remote sensing

MDPI

Article

Synergistic Use of Multi-Temporal Radar and Optical Remote Sensing for Soil Organic Carbon Prediction

Sara Dahhani [1,*], Mohamed Raji [1] and Yassine Bouslihim [2]

[1] Faculty of Sciences Ben M'sik, Hassan II University of Casablanca, Sidi Othmane, Casablanca P.O. Box 7955, Morocco
[2] National Institute of Agricultural Research (INRA), CRRA Tadla, Rabat P.O. Box 415, Morocco; yassine.bouslihim@inra.ma or yassine.bouslihim@gmail.com
* Correspondence: s.dahhani.gis@gmail.com

Abstract: Exploring soil organic carbon (SOC) mapping is crucial for addressing critical challenges in environmental sustainability and food security. This study evaluates the suitability of the synergistic use of multi-temporal and high-resolution radar and optical remote sensing data for SOC prediction in the Kaffrine region of Senegal, covering over 1.1 million hectares. For this purpose, various scenarios were developed: Scenario 1 (Sentinel-1 data), Scenario 2 (Sentinel-2 data), Scenario 3 (Sentinel-1 and Sentinel-2 combination), Scenario 4 (topographic features), and Scenario 5 (Sentinel-1 and -2 with topographic features). The findings from comparing three different algorithms (Random Forest (RF), XGBoost, and Support Vector Regression (SVR)) with 671 soil samples for training and 281 samples for model evaluation highlight that RF outperformed the other models across different scenarios. Moreover, using Sentinel-2 data alone yielded better results than using only Sentinel-1 data. However, combining Sentinel-1 and Sentinel-2 data (Scenario 3) further improved the performance by 6% to 11%. Including topographic features (Scenario 5) achieved the highest accuracy, reaching an R^2 of 0.7, an RMSE of 0.012%, and an RPIQ of 5.754 for the RF model. Applying the RF and XGBoost models under Scenario 5 for SOC mapping showed that both models tended to predict low SOC values across the study area, which is consistent with the predominantly low SOC content observed in most of the training data. This limitation constrains the ability of ML models to capture the full range of SOC variability, particularly for less frequent, slightly higher SOC values.

Keywords: soil organic carbon; Sentinel-1; Sentinel-2; multi-temporal data; radar imagery; optical imagery

check for updates

Citation: Dahhani, S.; Raji, M.; Bouslihim, Y. Synergistic Use of Multi-Temporal Radar and Optical Remote Sensing for Soil Organic Carbon Prediction. *Remote Sens.* **2024**, *16*, 1871. https://doi.org/10.3390/rs16111871

Academic Editors: Xiaoling Wu, Chong Luo, Liujun Zhu and Xiaoji Shen

Received: 25 March 2024
Revised: 11 May 2024
Accepted: 13 May 2024
Published: 24 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Soil organic carbon (SOC) constitutes an essential element within the global carbon cycle, playing an important role in mitigating climate change, improving soil health, and enhancing agricultural productivity. Quantifying and monitoring SOC content is essential for evaluating soil quality, orienting sustainable land management practices, and achieving international climate change mitigation commitments [1]. Consequently, SOC mapping has garnered global interest as a means of addressing environmental and food security challenges. Interest in SOC mapping has been particularly pronounced in Africa [2–4], which faces a unique combination of challenges and opportunities in soil management. The diverse climates and ecosystems of Africa present a varied soil landscape, where accurate SOC mapping can make a significant contribution to improving agricultural resilience, food security, and climate change adaptation efforts [5]. Furthermore, digital mapping of SOC in a sub-Saharan country like Senegal can make a significant contribution to achieving several Sustainable Development Goals (SDGs).

In this context, the integration of machine learning (ML) algorithms with Earth observation (EO) data has been recognized as a powerful approach for improving the accuracy and efficiency of SOC prediction and mapping [6,7]. According to Nenkam Mentho et al. [8],

among 110 studies conducted in Africa, 34 and 6 specifically focused on SOC and soil organic matter (SOM), respectively, both with and without the consideration of other soil attributes. For instance, Hengl et al. [9] demonstrated the utility of the Africa Soil Information Service (AfSIS) in conjunction with Moderate Resolution Imaging Spectroradiometer (MODIS) data for the mapping of various soil properties, including SOC and pH, at a resolution of 250 m. Utilizing the same data source, Vågen et al. [5] employed a Random Forest model for SOC mapping across the African continent. Furthermore, Hengl et al. [10] generated 30 m resolution pan-African maps detailing various soil nutrients, such as SOC, pH, total nitrogen (N), phosphorus (P), and potassium (K), among others, through the combination of diverse EO datasets and ensemble ML algorithms. Bouasria et al. [11] explored the feasibility of utilizing pan-sharpened Landsat-8 imagery (15 m resolution) for SOM mapping via multiple linear regression and artificial neural networks. Similarly, Bouslihim et al. [12] employed a Random Forest approach for SOM mapping using Landsat-8 imagery at a 30 m resolution.

Recent advances in remote sensing technologies have expanded the opportunities for digital soil mapping (DSM). Sentinel-1 (C-band synthetic aperture radar) and Sentinel-2 (multi-spectral optical data) satellites can provide unprecedented opportunities for detailed and frequent monitoring of the Earth's surface, including soil properties. While Sentinel-2 provides high-resolution optical images useful for capturing surface features and vegetation indices, Sentinel-1 radar data offer advantages by penetrating cloud cover and providing information on soil moisture, which is closely linked to SOC content [13,14]. Within the African context, out of 110 studies, 11 have utilized Sentinel-2 data for DSM purposes, yet only 2 have yielded SOC maps at a 10 m resolution [8]. In the first study, Mponela et al. [15] used Sentinel-2 data to determine soil fertility (including SOC, NPK, etc.) for a 0.45 ha area in Malawi. Additionally, Flynn et al. [16] predicted soil particle size distribution and SOC content at a 10 m resolution over a 366 ha area in South Africa. Despite the potential, the application of Sentinel data in Africa for SOC mapping remains underexploited. Predominantly, global studies have employed Sentinel data from a single date [17–21]. However, a limited number of investigations have harnessed multi-temporal data from Sentinel-1 or Sentinel-2 for enhanced analysis [22–24].

This study investigates several hypotheses related to DSM for SOC prediction. Firstly, we hypothesized that the combined use of multi-temporal Sentinel-1 and Sentinel-2 data would outperform the individual use of either data source in predicting SOC content. Secondly, we posited that incorporating topographic features as auxiliary environmental variables would further enhance the accuracy of SOC prediction models. Finally, we anticipated that different machine learning algorithms (RF, SVR, and XGBoost) would exhibit varying performance levels depending on the specific combination of input variables and the chosen scenario. To test these hypotheses, we evaluated the efficacy of these data sources and algorithms across various scenarios, aiming to identify the optimal approach for generating high-resolution SOC maps. This research contributes valuable insights into the synergistic potential of Sentinel data and the role of environmental variables and machine learning in advancing digital soil mapping techniques for SOC prediction. In addition, this paper supports SDG 13 (Climate Action) by providing crucial data for understanding and monitoring carbon sequestration capacities, thus informing climate change mitigation strategies, and SDG 15 (Life on Earth) through its potential to improve soil health, promote sustainable land use practices, and combat desertification, which is particularly important in arid and semi-arid regions. In addition, by enabling better-informed agricultural practices, this research indirectly contributes to SDG 2 (Zero Hunger) and SDG 1 (No Poverty) by improving food security and livelihoods through improved soil fertility and crop yields. Thus, digital mapping of soil organic carbon serves as a multi-disciplinary tool that cuts across various environmental and socio-economic aspects of sustainable development in the context of African countries.

2. Materials and Methods

2.1. Methodology

The flowchart presented in Figure 1 outlines the process for predicting SOC using Sentinel-1 (radar) and Sentinel-2 (multi-spectral) data, topographic features, and ML algorithms. The methodology is divided into three main stages.

