, kur, and λ present the number of observations, mean value, minimum value, maximum value, skewness, kurtosis, and Box-Cox transformation parameter, respectively

R_{S2+AL2}, which employed the Gaussian function. The sill values of all fitted experimental semivariogram models were positive (ranging from 4.94 to 249.12), indicating measurable spatial variability, likely influenced by inherent substrate properties and potential sampling errors. The higher nugget value for R_{S2+AL2} (12.33) compared to R_{S2} (8.65) and R_{AL2} (0.00) suggested that the transformed residuals from dual-source RS data exhibited the greater spatial randomness, which may be caused by random factors (e.g. sampling error, and undetectable and inherent variability). When dual-source RS data were combined with topographic and climatic variables, the nugget value decreased to 0.00, suggesting reduced spatial randomness in $R_{S2+AL2+T+C}$. When considering the N/S value, integrating S2 and AL2 data did not reduce the N/S value compared to single-source RS data. However, after the inclusion of topographic and climatic variables alongside S2 and AL2 variables, the N/S value was significantly reduced from 0.56 in R_{S2+AL2} to 0.00 in $R_{S2+AL2+T+C}$, indicating a reinforcement of the spatial autocorrelation structure. Furthermore, among the transformed residuals, $R_{S2+AL2+T+C}$ achieved the lowest MAE (1.36), lowest RMSE (1.72), smallest nugget (0.00), and N/S value (0.00) while obtaining the highest R^2 (0.33). These results suggest that $R_{S2+AL2+T+C}$ was the most suitable for kriging interpolation compared to the others. Additionally, the N/S values of R_{AL2}, R_{T+C}, and $R_{S2+AL2+T+C}$ were all 0.00, indicating the existence of strong spatial autocorrelation. Meanwhile, the N/S values of R_{S2} and R_{S2+AL2} were 0.32 and 0.56, respectively, suggesting the existence of moderate spatial autocorrelation. The strong to moderate spatial autocorrelation structures of the transformed residuals revealed that intrinsic sources of variability (e.g. soil texture, soil types, mineral composition, biological activity, and pedogenesis) highly affected the spatial dependence of SOCD across the entire study area.

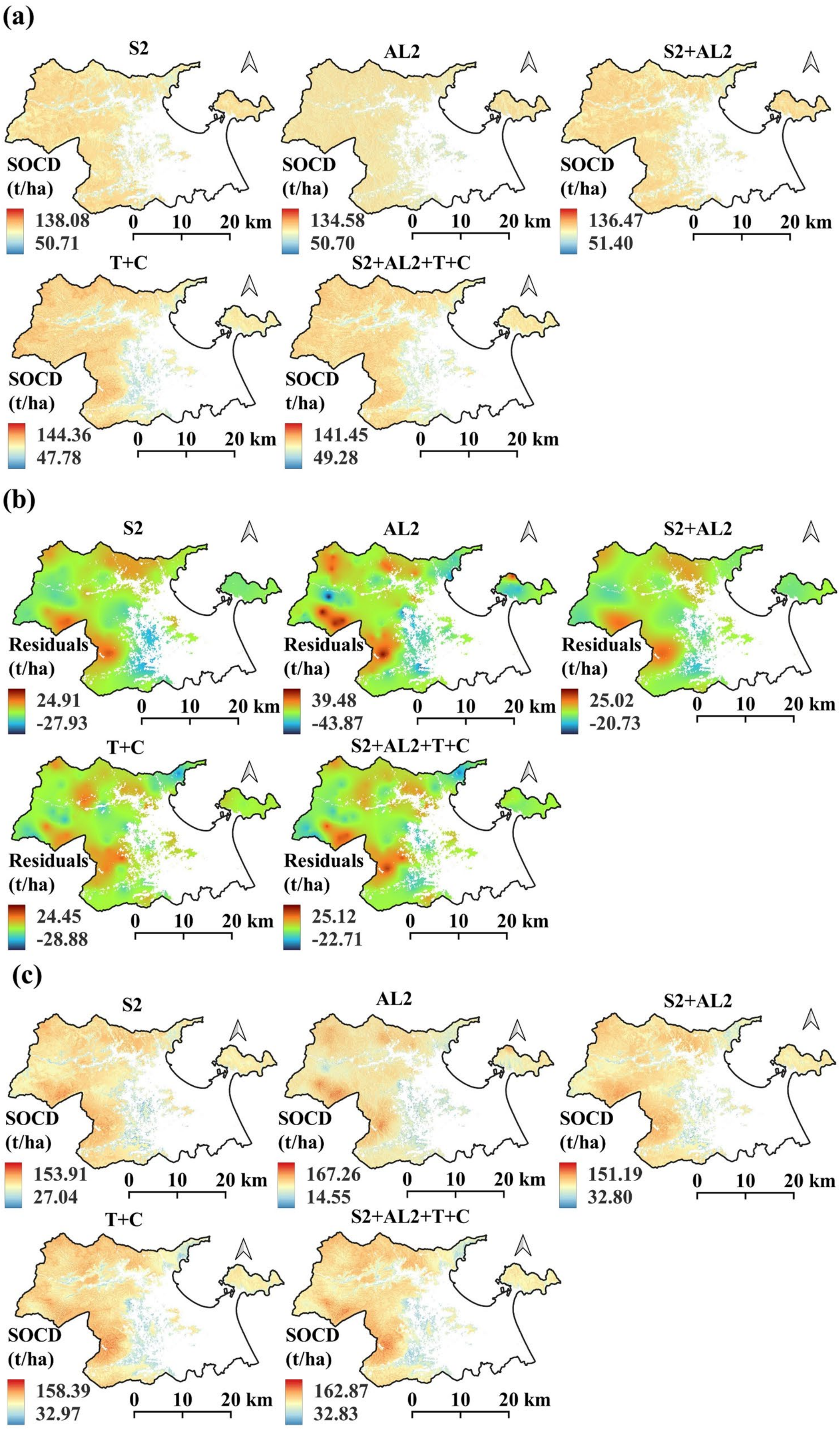
Based on the optimal parameters derived from the best-fitted experimental semivariograms, the OK method was used to estimate the spatial patterns of the Box-Cox transformed residuals of forest SOCD from five RF models, which were then back-transformed to the original scale (Fig. 6b). The range of SOCD residuals was – 27.93–24.91 t·ha^{-1} for Group One, –43.87–39.48 t·ha^{-1} for Group Two, – 20.73–25.02 t·ha^{-1} for Group Three, – 28.88–24.45 t·ha^{-1} for Group Four, and – 22.71–25.12 t·ha^{-1} for Group Five. Comparing absolute values of residuals, it was noted that increasing the number of data sources resulted in decreasing the values of residuals. Moreover, all maps of residuals showed the overestimation of low SOCD values located close to non-forest areas in the low-altitude terrain and the underestimation of high SOCD values in the western, northern, and northwestern parts of the study area in high-altitude terrain.

