Remote Sensing Applications: Society and Environment 40 (2025) 101736



Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

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



Estimating soil organic carbon using time series Band 11 (SWIR) of multispectral Sentinel-2 satellite images and machine learning algorithms

Mehdi Golkar Amoli [a], Mahdi Hasanlou [a,*], Farhad Samadzadegan [a], Ruhollah Taghizadeh-Mehrjardi [b], Farzaneh Dadrass Javan [c]

[a] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] Soil Science and Geomorphology, Department of Geoscience, University of Tübingen, Rümelinstrasse 19–23, 72070, Tübingen, Germany
[c] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, NB Enschede, 7522, the Netherlands

ARTICLE INFO

Keywords:
SOC
Time series
B11
XGBoost
Machine learning
PCA
Feature extraction

ABSTRACT

Soil Organic Carbon (SOC) is a critical soil property impacting food security and climate change. Traditional methods for SOC estimation are time-consuming, expensive, and unsuitable for large-scale application. Consequently, researchers have increasingly focused on utilizing Remote Sensing (RS) images for SOC estimation over the past two decades. However, achieving high SOC estimation accuracy (more than 80 %) remains challenging. This limitation often stems from a mismatch between the complexity of SOC and the information captured by traditional RS observations (e.g., reflectance bands or spectral indices), as conventional feature extraction methods from RS images may not be detailed enough to monitor the many factors influencing SOC concentration. One promising solution to enhance feature extraction is the use of time series observations, analyzing multiple images over time instead of relying on single-time images. This study proposes a novel approach leveraging time series of the Sentinel-2 satellite's B11 band (centered around 1610 nm, a region sensitive to SOC absorption features) along with Principal Component Analysis (PCA) and Independent Component Analysis (ICA) transformations to extract more meaningful temporal features. Specifically, ten new features based on temporal variations were derived by applying PCA and ICA to the B11 band time series images. These temporal features were then combined with features derived from the median of all Sentinel-2 images acquired during the summer of 2019, corresponding to the soil data collection period. Four machine learning algorithms (RF, GBRT, XGBoost, and LightGBM) were employed across four distinct scenarios to evaluate the novel feature extraction method and a feature selection algorithm. The scenarios were designed as follows: Scenario one (S#1) and Scenario two (S#2) did not utilize the time series features, while Scenario three (S#3) and Scenario four (S#4) did. A binary Genetic Algorithm (GA) for feature selection was implemented in S#2 and S#4, distinguishing them from S#1 and S#3 respectively. XGBoost performed best, achieving an R^2 of 0.891 in S#4 (time series features and GA). Incorporating time series features significantly improved accuracy by 0.11, while GA-based feature selection added another 0.05. The findings highlight the effectiveness of the developed feature extraction algorithm, using Sentinel-2's B11 time series and advanced transformations, for substantially improving SOC level estimation.

* Corresponding author.
E-mail addresses: golkaramoli@ut.ac.ir (M.G. Amoli), hasanlou@ut.ac.ir (M. Hasanlou), samadz@ut.ac.ir (F. Samadzadegan), ruhollah.taghizadeh-mehrjardi@mnf.uni-tuebingen.de (R. Taghizadeh-Mehrjardi), f.dadrassjavan@utwente.nl (F. Dadrass Javan).

https://doi.org/10.1016/j.rsase.2025.101736
Received 13 August 2024; Received in revised form 20 July 2025; Accepted 18 September 2025
Available online 22 September 2025
2352-9385/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Soil organic carbon (SOC) is one of the main parts of fertile soils. It constitutes a significant portion of Soil Organic Matter (SOM)(de Santana et al., 2019; Pouladi et al., 2023). This organic matter serves as a critical factor in enhancing agricultural productivity, offering numerous benefits such as increasing nutrient absorption, increasing water capacity retention, improving soil structure, and reducing soil erosion (Bagherifam et al., 2021; Jobbágy and Jackson, 2000; Melillo et al., 2002). There are two main ways to measure SOC: the traditional way and the Remote Sensing (RS) way(Li et al., 2025). The traditional way involves taking soil samples and testing them in a laboratory, which is time-consuming, expensive, and challenging to do over large areas. For this reason, RS has become more prevalent in recent years. RS uses the electromagnetic spectrum, a range of invisible and visible light. Scientists can detect certain chemical bonds in the soil, like those between carbon, hydrogen, and oxygen, by looking at specific wavelengths within the visible near-infrared (VNIR) and shortwave infrared (SWIR) range (Gholizadeh et al., 2021; Oliveira et al., 2024). These bonds create absorption peaks at specific wavelengths, which help RS scientists identify areas with higher SOC concentrations (Angelopoulou et al., 2019; Zeraatpisheh et al., 2021). Over the past two decades, the use of multispectral (MS) satellites for estimating SOC has gained significant attention, fueled by numerous research studies (Gomez et al., 2008; Luo et al., 2022). Among the key MS satellites employed in SOC estimation is the Sentinel-2 constellation, consisting of two satellites: Sentinel-2A and Sentinel-2B. Each satellite is equipped with a multispectral sensor instrument (MSI), capable of capturing imagery in 13 spectral bands within the VNIR and SWIR regions of the electromagnetic spectrum (Stromann et al., 2019; Xue et al., 2025). One notable advantage of the Sentinel-2 satellites is their frequent revisit rate, allowing for generating output products within a short interval, even as little as 5 day at the equator (He et al., 2021).

Recent research uses single-time MS satellite imagery (e.g., Landsat-8, Sentinel-2) alongside topographical and climatic data to estimate SOC concentration. However, these studies typically rely on single-time satellite images combined with various features extracted from climate and topographic data. This approach often misses the dynamic nature of SOC, and only a limited number of features in their large feature datasets have been found to have a significant relationship with SOC levels. Consequently, the accuracy of SOC estimation in these studies is often limited because the generated features can only partially capture the complex nature of SOC variation. Urbina-Salazar et al. Compared the impact of Sentinel-2 satellite images and topography parameters on estimating SOC. Sentinel-2 images showed higher accuracy ($R^2 = 0.46$) than topography parameters ($R^2 = 0.35$) (Urbina-Salazar et al., 2023). Dvorakova et al. utilized Sentinel-2 and APEX sensors to estimate SOC in a region with a wide range of SOC levels. The research aimed to examine the influence of crop residue on SOC estimation. Although the Cellulose Absorption Index (CAI) index partially differentiated crop residue from the soil, the model's overall accuracy with Sentinel-2 images was low ($R^2 = 0.49$) due to the limitations of the Partial Least Square Regression (PLSR) model and the relatively low contribution of SOC in the study area (Dvorakova et al., 2020). Zhang et al. utilized Sentinel-2 and MODIS images and various ML techniques to estimate SOM in the Songnen Plain, located in northern China. The findings revealed that Sentinel-2 demonstrated slightly higher accuracy ($R^2 = 0.67$) than MODIS ($R^2 = 0.63$) (Zhang et al., 2021).

Zhang et al. addressed SOC estimation in semi-arid agricultural and natural soils in central Iran, specifically focusing on the impacts of mining activities. While, their Support Vector Regression (SVR) model, integrating both Landsat-9 imagery and topography-based features, achieved an R^2 of 0.64 and confirmed mining's influence on SOC distribution (Zhang et al., 2024). Huang et al. used ground-based hyperspectral data with a single-time Sentinel-2 image and employed Iteratively Retaining Informative Variables (IRIV) for feature selection to estimate SOC in an oasis delta in Xinjiang, China, achieving an R^2 of 0.64 (Huang et al., 2025).

Due to the low accuracy of most studies that use single-time RS images, a practical solution is needed to increase the accuracy of SOC estimation. A highly effective approach to tackling this issue is using time series data from MS images. Using time series images compared to a single image in SOC estimation has many advantages, including (1) soil sample data gathering is a challenging and time-consuming process that typically spans several weeks or months. Therefore, single-time images cannot accurately capture the electromagnetic wave variations during data gathering (2). Single-date images may be affected by interference from crops or crop residues during the imaging (Dvorakova et al., 2020, 2023) (3) If no changes in cultivation type exist, SOC tends to have a moderately constant value. In addition, from a technical view, time series analysis enables the identification of anomalies in SOC concentrations that may not be apparent in a single image. By considering the temporal dimension, we can better account for the inherent soil variability that directly influences SOC concentration. Luo et al. utilized Sentinel-2 and Landsat-8 satellite images to estimate SOM from 2014 to 2020. They implemented Recursive Feature Elimination (RFE) and Boruta methods for feature selection, resulting in a 2 % accuracy improvement. Comparatively, the six-year time series of Landsat-8 images outperformed the one-year time series of Sentinel-2 images, achieving around 0.70 explanation accuracy (Luo et al., 2022). Gallo et al. evaluated Landsat-5 time series images for creating Bare Soil Composite Images (BSCI) to estimate clay content and SOM. Using the PLSR model, they achieved an R^2 value of 0.34 for SOM estimation (Gallo et al., 2018). Zhang et al. used a new architecture, combining Long-Short-Term-Memory (LSTM) and Convolutional Neural Network (CNN). The network incorporated the MODIS sensor's phenology product time series, topography parameters, and climate variables. The resulting accuracy was approximately $R^2 = 0.35$ (Zhang et al., 2022). Zeraatpisheh et al. used environmental and RS covariates to evaluate ML algorithms, including RF, Cubist, and Support Vector Regressor (SVR), in SOC estimation. They concluded that using a single-date RS covariate did not yield high accuracy. The study recommended incorporating a time series of RS images with climate and topography parameters to improve SOC estimation accuracy (Zeraatpisheh et al., 2022). Shafizadeh-Moghadam et al. compared the ability of the time series and single date images of Sentinel-2 in SOC estimation. Based on their results, multi-temporal data outperformed single-date data (Shafizadeh-Moghadam et al., 2022). Dvorakova et al. addresses the challenge of surface disturbances in SOC estimation by refining time-series composite methods for Sentinel-2 data. They specifically

focused on establishing thresholds (NDVI <0.25, NBR2 < 0.05) and spectral normalization to ensure the resulting 'exposed soil composite' more accurately represents optimal (dry, bare) soil conditions, aiming to improve upon less selective temporal compositing approaches and at least reach the RPD = 1.4.(Dvorakova et al., 2023). Luo et al. developed a method for optimizing SOM estimation in Northeast China by identifying the optimal time interval (20 days) and period (May, specifically DOY 120–140) for generating multi-year synthetic Sentinel-2 images. While using these optimized Sentinel-2 synthetic images alone yielded a top R^2 of 0.51, they demonstrated that incorporating environmental covariates, such as yearly average temperature, significantly improved prediction accuracy to $R^2 = 0.67$ (Luo et al., 2023). Zhao et al. investigated the utility of temporal composites from 850 DESIS hyperspectral satellite images for large-scale SOC estimation across Bavaria, Germany, employing deep learning models. They achieved an R^2 of 0.63 using only hyperspectral data, which improved to R^2 of 0.69 when topographical features (from DEM) were also incorporated (Zhao et al., 2025).

Leveraging raw time series data from satellite images offers valuable insights. However, extracting high-level features from these observations can achieve an even greater understanding of the SOC estimation problem. Feature extraction techniques transform the vast amount of spectral data from time series images into a smaller set of high-level and meaningful features. These features capture the significant trends and variations within the time series data. By focusing on these high-level features, ML models can be trained to achieve significantly higher accuracy in SOC estimation. This allows for estimating SOC levels based solely on RS observations and without using climate or topography data. Sun et al. used the Cross wavelet (XWT) transform as a feature extraction method from fractional vegetation and soil-related endmember time series data for SOM estimation. By analyzing these time series data using XWT, the researcher could extract features that capture the complex interplay between vegetation and soil type over time and achieved a great accuracy of $R^2 = 0.87$ with the RF algorithm (Sun et al., 2022). Golkar et al. used PCA/ICA transformations along with the HFC algorithm on PRISMA Hyperspectral images to improve the accuracy of SOC estimation. This approach increased the R^2 value from 0.65 to 0.82, demonstrating the promise of feature extraction methods in SOC estimation (Golkar Amoli et al., 2024). Pavlovic et al. performed a feature extraction strategy based on land use/cover classification of Sentinel-2 images using a U-Net image segmentation network. They used features extracted from a U-Net latent layer to train the ML model for SOC estimation. The ML model was trained to predict the SOC concentration based on the land cover information extracted by the U-Net's last layer and reached the accuracy of $R^2 = 0.73$ (Pavlovic et al., 2024). Saberioon et al. Leveraging the extensive LUCAS soil spectral library, this research demonstrated the power of deep learning for SOC estimation by employing a Stacked Autoencoder (SAE) for sophisticated feature extraction from Vis-NIR-SWIR spectra. Their SAE-1DCNN model, which utilized these learned deep features, achieved a notable R^2 of 0.78, outperforming models based on full spectra and showcasing an advanced method for spectral data processing in SOC prediction (Saberioon et al., 2024).

