Smart Agricultural Technology 13 (2026) 101935

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Monitoring soil organic carbon from earth observation in the era of cover cropping and no-tillage

Fabio Castaldi [a,*], Chien-Hui Syu [b], Miguel Conrado Valdez [c], Chun-Chien Yen [b,d], Chi-Farn Chen [c], Flavio Bertinaria [e], Piero Toscano [a]

[a] Institute of BioEconomy, National Research Council of Italy (CNR), Via Giovanni Caproni 8, 50145, Firenze, Italy
[b] Taiwan Agricultural Research Institute, Taichung City, Taiwan
[c] Center for Space and Remote Sensing Research, National Central University, Taoyuan City, Taiwan
[d] Department of Soil and Environmental Sciences, National Chung Hsing University, Taiwan
[e] Barilla G. e R. Fratelli S.p.A., Parma, Italy

ARTICLE INFO

Keywords:
Bare-soil
Regenerative agriculture
Data fusion
Multitemporal analysis
Alphaearth
Satellite embeddings

ABSTRACT

The estimation of Soil Organic Carbon (SOC) from optical image spectroscopy typically relies on the availability of bare-soil conditions, which are increasingly rare due to the widespread adoption of conservation agriculture practices. This study evaluates alternative strategies for SOC prediction under limited bare-soil availability by comparing four methodological approaches based on Sentinel-2 imagery and related products: (i) bare-soil multispectral composites, (ii) vegetation indices, (iii) AlphaEarth Satellite Embeddings, and (iv) a hybrid geostatistical–machine learning model (KpR–Cubist). These methods were tested across three cropland regions with contrasting pedoclimatic conditions: Italy, France, and Taiwan. The evaluation relied on >1800 topsoil samples collected between 2020 and 2024. Results show that bare-soil availability varies significantly by region, with cloud cover and vegetation/farm management being the main limiting factors. Models using Satellite Embeddings consistently achieved the highest predictive accuracy (RPIQ up to 2.24) , outperforming conventional bare-soil composite and vegetation-based models. Incorporating spatial coordinates further improved model performance, revealing strong spatial autocorrelation in SOC distribution. The hybrid kriging–Cubist approach achieved comparable accuracy to the embedding-based models, confirming the value of integrating spatial dependence into data-driven frameworks. Overall, the study demonstrates that deep-learning–derived satellite embeddings models provide effective alternatives for SOC estimation in croplands where bare-soil imagery is increasingly unavailable due to sustainable soil management practices.

1. Introduction

Soil imaging spectroscopy is grounded in the physical interactions between soil properties and electromagnetic radiation. Consequently, the estimation of soil organic carbon (SOC) from satellite imagery relies on the availability of bare soil scenes. In conventional agricultural systems, such conditions are generally restricted to narrow temporal windows, while truly optimal circumstances, such as seedbed preparation, dry soil, and cloud-free atmosphere, are even rarer. The widespread availability of optical satellite imagery, notably from the ESA Copernicus Sentinel-2 mission and the NASA Landsat constellation, has enabled the development of multitemporal approaches. These approaches increase the likelihood of observing soil under ideal conditions. By exploiting data accumulated over multiple years, these methods synthesize information into composite images. Such composites have consistently demonstrated strong potential for accurate SOC estimation [1–5]. Such approaches enhance temporal coverage and mitigate the effects of individual scene limitations. As a result, they provide more robust and reliable inputs for SOC modeling [6,7]. However, monitoring SOC changes remains particularly challenging, as it requires information for a specific year. For this reason, relying on broad multi-year satellite collections is less advisable, since they may not accurately capture the actual SOC level for the year of interest.

Moreover, in recent decades, the adoption of conservation agriculture practices has further reduced the frequency of bare soil occurrence. These practices include minimum tillage, no-tillage, and the use of cover

* Corresponding author.
E-mail address: fabio.castaldi@cnr.it (F. Castaldi).

https://doi.org/10.1016/j.atech.2026.101935
Received 12 December 2025; Received in revised form 19 February 2026; Accepted 27 February 2026
Available online 28 February 2026

crops between main cropping cycles. While they are beneficial for preserving soil fertility, biodiversity, and ecosystem functions, they also limit direct soil exposure. As a result, attempts to extract bare soil scenes from satellite collections may often prove unfeasible due to the increasing masking of soils by vegetation or crop residues. Detecting bare-soil conditions can be even more challenging in olive groves, vineyards, other tree crops, or agroforestry systems. This limitation reduces the effectiveness of satellite-based SOC and soil property assessments and highlights the need for alternative estimation approaches that do not rely on bare soil satellite imagery. This progressive and large-scale transition toward conservation-oriented management represents a structural shift in cropland systems, fundamentally altering the conditions under which soil properties can be monitored from space.