Evaluation of random forest regression kriging models

Based on Eq. 5, the predicted SOCD values from five RFRK models were calculated. Table 5 presents the accuracy metrics for RFRK in estimating SOCD across the validation set for five data groups. Similar to RF, increased data sources also found better performance in RFRK. The results showed that RFRK using Group Two (AL2) obtained lower accuracy for estimating forest SOCD than that using Group One (S2). The values of R^2, MAE, and RMSE were 0.33, 14.87 t·ha^{-1}, and 18.50 t·ha^{-1} for the former and were 0.49, 13.14 t·ha^{-1}, and 16.11 t·ha^{-1} for the latter, respectively. Integrating S2 and AL2 data in Group Three improved the prediction accuracy of RFRK, resulting in R^2 of 0.61, MAE of 11.45 t·ha^{-1}, and RMSE of 14.14 t·ha^{-1}, compared to using single-source RS data. The prediction accuracy of RFRK continued to elevate considerably after the addition of topographic and climatic variables to SAR-MSI data fusion in Group Five, with R^2 increased to 0.78, MAE decreased to 8.28 t·ha^{-1}, and RMSE decreased to 10.54 t·ha^{-1}. Notably, the influence of topographic attributes and climatic factors was more significant than RS data fusion in forest SOCD prediction using RFRK, according to the comparison of Group Three (S2 + AL2) and Group Four (T + C).

The comparison between the RF and RFRK models demonstrates that incorporating the kriged residuals via OK to account for spatial autocorrelation resulted in higher accuracy than relying solely on environmental variables. The superior performance of RFRK over RF was visually confirmed through graphical analysis. Figure 7 presents scatter plots of observed versus predicted values from the model validation process, with different colors used to distinguish the models. The RFRK scatter points (blue) were consistently closer to the 1:1 trend line than the RF scatter points (grey) across all data groups. In addition, the results from Table 5 showed that all hybrid models had positive values of RI_{MAE}, RI_{RMSE}, RI_{R^2} simultaneously, indicating a relative improvement in the RFRK models compared to the RF models for predicting forest SOCD. The improvement in the RFRK models was notable, with RI_{R^2} ranging from 8.20 to 65.00%, RI_{MAE} ranging from 8.18 to 21.07%, and RI_{RMSE} ranging from 6.76 to 18.18%. Among the five data groups, the most noticeable improvement in accuracy was observed in the RFRK models using single-source RS data, with RI_{R^2} reaching 28.95% in Group One and 65.00% in Group Two. Although the extent of improvements varied, RFRK based

Fig. 6 The SOCD maps by RF (**a**), residuals maps by OK (**b**), and SOCD maps by RFRK (**c**), based on five data groups