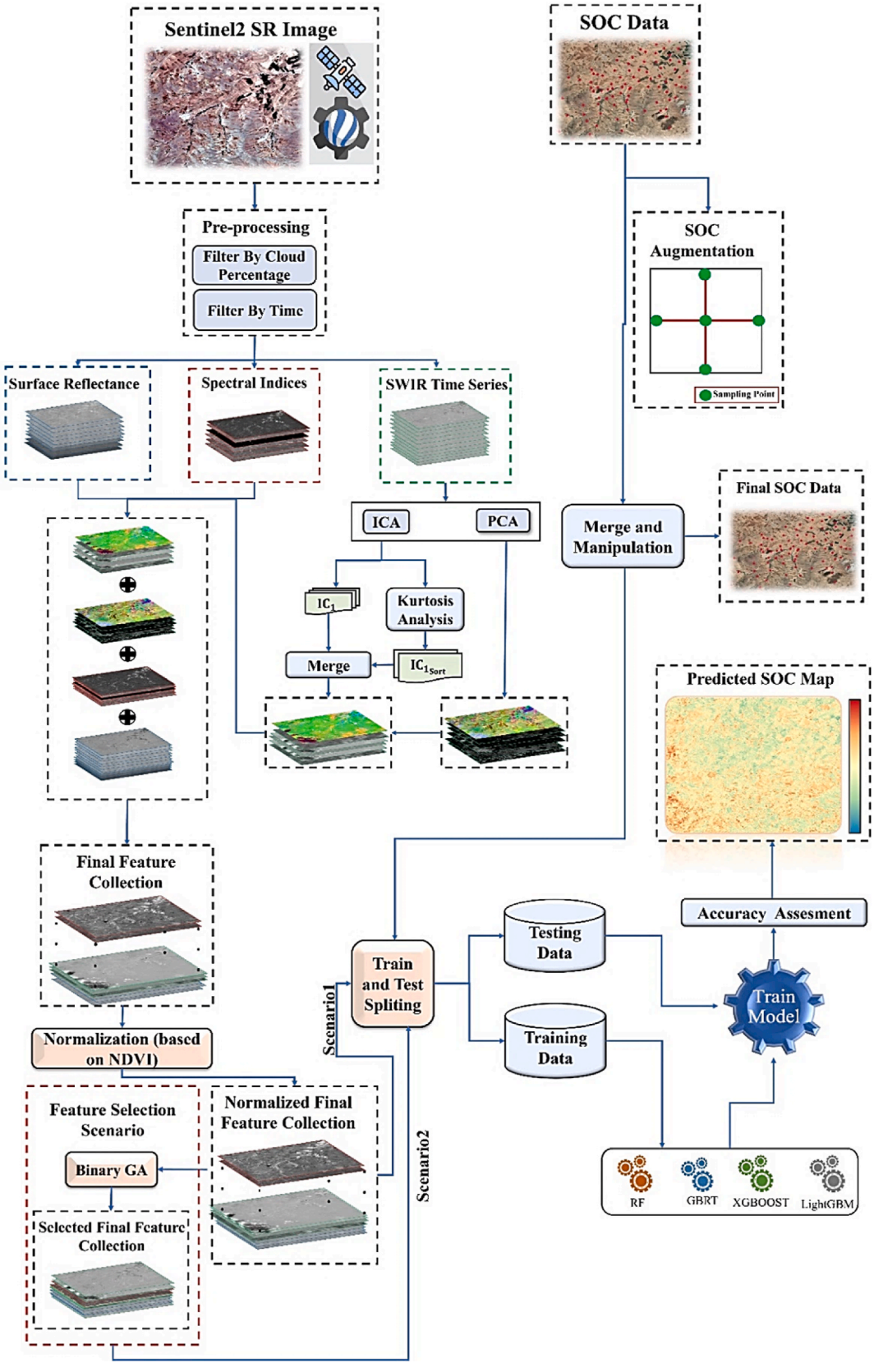
Previous studies show common RS methods for estimation often face limits, like using static images that miss SOC's temporal changes, or relying on broad, untargeted spectral bands. Many also need non-RS data for better accuracy. To overcome these issues, like needing extra data, static images, or unfocused bands, our main goal is to create an accurate SOC estimation method using only Sentinel-2 satellite images. We thus avoid other data like topography or climate. To work without these inputs, this study innovatively uses time-series data from Sentinel-2's Band B11. We apply dimensionality reduction (Principal Component Analysis, Independent Component Analysis) directly to the B11 time series to extract new, powerful temporal features. These newly created temporal features are central to our SOC estimation; in some modeling scenarios, they are used alongside standard Sentinel-2 features (like surface reflectance and spectral indices from a key summer period, as detailed in our methodology). Our evaluation centers on these B11 time-series features, as B11 is very sensitive to SOC. By using PCA/ICA on this data, plus feature selection, we aim to find the key features for SOC changes and greatly improve estimation's accuracy. The study has two main aims: first, to assess how B11 time-series features improve SOC estimation, and second, to find the most critical predictive features.

2. Materials and methods

Fig. 1 briefly overviews the proposed method applied to estimate the SOC in the study area. This method consists of six main sections: (1) soil sampling, (2) feature extraction from single-time images, (3) feature extraction from time series images, (4) incorporating feature selection, (5) model training and evaluation, and (6) spatially detailed SOC map generation.

2.1. Study area

The study area is located in Sanandaj city, Kurdistan province, Iran (Fig. 2), covering 102 km^2. Kurdistan province experiences the influence of two main air currents, hot and cold, which give rise to various climates. The province's western region, including Baneh City, receives the highest precipitation, averaging 800 mm annually. The central part of the province, Sanandaj, receives approximately 500 mm of precipitation per year. Due to the presence of the Zagros mountains in the study area, some elevation changes are observed. The maximum and minimum elevations in the study area are 2600 and 1840 m, respectively, with an average slope of about 6°. The study area maintains an average mean temperature of around 14°C, with minimum and maximum temperatures averaging around −6.5°C in January and 25°C in July, respectively. The study site encompasses six distinct soil groups, classified using the USDA Soil Taxonomy method. These soil groups include Haplarigids, Haplocarbids, Calixerepts, Haplocryepts, and Haploxerolls. Each group represents various soil characteristics and properties, such as sandy textures, carbonates (lime) content, lime accumulation horizons, clay accumulation, and mixed mineralogy. (Bockheim et al., 2014).



(caption on next page)

Fig. 1. The flowchart of the proposed SOC estimation method.

2.2. Soil sampling

A meticulous sampling approach is essential to accurately capture significant differences in soil properties across a wide geographical area. This involves selecting specific sampling locations based on anticipated variations in factors influencing soil formation, aiming to represent the diverse landscape characteristics within the study region. Using the constrained Latin hypercube sampling (CLHS) method with normalized difference vegetation index, wind effect, valley depth, Multiresolution index of Valley bottom flatness, normalized height, and topographic wetness index as input data, 91 soil samples were gathered from the 0–10 cm depth layer in the summer of 2019 (Wan et al., 2021). It is worth mentioning the soil samples were collected from wheat and pea farms right after the crops were harvested. This timing left the fields bare soil, making it easier for RS images to interact with the ground. To enhance the efficiency of ML algorithms due to the limited sample data, artificial points were generated around each sample in four main directions, 10 m apart. While these points maintained the same SOC levels as the central point, they exhibited different RS observations. This technique constitutes a form of data augmentation. After collection, the samples were air-dried at room temperature (approximately 25 °C), crushed, and sieved before undergoing analysis for SOC using the Walkley-Black method.

2.3. Sentinel-2 image datasets

2.3.1. Sentinel-2 single time image

The features used in this study are exclusively derived from Sentinel-2 satellite images. The Sentinel-2 satellite has the MSI sensor, which captures the Earth's surface in 13 spectral bands ranging from VNIR to SWIR (as shown in Table 1). During the soil sampling process in the summer of 2019, we generated a synthetic median image using all available surface reflectance images captured by Sentinel-2A and Sentinel-2B. This was made possible by the advanced technology of Google Earth Engine (GEE), a cloud computing infrastructure specifically designed for large-scale RS image processing (Zurqani, 2025). The synthetic median images were created using Level-2A images from Sentinel-2, which have undergone geometric and radiometric corrections to represent surface reflectance

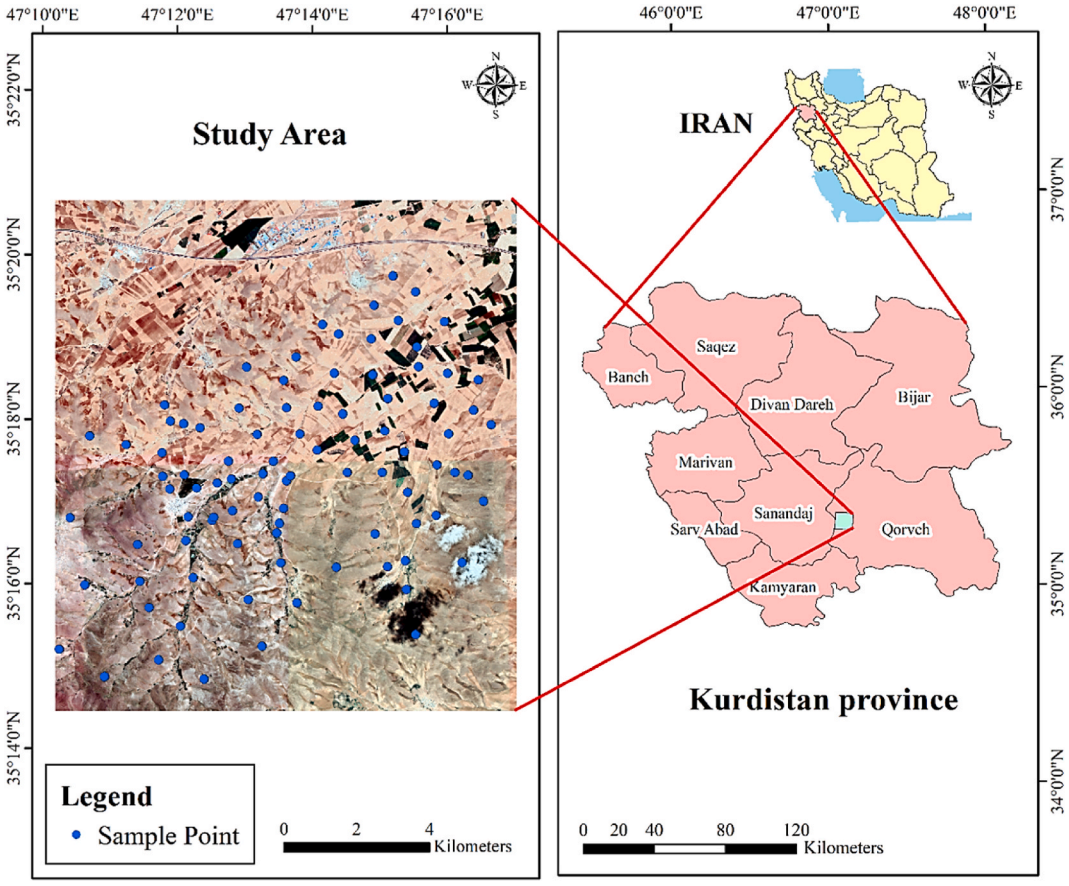


Fig. 2. Study area and distribution of sampling points.

Table 1
Characteristics of Sentinel-2 satellite.

Band Names	Wavelength(nm)	Bandwidth(nm)	Resolution(m)	Description
B1	443	21	60	Coastal Aerosol
B2	490	66	10	Blue
B3	560	36	10	Green
B4	665	31	10	Red
B5	705	16	20	Red Edge1
B6	740	15	20	Red Edge2
B7	783	20	20	Red Edge3
B8	842	106	10	Nir
B8A	865	22	20	Nir narrow
B9	945	21	60	Water vapor
B11	1610	94	20	SWIR1
B12	2190	185	20	SWIR2

values accurately. Additionally, to remove the influence of cloud cover on the Sentinel-2 images, only Level-2A images with less than 10 % cloud cover were used. In addition to the reflectance bands extracted from median images, six spectral indices have been utilized to characterize the condition of plants and soil. These indices include Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Soil Adjusted Vegetation Index (SAVI), Normalized Difference Moisture Index (NDMI), Global Vegetation Moisture Index (GVMI), and Bare Soil Index (BSI). The first three indices primarily focus on plant vegetation health and chlorophyll concentration (Bagheri and Kafashan, 2025). Their foundation lies in the relationship between the red and Near-Infrared bands. In simpler terms, a healthy plant exhibits higher absorption in the red region and higher reflection in the NIR region of the electromagnetic wave, signifying its vigor (Verstraete and Pinty, 1996). The following two indices (NDMI and GVMI) focus on measuring the moisture in the vegetation. They are based on a relationship between the reflectance of healthy vegetation in the range of SWIR and NIR bands. Unlike other indices, the last index identifies bare soil (Table 2). Also, Fig. 3 shows a scatter plot between the spectral index and SOC.

2.3.2. Sentinel-2 time series images

The B11 band of the Sentinel-2 satellite operates within the 1610 nm range of the electromagnetic spectrum, which is particularly significant for identifying soil chemistry parameters like organic carbon content (Gholizadeh et al., 2021; Norouzi et al., 2021). To provide a stronger theoretical foundation for its selection and use, the spectral reflectance of SOC sampling points was generated over the study areas (Fig. 4). A plot of this data significantly showed that the B11 band has the highest spectral reflectance among the different bands, with mean value of reflectance about 0.346, reinforcing its importance for SOC analysis. The three-month soil sampling duration, during which there were no changes in cultivation or other soil-affecting activities, makes the B11 band (SWIR) time series a unique tool for a detailed analysis of the behavior of SOC. By utilizing feature extraction methods (section 2.4), the information contained in this time series can yield higher-level features, revealing the complex relationship between electromagnetic waves and SOC, ultimately leading to improved estimation accuracy. For this purpose, this study employed 20 sets of SWIR band (B11) images from Sentinel-2 satellite data collected during soil sampling (Table 3 and Fig. 5). Feature extraction methods ICA and PCA were employed to extract significant information and identify anomalies during the soil sampling (more detail explained in Section 2.4).

2.4. Feature extraction

Considering that SOC is a complex phenomenon influenced by several parameters in its formation and changes, having a definitive opinion about the parameters that influenced SOC creation becomes challenging. Therefore, in this paper, both ICA and PCA methods were utilized to reduce dimensions and extract high-level features from Band B11 time series data. Using both ICA and PCA, this study uses their strengths in handling complex data. The ICA method enables the identification of independent components (ICs), which can

Table 2
Characteristics of spectral indices.