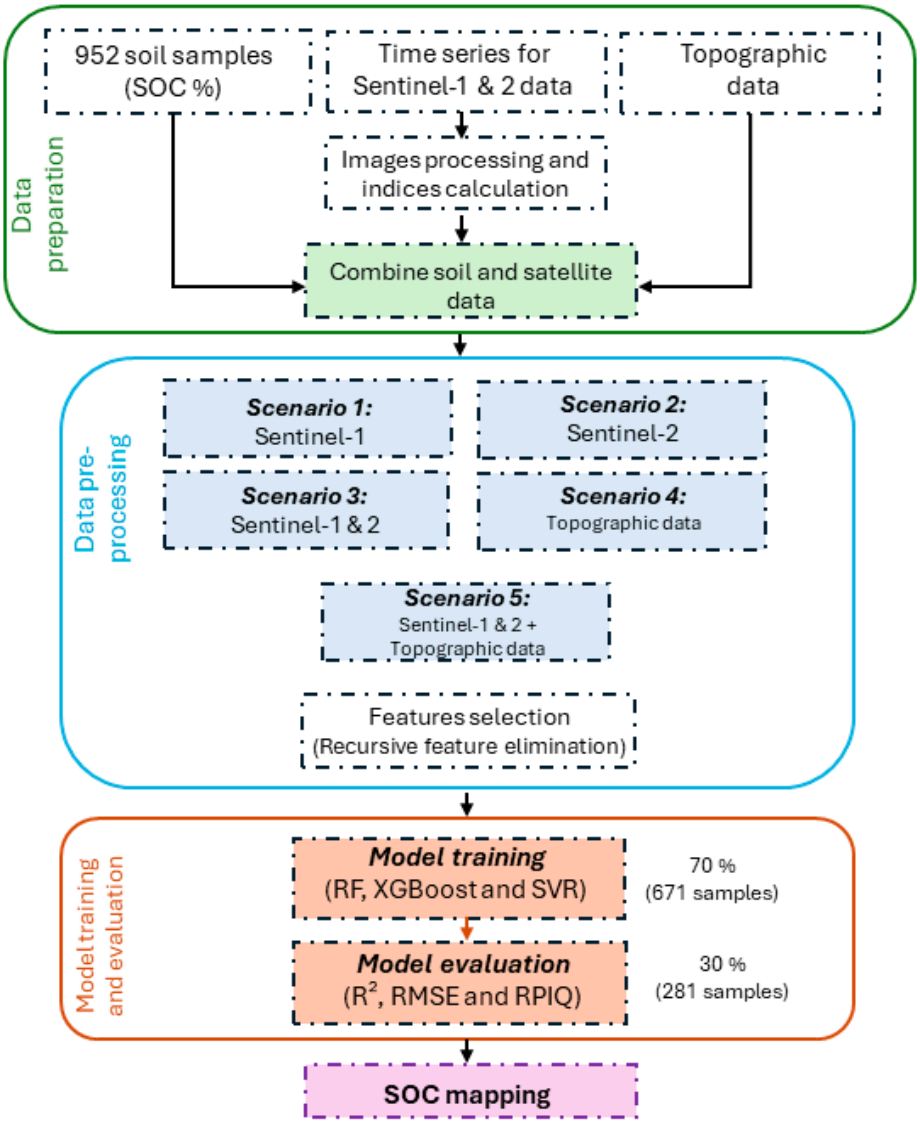

Figure 1. Methodological flowchart adopted to predict SOC under different scenarios.

(1) Data preparation: Multi-temporal data from Sentinel-1 and Sentinel-2 were processed for various radiometric and geometric image corrections, Sentinel-2 bands were used to extract various remote sensing indices, and topographic features were prepared, and all these data were combined with SOC content ground samples (952).

(2) Data pre-processing: The prepared data were considered under four scenarios to evaluate the suitability of Sentinel products for SOC prediction: Scenario 1 (only Sentinel-1 data), Scenario 2 (only Sentinel-2 data), Scenario 3 (Sentinel-1 and -2 combination), Scenario 4 (topographic features), and Scenario 5 (Scenario 3 and Scenario 4). Also, feature selection was used to identify the most relevant variables for SOC prediction.

(3) Modeling and evaluation: Three ML algorithms were applied, Random Forest (RF), XGBoost, and Support Vector Regression (SVR), using a 70/30% split for training and

testing. All models were evaluated using the coefficient of determination (R^2), the Root Mean Square Error (RMSE), and the Ratio of Performance to Inter-Quartile Range (RPIQ).

(4) Finally, the best models were used for SOC spatial prediction.

2.2. Study Area Description

The Kaffrine region covers an area of 11,181 km^2 (≈1.1 million ha), representing approximately 5.6% of Senegal. It is situated in central Senegal, bounded by the coordinates 14°43′46.6″N 15°51′40.2″W and 13°45′31.4″N 14°34′00.7″W (Figure 2). The region serves as a transitional zone between the Sahelian and Sudanian climatic domains. The topography is predominantly flat, with a gentle slope descending from north to south. The area is characterized by three primary soil types: tropical ferruginous, hydromorphic, and holomorphic soils. Climatically, Kaffrine experiences high temperatures throughout the year, with notable fluctuations, and has a distinct seasonal pattern comprising a short rainy season from July to October and a prolonged dry season lasting from eight to nine months. The average annual rainfall recorded for the period from 2016 to 2021 was approximately 702.6 mm.

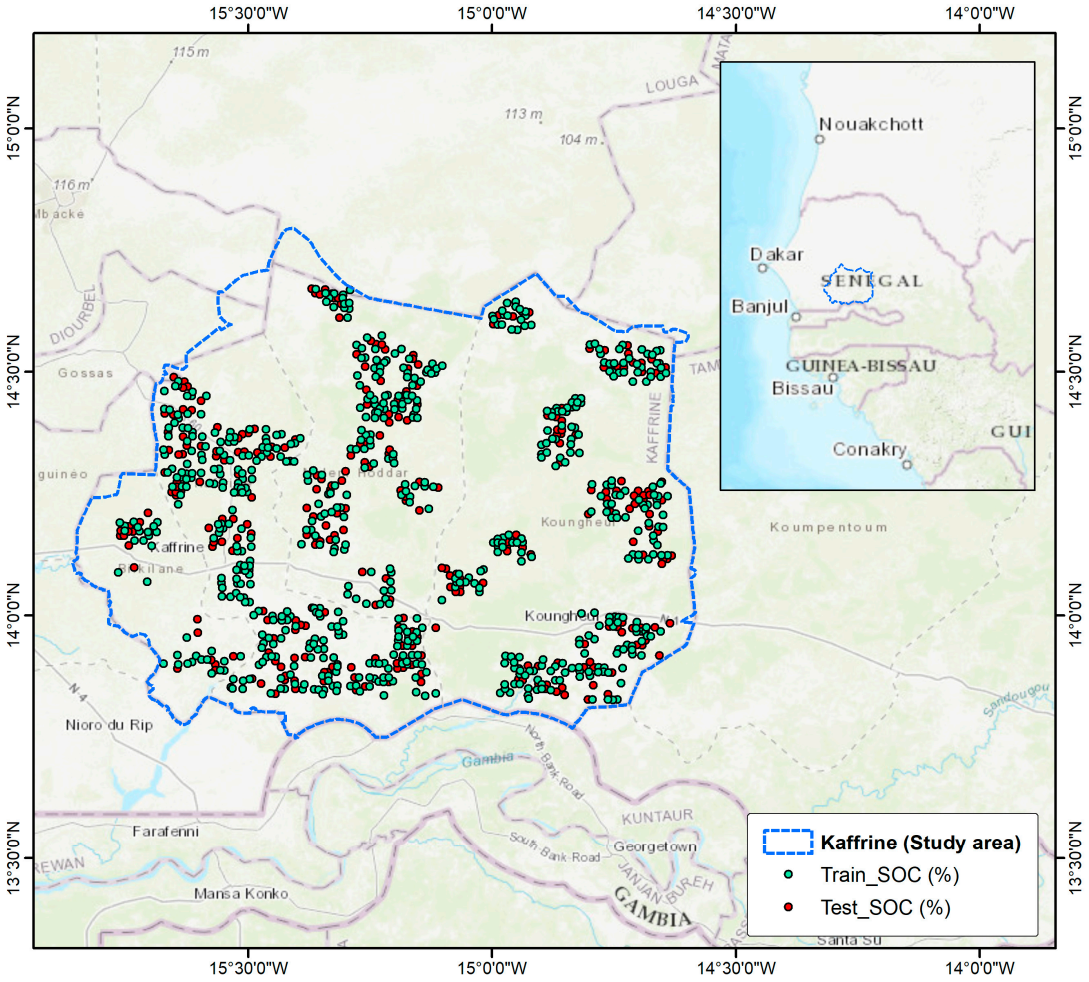

Figure 2. Limits of Kaffrine region and geographical localization of soil samples.

2.3. Soil and Remote Sensing Data Preparation

Soil data: As a first step, a soil sampling design was structured to ensure that the sampling accurately represented the soil properties across the study area. For that, a stratified random sampling design was applied, and the entire study area was partitioned into different blocks (10 × 10 km), which enabled a systematic organization of the sampling effort and ensured extensive coverage of the study area. Out of these blocks, 45 were

selected through a random selection process to ensure that our sampling represented the various landforms and soil types within the study area, thereby minimizing any potential bias that could arise from selectively choosing specific blocks. Subsequently, in each of these 45 randomly selected blocks, soil sampling was conducted at 23 distinct sites, and some sites were eliminated due to access constraints. Between 2018 and 2019, soil samples were collected at each site from the top 20 cm. After collection, the soil samples were transported to the laboratory for preparation and analysis. The preparation involved drying the soil, removing all plant debris, and sieving through a 2 mm mesh to achieve a uniform soil fraction for analysis, and the SOC content was measured using the Walkley–Black method [25].

Remote sensing data: The multi-temporal dataset included images from Sentinel-1, obtained from https://search.asf.alaska.edu/ (accessed on 22 Decembre 2023), and Sentinel 2, obtained from the Copernicus Data Space Ecosystem (https://browser.dataspace.copernicus.eu/, accessed on 15 January 2024). For Sentinel-1, the dataset included a series of synthetic aperture radar (SAR) images extending from May 2018 to March 2019, including 4 scenes to cover the study area. These images featured dual polarization modes (VH and VV) and were all captured in an ascending orbit. The pre-processing steps for Sentinel-1 imagery were performed using SNAP (8.0.0) software, encompassing calibration to convert digital number (DN) values into backscatter coefficients, multi-looking to reduce speckle noise, and filtering to further improve image quality, and, since SAR images have side view imaging characteristics, SAR image geometric misrepresentation may appear in relief displacement. The Radar Geometric Terrain Correction tool was chosen to apply the Range Doppler method for image registration [26,27]. In total, we obtained 22 images (11 for VH polarization and 11 for VV polarization).