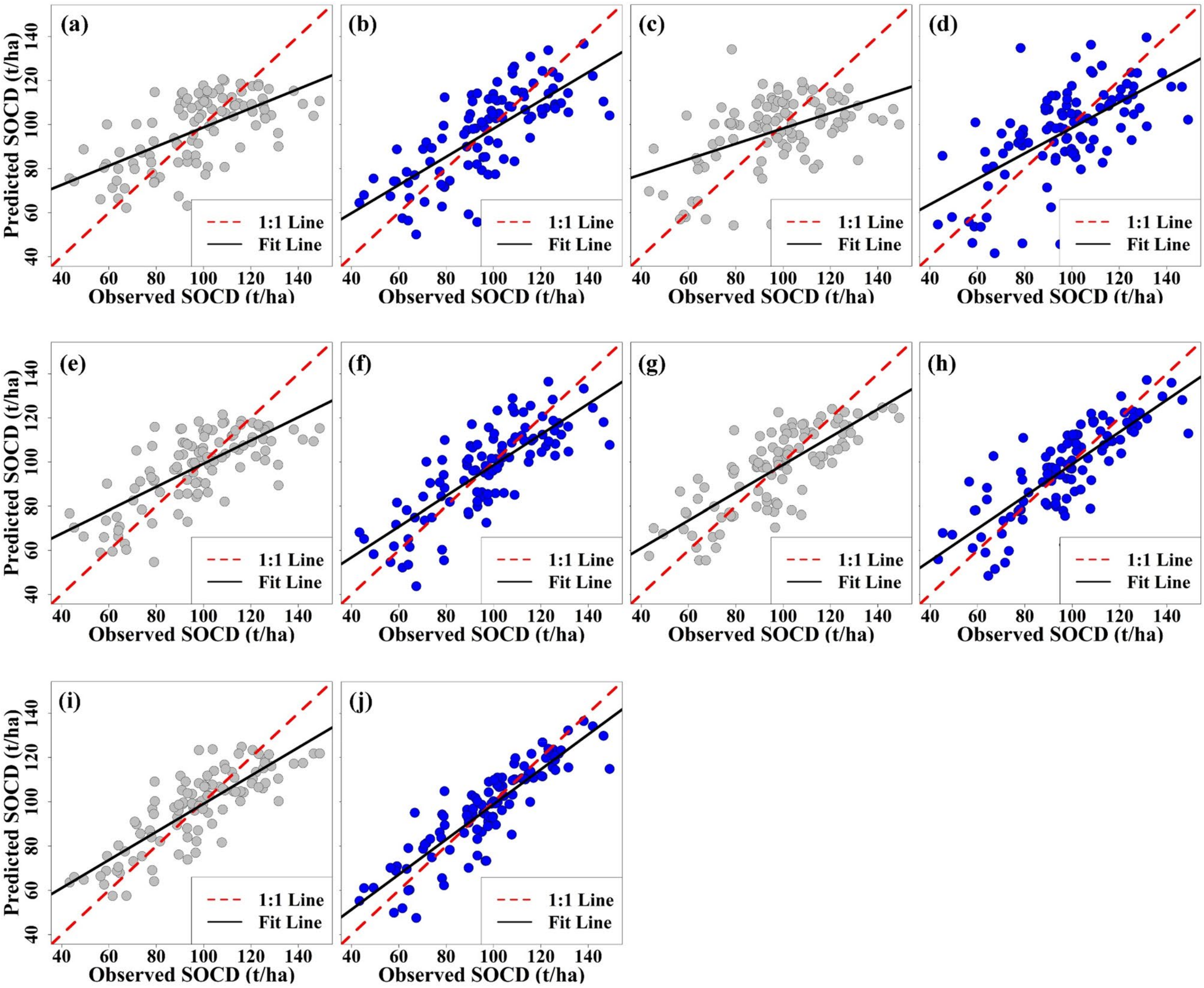


Fig. 7 Scatter plots of predicted versus observed SOCD from validation data by RF (grey dots) and RFRK (blue dots) based on Group One (**a**, **b**), Group Two (**c**, **d**), Group Three (**e**, **f**), Group Four (**g**, **h**), and Group Five (**i**, **j**)

on all environmental variables provided the highest accuracy for estimating forest SOCD in our study.

Spatial distribution of SOCD

The distribution of SOCD based on the RFRK models (Fig. 6c) was acquired by combining Fig. 6a and b. Five maps had a similar distribution of SOCD with the RFRK models, while values changed a lot. Forest SOCD maps generated by the RFRK models showed a clumped distribution compared to those produced by the RF models. By combining the interpolated values of residuals, the spatial distribution of forest SOCD was smoothed, and spatial randomness was reduced. Apart from accuracy evaluation, the generalization capability of the models was also important. The variability of SOCD values in the predicted maps could reflect the adaptability of the models to new samples to some extent. The range of SOCD predictions obtained from the RF models was 50.71–138.08 t·ha^{-1} for Group One, 50.70–134.58 t·ha^{-1} for Group Two, 51.40–136.47 t·ha^{-1} for Group Three, 47.78–144.36 t·ha^{-1} for Group Four, and 49.28–141.45 t·ha^{-1} for Group Five. The range for SOCD from the RFRK models was 27.04–153.91 t·ha^{-1} for Group One, 14.55–167.26 t·ha^{-1} for Group Two, 32.80–151.19 t·ha^{-1} for Group Three, 32.97–158.39 t·ha^{-1} for Group Four, and 32.83–162.87 t·ha^{-1} for Group Five. The ranges of predicted SOCD using the RFRK models were more extensive than those using the RF models, showing that RFRK had enhanced generalization capabilities compared to RF. Moreover, the RFRK maps showed that the spatial distribution of SOCD in forest areas of the study site gradually decreased from west to east, roughly in line with the topographic features. The low SOCD was majorly distributed in the eastern and southeastern parts of our study site because of lower vegetation coverage and intensive human activities. In contrast, high SOCD areas mainly appeared in the western, northern, and northwestern parts of the study site, where a large area of the primary forests was under the protection and conservation of the Danang Department of Forest Protection.

Discussion

Influence of multi-source environmental data on random forest regression kriging for predicting forest SOCD

The comparison of prediction accuracy in this study revealed that choosing environmental covariates from various data sources and their combinations greatly impacted the performance of the predictive model for SOCD estimation in forest landscapes. This finding is in line with the result of Zhou et al. (2020b), who reported that the prediction accuracy of predictive models in estimating forest SOCD could be susceptible to the selection of different data sources. This comparable result was also supported by Shafizadeh-Moghadam et al. (2022). Our study shows that selecting an optimal data combination could lead to an increase in predictive power of RFRK as high as up to 136.36% in R^2 values, according to comparing Group Two (AL2) and Group Five (S2 + AL2 + T + C). In addition, better performance in RFRK was found with increasing data sources.

The influence of different data groups on the prediction accuracy of RFRK for forest SOCD estimation was illustrated through attribute importance analysis (Fig. 4) and the implementation of the RFRK models (Table 5). RFRK based on AL2-derived variables demonstrated less accuracy in predicting SOCD than that based on S2-derived variables, according to our findings, caused by the saturation problem of SAR backscatters in forested areas. Within forest ecosystems, soils are typically obscured by dense vegetation, which poses challenges for the application of RS in soil mapping, as sensors cannot directly detect soil surfaces (Zhou et al. 2020b). Given this background, many previous DSM studies for SOCD mapping have incorporated vegetation indices derived from MSI imagery and backscatter data from SAR imagery into soil prediction models for vegetated areas (Odebiri et al. 2020; Sothe et al. 2022; Shafizadeh-Moghadam et al. 2022; Kumar et al. 2022). Theoretically, vegetation structural parameters, such as biomass, basal area, tree density, tree height, and trunk diameter, are particularly susceptible to the capability of SAR backscatters in vegetation estimation (Joshi et al. 2017). As these vegetation parameters increase, SAR backscatter signals tend to reach their saturation point more quickly, limiting the effectiveness of SAR data in forest landscapes (Yunjin Kim and van Zyl 2001). Contrary to SAR backscatters, many MSI vegetation indices are sensitive to canopy structure variations over all of the forested biomes with minimal saturation problems (Huete et al. 1997). An improved accuracy of RFRK was observed when S2 MSI variables were combined with L-band AL2 variables. This was not surprising as more helpful information related to the horizontal vegetation structures captured by the MSI sensor (Wood et al. 2012) and the vertical vegetation structures captured by the SAR sensor (Treuhaft and Siqueira 2000; Tello et al. 2018) was aggregated to estimate forest SOCD. Besides, this RS data fusion can harness the spatial and temporal advantages of different sensors, rather than relying solely on a single sensor, to address the saturation problem in vegetation estimation (Mutanga et al. 2023). Given the lack of studies integrating S2 MSI and L-band AL2 imageries for forest SOCD estimation, we used Landsat 8 imagery as a commonly employed alternative to S2 imagery for evaluating the benefits of combining MSI with L-band SAR. Similar to our study, Ceddia et al. (2017) found that under forest coverage, the integration of vegetation indices derived from Landsat 8 and ALOS PALSAR backscattering coefficients improved the prediction accuracy of soil carbon stock. Wang et al. (2020c) also examined the fusion of Landsat 8 and ALOS PALSAR data in predicting SOC content in entire Spain land covers, including forest land, and pointed out that the integrated approach held better prediction accuracy than only the Landsat 8 image and only the ALOS PALSAR image.