Index Name	Equation	MSI Band	Goal
NDVI	$NDVI = \frac{NIR - Red}{NIR + Red}$	B4, B8	Vegetation Growth
EVI	$EVI = \frac{2.5 \times (NIR - Red)}{(NIR + 6 \times Red - 7.5 \times Blue + 1)}$	B4, B8, B2	Vegetation Growth
SAVI	$SAVI = \frac{NIR - Red}{NIR + Red + 1.5} \times 1.5$	B4, B8	Vegetation Growth
GVMI	$GVMI = \frac{(NIR + 0.1) - (SWIR + 0.2)}{(NIR + 0.1) + (SWIR + 0.2)}$	B8, B11	Soil Moisture
NDMI	$NDMI = \frac{NIR - SWIR}{NIR + SWIR}$	B8, B11	Soil Moisture
BSI	$BSI = \frac{(NIR + SWIR) - (NIR + Blue)}{(NIR + SWIR) + (NIR + Blue)}$	B2, B3, B8, B11	Bare Soil Identification

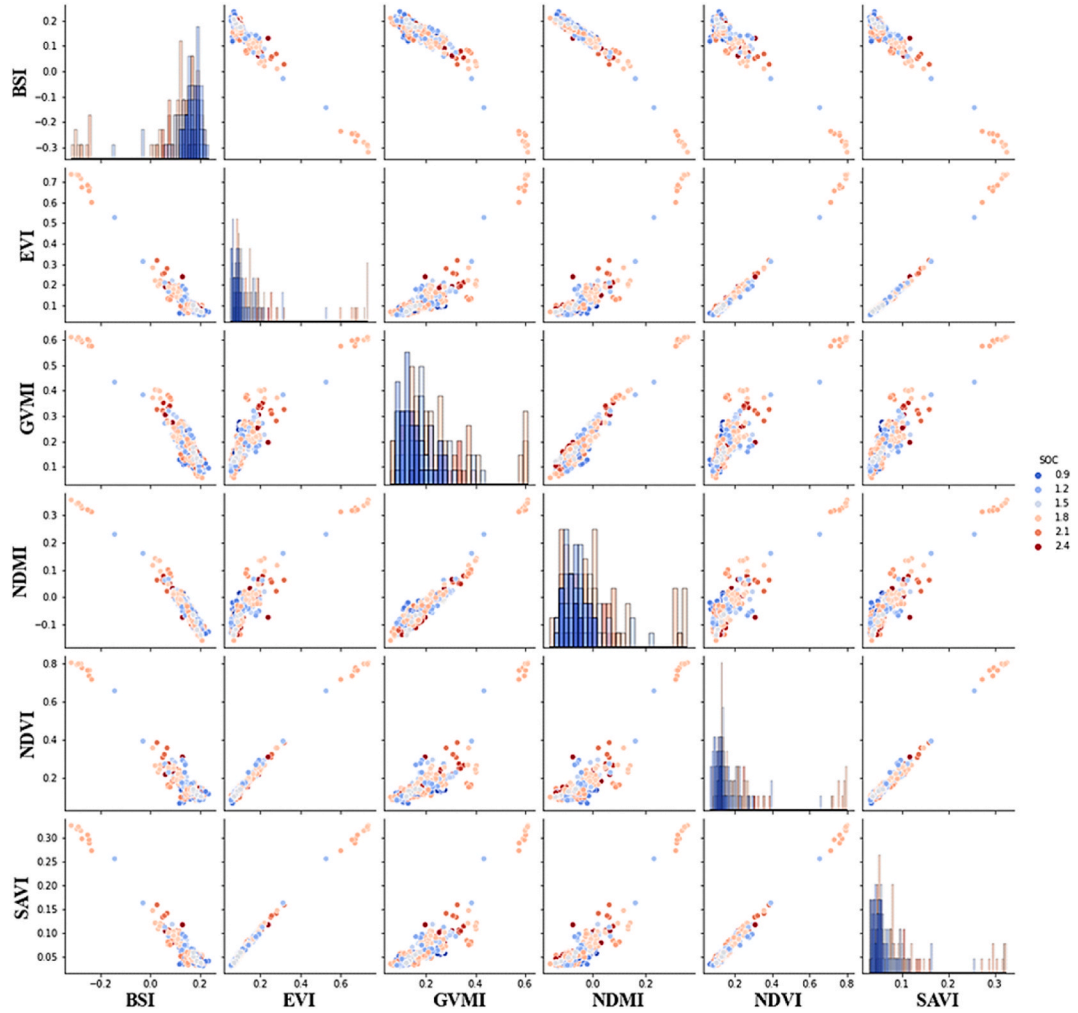


Fig. 3. The scatterplot between spectral indices and SOC.

help uncover underlying patterns and sources of variation in the SOC data. On the other hand, PCA's ability to capture variance allows for a deeper understanding of the data's overall structure. Combining these techniques offers a robust approach to estimating SOC.

2.4.1. Principal Component Analysis

PCA is a mathematical algorithm that simplifies the exploration of large datasets by reducing their dimensionality while retaining most of the data's variation. It achieves this by identifying principal components (PCs), which are directions capturing the highest variability in the data (Roweis, 1998). These components allow for the identification of patterns and the grouping of similar samples. The first part of the PCA algorithm is related to the calculation correlation matrix (Eq. (1) and Eq. (2)):

$$X_C = X - mu \tag{1}$$

$$C = \frac{1}{n} X_C^T X_C \tag{2}$$

in Eq. (1), X means dataset and can be represented as an n × m matrix, where each row corresponds to a data sample and each column represents a feature. In Eq. (1), the data is mean-centered by subtracting the mean of each feature from the corresponding feature values across all data samples. The mean-centered data is denoted by X_c, and it is obtained by subtracting the mean vector, mu, from the original data matrix. In Eq. (2), the covariance matrix (C) is calculated.

$$C \times V = V \times \lambda \tag{3}$$

$$Z = X_C \times V \tag{4}$$

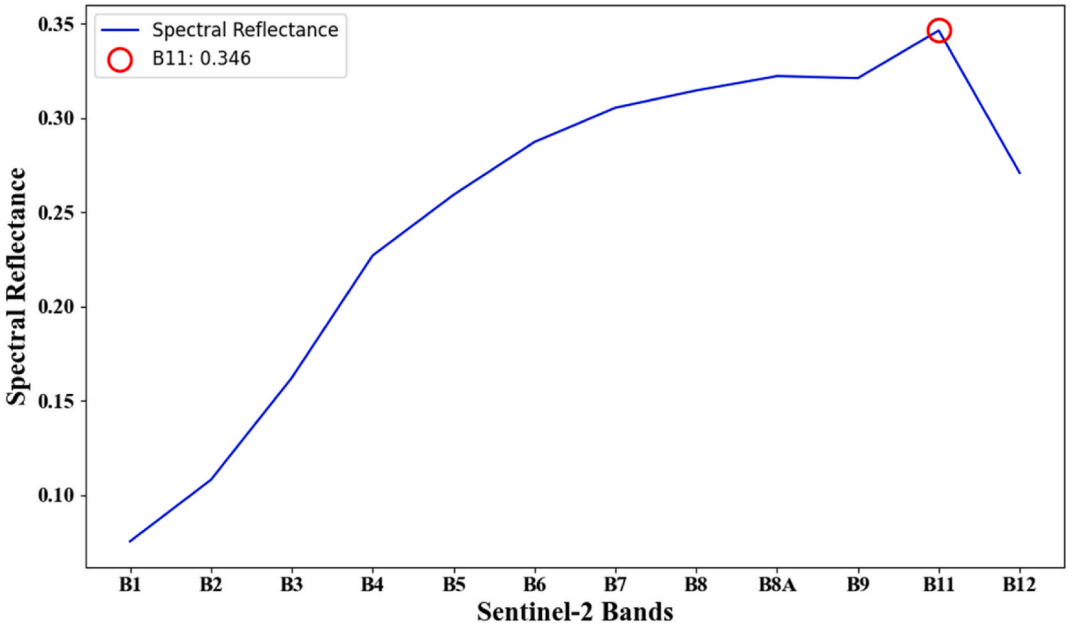


Fig. 4. Spectral reflectance of SOC sampling points over the study area and different spectral bands of Sentinel-2.

Table 3
Date and cloud coverage of Sentinel-2 time series images.

Product Number	Date	Cloud Percentage (%)	Product Number	Date	Cloud Percentage (%)
1	2019-06-23	0.03	11	2019-08-14	0.05
2	2019-06-28	0.09	12	2019-08-19	0.01
3	2019-07-03	0.03	13	2019-08-22	0.01
4	2019-07-08	0.04	14	2019-08-27	2.06
5	2019-07-13	0.03	15	2019-09-03	0.03
6	2019-07-18	0.02	16	2019-09-08	0.08
7	2019-07-23	3.31	17	2019-09-13	0.04
8	2019-07-28	0.05	18	2019-09-18	0.01
9	2019-08-02	2.18	19	2019-09-26	1
10	2019-08-07	0.03	20	2019-09-28	0.05

After obtaining the covariance matrix, the eigendecomposition is performed to find eigenvectors and eigenvalues of the covariance matrix Eq. (3). The eigenvectors represent the PCs, and the corresponding eigenvalues measure the variance captured along each PC. In Eq. (3), V is the matrix of eigenvectors of size m × m, and λ is the diagonal matrix of eigenvalues, also of size m × m. The eigenvectors in matrix V are sorted in descending order based on their corresponding eigenvalues in the diagonal matrix λ. The PCs are chosen from the top-k eigenvectors that capture the highest variance in the data. Typically, k is determined based on a certain variance threshold or the desired dimensionality reduction (Roweis, 1998). Finally, the original data is transformed into the new coordinate system represented by the selected PCs (Eq. (4)). In Eq. (4), Z is the transformed data matrix of size n × k, and X_C is the mean-centered original data of size n × m. In the end, the resulting transformed data matrix, Z, has reduced dimensionality, and the columns of Z represent the new features that capture the most significant variability in the original data. This research utilizes PCA transformation for feature extraction and dimension reduction from the SWIR time series dataset. The first five PCs, sorted automatically based on variance, are selected as feature sets representing SWIR time series dataset anomalies during soil sampling (Fig. 6).

2.4.2. Independent Component Analysis

ICA is a statistical technique used to separate mixed signals into their independent components (IC). It is employed in dimension reduction to identify, allowing for efficient representation and analysis of complex data by capturing the underlying sources of variation (Calhoun et al., 2009; Jiang et al., 2025). In this research, the ICA transformation was utilized to investigate the ICs of the SWIR time series. Applying the ICA transformation to the SWIR time series, the first three ICs were selected as important features for representing anomalies (Fig. 7).

2.4.3. Kurtosis analysis

In the preceding section, we choose the initial three ICs from the FastICA algorithm. This choice stems from the algorithm's

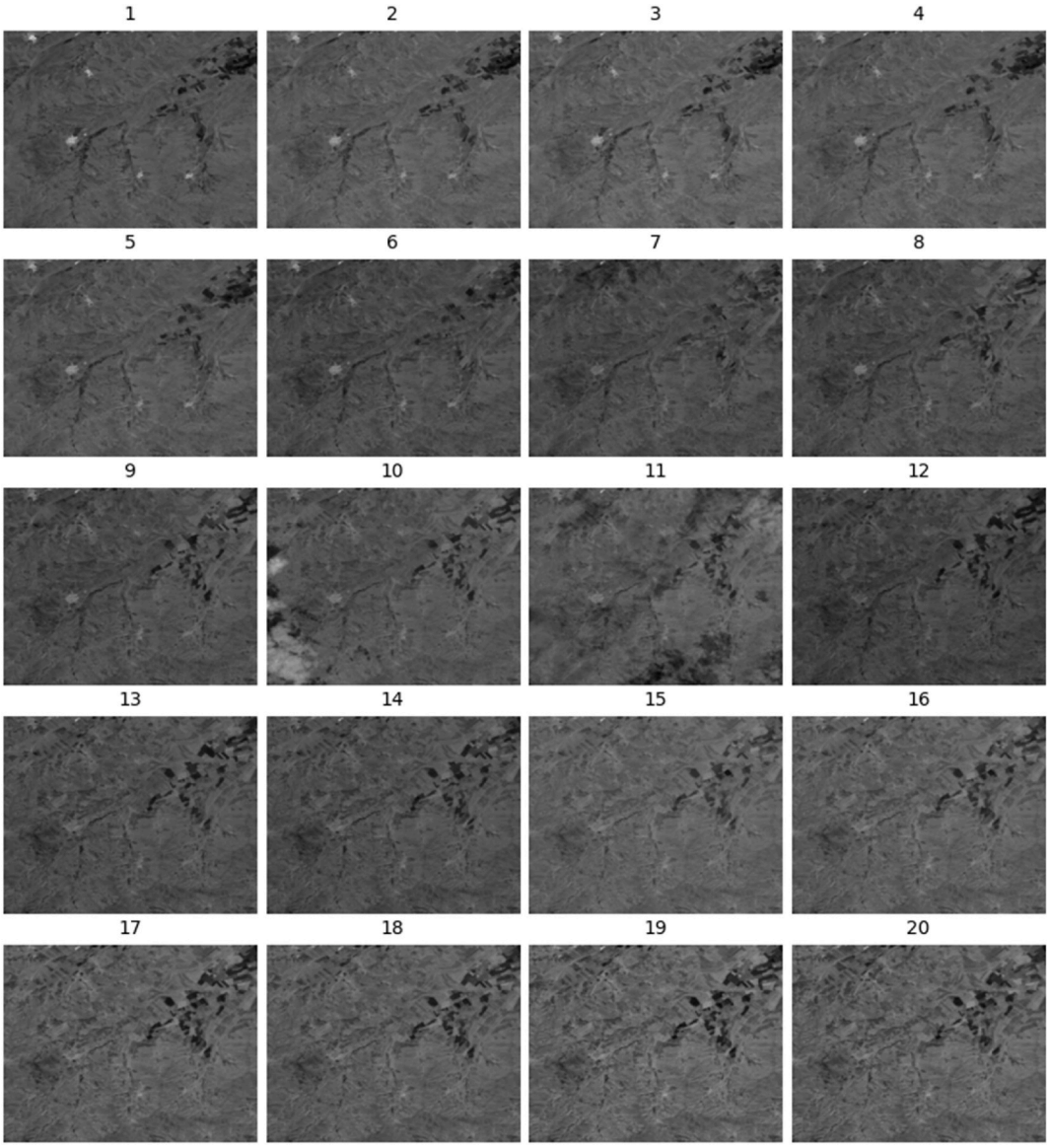


Fig. 5. SWIR band time series images. The information of each image is based on its top number in Table 3.

emphasis on identifying primary information sources within the image. If we were to include all five ICs derived from FastICA, there exists the potential for specific nuances of information to become obscured. Conversely, the ICA algorithm prioritizes extracting independent and non-normally distributed components. A pivotal yardstick for gauging non-normality resides in its kurtosis parameters. With this in mind, we have elected to incorporate the two ICs showcasing the highest kurtosis rates, augmenting them as additional features alongside the original three ICs (Wang and Chang, 2006).

2.5. Machine learning algorithms

This research delves into the domain of ensemble learning for SOC estimation, employing four ML algorithms: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Gradient Boosting Regression Tree (GBRT), and LightGradient Boosting Machine (LightGBM), over all different scenarios (Section 2.7). Among these, RF stands out as one of the most popular ML algorithms in soil digital mapping, exhibiting superior accuracy compared to traditional ML methods. However, the remaining three techniques are relatively novel in ML and seldom employed for SOC estimation. The details of each algorithm will be elucidated in the following sections. All of the feature extraction methods and ML algorithms were implemented using Python programming language libraries, including Scikit-learn, the LightGBM Python package, and the XGBoost Python package.