Furthermore, Sentinel-2 L1C multi-spectral images were acquired from May 2018 to March 2019 (July and September were excluded due to unfavorable weather conditions). A total of 9 acquisition dates were obtained and atmospherically corrected using the sen2cor processor in the SNAP (8.0.0) software [28]. Sentinel-2 bands at each date were used to calculate various remote sensing indices, such as Brightness Index (BI), Coloration Index (CI), Modified Normalized Difference Water Index (MNDWI), MERIS Terrestrial Chlorophyll Index (MTCI), Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Redness Index (RI), and Soil-Adjusted Vegetation Index (SAVI) values, and the formulas used for the index calculations are detailed in Table 1. The index labels were coded as follows: *Index_Month_Year*; for instance, NDVI_5_18 refers to the NDVI for May 2018.

Table 1. List of remote sensing indices calculated from Sentinel-2 bands.

Index	Full Name	Formula	Reference
BI	Brightness Index	sqrt ($(Red^2/Green^2)/2$)	[29]
CI	Coloration Index	(Red − Blue)/Red	[30]
MNDWI	Modified Normalized Difference Water Index	(Green − SWIR)/(Green + SWIR)	[31]
MTCI	MERIS Terrestrial Chlorophyll Index	(Red Edge 2 − Red Edge 1)/(Red Edge 1 − Red)	[32]
NDVI	Normalized Difference Vegetation Index	((NIR − Red)/(NIR + Red))	[33]
NDWI	Normalized Difference Water Index	(Green − NIR)/(Green + NIR)	[34]
RI	Redness Index	(Red − Green)/(Red + Green)	[33]
SAVI	Soil-Adjusted Vegetation Index	((NIR − Red)/(NIR + Red + L)) × (1 + L)	[35]

The digital elevation model was obtained from ASTGTM (version 3 with a 30 m resolution) and used to extract different topographic features, such as elevation, slope, aspect, Topographic Wetness Index (TWI), profile curvature, plan curvature, and Multi-Resolution Index of Valley Bottom Flatness (MRVBF), using the SAGA program (version 9.1.2). The

elevation band was resampled to a 10 m resolution using the bilinear interpolation method in QGIS Desktop (version 3.34.0) before the calculation of other topographic features.

2.4. Data Pre-Processing and Machine Learning Algorithms

The Recursive Feature Elimination (RFE) method was employed, utilizing a Random Forest Regressor as the estimator [36]. RFE selects features by recursively considering progressively smaller sets of features; it initiates with all predictors in the dataset and sequentially removes the least significant feature at each iteration. A *Random Forest Regressor* was configured with 100 trees and a fixed random state (42) to ensure reproducibility and was instructed to select the top 10 features for Scenario 1 and Scenario 2 and the top 20 features for Scenario 3 and Scenario 5 based on the training dataset. For Scenario 4, all seven topographic features were used.

Furthermore, three different ML algorithms were compared: (1) Random Forest (RF), an ensemble learning method that constructs a multitude of decision trees at the training time and outputs the average prediction (for the regression task) of the individual trees [37]. RF is highly recommended for remote sensing applications due to its ability to handle large datasets and its robustness against overfitting, which makes it a powerful tool for land cover classification [38], estimation of soil properties [39], and biomass prediction [40], among other applications. (2) XGBoost (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient-boosted decision trees, designed for speed and performance. Developed by Chen and Guestrin [41], XGBoost has gained popularity through its performance in ML challenges and has been noted for its ability to handle sparse data and for its scalability and regularized boosting technique that helps prevent overfitting. (3) Support Vector Regression (SVR) applies the principles of support vector machines (SVMs) to regression problems. The SVR model aims to fit the best line within a predefined or epsilon margin of tolerance with the goal of minimizing error and fitting the model within the defined threshold [42]. RF and SVR models were selected due to their widespread application in DSM applications and their history of yielding diverse results. Comparing these established methods allowed us to assess which is more suitable for this specific case. Additionally, we included XGBoost, which can be considered a newer ML algorithm, to explore its potential benefits in SOC prediction.

Each model was developed using 70% of the data (n = 671), and the remaining 30% (n = 281) was used for model testing. Also, the hyperparameter tuning approach was applied for every model based on its parameters, and all these parameters are listed in Table 2 with descriptions. The Google Colab platform [43] was used for all steps related to data pre-processing, predictive modeling, and SOC mapping. Due to the limited performance of the free version of Google Colab, the layer stack prepared for SOC mapping was divided into 10 parts, and each part was used with the desired model to predict SOC; afterwards, the 10 parts were mosaicked to return a single SOC raster. The complete script utilized in Google Colab for this research is accessible on GitHub (links are available in the Data Availability Statement).

All developed models were evaluated using three metrics: the coefficient of determination (R^2) (Equation (1)), the Root Mean Square Error (RMSE) (Equation (2)), and the Ratio of Performance to Inter-Quartile Range (RPIQ) (Equation (3)). R^2 provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of the total variation in outcomes explained by the model [44]. An R^2 of 1 indicates perfect correlation, while an R^2 of 0 indicates that the model does not explain any of the variability in the response data around its mean. The RMSE is a standard way to measure the error of a model in predicting quantitative data [45]. It represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. RMSE is particularly useful when large errors are particularly undesirable, as it squares the errors before averaging, thus giving a relatively high weight to large errors. The RPIQ is calculated by dividing the interquartile distance (IQR) by the RMSE [46]; higher RPIQ values indicate better model performance, as they

suggest that the model's predictions are accurate relative to the natural variability of the data, and lower RPIQ values suggest that the model's predictions are less accurate, with prediction errors that are large in comparison to the variability of the dataset.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum\left(y_i - \underline{y}\right)^2} \quad (1)$$

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

$$RPIQ = {}^{IQR}/_{RMSE} \quad (3)$$

where y_i is the actual value of the dependent variable for the ith observation, $\hat{y}_i$ is the predicted value of the dependent variable for the ith observation, and $\underline{y}$ is the mean value of the dependent variable. The IQR represents the range between the first (25th percentile) and third quartiles (75th percentile) of the observed data.

Table 2. List of hyperparameters used for RF, SVR, and XGBoost model tuning.

Model	Hyperparameter	Description
RF	n_estimators	The number of trees in the forest
	max_features	The number of features to consider when looking for the best split
	max_depth	The maximum depth of the tree
	min_samples_split	The minimum number of samples required to split an internal node
	min_samples_leaf	The minimum number of samples required to be at a leaf node
SVR	C	Regularization parameter
	epsilon	Specifies the epsilon tube
	gamma	Kernel coefficient for r'bf', p'oly', and s'igmoid'.
XGBoost	learning_rate	(or eta in XGBoost documentation) Step size shrinkage used to prevent overfitting
	max_depth	Maximum depth of a tree
	gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree
	colsample_bytree	Control the subsample ratio of columns for the tree building at different levels of tree building
	min_child_weight	Minimum sum of instance weight (hessian) needed in a child
	subsample	Subsample ratio of the training instances
	n_estimators	Number of gradient boosted trees, equivalent to the number of boosting rounds

3. Results

3.1. Statistical Description

For the training dataset with 671 samples (Table 3), the SOC content ranges from a minimum of 0.11% to a maximum of 0.72%, with a mean value of approximately 0.22%. The standard deviation is 0.0725, indicating a moderate spread around the mean. The 25th, 50th (median), and 75th percentiles are 0.175%, 0.21%, and 0.26%, respectively, showing a slight skew towards lower SOC values (Figure 3). Comparatively, the test dataset (281 samples) shows a slightly tighter range of SOC values, from 0.12% to 0.57%, with a mean value very close to that of the training set, at about 0.22%. The standard deviation in the test set is slightly lower at 0.0692, suggesting a slightly less varied set of SOC percentages than in the training dataset. Percentile values are also similar to those of the training set, with the 25th, 50th, and 75th percentiles at 0.18%, 0.21%, and 0.26%, respectively. Overall, both datasets show a relatively consistent range of SOC percentages, with a central tendency around 0.22%. The slight differences in spread and range between the training and test datasets

suggest minor variations in soil organic carbon content across the two datasets, but, overall, they exhibit similar statistical properties, with a low SOC content.

Table 3. Summary statistics for train and test SOC (%) data.