Likewise, an accuracy improvement was observed in RFRK when topographic and climatic variables were integrated into the fusion of S2 and AL2 variables. It was revealed that the limits of using the dual-source RS data could be significantly reduced by the addition of topographic and climatic data, resulting in a greater accuracy of RFRK for SOCD prediction in forested areas. In other words, topography and climate were more effective than RS data for predicting forest SOCD with RFRK in our study site. This is attributed to stronger influences of topographic and climatic variables on the spatial variability of forest SOCD compared to RS data and RS data fusion in our study site (Fig. 4). According to Mulder et al. (2011), in densely vegetated areas, soil properties, including SOC, mineralogy, texture, soil iron, soil moisture, soil salinity and carbonate content typically rely on indirect retrievals using soil indicators, such as plant functional groups and productivity changes, rather than being assessed through the use of RS data. As two of the five soil-forming factors (Jenny 1994), topography and climate exert direct influences on these soil indicators at the regional scale (Lamichhane et al. 2019), which in turn affect the variation in SOC content. That is to say, topography affects the water balance by promoting lateral water exchange through surface runoff, which helps to create areas with favorable moisture conditions for vegetation growth and subsequent accumulation of plant litter but promotes soil erosion by moving soil material to different places, thereby altering the patterns of SOC storage (Schwanghart and Jarmer 2011). Similarly, climate determines not only the nature and intensity of the weathering of parent materials that occur over large geographic areas but also net primary productivity, which is more significant to SOC accumulation (Lamichhane et al. 2019). Specifically, temperatures ranging from 10 to 30 °C generally facilitate the decomposition process of soil organic matter due to strong microbial activity, thereby accounting for a decrease of SOC stock at the surface soil layer (Rasel et al. 2017; Wang et al. 2022; Kumar et al. 2022). Precipitation not only encourages plant growth, resulting in SOC enrichment of the topsoil through the decomposition of leaves, branches, and other organic materials (Leff et al. 2012; Huang and Zhang 2016), but it also causes the soil to absorb soil organic matter from the litter layer (Cleveland et al. 2006). The greater effectiveness of topography and climate compared to RS data in estimating SOCD in mountainous forests, as shown in our results, aligns with findings from previous studies (Wang et al. 2017, 2018; Zhou et al. 2020b). Contrary to our study, the result of Wang et al. (2020a, 2023) considered that SOC stocks were primarily determined by MSI variables, followed by topographic and climatic variables in forest ecosystems of dense natural vegetation. The discrepancy is possibly due to the local variations in SOC dynamics, where the dominant driving factors likely differ between study sites (Lamichhane et al. 2019).

Enhancement of random forest regression kriging over random forest

The results in Table 5 demonstrate that RFRK improved the prediction accuracy by incorporating interpolated values of residuals to RF in all data groups. Theoretically, the RF algorithm can effectively exploit the non-linear relationship between the response variable and the predictor covariates (Lamichhane et al. 2019). However, the implementation of this tree-based ML model for spatial estimation is a challenging task, as it assumes that the data are independent and spatially uncorrelated (Erdogan Erten et al. 2022). The detection of the spatial autocorrelation structures of the residuals through semivariogram analysis in Table 6 confirmed that the RF models could not adequately extract structured information of forest SOCD in our study. According to Guo et al. (2015), if the ML residuals demonstrate spatial autocorrelation, then the performance of the ML model could be enhanced by interpolating residuals using the kriging method and then adding these kriged residuals back to the ML prediction. Thus, the accuracy improvement of the RFRK models over the RF models in this study for forest SOCD estimation can be explained by the successful integration of the predictive power of the tree-based ML approach in exploiting non-linear relationships and the capability of geostatistics in accounting for unexplained information in residuals component. The outperformance of this two-step hybrid approach over RF in our findings is consistent with previous studies (Guo et al. 2015; Tziachris et al. 2019; Silatsa et al. 2020).

The relative improvement values of the RFRK models compared to the RF models were notable (Table 5). The notable enhancement is attributed to the strong to moderate spatial autocorrelation of residuals from the RF models (Silatsa et al. 2020), indicated by the N/S value of the experimental semivariogram models (Table 6). Two main reasons can explain those high spatial autocorrelations in our study. Firstly, in areas with complex topography (e.g. mountainous regions) with little impact of anthropogenic activities, SOCD often shows strong spatial autocorrelation (Cambardella et al. 1994; Long et al. 2018; Yao et al. 2020). This spatial structure was passed through the RF models to the residuals of SOCD due to the RF's limitation, causing the residuals to still retain considerable spatial autocorrelation in space. Secondly, the number of sampling points in our study was likely sufficient for the OK interpolation. Several studies have reported that the size of sampling points will affect the accuracy of the kriging interpolation (Zhu and Lin 2010; Yao et al. 2020). A large enough number of sampling points is vital for detecting underlying spatial autocorrelation structures by reducing the sampling density (Brus and Heuvelink 2007; Dlugoß et al. 2010). According to Webster and Oliver (1992), at least 100 data points are required to estimate reliable semivariograms. In this study, 104 data points were utilized to assess the spatial autocorrelation of the residuals, ensuring the robustness of the spatial analysis. On the one hand, the notable enhancement of RFRK over RF in our study is consistent with the results of Guo et al. (2015) and Tziachris et al. (2019). On the other hand, our finding is contrary to that of Silatsa et al. (2020), who pointed out a limited increase in the accuracy of RFRK. The contradictory result can be explained by the much lower point data spacing density for SOCD estimation at the national scale in the study of Silatsa et al. (2020), which caused the poor capture of spatial autocorrelation in the semivariogram analysis. Besides, compared to RF, the relative improvement of RFRK using single-source RS data was more noticeable in prediction accuracy than that using multi-source environmental data. It was revealed that RFRK can significantly lower the limitations of using single-sensor RS data in RF for forest SOCD prediction.