An alternative strategy might consist of exploiting the indirect association between SOC and crop growth characteristics [8]. Nevertheless, the relationship between SOC, or more generally, soil fertility and remotely sensed vegetation metrics, is typically weak and highly inconsistent. This association is strongly modulated by year-to-year variations in climate, water availability, and biotic stresses, including pests and diseases. Agronomic interventions, such as tillage, fertilization, cultivar selection interested and crop rotation, can further influence crop performance independently of SOC content [9]. Collectively, these interacting factors generate considerable uncertainty. As a consequence, the reliable estimation of SOC from satellite-derived vegetation indicators alone remains challenging [10]. However, some vegetation indices have shown a good correlation with SOC [3], although their ability to reliably estimate SOC still needs to be thoroughly evaluated.

Alternatively, geostatistical approaches can provide effective prediction models that account for the spatial autocorrelation of SOC. However, these methods require dense and spatially extensive sampling networks to adequately capture the spatial variability of the target variable. In some cases, spatial autocorrelation may be weak or even absent. This is particularly evident in cropland areas characterized by highly fragmented landscapes, where different land uses and management practices alternate over short distances. Moreover, geostatistical models typically provide static estimates that depend solely on spatial position, whereas SOC monitoring frameworks should also be capable of detecting temporal changes. In this regard, hybrid approaches that combine geostatistics and machine learning offer a promising strategy to overcome these limitations. Their effectiveness is further enhanced when integrated with satellite-derived information [11,12]. Another approach to account for spatial information is to include the geographic coordinates of soil samples as additional covariates in satellite-based machine learning models. In Castaldi et al. [6], this addition led to a significant increase in predictive performance across a wide range of soil and climatic conditions compared with models using only Sentinel-2 bands.

A further promising strategy entails the integration of multiple Earth Observation (EO) data sources. This includes multi-temporal and multi-sensor datasets such as optical imagery (both multispectral and hyperspectral), Synthetic Aperture Radar (SAR), thermal data, and other satellite-derived covariates [13–16]. Such data fusion approaches offer the advantage of capturing complementary information across diverse modalities; however, they often involve a trade-off in terms of reduced spatial and spectral resolution. Recent developments in deep learning–based super-resolution techniques have shown substantial promise in mitigating this limitation [17–19].

Among the most noteworthy advancements is the introduction of Satellite Embedding V1 on the Google Earth Engine (GEE) platform. This global dataset provides learned geospatial embeddings derived from multiple EO sources, delivering 10-meter spatial resolution and covering the period 2017–2024. The dataset encodes complex relationships among various input sources, including Sentinel-1, Sentinel-2, Landsat 8 and 9, PALSAR-2, ERA5-Land, and digital elevation models into 64-band feature vectors [20]. Satellite Embedding V1 holds considerable potential to address limitations related to spatial resolution and image availability. However, its efficacy for the quantitative prediction of soil properties remains to be systematically evaluated.

Consequently, in this study, we compare four methodological approaches for SOC estimation from Sentinel-2 images and products. The first approach entails the use of multitemporal Sentinel-2 imagery restricted to bare soil scenes, therefore excluding all the dates affected by green and dry vegetation. The covariates are the Sentinel-2 bands. This approach is currently widely used and can be considered a reference standard. The second approach is based on the use of three vegetation indices derived from multitemporal Sentinel-2 data, thereby simulating a scenario in which bare soil observations are not available. The third approach makes use of Satellite Embedding V1 data developed by the AlphaEarth foundation model as covariates to address the scarcity of bare soil images. To the best of our knowledge, this is the first time this kind of dataset has been used for soil properties estimation. The fourth combine geostatistics and the above described covariate sets by hybrid models.

These four approaches are tested across three croplands soil datasets originating from France, Italy, and Taiwan collected between 2020 and 2024, representing contrasting pedoclimatic contexts.

The specific objectives of the study are to:

- Identify the primary factors limiting the availability of bare soil imagery (atmospheric versus vegetation constraints);
- determine which covariates exhibit the strongest correlations with SOC and are more influential for its estimation;
- compare the accuracy of SOC prediction models obtained using the four approaches in the three different regions;
- provide insights into alternative strategies for SOC estimation in scenarios where conservation-oriented soil management increasingly limits bare soil availability.