Data	Count	Min	Max	Mean	Standard Deviation
Train	671	0.11	0.72	0.224	0.072
Test	281	0.12	0.57	0.223	0.069

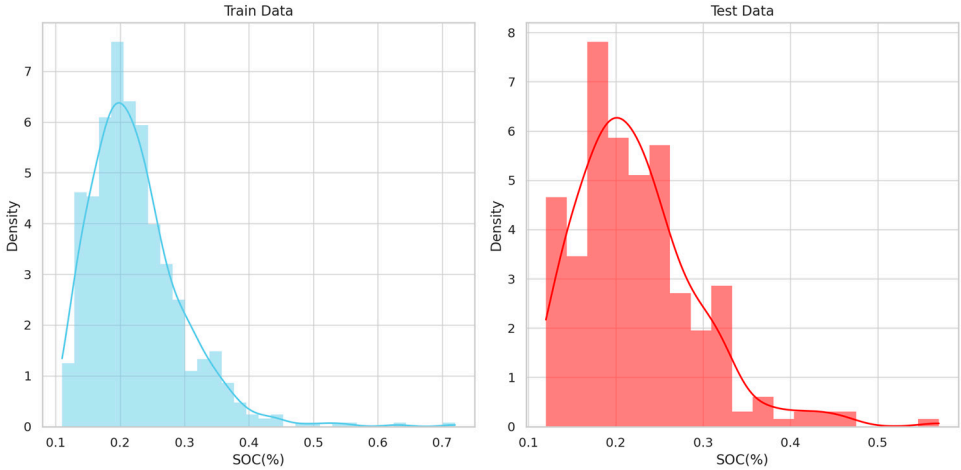

Figure 3. Distribution of SOC content for train and test data.

3.2. Feature Selection and Correlation Analysis

The Recursive Feature Elimination (RFE) method was used to select the most influential features across four scenarios. The important variables identified for different scenarios are listed in Table 4. For Sentinel-1 data in Scenario 1, out of 10 selected features, 5 VH features were selected for different months (VH_5_18, VH_6_18, VH_8_18, VH_9_18, and VH_3_19) and 5 VV features were selected for July, September, and December 2018 and for February and March 2019. For Sentinel-2 data in Scenario 2, three MNDWI features (MNDWI_6_18, MNDWI_12_18, and MNDWI_3_19), two SAVI features (SAVI_8_18 and SAVI_12_18), two MTCI features (MTCI_11_18 and MTCI_3_19), and two CI features (CI_5_18 and CI_3_19) were selected with one BI for June 2018. In the third scenario, which combined Sentinel-1 and Sentinel-2 data, 20 features were selected to equally represent both datasets and to see if the same features would be selected. The results revealed that, from the 20 features, 15 were selected from Sentinel-2 data and only 5 from Sentinel-1, including 3 VH variables and 2 VV variables. For Scenario 4, all seven topographic features were used, and the high importance of elevation was noticed. For the last scenario, which combined Sentinel-1, Sentinel-2, and topographic features, 1 topographic feature was selected (elevation) and 4 and 15 features were selected from Sentinel-1 and Sentinel-2, respectively.

The correlation analysis revealed that the backscatter coefficient in VV polarization for March 2019 (VV_3_19) showed the highest correlation with SOC, with a value of 0.202, followed by VV_12_18 polarization with a negative correlation of −0.17 and VH_6_18 with a positive correlation of 0.16. This indicates that the relationship between SOC content and radar backscatter in both VV and VH polarizations is not uniformly positive or negative; rather, it is variable. Both polarizations exhibit correlations with SOC, albeit with differing magnitudes and directions (positive and negative). For Sentinel-2 multi-spectral data, the MNDWI exhibited a generally positive correlation with SOC, with coefficients ranging from 0.15 to 0.27. Conversely, a negative correlation was observed between SOC content and SAVI, with coefficients from −0.25 to −0.28. Similarly, the CI demonstrated a negative correlation with SOC, with values between −0.27 and −0.36. Lastly, elevation showed the highest correlation value of 0.42.

Table 4. List of selected features (bands) across different scenarios.

Scenario	Selected Features
Scenario 1 (Sentinel-1)	VH_5_18, VH_6_18, VH_8_18, VH_9_18, VH_3_19, VV_7_18, VV_9_18, VV_12_18, VV_2_19, VV_3_19
Scenario 2 (Sentinel-2)	BI_6_18, CI_5_18, CI_3_19, MNDWI_6_18, MNDWI_12_18, MNDWI_3_19, MTCI_11_18, MTCI_3_19, SAVI_8_18, SAVI_12_18
Scenario 3 (Sentinel-1 + Sentinel-2)	BI_5_18, BI_6_18, CI_5_18, CI_3_19, MNDWI_5_18, MNDWI_6_18, MNDWI_12_18, MNDWI_3_19, MTCI_10_18, MTCI_11_18, MTCI_3_19, NDWI_8_18, SAVI_8_18, SAVI_12_18, SAVI_3_19, VH_5_18, VH_6_18, VH_9_18, VV_9_18, VV_3_19
Scenario 4 (Topography)	Elevation, slope, aspect, TWI, profile curvature, plan curvature, MRVBF
Scenario 5 (Sentinel-1 + Sentinel-2 + Topography)	Elevation, BI_5_18, BI_6_18, CI_5_18, CI_3_19, MNDWI_5_18, MNDWI_6_18, MNDWI_12_18, MNDWI_3_19, MTCI_10_18, MTCI_11_18, MTCI_3_19, NDVI_8_18, SAVI_8_18, SAVI_12_18, SAVI_3_19, VH_5_18, VH_6_18, VH_9_18, VV_3_19

3.3. Machine Learning Performance

Table 5 lists the fitted values of the different hyperparameters used and the R^2, RMSE, and RPIQ results for the different models in four scenarios. The RF model consistently outperformed the other models across all scenarios, indicating its superior predictive ability for this dataset (Table 6 and Figure 4). In Scenario 1, the RF model achieved an R^2 of 0.36, an RMSE of 0.042, and an RPIQ of 1.644, surpassing XGBoost (R^2: 0.34, RMSE: 0.046, RPIQ: 1.501) and SVR (R^2: 0.21, RMSE: 0.054, RPIQ: 1.279). This trend continued, with the RF model exhibiting the highest performance in Scenario 2 (R^2: 0.49, RMSE: 0.037, RPIQ: 1.866), Scenario 3 (R^2: 0.61, RMSE: 0.024, RPIQ: 2.877), Scenario 4 (R^2: 0.65, RMSE: 0.02, RPIQ: 3.45), and Scenario 5 (R^2: 0.70, RMSE: 0.012, RPIQ: 5.754). Comparatively, the SVR model consistently showed the lowest performance metrics across all scenarios, indicating that it might be less suitable for SOC prediction under these study conditions. Scenario 5 represented the best outcome for all models, suggesting that combining all data (Sentinel-1, Sentinel-2, and topography) was most conducive to predictive modeling. The RF model's superior performance in Scenario 5, with an R^2 of 0.70, an RMSE of 0.012, and a higher RPIQ of 5.754, demonstrated its robustness and efficiency in handling the relationship between different predictors and SOC content.

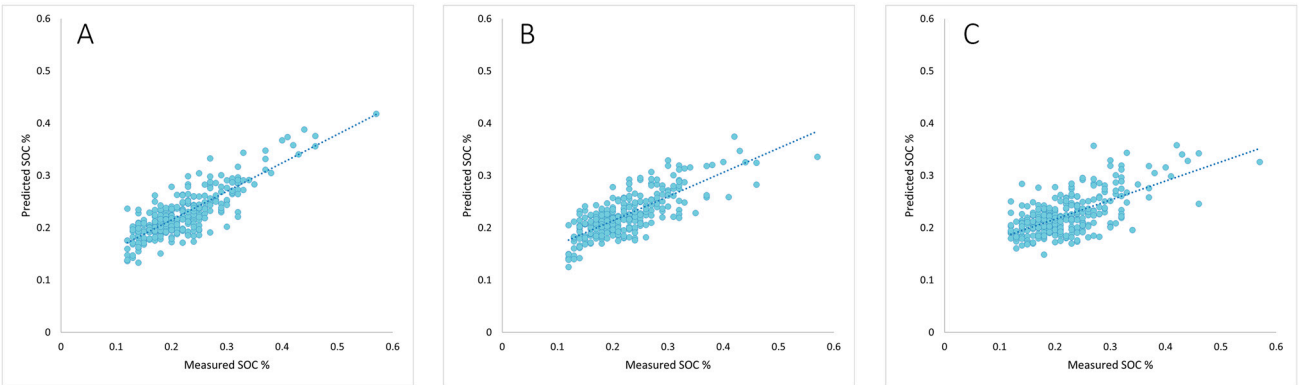

Figure 4. Scatter plots of measured vs. predicted SOC % for RF (**A**), XGBoost (**B**), and SVR (**C**) models under Scenario 5.

In Scenario 5, which exhibited the highest performance among all scenarios, the importance of various predictors was analyzed for three models: RF, XGBoost, and SVR (Figure 5). Elevation stands out as the most influential variable within all three models. Following elevation, the CI for the date 3_19 is the next most prominent variable for the

RF and XGBoost models, suggesting its repeated importance. As we delve further into the hierarchy, VV and VH radar bands from Sentinel-1, acquired at different time points, consistently rank high in importance, particularly for the RF and XGBoost models. This pattern also holds true for MNDWI and MTCI, where different dates yield a consistently high ranking across these two models, reflecting their key roles as predictors. In contrast to the RF and XGBoost models, which display a concurrence in the importance ranking of these variables, the SVR model also assigns high importance to elevation, indicating its cross-model relevance. However, its pattern of importance for other variables differs, allocating varying degrees of importance to the radar bands and spectral indices.