Uncertainties and limitations

This study had several potential uncertainties, even though the RFRK models performed well in predicting the spatial distribution of forest SOCD. Firstly, some sampling locations randomly predetermined with a 2.5 × 2.5 km grid could not be approached due to the obstructions of surface dense vegetation and complex terrain in the forests, so there might be sampling errors. Secondly, the bilinear interpolation approach was used to resample environmental variables to a spatial resolution of 10 m, which would cause method errors. Thirdly, the environmental variables included as input for RFRK were acquired from various data sources, possibly contributing to modeling errors. Fourthly, it is worth noting that the upper 30 cm soil layer is known to contribute to about 50% of the total SOC in the top 100 cm of the soil (Kumar et al. 2022). However, our study focused exclusively on estimating the SOC in the topsoil (0–30 cm) of the forest, possibly leading to the underestimation issue. Finally, the used RS data encountered issues that need to be considered for future research. These included: (1) The cloud-free S2 mosaic contained inconsistency in atmospheric conditions due to different image capture times; (2) The average tree height of 17.5 m in the study site (Huy et al. 2016) could increase shadow distribution within each pixel, negatively effect on variables derived from RS imageries (Alavipanah et al. 2022); (3) In forest areas, AL2 data was influenced by the topography of the surface beneath the vegetation (Van Zyl 1992).

Conclusions

We applied the RF plus residuals kriging approach to predict SOCD in Central Vietnamese forests using multi-source environmental data. The following main conclusions can be drawn from this study: (1) The selection of environmental covariates affected the performance of RFRK. The prediction accuracy of RFRK increased with the inclusion of additional environmental data sources, with RFRK based on all data sources (S2 + AL2 + T + C) achieving high accuracy ($R^2 = 0.78$, MAE = 8.28 $t \cdot ha^{-1}$, and RMSE = 10.54 $t \cdot ha^{-1}$). (2) The RFRK models consistently outperformed the RF models in forest SOCD prediction across five data groups, with a notable relative improvement. Our study highlighted the potential of combining the strengths of the RF algorithm and the OK technique based on environmental covariates derived from S2, AL2, topographic, and climatic data to create a reliable framework for accurately analyzing the spatial distribution of SOCD in forest ecosystems.

Acknowledgements The authors thank the staff of the Danang Forest Protection Department and the colleagues of Hue University of Agriculture and Forestry who participated in the field surveys and data collection. We also want to show our gratitude to Prof. Dr. Morihiro Maeda and Prof. Dr. Kazuto Shima for allowing us to use the laboratories at the Okayama University, Japan to conduct experiments.

Author contributions All authors contributed to the study conception and design. All authors read and approved the final manuscript.

Funding Open Access funding provided by Okayama University. No funding was received to assist with the preparation of this manuscript.

Data availability The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Ethical approval The authors declare that this manuscript is not under consideration for publication anywhere else. The submitted work is original and has not been published previously in any form or language, either partially or in full. The results were presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

References

Abbaszad P, Asadzadeh F, Rezapour S et al (2024) Evaluation of Landsat 8 and Sentinel-2 vegetation indices to predict soil organic carbon using machine learning models. Model Earth Syst Environ 10:2581–2592. https://doi.org/10.1007/s40808-023-01916-x

Alavipanah SK, Karimi Firozjaei M, Sedighi A et al (2022) The Shadow Effect on Surface Biophysical variables derived from Remote sensing: a review. Land 11:2025. https://doi.org/10.3390/land11112025

Ameray A, Bergeron Y, Valeria O et al (2021) Forest carbon management: a review of silvicultural practices and management strategies across boreal, temperate and tropical forests. Curr Rep 7:245–266. https://doi.org/10.1007/s40725-021-00151-w

Asuero AG, Sayago A, González AG (2006) The correlation coefficient: an overview. Crit Rev Anal Chem 36:41–59. https://doi.org/10.1080/10408340500526766

Ayele GT, Demissie SS, Jemberrie MA et al (2019) Terrain effects on the spatial variability of soil physical and chemical properties. Soil Syst 4:1. https://doi.org/10.3390/soilsystems4010001

Balzter H (2001) Forest mapping and monitoring with interferometric synthetic aperture radar (InSAR). Prog Phys Geogr 25:159–177. https://doi.org/10.1191/030913301666986397

Benesty J, Chen J, Huang Y (2008) On the importance of the pearson correlation coefficient in noise reduction. IEEE Trans Audio Speech Lang Process 16:757–765. https://doi.org/10.1109/TASL.2008.919072

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Brownstein G, Steel JB, Porter S et al (2012) Chance in plant communities: a new approach to its measurement using the nugget from spatial autocorrelation. J Ecol 100:987–996. https://doi.org/10.1111/j.1365-2745.2012.01973.x

Brus DJ, Heuvelink GBM (2007) Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138:86–95. https://doi.org/10.1016/j.geoderma.2006.10.016

Cambardella CA, Moorman TB, Novak JM et al (1994) Field-scale variability of soil properties in central Iowa soils. Soil Sci Soc Am J 58:1501–1511. https://doi.org/10.2136/sssaj1994.03615995005800050033x

Camera C, Zomeni Z, Noller JS et al (2017) A high resolution map of soil types and physical properties for Cyprus: a digital soil mapping optimization. Geoderma 285:35–49. https://doi.org/10.1016/j.geoderma.2016.09.019

Ceddia M, Gomes A, Vasques G, Pinheiro É (2017) Soil carbon stock and particle size fractions in the Central Amazon predicted from remotely sensed relief, multispectral and radar data. Remote Sens 9:124. https://doi.org/10.3390/rs9020124

Cleveland CC, Reed SC, Townsend AR (2006) Nutrient regulation of organic matter decomposition in a tropical rain forest. Ecology 87:492–503. https://doi.org/10.1890/05-0525

Dlugoß V, Fiener P, Schneider K (2010) Layer-specific analysis and spatial prediction of Soil Organic Carbon using terrain attributes and Erosion modeling. Soil Sci Soc Am J 74:922–935. https://doi.org/10.2136/sssaj2009.0325

Don A, Schumacher J, Freibauer A (2011) Impact of tropical land-use change on soil organic carbon stocks - a meta-analysis. Glob Chang Biol 17:1658–1670. https://doi.org/10.1111/j.1365-2486.2010.02336.x

Drusch M, Del Bello U, Carlier S et al (2012) Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens Environ 120:25–36. https://doi.org/10.1016/j.rse.2011.11.026