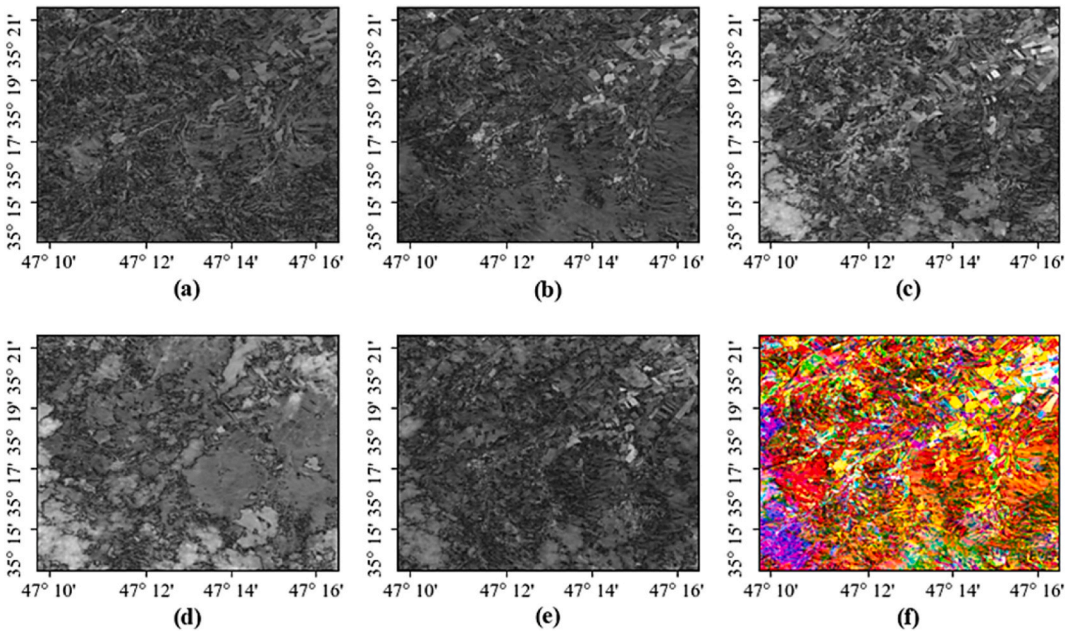


Fig. 6. The result of applying PCA transformation on SWIR time series. (a) first PC, (b) second PC, (c) third PC, (d) fourth PC, (e) fifth PC, (f) false color composite from the first three PCs (PC1, PC2, PC3).

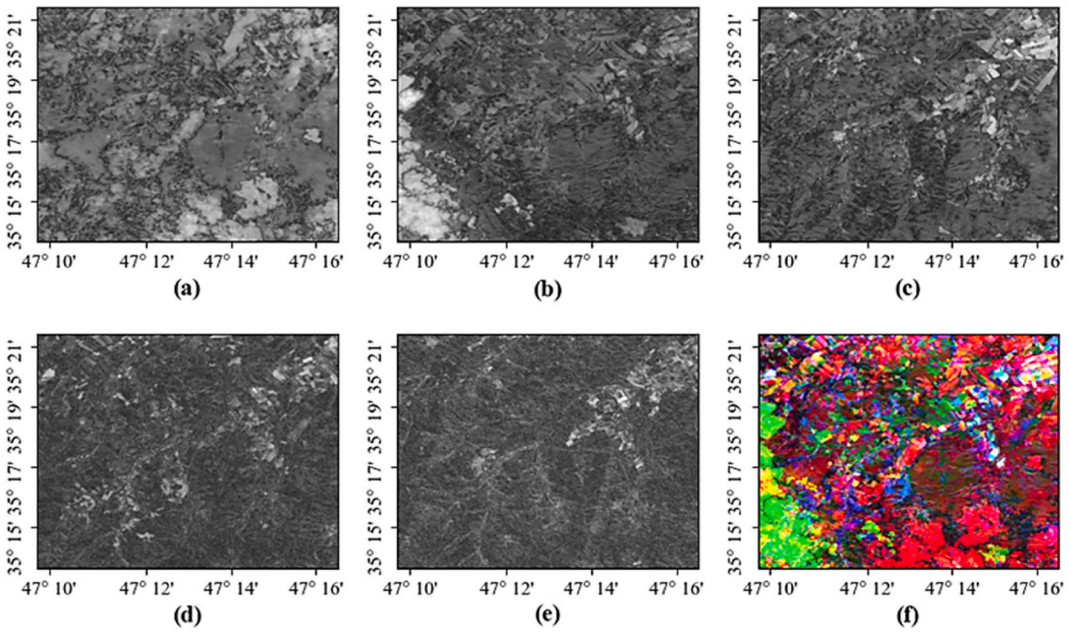


Fig. 7. The result of applying ICA transformation on SWIR time series. (a) first IC, (b) second IC, (c) third IC, (d) fourth IC, (e) fifth IC, (f) false color composite from the first three ICs (IC1, IC2, IC3).

2.5.1. Random Forest

RF algorithm is an ensemble learning method originally proposed by Leo Breiman and Adele Culter in 2006. It is widely employed for various purposes, including classification and regression tasks. The RF algorithm possesses several notable advantages. One of its key strengths lies in the implementation of the bagging technique. Unlike using the entire dataset for each decision tree, the RF algorithm employs a unique subset of the data, thereby ensuring the independence of each tree. Consequently, the decision trees in the forest remain uncorrelated. Moreover, the RF algorithm effectively addresses the overfitting issue using the Out of Bag (OOB) technique. This technique involves the allocation of one-third of the training data as test data for evaluating each decision tree. Generally,

owing to its scale and data-type independence, RF has found extensive application in various research studies concerning SOC estimation. It can be asserted that RF, in a broad sense, demonstrates superior accuracy when compared to other traditional ML algorithms, such as PLSR, SVR, and decision trees (Yang et al., 2016).

2.5.2. Gradient Boosting Regression Tree

GBRT is a powerful algorithm that combines gradient boosting and decision trees for regression tasks. It iteratively builds an ensemble of trees, with each tree correcting the errors made by the previous ones. By minimizing a loss function, it adds new trees to the ensemble. GBRT can handle data types and missing values (Prettenhofer and Louppe, 2014). It is flexible and captures complex relationships between input variables and the target. Unlike Random Forest, GBRT builds trees sequentially, considering previous errors. Overall, GBRT is a robust and versatile algorithm for regression tasks (Ye et al., 2021).

2.5.3. eXtreme gradient boosting

The XGBoost algorithm is not just an ordinary algorithm but a meticulously designed software package. Initially developed by researchers at the University of Washington, XGBoost was written in C++ and introduced as the LIBSVM library. While the foundation of this algorithm is similar to the GBRT algorithm, there are two key distinctions in its structure and implementation. One major difference lies in the optimizer function. XGBoost employs the regularized gradient boosting algorithm instead of the gradient descent algorithm used in traditional approaches(Shao et al., 2024). This modification enhances the algorithm's performance. Another significant divergence is its utilization of parallel computing capabilities. This is achieved through the column-blocking technique, resulting in a remarkable speed improvement over previous algorithms (Chen et al., 2015). By incorporating these enhancements, XGBoost stands out as a powerful and efficient tool in ML.

2.5.4. Light gradient Boosting Machine

LightGBM is an ensemble learning algorithm developed by Microsoft, known for its high training speed compared to other boosting algorithms. This speed is achieved by utilizing the leaf-wise approach, which differs from the traditional level-wise approach employed by decision trees. The leaf-wise approach aims to select features that minimize the loss function rapidly rather than increasing the tree depth. In the leaf-wise approach, LightGBM calculates histograms of features to enhance the computational efficiency within optimization loops. These histograms assist in identifying optimal split points during tree construction. By focusing on the most informative features, LightGBM accelerates the learning process (Ke et al., 2017). Compared to other boosting algorithms, LightGBM demonstrates superior performance in speed and RAM consumption. Its optimized implementation allows for faster training, making it a popular choice for handling large datasets efficiently.

2.6. Feature selection

This research utilized 28 RS features extracted from Sentinel-2 for SOC estimation. As SOC is a complex phenomenon and its changes are strongly influenced by location and time, developing a universal explicit formula to relate it with RS features across all places and times becomes challenging. Moreover, distinctive features might not positively contribute to the SOC concentration; instead, they increase computational costs and can lead to reduced accuracy. Hence, employing an effective feature selection method becomes crucial to reduce complexity and enhance model accuracy. A binary genetic algorithm (GA) is used to address this. GA is an evolutionary algorithm based on the population, inspired by Darwin's theory of survival. (Lambora et al., 2019). The GA algorithm's main performance is finding a suitable solution for an optimization problem within a limited number of iterations without exhaustively checking all possible solutions. The algorithm involves creating an initial population of binary strings, applying selection, crossover, and mutation operators to generate offspring, evaluating their fitness, and repeating this process until a satisfactory solution is found. The fitness function determines the quality of each solution, and the algorithm evolves the population over generations, favoring better solutions based on their fitness values (Eq. (5)).

$$fitness\ function = \frac{R^2}{RMSE} \tag{5}$$

2.7. Different scenarios of training

The main goal of this research is to explore the potential of feature extraction from band B11 time series images to improve the

Table 4
Details of parameters set in the binary GA algorithm.

Parameters Name	Value
Maximum number of iterations	300
Population size	200
Mutation probability	0.07
Cross-over probability	0.8
Cross-over type	Two-point crossover
Elite ratio	0.01

accuracy of SOC estimation and to develop a pipeline for SOC estimation using only RS observations. To achieve this goal, the SOC estimation process is conducted in four distinct scenarios (Table 5). Conducted in four different scenarios (Table 5), the SOC estimation process is presented as follows: Scenario 1 (S#1) involves SOC estimation without using features from time series data, while Scenario 2 (S#2) is similar to S#1 but employs a binary GA for feature selection. In Scenario 3 (S#3), SOC estimation is performed using all 28 time series features, and in Scenario 4 (S#4), the GA algorithm is applied to select the best features for SOC estimation. By conducting SOC estimation under these different scenarios, valuable insights into the impact of feature extraction and feature selection algorithms on accuracy are aimed to be gained, thereby enhancing the understanding of SOC estimation using RS observations. In this research, S#2 and S#4 used Binary Elitism GA for feature selection. The details of the parameters used in the GA algorithm are shown in Table 4.

2.8. Model evaluation

Splitting the dataset into training and testing sets is one of the common approaches in ML to evaluate model performance. However, this method has a potential drawback: it selects only one specific part of the dataset as the test data, and this selection is random. As a result, overfitting or underfitting may occur, particularly with small datasets. To address these issues and enhance the model's reliability, it is preferable to employ the concept of cross-validation. Cross-validation involves dividing the dataset into multiple folds and iteratively using different folds for training and testing (Emadi et al., 2020; Zhang et al., 2021, 2022), This process helps to provide a more reliable evaluation of the model. This research employed K-fold cross-validation, with K set to 10. The model's performance is evaluated using several metrics, including the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and the Ratio of Performance to Deviation (RPD). RPD measures how well the model's predictions align with the observed values. A higher RPD indicates that the model's predictions have smaller variations around the observed values, suggesting a more reliable and accurate result (Chang et al., 2001). According to the DSM guidelines, the RPD ranges can be interpreted as follows: when the RPD falls between 1.0 and 1.4, the model possesses a limited capability, merely allowing the identification of extreme values; when the RPD lies between 1.4 and 2.0, the model is deemed fair and can be employed for evaluation and correlation purposes; when the RPD ranges from 2.0 to 2.5, the model exhibits very good quantitative performance; and when the RPD exceeds 2.5, the model demonstrates excellent predictive ability (Yang et al., 2022). The R^2 value indicates the fraction of the variance in the dependent variable that is accounted for by the independent variables used in the model. R^2 is a value between 0.0 and 1.0, with higher values indicating a better ability of the model to capture the variability in the dependent variable (Nagelkerke, 1991; Taghadosi et al., 2019). It is important to note that each metric's results are reported based on the average results obtained from ten-fold cross-validation, ensuring a robust assessment of the model's performance. These statistical indicators were calculated with the following equations:

$$MAE = \frac{\sum_{i=1}^{n} |o_i - p_i|}{n} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (o_i - p_i)^2}{n}} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (p_i - o_i)^2}{\sum_{i=1}^{n} (o_i - \overline{o})^2} \tag{8}$$

$$RPD = \frac{Std}{RMSE} \tag{9}$$

where O_i and P_i denote the observed and predicted SOC values, respectively, and n denotes the number of soil samples used for validation $\overline{O}$ and denotes the means of the observed SOC values. Std is the standard deviation of measured SOC.

Table 5
The details of different scenarios of learning.

Scenario Name	Features Used	Features Name	Number of Features Used
S#1	Sentinel-2 Surface Reflectance and Spectral Indices	B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, B12, NDVI, EVI, SAVI, NDMI, GVMI, BSI	18
S#2	GA Algorithm Apply on Scenario One Features	B1, B2, B5, B6, B9, B11, B12, EVI	8
S#3	Sentinel-2 Surface Reflectance, Spectral Indices, and SWIR Time Series Features	B1, B2, B3, B4, B5, B6, B7, B8, B8A,9, B11, B12, NDVI, EVI, SAVI, NDMI, GVMI, BSI, PC1, PC2, PC3, PC4, PC5, IC1, IC2, IC3, IC_{1sort}, IC_{2sort}	28
S#4	GA Algorithm Apply on Scenario Three Features	B1, B5, B6, B8, B8A, B9, B11, B12, NDVI, EVI, SAVI, NDMI, PC1, PC2, PC3, PC4, IC1, IC2, IC3, IC_{1sort}	20

3. Result

3.1. Descriptive analysis of soil organic carbon contents

The chemical experiment result has shown the SOC range between 0.74 and 2.42 %, such that the mean and standard deviation of SOC are 1.59 % and 0.43 %, respectively. That can be said that the study area has a good situation from SOC (Table 6).

After analyzing the probability distribution function (PDF) and density histogram of the SOC data, we discovered that the PDF exhibits a bimodal distribution (Fig. 8). Unlike a normal distribution, which typically has a single peak, the bimodal distribution displays two distinct peaks. These two asymmetric peaks correspond to SOC values of 1.4 % and 1.9 %, respectively.

The presence of two distinct peaks in the PDF when measuring SOC in a specific area suggests the existence of different carbon sources or processes at play. These peaks could indicate diverse sources such as natural vegetation and human-induced activities like agriculture or urbanization. Fig. 9 investigates this bimodal distribution by analyzing the relationship between SOC and the SWIR band of Sentinel-2 images. The SWIR band was chosen due to its absorption points suitable for SOC in this spectral range. Fig. 9 shows that for the same SWIR band reflection values, the probability density function exhibits a two-peaked behavior. This indicates a non-linear and conditional relationship between SOC and the SWIR band. The presence of such inclusion implies that the amount of organic carbon in the soil is conditionally dependent on the reflectance of the SWIR band. This conditional relationship might be significantly related to differences in soil types and cultivation across various parts of the region (Ladyman et al., 2013). Complex systems often exhibit conditional relationships between variables, resulting in diverse patterns and behaviors (Ladyman et al., 2013). Overall, the bimodal distribution of SOC highlights the complexity of carbon dynamics in the study area.