2. Materials and methods

2.1. Study areas

2.1.1. Italy

The Italian region of interest corresponds to the flat plains of the river Po Basin (Fig. 1), one of Italy's most densely populated and intensively exploited regions [21]. According the Köppen–Geiger classification [22] the region interests mainly Cfa (temperate with no dry season, hot summer) and Csa (temperate, dry and hot summer) climatic zones. The basin supports a wide spectrum of agricultural activities, including intensive crop cultivation and livestock farming, with high nitrogen inputs. The agricultural phenology is characterized by the coexistence of autumn–winter and spring–summer cropping systems typical of temperate climates, resulting in distinct seasonal trajectories in vegetation and spectral responses over the year.

The basin exhibits marked variability in soil fertility, driven by differences in land use practices, management techniques, climate, soil types, and topography. At the same time, soil degradation processes are widespread, highlighting ongoing challenges for sustainable land management [23]. Based on FAO soil classification [24] and the European Soil Database [25], the predominant soil types in the basin are Cambisols and Luvisols in the western areas, Fluvisols toward the eastern coast, and Regosols in the hilly right-bank regions of the river.

2.1.2. France

The French study areas are located across three administrative regions: Bourgogne–Franche-Comté in central France, Auvergne–Rhône-Alpes in the south-central part of the country, and Normandie in the north, near the North Sea (Fig. 1). These regions encompass a variety of agro-climatic and soil conditions representative of major French croplands. According to the Köppen–Geiger classification [22], the study sites fall within two main climate zones: Cfa (temperate with no dry