Emadi M, Taghizadeh-Mehrjardi R, Cherati A et al (2020) Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. Remote Sens 12:2234. https://doi.org/10.3390/rs12142234

Erdogan Erten G, Yavuz M, Deutsch CV (2022) Combination of machine learning and kriging for spatial estimation of geological attributes. Nat Resour Res 31:191–213. https://doi.org/10.1007/s11053-021-10003-w

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol 37:4302–4315. https://doi.org/10.1002/joc.5086

Forkuor G, Hounkpatin OKL, Welp G, Thiel M (2017) High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: a comparison of machine learning and multiple Linear regression models. PLoS ONE 12:e0170478. https://doi.org/10.1371/journal.pone.0170478

Gitelson AA, Merzlyak MN (1998) Remote sensing of chlorophyll concentration in higher plant leaves. Adv Sp Res 22:689–692. https://doi.org/10.1016/S0273-1177(97)01133-2

Guan J-H, Deng L, Zhang J-G et al (2019) Soil organic carbon density and its driving factors in forest ecosystems across a northwestern province in China. Geoderma 352:1–12. https://doi.org/10.1016/j.geoderma.2019.05.035

Guo P-T, Li M-F, Luo W et al (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

Hoang Khanh Linh N, Van Chuong H (2015) Assessing the impact of urbanization on urban climate by remote satellite perspective: a case study in Danang city, Vietnam. Int Arch Photogramm Remote Sens Spat Inf Sci XL –7/W3:207–212. https://doi.org/10.5194/isprsarchives-XL-7-W3-207-2015

Hohn ME (1991) An introduction to applied geostatistics. Comput Geosci 17:471–473. https://doi.org/10.1016/0098-3004(91)90055-I

Huang L, Zhang Z (2016) Effect of rainfall pulses on plant growth and transpiration of two xerophytic shrubs in a revegetated desert area: Tengger Desert, China. CATENA 137:269–276. https://doi.org/10.1016/j.catena.2015.09.020

Huete A (1988) A soil-adjusted vegetation index (SAVI). Remote Sens Environ 25:295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Huete AR, HuiQing L, van Leeuwen WJD (1997) The use of vegetation indices in forested regions: issues of linearity and saturation. In: IGARSS'97. 1997 IEEE International Geoscience and Remote Sensing Symposium Proceedings. Remote Sensing - A Scientific Vision for Sustainable Development. IEEE, pp 1966–1968

Huy B, Poudel KP, Temesgen H (2016) Aboveground biomass equations for evergreen broadleaf forests in South Central Coastal ecoregion of Viet Nam: selection of eco-regional or pantropical models. Ecol Manage 376:276–283. https://doi.org/10.1016/j.foreco.2016.06.031

IPCC (2006) 2006 IPCC guidelines for national greenhouse gas inventories. In Eggleston HS, Buendia L, Miwa K, Ngara T and Tanabe K (eds) Prepared by the National Greenhouse Gas Inventories Programme. Japan

Jackson RB, Lajtha K, Crow SE et al (2017) The Ecology of Soil Carbon: pools, vulnerabilities, and biotic and abiotic controls. Annu Rev Ecol Evol Syst 48:419–445. https://doi.org/10.1146/annurev-ecolsys-112414-054234

Jenny H (1994) Factors of soil formation: a system of quantitative pedology. Courier Corporation

Jha N, Tripathi NK, Barbier N et al (2021) The real potential of current passive satellite data to map aboveground biomass in tropical forests. Remote Sens Ecol Conserv 7:504–520. https://doi.org/10.1002/rse2.203

John K, Abraham Isong I, Michael Kebonye N et al (2020) Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. Land 9:487. https://doi.org/10.3390/land9120487

Joshi N, Mitchard ETA, Brolly M et al (2017) Understanding 'saturation' of radar signals over forests. Sci Rep 7:3505. https://doi.org/10.1038/s41598-017-03469-3

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

Kim Y, van Zyl J (2001) Comparison of forest parameter estimation techniques using SAR data. In: IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217). IEEE, pp 1395–1397

Kravchenko A, Bullock DG (1999) A comparative study of interpolation methods for Mapping Soil Properties. Agron J 91:393–400. https://doi.org/10.2134/agronj1999.00021962009100030007x

Kumar P, Sajjad H, Tripathy BR et al (2018) Prediction of spatial soil organic carbon distribution using Sentinel-2A and field inventory data in Sariska Tiger Reserve. Nat Hazards 90:693–704. https://doi.org/10.1007/s11069-017-3062-5

Kumar M, Kumar A, Thakur TK et al (2022) Soil organic carbon estimation along an altitudinal gradient of chir pine forests in the Garhwal Himalaya, India: a field inventory to remote sensing approach. L Degrad Dev 33:3387–3400. https://doi.org/10.1002/ldr.4393

Lal R (2016) Soil health and carbon management. Food Energy Secur 5:212–222. https://doi.org/10.1002/fes3.96

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

Lausch A, Baade J, Bannehr L et al (2019) Linking Remote Sensing and Geodiversity and their traits relevant to Biodiversity—Part I: soil characteristics. Remote Sens 11:2356. https://doi.org/10.3390/rs11202356

Leff JW, Wieder WR, Taylor PG et al (2012) Experimental litterfall manipulation drives large and rapid changes in soil carbon cycling in a wet tropical forest. Glob Chang Biol 18:2969–2979. https://doi.org/10.1111/j.1365-2486.2012.02749.x

Li Z, Bi S, Hao S, Cui Y (2022) Aboveground biomass estimation in forests with random forest and Monte Carlo-based uncertainty analysis. Ecol Indic 142:109246. https://doi.org/10.1016/j.ecolind.2022.109246

Liu Z, Deng Z, Davis SJ et al (2022) Monitoring global carbon emissions in 2021. Nat Rev Earth Environ 3:217–219. https://doi.org/10.1038/s43017-022-00285-w

Long J, Liu Y, Xing S et al (2018) Effects of sampling density on interpolation accuracy for farmland soil organic matter concentration in a large region of complex topography. Ecol Indic 93:562–571. https://doi.org/10.1016/j.ecolind.2018.05.044

Luo Z, Feng W, Luo Y et al (2017) Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. Glob Chang Biol 23:4430–4439. https://doi.org/10.1111/gcb.13767