3.2. Hyperparameter tuning

In S#2 and S#4, the hyperparameters of the ML algorithms were fine-tuned using the grid search method. But, in S#1 and S#3, the same default values were applied (Table 7).

3.3. Scenario one (S#1)

In this scenario, all the features extracted from the synthetic median image of the Sentinel-2 were utilized (see Section 2.3.1). The algorithms employed in this scenario demonstrated R^2 score from 0.70 to 0.74 (Table 8). LightGBM exhibited the highest R^2 score among these algorithms, followed by XGBoost, RF, and GBRT algorithms, as shown in Fig. 10. Despite its simplicity compared to the other scenarios, this scenario is notable for yielding satisfactory results across all methods, with the RPD being approximately two.

3.4. Scenario two (S#2)

In S#2, the capabilities of the GA for feature selection were investigated, and its impact on SOC estimation was evaluated. The features selected by the GA include B1, B2, B5, B6, B9, B11, B12, and EVI. This selection underscores the pivotal role played by the electromagnetic spectrum's Blue, Red Edge, and SWIR regions in SOC estimation. Within this scenario, the models XGBoost and GBRT showcase the highest accuracy levels, achieving R^2 scores of 0.8211 and 0.8154, respectively (Table 8 and Fig. 11). A noteworthy aspect of this scenario is the substantial increase in estimation accuracy compared to S#1. The R^2 score experiences an average increase of 0.07 and an RMSE decrease of about 0.04 compared to S#1 (Table 8). These improvements reflect the tangible impact of selecting the most pertinent features.

3.5. Scenario three (S#3)

S#3 incorporated all available features, utilizing time series and synthetic median image features. Based on the results, time series features substantially improved accuracy by approximately 11 %. Also, RMSE decreased by about 0.07, and RPD increased by about 0.6 compared to the corresponding scenario (S#1). This undeniable enhancement shows the effectiveness of features extracted from time series data and the correct selection of the B11 band to monitor changes in soil condition. Within this scenario, similar to S#1, the LightGBM and XGBoost algorithms demonstrated the best performance, achieving an R^2 score of 0.8516 and 0.8382, respectively.

3.6. Scenario four (S#4)

S#4 is the most comprehensive in this research as it utilizes both time series features and GA based feature selection. In this scenario, the GA identified 20 important features. Notably, of the [e.g., 10] initial time series features considered by the GA, eight were

Table 6
Descriptive statistics for SOC (%).

Data Type	Number	Mean (%)	Median (%)	SD (%)	Min (%)	Max (%)	Range (%)	Skew
Total Data	91	1.59	1.59	0.43	0.74	2.42	1.68	0.047

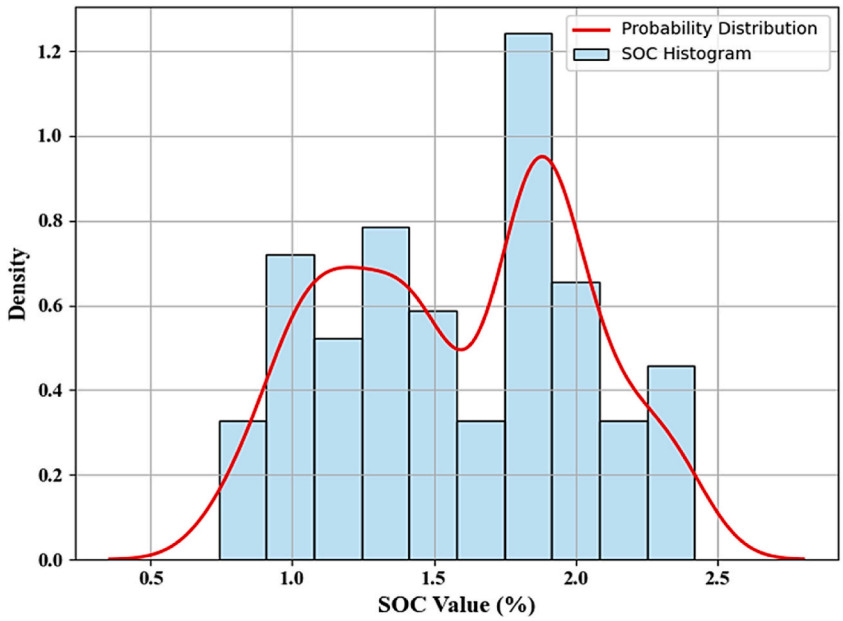


Fig. 8. Plots of the density histogram and probability density function of SOC data.

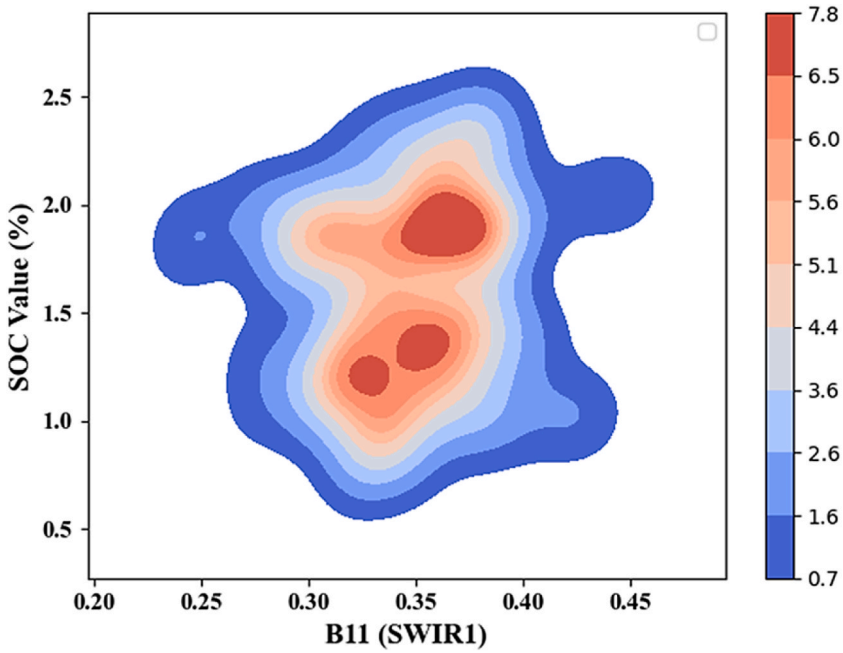


Fig. 9. Probability distribution function contour estimation between SOC and SWIR band.

selected, meaning the GA deemed 80 % of these specific time series features as important. As expected, this scenario achieved the highest R^2 score compared to the others. These results highlight the superiority of higher-level features extracted from time series data over conventional features from synthetic median images for estimating SOC. Additionally, the XGBoost and GBRT algorithms demonstrated the highest predictive performance, achieving impressive R^2 scores of 0.891 and 0.8764, respectively, which are considered very high in SOC studies (Table 9). In contrast, the RF algorithm yielded the lowest R^2 value (0.8462), indicating the superiority of boosting-based algorithms over bagging-based algorithms in SOC estimation (Table 9 and Figs. 12 and 13)

Table 7
The optimum values of the tuning parameters of different models in S#2 and S#4. Lr (Learning rate), N_estimators (Number of estimators), and Max_depth (Maximum depth).

Method	S#2	S#4
RF	N_estimators = 1000 Max_depth = 20	N_estimators = 280 Max_depth = 15
XGBoost	N_estimators = 2000 Lr = 0.018 Max_depth = 5	N_estimators = 2100 Lr = 0.019 Max_depth = 5
LightGBM	N_estimators = 1400 Lr = 0.033 Max_depth = 10	N_estimators = 1100 Lr = 0.031 Max_depth = 10
GBRT	N_estimators = 1000 Lr = 0.01 Max_depth = 5	N_estimators = 1000 Lr = 0.02 Max_depth = 5

Table 8
Evaluation of the performance of the different machine learning algorithms over S#1 and S#2.

Methods		S#1				S#2			
		R^2	RMSE	MAE	RPD	R^2	RMSE	MAE	RPD
RF	Train	0.9628	0.028	0.059	5.20	0.9734	0.025	0.047	6.14
	Test	0.7317	0.046	0.158	2.01	0.7867	0.044	0.129	2.31
XGBoost	Train	0.9999	0.002	0.003	837	0.9999	0.014	0.012	9.46
	Test	0.7416	0.046	**0.136**	2.07	**0.8211**	**0.042**	**0.09**	**2.58**
GBRT	Train	0.9999	0.004	0.001	199.9	0.9999	0.004	0.001	10.2
	Test	0.7191	**0.044**	0.126	1.94	0.8154	0.042	0.1	2.55
LightGBM	Train	0.9767	0.025	0.046	6.51	0.9968	0.015	0.016	8.05
	Test	**0.7492**	0.046	0.146	**2.08**	0.7849	0.044	0.117	2.30

4. Discussion

This research demonstrates that accurate SOC estimation is indeed feasible using solely RS data, particularly through a synergistic approach that combines the temporal features of Sentinel-2's B11 band time series and rigorous feature selection. This multi-stage process was pivotal in enhancing SOC estimation accuracy by effectively bridging the gap between conventional RS observations and the complex spectral responses of SOC. The foundation of this improvement lies in leveraging the B11 (SWIR, 1610 nm) band time series. The inherent sensitivity of this band to soil chemistry, including organic carbon, makes its temporal data particularly valuable for continuously monitoring soil changes. This rich temporal information was then transformed into more potent, high-order features using PCA and ICA. PCA was instrumental in capturing hidden patterns and dominant modes of variance within the B11 time series, which can relate to underlying SOC influencing factors like subtle changes in soil moisture or vegetation effects over time. Concurrently, ICA helped disentangle mixed signals by separating the time series into statistically independent components, potentially isolating influences such as vegetation phenology from agricultural practices, thereby allowing models to focus on components more directly related to SOC. The significant impact of this feature engineering step is evident in the substantial accuracy improvement of approximately 11 % when comparing scenarios relying on these high-order features (S#3) against those using raw spectral data (S#1), as illustrated in Figs. 14 and 15.

As detailed in Section 2.6, a GA was employed for feature selection following the extraction of variables from time-series data. Our results indicate that this GA-based feature selection significantly enhanced the accuracy of SOC estimation, leading to an average improvement of approximately 5 % (Table 8). This enhancement is primarily attributed to the algorithm's effectiveness in removing irrelevant features and mitigating multicollinearity among predictors. Further analysis of the GA's output (Table 5) reveals that eight specific features from the initial set in S#1 were identified as most influential and thus selected for S#2. These critical features include spectral bands B1, B2, B5, B6, B9, B11, B12, and the EVI. Collectively, these selected bands represent key spectral regions highly effective for SOC identification, namely the SWIR, red-edge, and coastal aerosol portions of the electromagnetic spectrum. Notably, among all the spectral indices considered, only the EVI. EVI is recognized as an advanced version of the NDVI, offering a clearer assessment of plant health by minimizing the influence of soil background. The application of the GA in the fourth scenario yielded significant outcomes. From the initial 28 features in S#3, the GA selected 20. Notably, eight of these 20 selected features were derived from time-series data. This retention of 80 % of the time-series features underscores their considerable importance and validates the effectiveness of our feature extraction strategy. Furthermore, the re-selection of seven out of the eight features previously identified as important in S#2 for S#4 reinforces the robustness and consistent significance of these predictors.

By examining the results of different scenarios, it can be seen that using time series features has improved SOC estimation accuracy. However, more interpretable analyses are needed to better understand the performance of features extracted from time series data. In this regard, the permutation importance method was used to check the importance of features. In this method, reducing an algorithm's accuracy after shuffling the order of a specific feature indicates its importance. This method was applied to the fourth scenario using

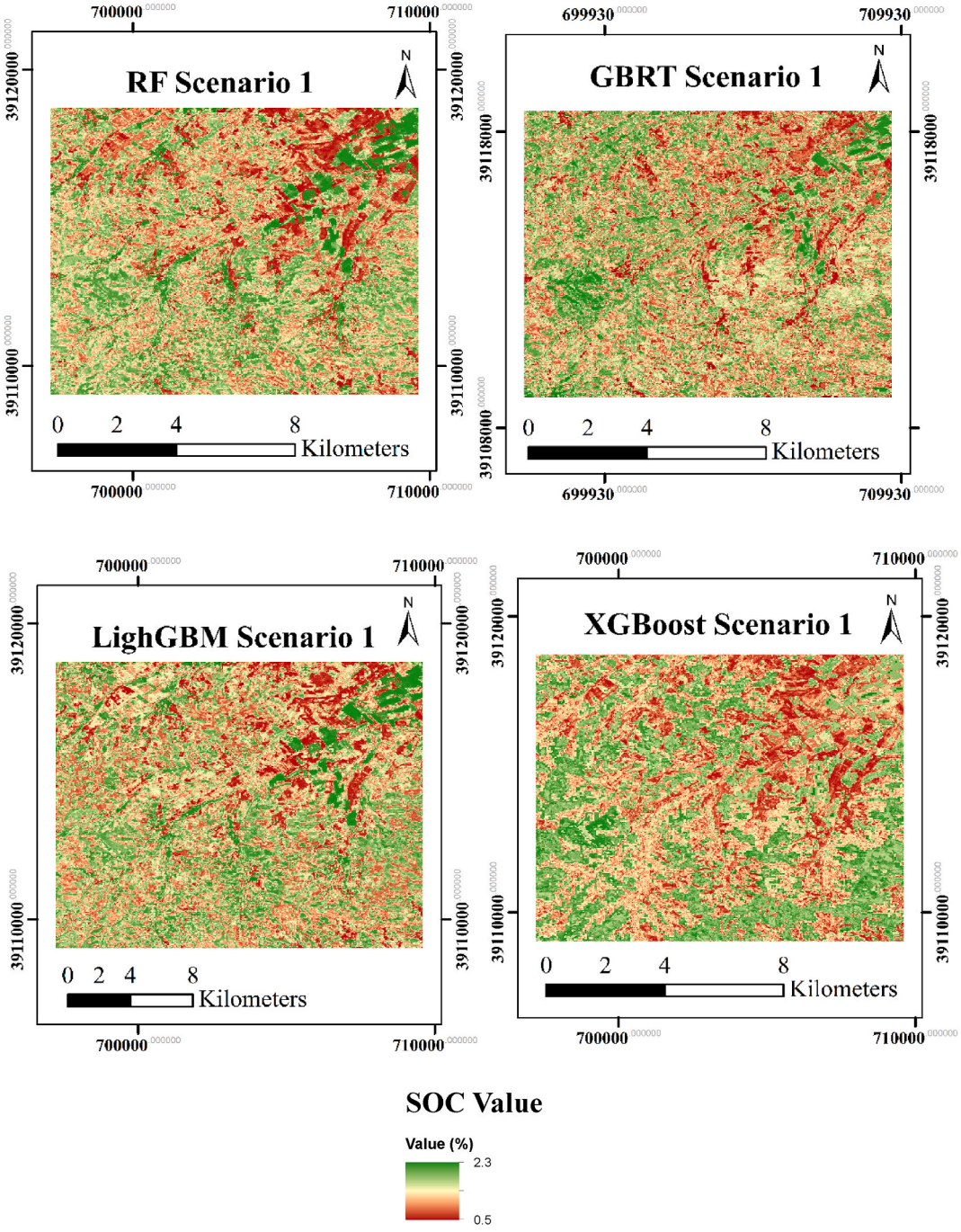


Fig. 10. SOC distributed map resulted from different machine learning algorithms in S#1. It is noteworthy that the grid coordinates in maps are represented based on UTM coordinates in Zone 38N.