Table 5. Fitted values of different hyperparameters.

Model	Hyperparameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
	Number of selected features	10	10	20	7	20
RF	n_estimators	100	500	100	100	100
	max_features	Log2	Log2	Log2	Log2	Log2
	max_depth	10	10	15	5	16
	min_samples_split	2	2	5	2	5
	min_samples_leaf	2	2	2	2	2
SVR	C	0.1	10	1	0.1	0.5
	epsilon	0.01	0.01	0.01	0.01	0.01
	gamma	0.01	1	0.01	0.01	0.01
XGBoost	learning_rate	0.1	0.05	0.1	0.05	0.1
	max_depth	7	7	4	5	5
	gamma	0	0	0	0	0
	colsample_bytree	0.5	0.5	1	0.5	1
	min_child_weight	5	10	10	5	5
	subsample	1	0.5	0.5	0.5	0.5
	n_estimators	50	50	50	50	50

Table 6. Validation accuracy for the three models across different scenarios.

Scenario	Model	R2	RMSE	RPIQ
Scenario 1	RF	0.36	0.042	1.644
	XGBoost	0.34	0.046	1.501
	SVR	0.21	0.054	1.279
Scenario 2	RF	0.49	0.037	1.866
	XGBoost	0.45	0.039	1.770
	SVR	0.35	0.049	1.409
Scenario 3	RF	0.61	0.024	2.877
	XGBoost	0.51	0.028	2.466
	SVR	0.38	0.047	1.469

Table 6. *Cont.*

Scenario	Model	R2	RMSE	RPIQ
Scenario 4	RF	0.65	0.02	3.45
	XGBoost	0.62	0.023	3
	SVR	0.47	0.035	1.971
Scenario 5	RF	**0.7**	**0.012**	**5.754**
	XGBoost	0.64	0.017	4.061
	SVR	0.56	0.023	3.002

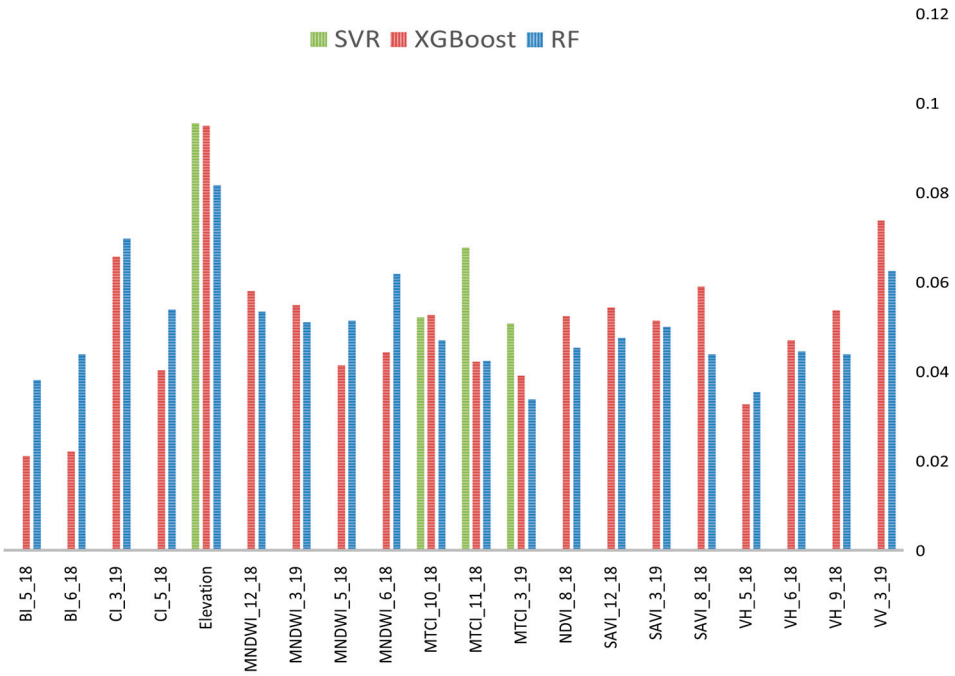

Figure 5. Feature importance for RF, XGBoost, and SVR models under Scenario 5.

3.4. Soil Organic Carbon Mapping

Figure 6 shows the spatial distribution of SOC using the RF and XGBoost algorithms with Scenario 6, which has been defined as the optimal combination of model and scenario configurations. The RF algorithm predicted SOC concentrations ranging from 0.12% to 0.42%, corresponding to the dominant low SOC content in the majority of the 671 soil samples used for model training. In this dataset, a small number of samples (n = 35) had SOC levels above 0.35%, and only seven samples exceeded the 0.45% value. Consequently, the RF prediction model satisfactorily captured the general pattern of SOC values. In contrast, the XGBoost algorithm predicted a more restricted range of SOC values, from 0.15% to 0.32%, indicating a lower degree of heterogeneity in its predictions than RF. Despite these variations, both algorithms generally reflected the bias of the training data towards the lowest SOC values, reducing the ability of the models to reflect the full range of SOC variability, particularly less frequent samples. The restricted predictive range of the models suggests a reduced sensitivity to the complex relationships between SOC and influencing covariates, leading to a potential underestimation of SOC levels in areas where they naturally exceed the dominant range of the training dataset.

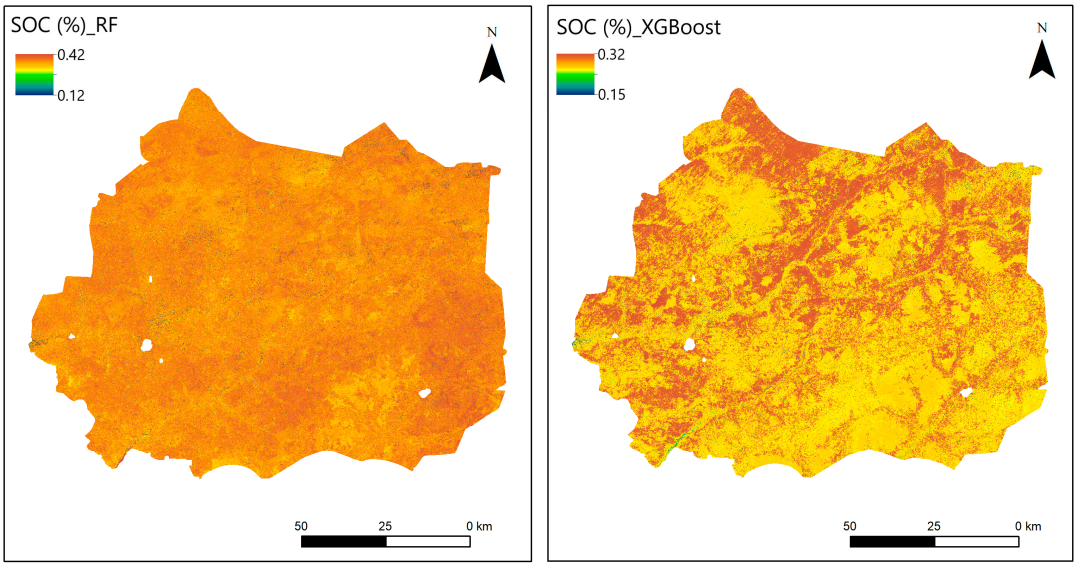

Figure 6. Spatial distribution of SOC content (%) for RF and XGBoost models.

4. Discussion

To thoroughly discuss the findings of this study, three main aspects were considered: (i) feature importance in SOC prediction, (ii) the performance of the various scenarios using Sentinel-1 and Sentinel-2 and topographic data, and (iii) the effectiveness and comparative analysis of the three ML algorithms.

Firstly, the RFE method was used to select the most important variables/features for SOC prediction. For that, 10 variables were identified for Scenarios 1 and 2, 20 variables were identified for Scenarios 3 and 5, and 7 variables were identified variables for Scenario 4. The number of variables for Scenarios 3 and 5 was increased to assess whether the RFE model would extract identical variables from Sentinel-1, Sentinel-2, and topographic data, or if one dataset would predominate over the others. The variables identified as being significant were MNDWI, SAVI, and MTCI, each with more than three variables from different months, indicating their relevance over different time periods. The importance of these variables is explained by the fact that SAVI and MTCI reflect vegetation [47,48], which is indirectly correlated with soil health and fertility [5,49] and consequently serves as a proxy for soil organic matter content [50]. This association has been supported by numerous studies that have identified vegetation indices, such as SAVI, NDVI, and others, to predict SOC or SOM [51–56]. The link between MNDWI and SOC is more indirect and complex. Similarly, SOC affects soil physical and chemical properties, including color, texture, and moisture retention capacity. These properties can influence soil reflectance characteristics in different spectral bands, including green and SWIR bands, and may indirectly highlight the importance of soil moisture parameters in SOC prediction [57–59], as moisture-rich environments can facilitate the preservation and accumulation of organic carbon in soil [1,60,61]. Furthermore, our results align with those of Lu et al. [62], who highlighted the importance of MNDWI alongside other soil moisture indices such as the Topographic Wetness Index (TWI) for SOC prediction. CI and BI showed a significant contribution to SOC prediction due to their ability to capture variations in soil color, which are often indicative of SOM content and other soil properties [63,64]. The correlation between SOC and CI and BI was already highlighted in previous studies, such as Saha et al. [65], which demonstrated that different spectral color indices, especially CI, are important for SOC prediction and mapping.