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

Matinfar HR, Maghsodi Z, Mousavi SR, Rahmani A (2021) Evaluation and prediction of Topsoil organic carbon using machine learning and hybrid models at a field-scale. CATENA 202:105258. https://doi.org/10.1016/j.catena.2021.105258

Matsushita B, Yang W, Chen J et al (2007) Sensitivity of the enhanced Vegetation Index (EVI) and normalized difference Vegetation Index (NDVI) to Topographic effects: a case study in high-density Cypress Forest. Sensors 7:2636–2651. https://doi.org/10.3390/s7112636

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

Meul M, Van Meirvenne M (2003) Kriging soil texture under different types of nonstationarity. Geoderma 112:217–233. https://doi.org/10.1016/S0016-7061(02)00308-7

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 VL, de Bruin S, Schaepman ME, Mayr TR (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

Mutanga O, Masenyama A, Sibanda M (2023) Spectral saturation in the remote sensing of high-density vegetation traits: a systematic review of progress, challenges, and prospects. ISPRS J Photogramm Remote Sens 198:297–309. https://doi.org/10.1016/j.isprsjprs.2023.03.010

National Institute of Agricultural Planning and Projection of Vietnam (2005) Soil map of Danang city. National Institute of Agricultural Planning and Projection of Vietnam, Hanoi

Nellis MD, Briggs JM (1992) Transformed Vegetation Index for measuring spatial variation in Drought Impacted Biomass on Konza Prairie, Kansas. Trans Kans Acad Sci 95:93. https://doi.org/10.2307/3628024

Nembrini S, König IR, Wright MN (2018) The revival of the Gini importance? Bioinformatics 34:3711–3718. https://doi.org/10.1093/bioinformatics/bty373

Nyamekye C, Kwofie S, Agyapong E et al (2021) Integrating support vector machine and cellular automata for modelling land cover change in the tropical rainforest under equatorial climate in Ghana. Curr Res Environ Sustain 3:100052. https://doi.org/10.1016/j.crsust.2021.100052

Odebiri O, Mutanga O, Odindi J et al (2020) Predicting soil organic carbon stocks under commercial forest plantations in KwaZulu-Natal province, South Africa using remotely sensed data. GIScience Remote Sens 57:450–463. https://doi.org/10.1080/15481603.2020.1731108

Pearson TRH, Brown SL, Birdsey RA (2007) Measurement guidelines for the sequestration of forest carbon. U.S. Department of Agriculture, Forest Service, Northern Research Station

Pouladi N, Møller AB, Tabatabai S, Greve MH (2019) Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342:85–92. https://doi.org/10.1016/j.geoderma.2019.02.019

Qi J, Chehbouni A, Huete AR et al (1994) A modified soil adjusted vegetation index. Remote Sens Environ 48:119–126. https://doi.org/10.1016/0034-4257(94)90134-1

Radočaj D, Gašparović M, Jurišić M (2024) Open remote sensing data in digital soil organic carbon mapping: a review. Agriculture 14:1005. https://doi.org/10.3390/agriculture14071005

Rasel SMM, Groen TA, Hussin YA, Diti IJ (2017) Proxies for soil organic carbon derived from remote sensing. Int J Appl Earth Obs Geoinf 59:157–166. https://doi.org/10.1016/j.jag.2017.03.004

Richardsons AJ, Wiegand A (1977) Distinguishing vegetation from soil background information. Photogramm Eng Remote Sens 43:1541–1552

Satdichanh M, Dossa GGO, Yan K et al (2023) Drivers of soil organic carbon stock during tropical forest succession. J Ecol 111:1722–1734. https://doi.org/10.1111/1365-2745.14141

Scharlemann JPW, Tanner EVJ, Hiederer R, Kapos V (2014) Global soil carbon: understanding and managing the largest terrestrial carbon pool. Carbon Manag 5:81–91. https://doi.org/10.4155/cmt.13.77

Schwanghart W, Jarmer T (2011) Linking spatial patterns of soil organic carbon to topography—a case study from south-eastern Spain. Geomorphology 126:252–263. https://doi.org/10.1016/j.geomorph.2010.11.008

Shafizadeh-Moghadam H, Minaei F, Talebi-khiyavi H et al (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

Shimada M, Isoguchi O, Tadono T, Isono K (2009) PALSAR Radiometric and geometric calibration. IEEE Trans Geosci Remote Sens 47:3915–3932. https://doi.org/10.1109/TGRS.2009.2023909

Silatsa FBT, Yemefack M, Tabi FO et al (2020) Assessing countrywide soil organic carbon stock using hybrid machine learning modelling and legacy soil data in Cameroon. Geoderma 367:114260. https://doi.org/10.1016/j.geoderma.2020.114260

Sothe C, Gonsamo A, Arabian J, Snider J (2022) Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. Geoderma 405:115402. https://doi.org/10.1016/j.geoderma.2021.115402

Suleymanov A, Tuktarova I, Belan L et al (2023) Spatial prediction of soil properties using random forest, k-nearest neighbors and cubist approaches in the foothills of the Ural Mountains, Russia. Model Earth Syst Environ 9:3461–3471. https://doi.org/10.1007/s40808-023-01723-4

Tello M, Cazcarra-Bes V, Pardini M, Papathanassiou K (2018) Forest structure characterization from SAR Tomography at L-Band. IEEE J Sel Top Appl Earth Obs Remote Sens 11:3402–3414. https://doi.org/10.1109/JSTARS.2018.2859050

Treuhaft RN, Siqueira PR (2000) Vertical structure of vegetated land surfaces from interferometric and polarimetric radar. Radio Sci 35:141–177. https://doi.org/10.1029/1999RS900108

Truong VT, Hoang TT, Cao DP et al (2019) JAXA Annual Forest Cover maps for Vietnam during 2015–2018 using ALOS-2/PALSAR-2 and Auxiliary Data. Remote Sens 11:2412. https://doi.org/10.3390/rs11202412

Tucker CJ (1979) Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ 8:127–150. https://doi.org/10.1016/0034-4257(79)90013-0

Tziachris P, Aschonitis V, Chatzistathis T, Papadopoulou M (2019) Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. CATENA 174:206–216. https://doi.org/10.1016/j.catena.2018.11.010