different ML algorithms. To better interpret the results, the features were divided into four categories: PCA features, ICA features, spectral bands, and spectral indices. The results indicated that PCA features held the highest importance, with an average contribution of 12.05 % across the four ML algorithms. This was followed by spectral bands (7.1 %), ICA features (4.5 %), and spectral indices (2.3 %). Notably, features derived from the PCA transformation were almost twice as effective as the original spectral reflectance bands alone (Fig. 16). This finding demonstrates the superiority of features extracted from time-series data compared to conventional, single-date spectral data for SOC estimation in this study. Furthermore, the higher importance of PCA features over ICA features suggests that, in this case, creating features based on the criterion of highest variance was more efficient than creating features based on statistical independence. Additionally, the feature permutation results for S#4 showed that specific selected spectral bands, B1 (coastal aerosol),

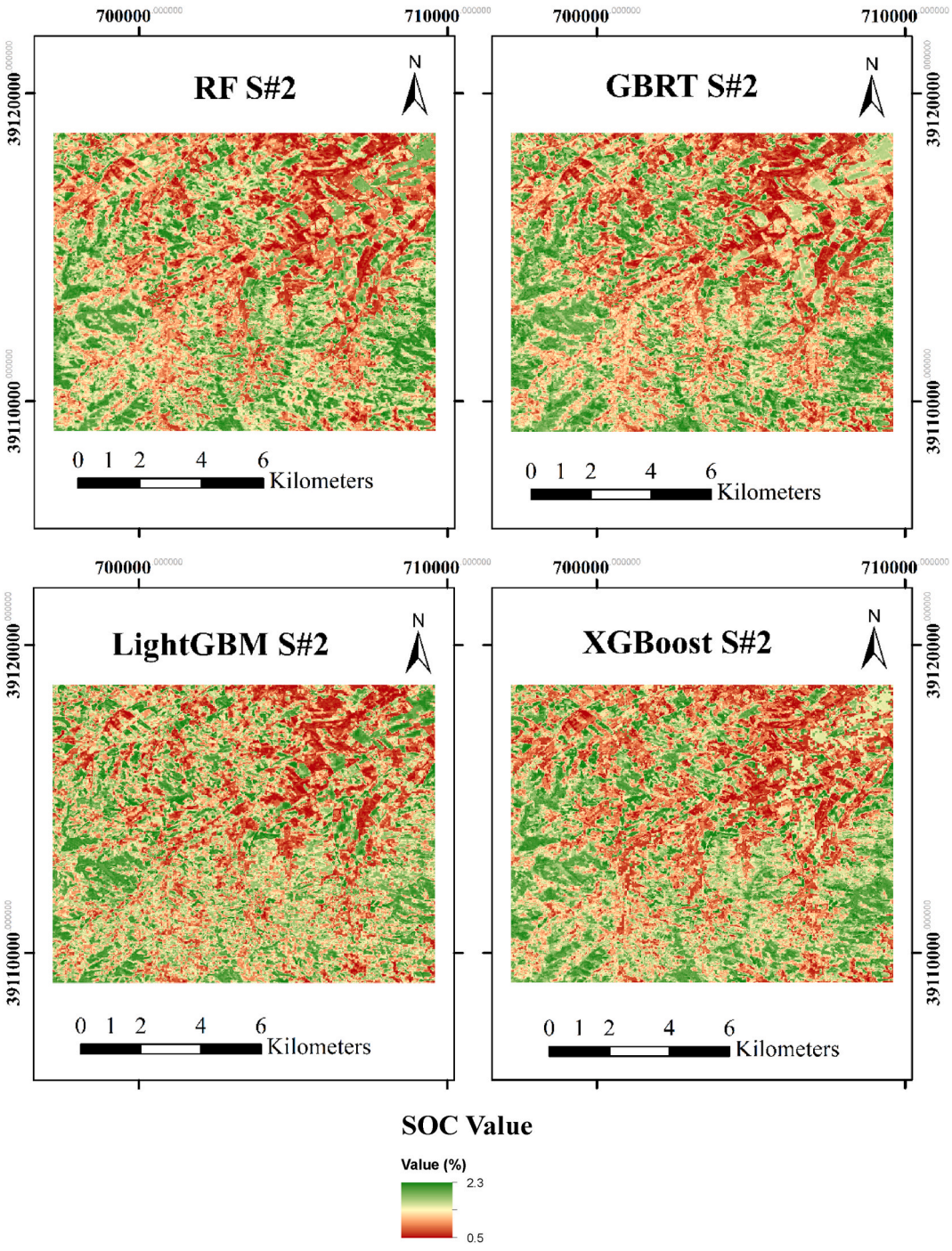


Fig. 11. SOC distributed map resulted from different machine learning algorithms in S#2. It is noteworthy that the grid coordinates in maps are represented based on UTM coordinates in Zone 38N.

B8A (NIR), and B11 (SWIR), had the most significant impact on SOC estimation, with importance scores of 26 %, 7 %, and 4 %, respectively. In other words, the coastal aerosol (centered around 440 nm), near-infrared (865 nm), and SWIR (1610 nm) regions of the electromagnetic spectrum were identified as the most critical for SOC estimation in this study area.

In this research, the abilities of four ML algorithms have been comprehensively compared (Fig. 17). In general, the XGBoost algorithm has a higher ability to estimate SOC, followed by GBRT, LightGBM, and RF algorithms; this trend has been seen in other studies. Ye et al. employed the LightGBM and XGBoost algorithms to estimate SOC concentration using GF-6 satellite data. The study reported that XGBoost achieved an R^2 value of 0.63, while LightGBM achieved an R^2 value of 0.62 (Ye et al., 2021). Xie et al.

Table 9
Evaluation of the performance of the different machine learning algorithms over S#3 and S#4.

Methods		S#3				S#4			
		R^2	RMSE	MAE	RPD	R^2	RMSE	MAE	RPD
RF	Train	0.9749	0.026	0.046	6.32	0.9779	0.025	0.043	6.73
	Test	0.8279	0.042	0.123	2.50	0.8461	0.041	0.115	2.67
XGBoost	Train	0.9999	0.001	0.000	900	0.9997	0.000	0.003	15.1
	Test	0.8382	0.041	**0.097**	2.66	**0.891**	**0.037**	**0.08**	**3.34**
GBRT	Train	0.9611	0.029	0.064	5.08	0.9998	0.004	0.003	12.3
	Test	0.8206	0.042	0.13	2.45	0.8764	0.038	0.085	3.15
LightGBM	Train	0.9932	0.018	0.023	12.2	0.9999	0.002	0.00	11.3
	Test	**0.8516**	**0.04**	0.1	**2.84**	0.87	0.038	0.09	3.07

investigated the potential of Sentinel-2 and Landsat-8 satellite imagery for estimating SOC content. They employed three distinct ML models to achieve promising results: XGBoost, GBRT, and RF. The R^2 for these models was reported to be 0.80, 0.77, and 0.74, respectively (Xie et al., 2022). Zhang et al. explored the use of Sentinel-1 and -2 images, along with features extracted from a DEM, for estimating SOC content in swamp wetlands of North China. Their findings indicate that the XGBoost algorithm achieved superior accuracy ($R^2 = 0.65$) compared to the RF algorithm ($R^2 = 0.54$) (Zhang et al., 2023). Overall, this study's high accuracy results from various factors, including the extraction of high-level features from time series data, proper feature selection, use of powerful ML models, and correct setting of hyperparameters. Thus, the accuracy rate has increased from 0.71 in S#1 to 0.89 in S#4.

A notable point for discussion is the more limited range of prediction values observed after incorporating time-series features (S#3 and S#4), as depicted in Fig. 18. Models particularly struggled to accurately predict low SOC concentrations (below 0.8 %). This underfitting is primarily attributed to the inherently weak spectral signature of low SOC (below 1 %) (Li et al., 2024; Pouladi et al., 2023). The PCA/ICA transformations applied to the time-series data may also contribute to this challenge. While PCA and ICA aim to extract the most influential data components, they inherently involve a trade-off between dimensionality reduction and information preservation. Consequently, retaining only the top 5 PCs and ICs might lead to underrepresentation or loss of information concerning low SOC concentrations. PCA captures directions of maximum data variance, while ICA aims to separate mixed signals into statistically independent components; both methods, therefore, focus on extracting the most prominent sources of variation. Thus, subtle spectral variations from low SOC, potentially masked by stronger influences like vegetation or soil moisture, may not be captured among these components and could be excluded. To address this underfitting, future work could explore retaining an increased number of components from PCA and ICA transformations (e.g., 10 instead of 5). This approach could enhance the model's ability to capture the subtle nuances associated with low SOC concentrations, as a more comprehensive feature space would be created, potentially better incorporating these subtle variations.

This study introduces an algorithm for estimating SOC based on Sentinel-2 imagery and underscores the capacity of time-series features to enhance estimation accuracy. Nevertheless, two key areas require further attention. Firstly, the optimal temporal window for creating time-series images (e.g., three months, six months, or more) needs precise definition, as current guidelines are lacking. Secondly, for reliable SOC estimation using RS, it is crucial that farmland be devoid of vegetation, with seedbed or post-harvest conditions being most suitable (Diaz-Gonzalez et al., 2022). This particular consideration merits further exploration in future studies. Additionally, subsequent research could investigate the utility of more advanced feature extraction algorithms to potentially supersede PCA and ICA, thereby aiming for greater SOC estimation accuracy.

5. Conclusion

SOC is a critical soil property impacting both food security and climate change. Due to its estimation significance, our study focuses on creating a reliable algorithm solely based on RS data to estimate SOC levels. This algorithm's importance becomes clear when we consider the difficulties of gathering extra data sources like geological maps, land use and cover maps, topography-related parameters, and other soil chemical parameters for SOC estimation. Gathering this supplementary data poses challenges such as high expenses, time-consuming procedures, and uncertainty. Most RS-based studies traditionally rely on single-time observations for SOC estimation. Often, these observations are inadequate to accurately capture SOC variations due to the intricate nature of carbon dynamics. In this research, we adopted a novel approach that involves feature extraction through PCA and ICA transformations applied to the time series of B11 band images from the Sentinel-2 satellite. This approach effectively addresses the need for higher-level features to model SOC concentration accurately. Consequently, the accuracy of SOC estimations has improved by approximately 0.11 compared to models using only conventional RS features. Also, using the GA feature selection method could increase the estimation accuracy by about 0.05. Overall, this research's achievements can be categorized into four subsections: (1) Utilizing features extracted from band B11 time series data (time series features) has significantly enhanced estimation accuracy. (2) Using the GA as a feature selection algorithm has also been effective in model improvement. (3) Using the newer ML algorithms based on boosting scenarios performs better than the RF algorithm. So, the XGBoost, GBRT, and LightGBM accuracy are 0.891, 0.8764, and 0.87, respectively, but the ultimate accuracy of RF is 0.8460. (4) The most important achievement of this research can be considered the presentation of an algorithm based on the observations of the Sentinel-2 satellite in estimating SOC.

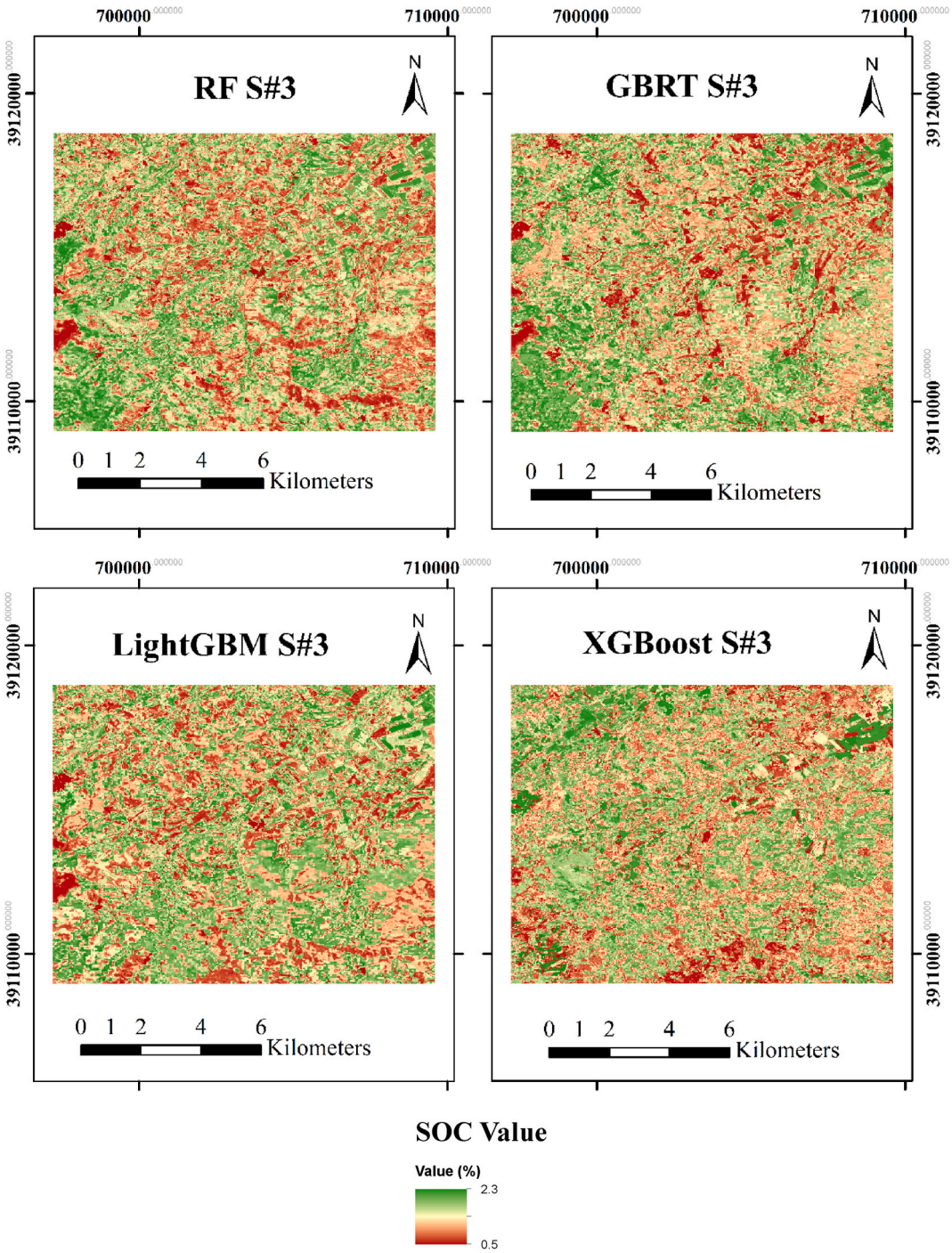


Fig. 12. SOC distributed map resulted from different machine learning algorithms in S#3. It is noteworthy that the grid coordinates in maps are represented based on UTM coordinates in Zone 38N.