The Sentinel-2-derived indices used in Scenario 2 contributed more significantly than the Sentinel-1 dual-polarization indices (VV and VH). This can be attributed to the superior ability of Sentinel-2 variables to predict SOC compared with Sentinel-1, which is reflected in the performance differences between the models. In detail, Scenario 2 showed higher

performances for RF (R^2 = 0.49, RMSE = 0.037%) and XGBoost (R^2 = 0.45, RMSE = 0.039%) compared to Scenario 1, for which the RF performance was R^2 = 0.36 and RMSE = 0.042% and the XGBoost performance was R^2 = 0.34 and RMSE = 0.046%. In addition, the combination of the two scenarios resulted in an even higher performance for RF (R^2 = 0.61, RMSE = 0.024%) and XGBoost (R^2 = 0.51, RMSE = 0.028%), with a significant contribution from Sentinel-2 variables. This advantage of Sentinel-2 has been confirmed by various studies, such as Nguyen et al. [54], who found that SOC prediction performance using Sentinel-2 was superior to that using Sentinel-1, with R^2 values of 0.44 versus 0.25. Zhang et al. [66] obtained similar results, with an R^2 of 0.47 for Sentinel-2 versus 0.26 for Sentinel-1. In addition, Fatholoumi et al. [67] and Wang and Zhou [68] pointed out that the use of multi-temporal variables improved prediction performance due to the dynamic relationship between SOC and vegetation across a longer period compared to using data from a single date. Furthermore, the improvement in performance observed from the combination of the two scenarios was further validated by Zhang et al. [66], who reported an improvement in accuracy ranging between 2% and 5%. Similarly, Zhou et al. [69] highlighted that combining Sentinel-1 and Sentinel-2 data led to an increase in SOC prediction accuracy by 5 to 6% and a reduction in error by 5% to 7%. Including topographical features increased the performance of all models, with a significant contribution from elevation, the highest performance being reached by the RF model with an R^2 of 0.7, an RMSE of 0.012%, and an RPIQ of 5.754. The importance and contribution of topographic features were highlighted by Zhou et al. [70], who showed that elevation, slope, and TWI contributed more than 27% to the model's explanation. Additionally, Li et al. [71] showed that relief and TWI were the most important variables controlling SOC. The same was demonstrated by Gibson et al. [72], indicating that topographic features have an impact on SOC modeling at different resolutions. Furthermore, the same reasoning for grouping environmental covariates was demonstrated by Duarte et al. [73], based on Landsat-8 and various other covariates, such as climate and topography, and yielded the best results for SOC stocks in forested land.

The comparison of ML algorithms revealed that RF and XGBoost outperformed the SVR model, mainly due to their ensemble nature, which offers greater adaptability in addressing complex, non-linear relationships within data. Across all scenarios, RF and XGBoost consistently demonstrated higher R^2 values compared to the SVR model, indicating a greater proportion of variance explained by the dependent variable, as well as lower RMSE values. These results are also reflected in other studies, such as that of Nguyen et al. [54], who highlighted that XGBoost and RF surpassed the SVR model in predicting SOC content using Sentinel-1 and Sentinel-2 data, achieving a higher performance with an R^2 value higher than 0.7. Similarly, Siewert [74] compared various algorithms for SOC prediction and identified a superior performance of RF models over others. Moreover, Zhang et al. [66] observed that RF could outperform XGBoost when using separate Sentinel data, which is in line with our findings of an RF with R^2 values of 0.61 and 0.7 for Scenarios 3 and 4, respectively, versus R^2 values of 0.51 and 0.64 for XGBoost and 0.38 and 0.56 for SVR. The performance results obtained in the present study are similar to those reported by Pouladi et al. [75], who used only Sentinel-2, and Nguyen et al. [54], with R^2 values around 0.72 for RF; however, these values were higher than those obtained in other studies that demonstrated low performance, such as Shafizadeh-Moghadam et al. [23] and Tajik et al. [76], with performance being characterized by R^2 values less than 0.5. The low performance in these studies can generally be attributed to factors such as high heterogeneity with an extensive study area size and the low density of sampling points [70]. In our case, the reasons for the low performance for Scenario 1 and Scenario 3 may be attributed to the low variability in SOC content (min = 0.11%, max = 0.72%), which could introduce complexity into the modeling process [12]. The SOC distribution also revealed that the XGBoost algorithm predicts a lower SOC value than the RF model. This could reflect more conservative estimation or potential underfitting where the XGBoost model does not fully capture the higher SOC values present in the training data, perhaps due to model complexity or regularization parameters. Clearly, both models have limitations in representing

the less frequent, slightly higher SOC values, which were few in the training data. This skew towards lower SOC values is a common problem in machine learning, where model performance is strongly influenced by the distribution of the training dataset. In practical applications, this could potentially mean that areas with naturally higher SOC levels could be underestimated.

5. Contributions, Limitations, and Future Research Directions

This study contributes significantly to the field of DSM by demonstrating the potential of combining Sentinel-1 and Sentinel-2 data for high-resolution (10 m) SOC prediction, offering valuable insights for stakeholders in African agriculture and beyond. Our scenario-based approach sheds light on the influence of different environmental variables on model performance, highlighting the importance of considering topography alongside remotely sensed data. Additionally, the comparative analysis of machine learning algorithms provides guidance for selecting the most suitable method based on specific data and objectives. However, some limitations were encountered. Despite achieving reasonable accuracy, the models exhibited a bias towards the dominant low SOC values within the training data, resulting in a reduced ability to capture the full range of SOC variability. This limitation suggests a potential underestimation of SOC, which could impact land management decisions. Additionally, computational limitations restricted the generation of uncertainty maps, hindering a more comprehensive assessment of model reliability.

Future research should prioritize addressing these limitations and exploring new avenues for improvement. Techniques like data augmentation or incorporating prior knowledge about SOC distribution could mitigate the bias towards low values and enhance the models' ability to represent the full spectrum of SOC variability. Exploring alternative feature selection methods, such as those based on expert opinion [77], as well as alternative machine learning approaches, such as ensemble methods or meta-learners that combine multiple algorithms with diverse structures, may improve prediction accuracy and overcome the problem of the limit of singular models in predicting SOC values outside the limit of dominant values. Furthermore, investigating computationally efficient methods for generating uncertainty maps remains crucial for enhancing the interpretability and reliability of SOC predictions. By addressing these challenges and building upon this study's foundation, future research can further advance DSM and provide increasingly accurate and reliable high-resolution SOC maps. These maps will be invaluable tools for stakeholders in African agriculture and other regions, supporting sustainable land management practices, soil conservation efforts, and informed decision making for improved agricultural productivity and environmental sustainability.

6. Conclusions

This study evaluated the suitability of time-series radar (Sentinel-1), optical (Sentinel-2), and topography data for SOC prediction across a variety of scenarios and predictive modeling frameworks. In conclusion, this research demonstrates the feasibility of integrating high-resolution EO data with ML algorithms to predict SOC in case of low-value content. The key findings are as follows:

- Combining multi-temporal Sentinel-1 and Sentinel-2 data enhances the precision of SOC prediction, with an improvement of R^2 values and reduced error compared to using single-source data. This underscores the benefit of multi-sensor data fusion for DSM applications.
- Including topographic data improves the accuracy of different models and signifies that the integration of all data inputs culminates in optimal model efficacy.
- RF and XGBoost algorithms outperform SVR in SOC prediction across different scenarios, highlighting the effectiveness of ensemble learning techniques in handling complex spatial datasets.
- Despite the overall success, the models predominantly predict low SOC values, reflecting the inherent limitations in capturing the full range of SOC variability, which

suggests the need for further refinement of modeling approaches to better address less frequent, high-concentration samples.

Finally, the generated SOC maps are crucial for informing sustainable land management practices and climate change mitigation strategies. Furthermore, in future studies, it will be interesting to test radar and optical data for other soil fertility parameters, or to evaluate time series for other satellite products such as hyperspectral data.

Author Contributions: Conceptualization, S.D., M.R. and Y.B.; Data curation, S.D. and Y.B.; Formal analysis, S.D. and Y.B.; Methodology, S.D., M.R. and Y.B.; Supervision, M.R.; Validation, S.D. and Y.B.; Writing—original draft, S.D., M.R. and Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request. The scripts used for this paper can be accessed at https://github.com/yassinebos/SOC_prediction-mapping (accessed on 28 April 2024).