Vågen T-G, Winowiecki LA (2013) Mapping of soil organic carbon stocks for spatially explicit assessments of climate change mitigation potential. Environ Res Lett 8:015011. https://doi.org/10.1088/1748-9326/8/1/015011

Van Zyl JJ (1992) The Effect of Topography on Radar Scattering from Vegetated Areas. In: [Proceedings] IGARSS '92 International Geoscience and Remote Sensing Symposium. IEEE, pp 1132–1134

Vatandaşlar C, Abdikan S (2022) Carbon stock estimation by dual-polarized synthetic aperture radar (SAR) and forest inventory data in a Mediterranean forest landscape. J Res 33:827–838. https://doi.org/10.1007/s11676-021-01363-3

Veronesi F, Schillaci C (2019) Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecol Indic 101:1032–1044. https://doi.org/10.1016/j.ecolind.2019.02.026

Wang S, Zhuang Q, Wang Q et al (2017) Mapping stocks of soil organic carbon and soil total nitrogen in Liaoning Province of China. Geoderma 305:250–263. https://doi.org/10.1016/j.geoderma.2017.05.048

Wang S, Adhikari K, Wang Q et al (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 Indic 84:263–272. https://doi.org/10.1016/j.ecolind.2017.08.046

Wang S, Gao J, Zhuang Q et al (2020a) Multispectral remote sensing data are effective and robust in mapping regional forest soil organic carbon stocks in a Northeast Forest Region in China. Remote Sens 12:393. https://doi.org/10.3390/rs12030393

Wang S, Zhuang Q, Jin X et al (2020b) Predicting Soil Organic Carbon and Soil Nitrogen Stocks in Topsoil of Forest Ecosystems in Northeastern China using Remote Sensing Data. Remote Sens 12:1115. https://doi.org/10.3390/rs12071115

Wang X, Zhang Y, Atkinson PM, Yao H (2020c) 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

Wang X, Li J, Xing G et al (2022) Soil Organic Carbon distribution, enzyme activities, and the temperature sensitivity of a tropical rainforest in Wuzhishan, Hainan Island. Forests 13:1943. https://doi.org/10.3390/f13111943

Wang T, Zhou W, Xiao J et al (2023) Soil Organic Carbon Prediction using Sentinel-2 data and Environmental Variables in a Karst Trough Valley Area of Southwest China. Remote Sens 15:2118. https://doi.org/10.3390/rs15082118

Webster R, Oliver MA (1992) Sample adequately to estimate variograms of soil properties. J Soil Sci 43:177–192. https://doi.org/10.1111/j.1365-2389.1992.tb00128.x

Webster R, Oliver MA (2007) Geostatistics for environmental scientists. Wiley

Wiesmeier M, Urbanski L, Hobley E et al (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

Wood EM, Pidgeon AM, Radeloff VC, Keuler NS (2012) Image texture as a remotely sensed measure of vegetation structure. Remote Sens Environ 121:516–526. https://doi.org/10.1016/j.rse.2012.01.003

Wu X, Washaya P, Liu L et al (2020) Rice Yield Estimation based on Spaceborne SAR: A Review from 1988 to 2018. IEEE Access 8:157462–157469. https://doi.org/10.1109/ACCESS.2020.3020182

Xia Y, McSweeney K, Wander MM (2022) Digital Mapping of Agricultural Soil Organic Carbon using soil forming factors: a review of current efforts at the Regional and National scales. Front Soil Sci 2:1–19. https://doi.org/10.3389/fsoil.2022.890437

Xu L, Shi Y, Fang H et al (2018) Vegetation carbon stocks driven by canopy density and forest age in subtropical forest ecosystems. Sci Total Environ 631–632:619–626. https://doi.org/10.1016/j.scitotenv.2018.03.080

Yang X, Xiao X, Qin Y et al (2021) Mapping forest in the southern Great Plains with ALOS-2 PALSAR-2 and Landsat 7/8 data. Int J Appl Earth Obs Geoinf 104:102578. https://doi.org/10.1016/j.jag.2021.102578

Yao X, Yu K, Deng Y et al (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

Yao X, Yu K, Deng Y et al (2020) Spatial variability of soil organic carbon and total nitrogen in the hilly red soil region of Southern China. J Res 31:2385–2394. https://doi.org/10.1007/s11676-019-01014-8

Yue J, Feng H, Yang G, Li Z (2018) A comparison of regression techniques for estimation of above-ground Winter Wheat Biomass using Near-Surface Spectroscopy. Remote Sens 10:66. https://doi.org/10.3390/rs10010066

Zhang R, Tang X, You S et al (2020) A Novel feature-level Fusion Framework using Optical and SAR Remote sensing images for Land Use/Land Cover (LULC) classification in cloudy mountainous area. Appl Sci 10:2928. https://doi.org/10.3390/app10082928

Zhao J, Xie H, Ma J, Wang K (2021) Integrated remote sensing and model approach for impact assessment of future climate change on the carbon budget of global forest ecosystems. Glob Planet Change 203:103542. https://doi.org/10.1016/j.gloplacha.2021.103542

Zhou Y, Hartemink AE, Shi Z et al (2019) Land use and climate change effects on soil organic carbon in North and Northeast China. Sci Total Environ 647:1230–1238. https://doi.org/10.1016/j.scitotenv.2018.08.016

Zhou T, Geng Y, Chen J et al (2020a) High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Sci Total Environ 729:138244. https://doi.org/10.1016/j.scitotenv.2020.138244

Zhou T, Geng Y, Chen J et al (2020b) Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China. Ecol Indic 114:106288. https://doi.org/10.1016/j.ecolind.2020.106288

Zhou Y, Zhao X, Guo X, Li Y (2022) Mapping of soil organic carbon using machine learning models: combination of optical and radar remote sensing data. Soil Sci Soc Am J 86:293–310. https://doi.org/10.1002/saj2.20371

Zhu Q, Lin HS (2010) Comparing ordinary Kriging and regression kriging for Soil properties in contrasting landscapes. Pedosphere 20:594–606. https://doi.org/10.1016/S1002-0160(10)60049-5

Zribi M, Muddu S, Bousbih S et al (2019) Analysis of L-Band SAR Data for Soil Moisture Estimations over Agricultural areas in the tropics. Remote Sens 11:1122. https://doi.org/10.3390/rs11091122

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.