CRediT authorship contribution statement

Mehdi Golkar Amoli: Formal analysis, Data curation. **Mahdi Hasanlou:** Supervision, Project administration, Methodology. **Farhad Samadzadegan:** Supervision, Methodology, Investigation, Formal analysis. **Ruhollah Taghizadeh-Mehrjardi:** Software, Resources, Funding acquisition. **Farzaneh Dadrass Javan:** Writing – original draft, Visualization, Data curation.

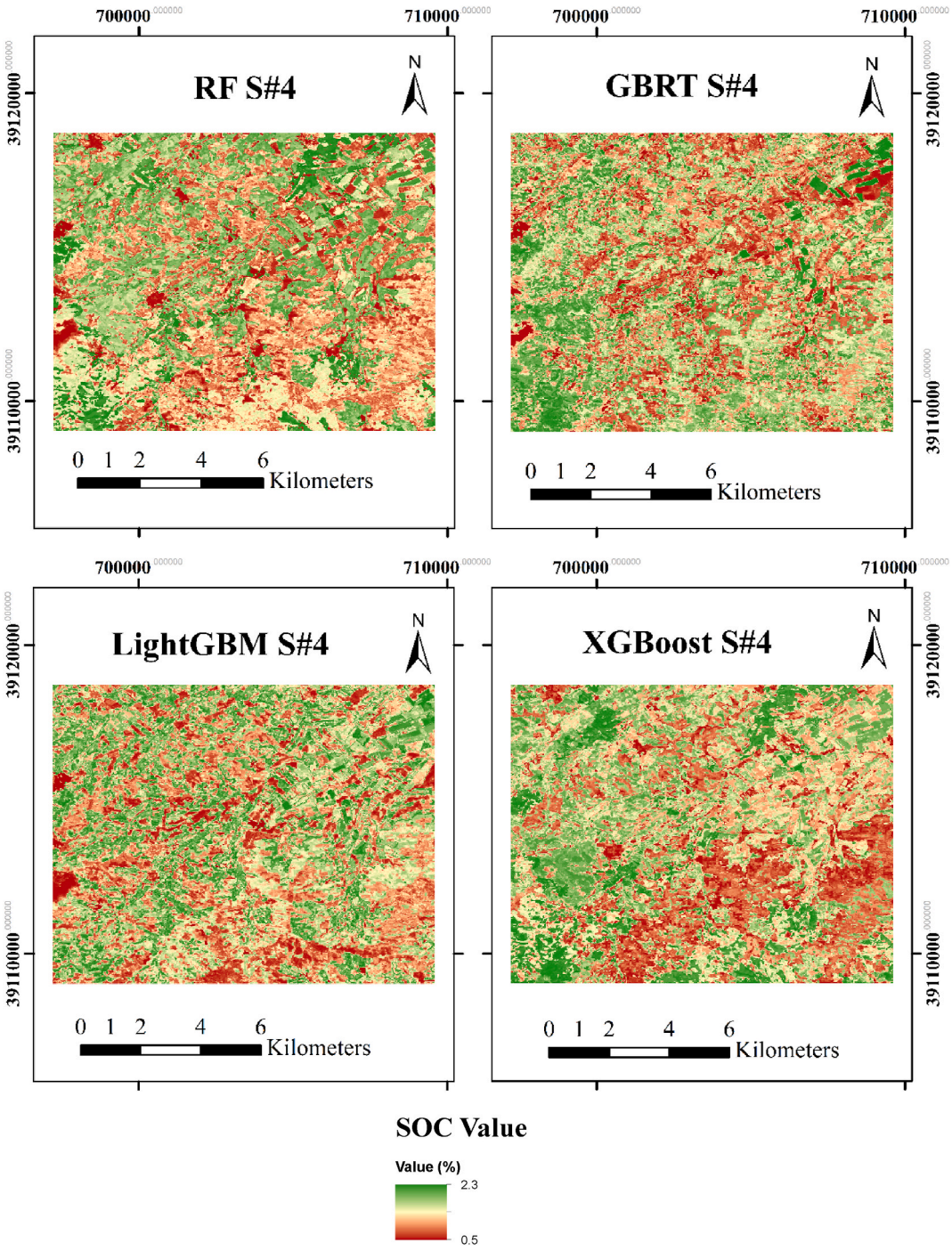


Fig. 13. SOC distributed map resulted from different machine learning algorithms in S#4. It is noteworthy that the grid coordinates in maps are represented based on UTM coordinates in Zone 38N.

Ethical Statement for Remote Sensing Applications: Society and Environment

Hereby, I, Mahdi Hasanlou consciously assure that for the manuscript "Estimating soil organic carbon using time series Band 11 (SWIR) of multispectral Sentinel-2 satellite images and machine learning algorithms," by Mehdi Golkar Amoli, Mahdi Hasanlou, Farhad Samadzadegan, Ruhollah Taghizadeh-Mehrjardi, and Farzaneh Dadrass Javan the following is fulfilled:

1) This material is the authors' own original work, which has not been previously published elsewhere.

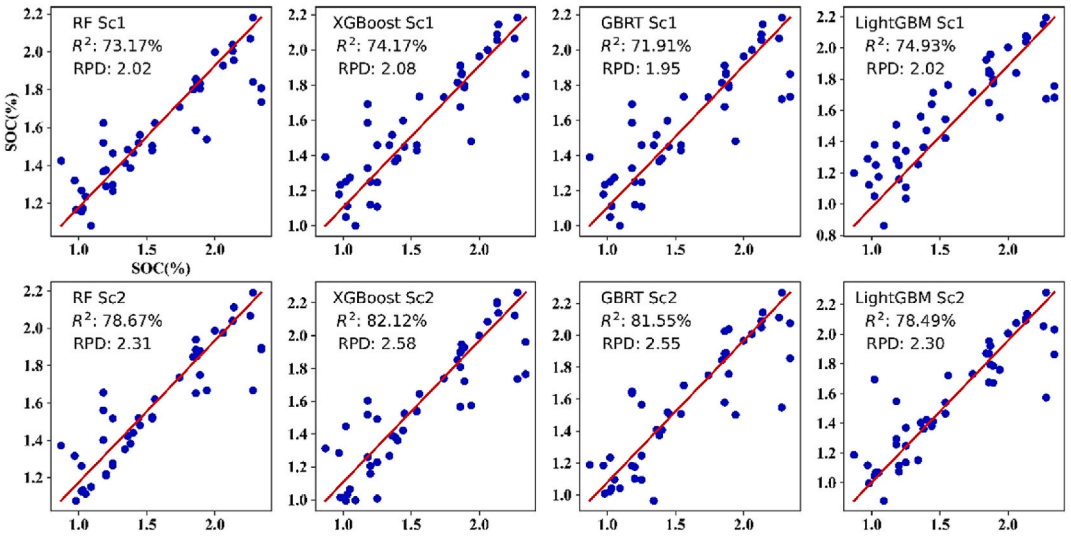


Fig. 14. Scatterplots and corresponding fitting curves of measured SOC and predicted SOC in S#1 (first row) and S#2 (second row). The x-axis is equal to the measured SOC, and the y-axis is equal predicted SOC. Sc1 corresponds to S#1, and Sc2 corresponds to S#2.

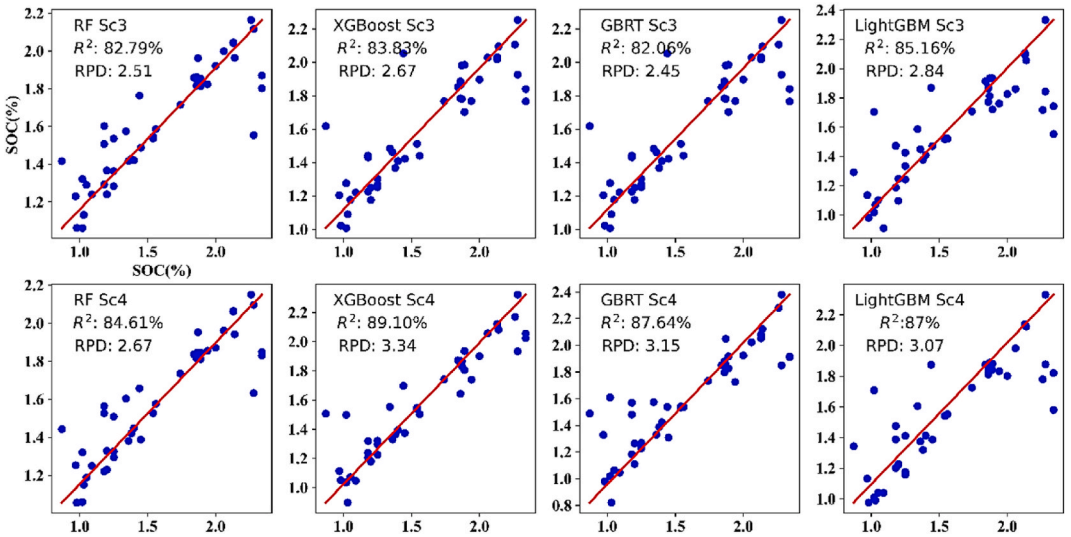


Fig. 15. Scatterplots and corresponding fitting curves of measured SOC and predicted SOC in S#3 (first row) and S#4 (second row). The x-axis is equal to the measured SOC, and the y-axis is equal predicted SOC. Sc3 corresponds to S#3, and Sc4 corresponds to S#4.

2) The paper is not currently being considered for publication elsewhere.
3) The paper reflects the authors' own research and analysis in a truthful and complete manner.
4) The paper properly credits the meaningful contributions of co-authors and co-researchers.
5) The results are appropriately placed in the context of prior and existing research.
6) All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference.
7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

The violation of the Ethical Statement rules may result in severe consequences.

I agree with the above statements and declare that this submission follows the policies of Remote Sensing Applications: Society and Environment as outlined in the Guide for Authors and in the Ethical Statement.

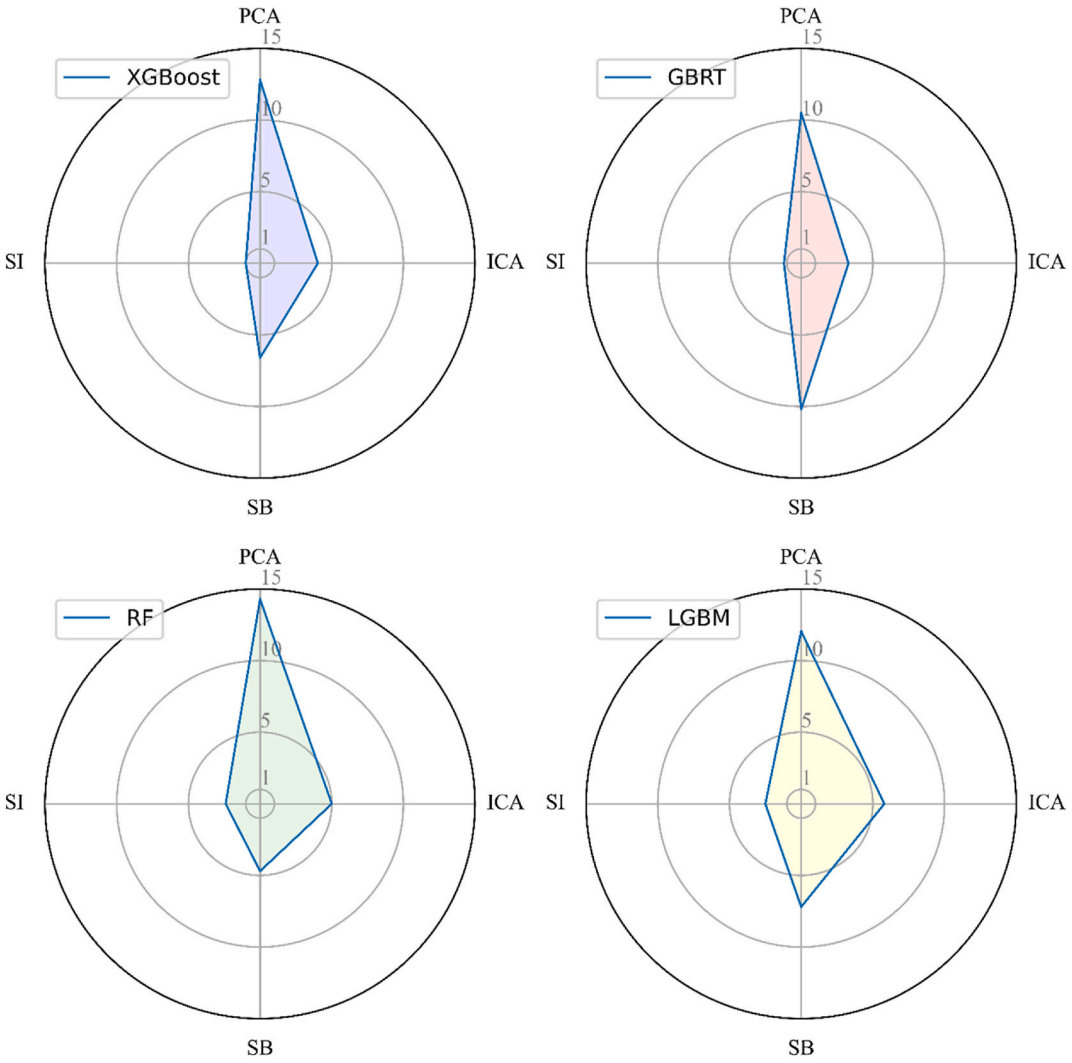


Fig. 16. Feature importance results are based on the permutation importance method for S#4 and across different ML algorithms. SB stands for Spectral Bands of Sentinel, and SI stands for Spectral Indices.