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

References

1. Lal, R. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. *Science* **2004**, *304*, 1623–1627. [CrossRef]
2. von Fromm, S.F.; Hoyt, A.M.; Lange, M.; Acquah, G.E.; Aynekulu, E.; Berhe, A.A.; Haefele, S.M.; McGrath, S.P.; Shepherd, K.D.; Sila, A.M.; et al. Continental-scale controls on soil organic carbon across sub-Saharan Africa. *Soil Discuss.* **2020**, *2020*, 1–39. [CrossRef]
3. Schulze, R.E.; Schütte, S. Mapping soil organic carbon at a terrain unit resolution across South Africa. *Geoderma* **2020**, *373*, 114447. [CrossRef]
4. Odebiri, O.; Mutanga, O.; Odindi, J.; Naicker, R. Modelling soil organic carbon stock distribution across different land-uses in South Africa: A remote sensing and deep learning approach. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 351–362. [CrossRef]
5. Vågen, T.G.; Winowiecki, L.A.; Tondoh, J.E.; Desta, L.T.; Gumbricht, T. Mapping of soil properties and land deg-radation risk in Africa using MODIS reflectance. *Geoderma* **2016**, *263*, 216–225. [CrossRef]
6. Al Masmoudi, Y.; Bouslihim, Y.; Doumali, K.; Hssaini, L.; Namr, K.I. Use of machine learning in Moroccan soil fertility prediction as an alternative to laborious analyses. *Model. Earth Syst. Environ.* **2022**, *8*, 3707–3717. [CrossRef]
7. Wadoux, A.M.-C.; Minasny, B.; McBratney, A.B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Sci. Rev.* **2020**, *210*, 103359. [CrossRef]
8. Nenkam Mentho, A.; Wadoux, A.M.C.; Minasny, B.; Silatsa, F.B.; Yemefack, M.; Ugbaje, S.; Akpa, S.; van Zijl, G.M.; Bouslihim, Y.; Chabala, L.; et al. Applications and Challenges of Digital Soil Mapping in Africa. Available online: https://ssrn.com/abstract=4725182 (accessed on 15 March 2024). [CrossRef]
9. Hengl, T.; Heuvelink, G.B.; Kempen, B.; Leenaars, J.G.; Walsh, M.G.; Shepherd, K.D.; Sila, A.; MacMillan, R.A.; de Jesus, J.M.; Tamene, L.; et al. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLoS ONE* **2015**, *10*, e0125814. [CrossRef] [PubMed]
10. Hengl, T.; Miller, M.A.E.; Križan, J.; Shepherd, K.D.; Sila, A.; Kilibarda, M.; Antonijević, O.; Glušica, L.; Dobermann, A.; Haefele, S.M.; et al. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Sci. Rep.* **2021**, *11*, 6130. [CrossRef]
11. Bouasria, A.; Namr, K.I.; Rahimi, A.; Ettachfini, E.M.; Rerhou, B. Evaluation of Landsat 8 image pansharpening in estimating soil organic matter using multiple linear regression and artificial neural networks. *Geo-Spat. Inf. Sci.* **2022**, *25*, 353–364. [CrossRef]
12. Bouslihim, Y.; John, K.; Miftah, A.; Azmi, R.; Aboutayeb, R.; Bouasria, A.; Razouk, R.; Hssaini, L. The effect of covariates on Soil Organic Matter and pH variability: A digital soil mapping approach using random forest model. *Ann. GIS* **2024**, *30*, 215–232. [CrossRef]
13. Sayedain, S.A.; Maghsoudi, Y.; Eini-Zinab, S. Assessing the use of cross-orbit Sentinel-1 images in land cover clas-sification. *Int. J. Remote Sens.* **2020**, *41*, 7801–7819. [CrossRef]
14. Urbina-Salazar, D.; Vaudour, E.; Baghdadi, N.; Ceschia, E.; Richer-de-Forges, A.C.; Lehmann, S.; Arrouays, D. Using sentinel-2 images for soil organic carbon content mapping in croplands of southwestern france. The usefulness of sentinel-1/2 derived moisture maps and mismatches between sentinel images and sampling dates. *Remote Sens.* **2021**, *13*, 5115. [CrossRef]
15. Mponela, P.; Snapp, S.; Villamor, G.B.; Tamene, L.; Le, Q.B.; Borgemeister, C. Digital soil mapping of nitrogen, phosphorus, potassium, organic carbon and their crop response thresholds in smallholder managed escarpments of Malawi. *Appl. Geogr.* **2020**, *124*, 102299. [CrossRef]
16. Flynn, T.; Rozanov, A.; Ellis, F.; de Clercq, W.; Clarke, C. Farm-scale digital soil mapping of soil classes in South Africa. *S. Afr. J. Plant Soil* **2022**, *39*, 175–186. [CrossRef]

17. Gholizadeh, A.; Žižala, D.; Saberioon, M.; Borůvka, L. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sens. Environ.* **2018**, *218*, 89–103. [CrossRef]
18. Castaldi, F.; Chabrillat, S.; Don, A.; van Wesemael, B. Soil Organic Carbon Mapping Using LUCAS Topsoil Database and Sentinel-2 Data: An Approach to Reduce Soil Moisture and Crop Residue Effects. *Remote Sens.* **2019**, *11*, 2121. [CrossRef]
19. Castaldi, F.; Hueni, A.; Chabrillat, S.; Ward, K.; Buttafuoco, G.; Bomans, B.; Vreys, K.; Brell, M.; van Wesemael, B. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS J. Photogramm. Remote Sens.* **2019**, *147*, 267–282. [CrossRef]
20. Wang, S.; Zhou, M.; Zhuang, Q.; Guo, L. Prediction Potential of Remote Sensing-Related Variables in the Topsoil Organic Carbon Density of Liaohekou Coastal Wetlands, Northeast China. *Remote Sens.* **2021**, *13*, 4106. [CrossRef]
21. Tripathi, A.; Tiwari, R.K. Utilisation of spaceborne C-band dual pol Sentinel-1 SAR data for simplified regres-sion-based soil organic carbon estimation in Rupnagar, Punjab, India. *Adv. Space Res.* **2022**, *69*, 1786–1798. [CrossRef]
22. Izurieta, J.E.A.; Santillán, C.A.J.; Márquez, C.O.; García, V.J.; Rivera-Caicedo, J.P.; Van Wittenberghe, S.; Delegido, J.; Verrelst, J. Improving the remote estimation of soil organic carbon in complex ecosystems with Sentinel-2 and GIS using Gaussian processes regression. *Plant Soil* **2022**, *479*, 159–183. [CrossRef] [PubMed]
23. Shafizadeh-Moghadam, H.; Minaei, F.; Talebi-Khiyavi, H.; Xu, T.; Homaee, M. Synergetic use of multi-temporal Sentinel-1, Sentinel-2, NDVI, and topographic factors for estimating soil organic carbon. *Catena* **2022**, *212*, 106077. [CrossRef]
24. Zhou, T.; Geng, Y.; Chen, J.; Pan, J.; Haase, D.; Lausch, A. 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.* **2020**, *729*, 138244. [CrossRef] [PubMed]
25. FAO. *Standard Operating Procedure for Soil Organic Carbon Walkley-Black Method Titration and Colorimetric Method*; Food & Agriculture Organization: Rome, Italy, 2019.
26. Dahhani, S.; Raji, M.; Hakdaoui, M.; Lhissou, R. Land cover mapping using sentinel-1 time-series data and ma-chine-learning classifiers in agricultural sub-saharan landscape. *Remote Sens.* **2022**, *15*, 65. [CrossRef]
27. Loew, A.; Mauser, W. Generation of geometrically and radiometrically terrain corrected SAR image products. *Remote Sens. Environ.* **2007**, *106*, 337–349. [CrossRef]
28. Main-Knorn, M.; Pflug, B.; Louis, J.; Debaecker, V.; Müller-Wilm, U.; Gascon, F. Sen2Cor for sentinel-2. In *Image and Signal Processing for Remote Sensing XXIII*; SPIE: Bellingham, WA, USA, 2017; Volume 10427, pp. 37–48.
29. Escadafal, R.; Girard, M.-C.; Courault, D. Munsell soil color and soil reflectance in the visible spectral bands of landsat MSS and TM data. *Remote Sens. Environ.* **1989**, *27*, 37–46. [CrossRef]
30. Escadafal, R.; Belghith, A.; Ben Moussa, H. Indices spectraux pour la télédétection de la dégradation des milieux naturels en Tunisie aride. In Proceedings of the 6th International Symposium on Physical Measurements and Signatures in Remote Sensing, Val d'Isère, France, 17–21 January 1994; pp. 17–21.
31. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]
32. Dash, J.; Curran, P.J. The MERIS terrestrial chlorophyll index. *Int. J. Remote Sens.* **2004**, *25*, 5403–5413. [CrossRef]
33. Bannari, A.; Morin, D.; Bonn, F.; Huete, A.R. A review of vegetation indices. *Remote Sens. Rev.* **1995**, *13*, 95–120. [CrossRef]
34. McFeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]
35. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
36. Darst, B.F.; Malecki, K.C.; Engelman, C.D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* **2018**, *19*, 65. [CrossRef] [PubMed]
37. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Bouslihim, Y.; Kharrou, M.H.; Miftah, A.; Attou, T.; Bouchaou, L.; Chehbouni, A. Comparing Pan-sharpened Landsat-9 and Sentinel-2 for Land-Use Classification Using Machine Learning Classifiers. *J. Geovisualization Spat. Anal.* **2022**, *6*, 1–17. [CrossRef]
39. John, K.; Bouslihim, Y.; Bouasria, A.; Razouk, R.; Hssaini, L.; Isong, I.A.; M'Barek, S.A.; Ayito, E.O.; Ambrose-Igho, G. Assessing the impact of sampling strategy in random forest-based predicting of soil nutrients: A study case from northern Morocco. *Geocarto Int.* **2022**, *37*, 11209–11222. [CrossRef]
40. Bouasria, A.; Bouslihim, Y.; Gupta, S.; Taghizadeh-Mehrjardi, R.; Hengl, T. Predictive performance of machine learning model with varying sampling designs, sample sizes, and spatial extents. *Ecol. Inform.* **2023**, *78*, 102294. [CrossRef]
41. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
42. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. In Proceedings of the Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, 2–5 December 1996.
43. Bisong, E. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Apress: Berkeley, CA, USA, 2019; pp. 59–64.
44. Piñeiro, G.; Perelman, S.; Guerschman, J.P.; Paruelo, J.M. How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecol. Model.* **2008**, *216*, 316–322. [CrossRef]
45. Smith, J.; Smith, P.; Addiscott, T. Quantitative methods to evaluate and compare soil organic matter (SOM) models. In *Evaluation of Soil Organic Matter Models: Using Existing Long-Term Datasets*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 181–199.