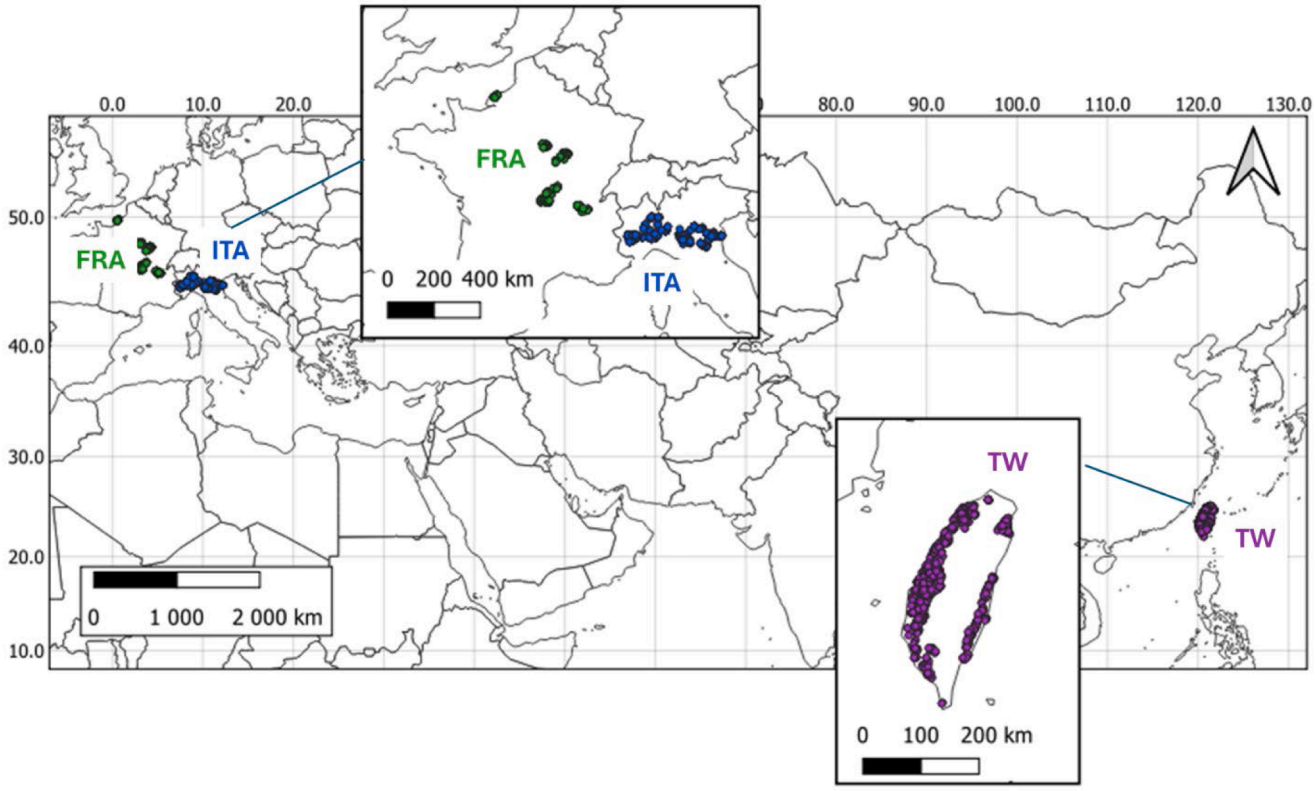

Fig. 1. Soil sampling location within the three regions of interest in France, Italy and Taiwan.

season and hot summer) in the southern part of Auvergne–Rhône-Alpes, and Cfb (temperate with no dry season and warm summer) in the remaining areas. Agricultural production across these regions is highly diversified, reflecting both climatic gradients and historical land-use patterns. Agricultural production across the French study areas is highly diversified, reflecting both climatic gradients and historical land-use patterns, and includes arable crops, permanent grasslands, livestock production, viticulture, and mixed farming systems [26–28].

The French study area exhibits a temperate agricultural phenology characterized by the coexistence of winter and summer crops, in line with the Italian case study. This cropping pattern results in marked seasonal dynamics in vegetation development and satellite-observed spectral responses.

SOC dynamics in these areas are influenced by contrasting management practices, ranging from conventional tillage to increasing adoption of conservation and integrated agriculture practices. Fertilization intensity and residue management vary with crop type and region, with generally higher nitrogen inputs in intensive cereal and forage systems. Based on FAO soil classification [24] and the European Soil Database [25], the predominant soil types across the study regions are Cambisols, Luvisols, Fluvisols, and Retisols.

2.1.3. Taiwan

The study area in Taiwan is located in the main rice-producing regions, where rice is the most extensively cultivated food crop. Among these regions, the rice cultivation area in western part is significantly larger than that in the eastern, northern, and southern parts of the Taiwan. Using township boundaries as the calculation unit, the soil sampling sites of this study cover at least the top 50 townships with the largest rice cultivation areas across Taiwan. The study area is predominantly located in lowland plains below 100 m in elevation, encompassing a humid subtropical climate with hot summers (Cfa) in the northern and eastern regions, a humid subtropical climate with dry winters and monsoonal influence (Cwa) in the western part, and a tropical monsoon climate (Aw) in the south. Fertilizer application rates and water management in paddy fields are influenced by local climate, soil type, and rice variety.

Crop phenology is governed by a subtropical climate with two main cropping seasons, resulting in multiple cropping cycles per year and frequent phenological transitions observable in satellite data.

According to the FAO classification, the soil samples collected in the study mainly include Acrisols, Anthrosols and Cambisols, while Fluvisols and Vertisols are less frequent.

2.2. Soil sampling and SOC laboratory measurements

2.2.1. Italy and France

Three soil sampling campaigns were conducted between 2020 and 2022 in Italy, and in 2023 in France. A total of 922 topsoil samples (0–20 cm) were collected, with 597 from Italy and 325 from France (Fig. 1). At each sampling location, a composite sample was prepared by combining five subsamples collected within a 5-meter radius. SOC content was then determined in the laboratory using the Walkley-Black method (ISO 14,235:1998).

2.2.2. Taiwan

Soil sampling in the study area of Taiwan was conducted between 2022 and 2024. The sampling periods mainly coincided with the rice harvesting seasons of the first (June–July) and second (November–December) cropping periods each year. A total of 1180 topsoil samples (0–20 cm) were collected. In each sampling field (approximately 0.2–0.5 ha), five subsamples were collected and composited into one mixed sample. The SOC content was determined using a TOC analyzer (solid TOC cube, Elementar) [29].

2.3. Remote sensing collections

Satellite data corresponding to the sampling locations in each region

of interest were retrieved using the rgee R package [30,31], which provides an interface to the Google Earth Engine (GEE) Python API, enabling the integration of Earth Engine functionalities within the R environment. For the bare soil and vegetation indices covariates, the Level 2A harmonized Sentinel-2 collection (COPERNICUS/S2_SR_HARMONIZED) available on GEE was employed, whereas for the Satellite Embedding approach, data were sourced from the Satellite Embedding V1 collection (GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL) (Fig. 2).