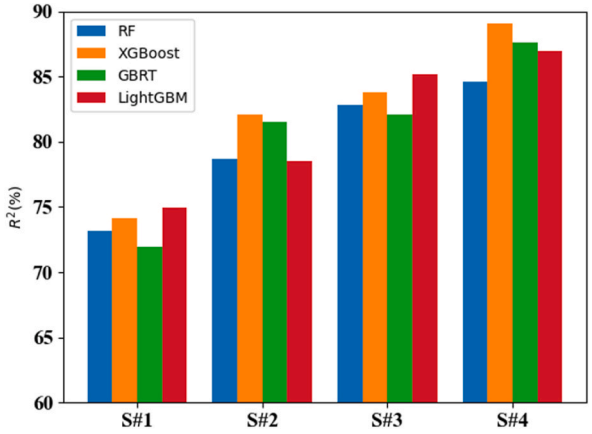


Fig. 17. Comparison of the accuracy of different machine learning algorithms in different scenarios.

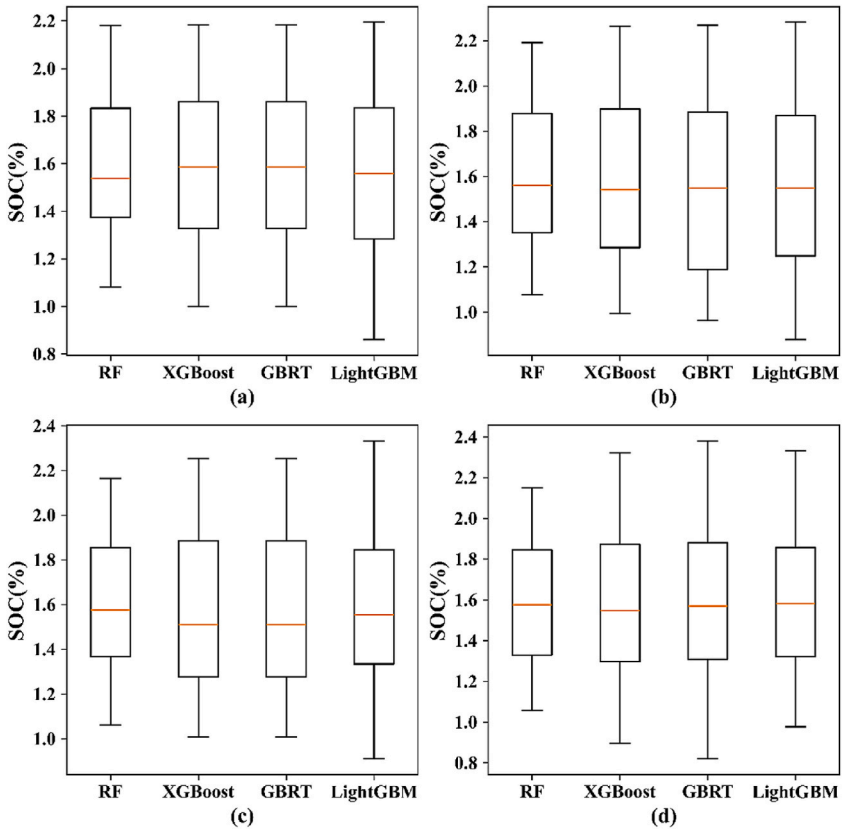


Fig. 18. The Boxplot of SOC prediction value with different machine learning in different scenarios.

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.

References

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D., 2019. Remote sensing techniques for soil organic carbon estimation: a review. Remote Sens. 11 (6), 676.

Bagheri, N., Kafashan, J., 2025. Appropriate vegetation indices and data analysis methods for orchards monitoring using UAV-based remote sensing: a comprehensive research. Comput. Electron. Agric. 235, 110356.

Bagherifam, S., Delavar, M.A., Keshavarz, P., Karimi, A., 2021. A review of soil carbon measurement methods: experimental considerations, advantages and disadvantages. Journal of Soil Management and Sustainable Production 10 (4), 1–30.

Bockheim, J., Gennadiyev, A., Hartemink, A., Brevik, E., 2014. Soil-forming factors and soil Taxonomy. Geoderma 226, 231–237.

Calhoun, V.D., Liu, J., Adalı, T., 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. Neuroimage 45 (1), S163–S172.

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65 (2), 480–490.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2 1 (4), 1–4.

de Santana, F.B., de Giuseppe, L.O., de Souza, A.M., Poppi, R.J., 2019. Removing the moisture effect in soil organic matter determination using NIR spectroscopy and PLSR with external parameter orthogonalization. Microchem. J. 145, 1094–1101.

Diaz-Gonzalez, F.A., Vuelvas, J., Correa, C.A., Vallejo, V.E., Patino, D., 2022. Machine learning and remote sensing techniques applied to estimate soil indicators–review. Ecol. Indic. 135, 108517.

Dvorakova, K., Heiden, U., Pepers, K., Staats, G., van Os, G., van Wesemael, B., 2023. Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties. Geoderma 429, 116128.

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 (12), 1913.

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., Scholten, T., 2020. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. Remote Sens. 12 (14), 2234.
Gallo, B.C., Demattê, J.A., Rizzo, R., Safanelli, J.L., Mendes, W.d.S., Lepsch, I.F., Sato, M.V., Romero, D.J., Lacerda, M.P., 2018. Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology. Remote Sens. 10 (10), 1571.
Gholizadeh, A., Neumann, C., Chabrillat, S., van Wesemael, B., Castaldi, F., Borůvka, L., Sanderman, J., Klement, A., Hohmann, C., 2021. Soil organic carbon estimation using VNIR–SWIR spectroscopy: the effect of multiple sensors and scanning conditions. Soil Tillage Res. 211, 105017.
Golkar Amoli, M., Hasanlou, M., Taghizadeh Mehrjardi, R., Samadzadegan, F., 2024. Exploring the potential of PRISMA satellite hyperspectral image for estimating soil organic carbon in marvdasht region, southern Iran. Remote Sens. 16 (12), 2149.
Gomez, C., Rossel, R.A.V., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: an Australian case study. Geoderma 146 (3–4), 403–411.
He, X., Yang, L., Li, A., Zhang, L., Shen, F., Cai, Y., Zhou, C., 2021. Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. Catena 205, 105442.
Huang, X., Wang, X., Guo, Y., An, B., 2025. Spatial inversion of soil organic carbon content based on hyperspectral data and sentinel-2 images. Land Degrad. Dev. 36, 3527–3544. https://doi.org/10.1002/ldr.5583.
Jiang, Z., Shi, G., Wu, S., Ding, X., Zhao, C., Wong, M.S., Lu, Z., 2025. Unveiling multimodal consolidation process of the newly reclaimed HKIA 3rd runway from satellite SAR interferometry, ICA analytics and Terzaghi consolidation theory. Rem. Sens. Environ. 318, 114561.
Jobbágy, E.G., Jackson, R.B., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10 (2), 423–436.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.
Ladyman, J., Lambert, J., Wiesner, K., 2013. What is a complex system? European Journal for Philosophy of Science 3, 33–67.
Lambora, A., Gupta, K., Chopra, K., 2019. Genetic algorithm-A literature review. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).
Li, T., Cui, L., Wu, Y., McLaren, T.I., Xia, A., Pandey, R., Liu, H., Wang, W., Xu, Z., Song, X., 2024. Soil organic carbon estimation via remote sensing and machine learning techniques: Global topic modeling and research trend exploration. Remote Sens. 16 (17), 3168.
Luo, C., Zhang, W., Zhang, X., Liu, H., 2023. Mapping soil organic matter content using Sentinel-2 synthetic images at different time intervals in Northeast China. International Journal of Digital Earth 16 (1), 1094–1107.
Luo, C., Zhang, X., Wang, Y., Men, Z., Liu, H., 2022. Regional soil organic matter mapping models based on the optimal time window, feature selection algorithm and Google Earth Engine. Soil Tillage Res. 219, 105325.
Melillo, J., Steudler, P., Aber, J., Newkirk, K., Lux, H., Bowles, F., Catricala, C., Magill, A., Ahrens, T., Morrisseau, S., 2002. Soil warming and carbon-cycle feedbacks to the climate system. Science 298 (5601), 2173–2176.
Nagelkerke, N.J., 1991. A note on a general definition of the coefficient of determination. Biometrika 78 (3), 691–692.
Norouzi, S., Sadeghi, M., Liaghat, A., Tuller, M., Jones, S.B., Ebrahimian, H., 2021. Information depth of NIR/SWIR soil reflectance spectroscopy. Rem. Sens. Environ. 256, 112315.
Oliveira, K.M.d., Gonçalves, J.V.F., Furlanetto, R.H., Oliveira, C.A.d., Mendonça, W.A., Haubert, D.d.F.d.S., Crusiol, L.G.T., Falcioni, R., Oliveira, R.B.d., Reis, A.S., 2024. Predicting particle size and soil organic carbon of soil profiles using VIS-NIR-SWIR hyperspectral imaging and machine learning models. Remote Sens. 16 (16), 2869.
Pavlovic, M., Ilic, S., Ralevic, N., Antonic, N., Raffa, D.W., Bandecchi, M., Culibrk, D., 2024. A deep learning approach to estimate soil organic carbon from remote sensing. Remote Sens. 16 (4), 655.
Pouladi, N., Gholizadeh, A., Khosravi, V., Borůvka, L., 2023. Digital mapping of soil organic carbon using remote sensing data: a systematic review. Catena 232, 107409.
Prettenhofer, P., Louppe, G., 2014. Gradient Boosted Regression Trees in Scikit-Learn. PyData 2014.
Roweis, S., 1998. Em algorithms for pca and spca. Advances in NeuralInformation Processing Systems. The MIT Press.
Saberioon, M., Gholizadeh, A., Ghaznavi, A., Chabrillat, S., Khosravi, V., 2024. Enhancing soil organic carbon prediction of LUCAS soil database using deep learning and deep feature selection. Comput. Electron. Agric. 227, 109494.
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.
Shao, Z., Ahmad, M.N., Javed, A., 2024. Comparison of random forest and xgboost classifiers using integrated optical and sar features for mapping urban impervious surface. Remote Sens. 16 (4), 665.
Stromann, O., Nascetti, A., Yousif, O., Ban, Y., 2019. Dimensionality reduction and feature selection for object-based land cover classification based on Sentinel-1 and Sentinel-2 time series using Google Earth Engine. Remote Sens. 12 (1), 76.
Sun, Q., Zhang, P., Jiao, X., Lun, F., Dong, S., Lin, X., Li, X., Sun, D., 2022. A remotely sensed framework for spatially-detailed dryland soil organic matter mapping: coupled cross-wavelet transform with fractional vegetation and soil-related endmember time series. Remote Sens. 14 (7), 1701.
Taghadosi, M.M., Hasanlou, M., Eftekhari, K., 2019. Retrieval of soil salinity from Sentinel-2 multispectral imagery. European Journal of Remote Sensing 52 (1), 138–154.
Urbina-Salazar, D., Vaudour, E., Richer-de-Forges, A.C., Chen, S., Martelet, G., Baghdadi, N., Arrouays, D., 2023. Sentinel-2 and sentinel-1 bare soil temporal mosaics of 6-year periods for soil organic carbon content mapping in Central France. Remote Sens. 15 (9), 2410.
Verstraete, M.M., Pinty, B., 1996. Designing optimal spectral indexes for remote sensing applications. IEEE Trans. Geosci. Rem. Sens. 34 (5), 1254–1265.
Wan, C., Kuzyakov, Y., Cheng, C., Ye, S., Gao, B., Gao, P., Ren, S., Yun, W., 2021. A soil sampling design for arable land quality observation by using SPCOSA–CLHS hybrid approach. Land Degrad. Dev. 32 (17), 4889–4906.
Wang, J., Chang, C.-I., 2006. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. IEEE Trans. Geosci. Rem. Sens. 44 (6), 1586–1600.
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 (7), 2685.
Xue, J., Zhang, X., Chen, S., Chen, Z., Lu, R., Liu, F., van Wesemael, B., Shi, Z., 2025. National-scale mapping topsoil organic carbon of cropland in China using multitemporal Sentinel-2 images. Geoderma 456, 117272.
Yang, J., Fan, J., Lan, Z., Mu, X., Wu, Y., Xin, Z., Miping, P., Zhao, G., 2022. Improved surface soil organic carbon mapping of SoilGrids250m using sentinel-2 spectral images in the qinghai–Tibetan plateau. Remote Sens. 15 (1), 114.
Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., Li, D.-C., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecol. Indic. 60, 870–878.
Ye, Z., Sheng, Z., Liu, X., Ma, Y., Wang, R., Ding, S., Liu, M., Li, Z., Wang, Q., 2021. Using machine learning algorithms based on GF-6 and Google Earth engine to predict and map the spatial distribution of soil organic matter content. Sustainability 13 (24), 14055.
Zeraatpisheh, M., Ayoubi, S., Mirbagheri, Z., Mosaddeghi, M.R., Xu, M., 2021. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. Geoderma Regional 27, e00440.
Zeraatpisheh, M., Garosi, Y., Owliaie, H.R., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., Xu, M., 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: a comparison of a group of environmental covariates. Catena 208, 105723.
Zhang, H., Wan, L., Li, Y., 2023. Prediction of soil organic carbon content using sentinel-1/2 and machine learning algorithms in swamp wetlands in Northeast China. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens.
Zhang, L., Cai, Y., Huang, H., Li, A., Yang, L., Zhou, C., 2022. A CNN-LSTM model for soil organic carbon content prediction with long time series of MODIS-based phenological variables. Remote Sens. 14 (18), 4441.

Zhang, M., Zhang, M., Yang, H., Jin, Y., Zhang, X., Liu, H., 2021. Mapping regional soil organic matter based on sentinel-2a and modis imagery using machine learning algorithms and google earth engine. Remote Sens. 13 (15), 2934.
Zhang, T., Li, Y., Wang, M., 2024. Remote sensing-based prediction of organic carbon in agricultural and natural soils influenced by salt and sand mining using machine learning. J. Environ. Manag. 352, 120107.
Zhao, X., Xiong, Z., Karlshöfer, P., Tziolas, N., Wiesmeier, M., Heiden, U., Zhu, X.X., 2025. Soil organic carbon estimation using spaceborne hyperspectral composites on a large scale. Int. J. Appl. Earth Obs. Geoinf. 140, 104504.
Zurqani, H.A., 2025. A multi-source approach combining GEDI LiDAR, satellite data, and machine learning algorithms for estimating forest aboveground biomass on Google Earth engine platform. Ecol. Inform., 103052