46. Castaldi, F.; Palombo, A.; Santini, F.; Pascucci, S.; Pignatti, S.; Casa, R. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. *Remote Sens. Environ.* **2016**, *179*, 54–65. [CrossRef]
47. Pastor-Guzman, J.; Brown, L.; Morris, H.; Bourg, L.; Goryl, P.; Dransfeld, S.; Dash, J. The Sentinel-3 OLCI Terrestrial Chlorophyll Index (OTCI): Algorithm Improvements, Spatiotemporal Consistency and Continuity with the MERIS Archive. *Remote Sens.* **2020**, *12*, 2652. [CrossRef]
48. Vani, V.; Mandla, V.R. Comparative study of NDVI and SAVI vegetation indices in Anantapur district semi-arid areas. *Int. J. Civ. Eng. Technol.* **2017**, *8*, 559–566.
49. Brevik, E.C.; Calzolari, C.; Miller, B.A.; Pereira, P.; Kabala, C.; Baumgarten, A.; Jordán, A. Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma* **2016**, *264*, 256–274. [CrossRef]
50. Ngatia, L.W.; Moriasi, D.; Grace, J.M., III; Fu, R.; Gardner, C.S.; Taylor, R.W. Land use change affects soil organic carbon: An indicator of soil health. In *Environmental Health*; Books on Demand: Norderstedt, Germany, 2021.
51. Crapart, C.; Finstad, A.G.; Hessen, D.O.; Vogt, R.D.; Andersen, T. Spatial predictors and temporal forecast of total organic carbon levels in boreal lakes. *Sci. Total Environ.* **2023**, *870*, 161676. [CrossRef] [PubMed]
52. Bian, Z.; Guo, X.; Wang, S.; Zhuang, Q.; Jin, X.; Wang, Q.; Jia, S. Applying statistical methods to map soil organic carbon of agricultural lands in northeastern coastal areas of China. *Arch. Agron. Soil Sci.* **2019**, *66*, 532–544. [CrossRef]
53. Kaya, F.; Keshavarzi, A.; Francaviglia, R.; Kaplan, G.; Başayiğit, L.; Dedeoğlu, M. Assessing Machine Learning-Based Prediction under Different Agricultural Practices for Digital Mapping of Soil Organic Carbon and Available Phosphorus. *Agriculture* **2022**, *12*, 1062. [CrossRef]
54. Nguyen, T.T.; Pham, T.D.; Nguyen, C.T.; Delfos, J.; Archibald, R.; Dang, K.B.; Hoang, N.B.; Guo, W.; Ngo, H.H. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Sci. Total Environ.* **2022**, *804*, 150187. [CrossRef]
55. Wang, S.; Zhuang, Q.; Jin, X.; Yang, Z.; Liu, H. Predicting Soil Organic Carbon and Soil Nitrogen Stocks in Topsoil of Forest Ecosystems in Northeastern China Using Remote Sensing Data. *Remote Sens.* **2020**, *12*, 1115. [CrossRef]
56. Wang, K.; Qi, Y.; Guo, W.; Zhang, J.; Chang, Q. Retrieval and Mapping of Soil Organic Carbon Using Sentinel-2A Spectral Images from Bare Cropland in Autumn. *Remote Sens.* **2021**, *13*, 1072. [CrossRef]
57. Liu, T.; Zhang, H.; Shi, T. Modeling and Predictive Mapping of Soil Organic Carbon Density in a Small-Scale Area Using Geographically Weighted Regression Kriging Approach. *Sustainability* **2020**, *12*, 9330. [CrossRef]
58. Sodango, T.H.; Sha, J.; Li, X.; Noszczyk, T.; Shang, J.; Aneseyee, A.B.; Bao, Z. Modeling the Spatial Dynamics of Soil Organic Carbon Using Remotely-Sensed Predictors in Fuzhou City, China. *Remote Sens.* **2021**, *13*, 1682. [CrossRef]
59. Pei, T.; Qin, C.-Z.; Zhu, A.-X.; Yang, L.; Luo, M.; Li, B.; Zhou, C. Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods. *Ecol. Indic.* **2010**, *10*, 610–619. [CrossRef]
60. Davidson, E.A.; Janssens, I.A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **2006**, *440*, 165–173. [CrossRef] [PubMed]
61. Scharlemann, J.P.; Tanner, E.V.; Hiederer, R.; Kapos, V. Global soil carbon: Understanding and managing the largest terrestrial carbon pool. *Carbon Manag.* **2014**, *5*, 81–91. [CrossRef]
62. Lu, W.; Lu, D.; Wang, G.; Wu, J.; Huang, J.; Li, G. Examining soil organic carbon distribution and dynamic change in a hickory plantation region with Landsat and ancillary data. *Catena* **2018**, *165*, 576–589. [CrossRef]
63. He, T.; Wang, J.; Lin, Z.; Cheng, Y. Spectral features of soil organic matter. *Geo-Spat. Inf. Sci.* **2009**, *12*, 33–40. [CrossRef]
64. Hossain, M.Z. Farmer's view on soil organic matter depletion and its management in Bangladesh. *Nutr. Cycl. Agroecosyst.* **2001**, *61*, 197–204. [CrossRef]
65. Saha, S.K.; Tiwari, S.K.; Kumar, S. Integrated use of hyperspectral remote sensing and geostatistics in spatial pre-diction of soil organic carbon content. *J. Indian Soc. Remote Sens.* **2022**, *50*, 129–141. [CrossRef]
66. Zhang, H.; Wan, L.; Li, Y. 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. Remote Sens.* **2023**, *16*, 5219–5230. [CrossRef]
67. Fathololoumi, S.; Vaezi, A.R.; Alavipanah, S.K.; Ghorbani, A.; Saurette, D.; Biswas, A. Improved digital soil mapping with multitemporal remotely sensed satellite data fusion: A case study in Iran. *Sci. Total Environ.* **2020**, *721*, 137703. [CrossRef] [PubMed]
68. Wang, L.; Zhou, Y. Combining Multitemporal Sentinel-2A Spectral Imaging and Random Forest to Improve the Accuracy of Soil Organic Matter Estimates in the Plough Layer for Cultivated Land. *Agriculture* **2022**, *13*, 8. [CrossRef]
69. Zhou, T.; Geng, Y.; Chen, J.; Liu, M.; Haase, D.; Lausch, A. Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China. *Ecol. Indic.* **2020**, *114*, 106288. [CrossRef]
70. Zhou, T.; Geng, Y.; Ji, C.; Xu, X.; Wang, H.; Pan, J.; Bumberger, J.; Haase, D.; Lausch, A. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. *Sci. Total Environ.* **2021**, *755*, 142661. [CrossRef] [PubMed]
71. Li, X.; McCarty, G.W.; Karlen, D.L.; Cambardella, C.A. Topographic metric predictions of soil redistribution and organic carbon in Iowa cropland fields. *Catena* **2018**, *160*, 222–232. [CrossRef]
72. Gibson, A.; Hancock, G.; Bretreger, D.; Cox, T.; Hughes, J.; Kunkel, V. Assessing digital elevation model resolution for soil organic carbon prediction. *Geoderma* **2021**, *398*, 115106. [CrossRef]

73. Duarte, E.; Zagal, E.; Barrera, J.A.; Dube, F.; Casco, F.; Hernández, A.J. Digital mapping of soil organic carbon stocks in the forest lands of Dominican Republic. *Eur. J. Remote Sens.* **2022**, *55*, 213–231. [CrossRef]
74. Siewert, M.B. High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: A case study in a sub-Arctic peatland environment. *Biogeosciences* **2018**, *15*, 1663–1682. [CrossRef]
75. Pouladi, N.; Møller, A.B.; Tabatabai, S.; Greve, M.H. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. *Geoderma* **2019**, *342*, 85–92. [CrossRef]
76. Tajik, S.; Ayoubi, S.; Zeraatpisheh, M. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. *Geoderma Reg.* **2020**, *20*, e00256. [CrossRef]
77. Pullanagari, R.R.; Cavalli, D. Advances and applications of multivariate statistics and soil-crop sensing to improve nutrient use efficiency and monitor carbon cycling. *Nutr. Cycl. Agroecosyst.* **2023**, *127*, 97–99. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.