2.3.1. Sentinel-2 bare soil composite

For each sampling location, a pixel-wise temporal mosaicking strategy was employed to extract bare soil imagery from the Sentinel-2 collections corresponding to the sampling year, producing a composite bare soil layer. Satellite images were selected among those acquired with a mean sun zenith angle lower than 70° and falling within the period corresponding to soil sampling for each study area. These solar conditions are generally met from late spring to early autumn in Italy and France, and over a wider time window in Taiwan, extending from early spring to mid-autumn. These periods also correspond to the highest probability of observing exposed soil between successive cropping cycles. The workflow involved three main steps. First, a set of filters was applied to each pixel across all acquisition dates to exclude those affected by atmospheric conditions (ATM_Mask). Second, pixels were screened to identify bare soil conditions, removing observations influenced by photosynthetically active or senescent vegetation, as well as soils with elevated moisture levels (BS_Mask) [32].

Satellite data affected by photosynthetically active vegetation were excluded using an NDVI threshold, retaining only observations with NDVI values lower than 0.35. To exclude senescent vegetation, a threshold of 0.125 on the Normalized Burn Ratio index (NBR2), calculated as the normalized difference between bands B11 and B12, was applied, leaving only data with NBR2 < 0.125 [33]. The combination of these two indices proved to be effective for selecting bare soil conditions in different soil type and climate conditions [6].

In the third step, for pixels retaining at least three valid bare soil observations after filtering, a mosaicking process was performed. The median (R50) of the reflectance values was then computed for each spectral band of the Sentinel-2 MultiSpectral Instrument (MSI) at the sampling sites in order to reduce the influence of outliers or anomalous observations arising from satellite data collection and ensuring more robust estimates [6]. A spatial resolution of 10 m was used, selected to align the satellite observations with the spatial extent of the field sampling units.

The number of available sampling locations, according to the workflow described above, was evaluated at each step:

1. **After ATM_Mask:** only samples not permanently affected by atmospheric conditions (i.e., where the pixel was not always cloud-covered) were retained.
2. **After BS_Mask:** from the remaining samples, those for which bare soil could not be detected (because they were consistently covered by vegetation, either green or dry, or were too humid) were excluded.
3. **After R50:** only samples with at least three bare soil observations were retained, since the median value could only be computed under this condition. Locations with fewer than three valid bare soil images were discarded.

For the samples retained after the third step, we calculated, for each dataset and sampling year:

- the average percentage of cloud-free images relative to the total number of available Sentinel-2 acquisitions at each location,
- the average percentage of bare soil images relative to the total number of available Sentinel-2 acquisitions at each location, and
- the frequency of bare soil images relative to the cloud-free images.

2.3.2. Sentinel-2 vegetation indices

After applying the ATM_MASK, only satellite data showing a NDVI values higher than 0.35 has been retained, therefore excluding bare soil conditions. After that the median values of three vegetation indices were computed for each sampling location and for the corresponding sampling year:

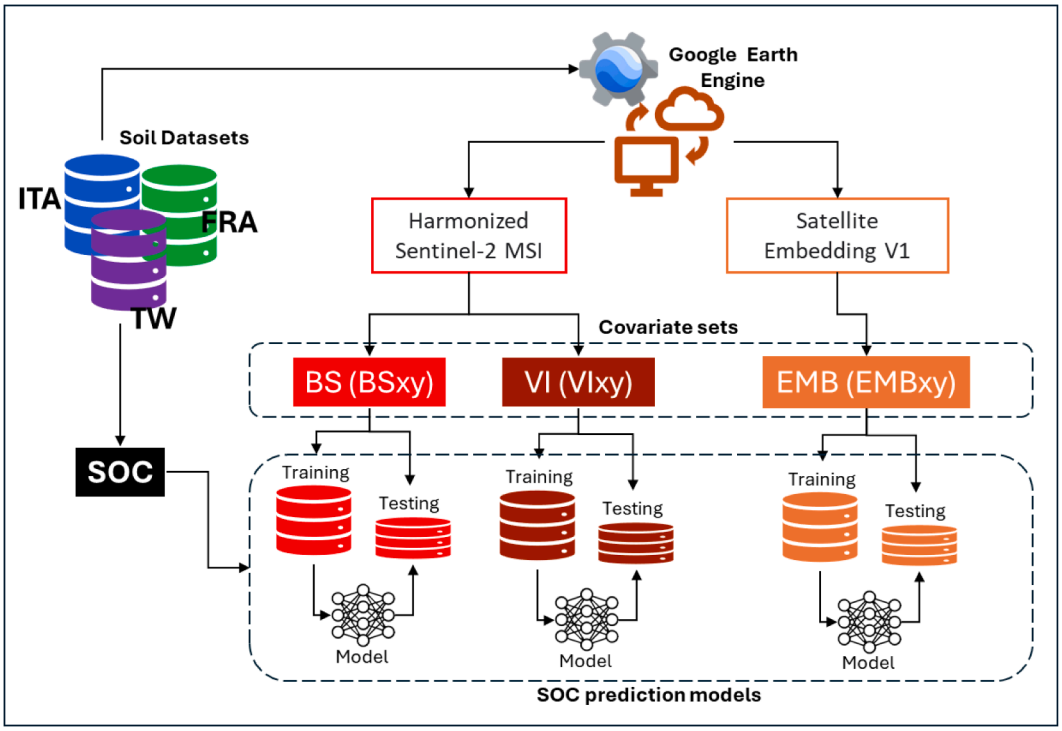

Fig. 2. General workflow of the satellite-derived covariate extraction and SOC prediction models.

- Normalized Difference Vegetation Index (NDVI)

$$NDVI = \frac{B8 - B4}{B8 + B4} \quad (1)$$

- Green Normalized Difference Vegetation Index (GNDVI)

$$GNDVI = \frac{B8 - B3}{B8 + B3} \quad (2)$$

- Normalized Difference Red Edge Index (NDRE)

$$NDRE = \frac{B8A - B5}{B8A + B5} \quad (3)$$

where B3, B4, B5, and B8A correspond to the Sentinel-2 bands centered at 560, 665, 705, and 865 nm, respectively.

As with bare-soil data, the median values of the three indices were used to reduce the influence of anomalous observations and to obtain values less affected by differences in crop type or rotation, representing the average fertility potential of each location.

2.3.3. Satellite Embedding dataset

The *Satellite Embedding* dataset was generated using AlphaEarth Foundations, a geospatial AI model developed by Google DeepMind [20]. The model integrates a wide range of Earth observation sources, including optical imagery (e.g., Sentinel-2), radar data (e.g., Sentinel-1), thermal observations from Landsat-8 and Landsat-9, digital elevation models, climate information (ERA5), and descriptive textual information.

AlphaEarth Foundations produces 64-dimensional geospatial embeddings at a 10-meter spatial resolution, designed to be directly compatible with GEE's analytical tools. These embeddings are analysis-ready, meaning that no atmospheric correction, cloud masking, or spectral transformation is required. For each sampling site and year, the median yearly value of the 64 embedding dimensions was computed to represent local surface conditions.

2.4. SOC estimation models

2.4.1. Covariates and dataset preparation

Six covariate sets, split into two groups were used to train the SOC prediction models (Fig. 2):

Satellite-based covariate sets:

- **BS:** included the R50 values of the eleven Sentinel-2 bands (B1–B12, excluding B9 and B10), representing median bare soil conditions.
- **VI:** included the median values of three vegetation indices (NDVI, GNDVI, and NDRE) derived from the Sentinel-2 image collection.
- **EMB:** included the 64 embeddings produced by the AlphaEarth Foundation geoAI model.

Satellite-based + Longitude and Latitude:

- **BSxy:** included the same covariates as **BS**, plus Longitude and Latitude.
- **VIxy:** included the same covariates as **VI**, plus Longitude and Latitude.
- **EMBxy:** included the same covariates as **EMB**, plus Longitude and Latitude.

For each soil dataset, the covariates were paired with the measured SOC values. Due to the BS mask, the number of available samples was reduced only for the BS and BSxy datasets. A reduced version of the other datasets was then created by selecting the same samples as in the BS dataset.

The Pearson's correlation coefficient was estimated for each covariate versus the SOC values as exploratory analysis.

Each dataset was then divided into training and testing subsets using an 80/20 split, ensuring that both subsets maintained the same SOC value range and distribution. The dataset was first sorted in ascending order according to the SOC. A validation subset was then constructed using a systematic sampling strategy, whereby one sample every five along the ordered SOC sequence was selected for validation, while the remaining samples were used for training.

2.4.2. Modeling approach

To evaluate the predictive performance of SOC for each dataset, we implemented a bootstrap-based modeling approach using the Cubist regression algorithm [34] through the caret package in R [35]. The training procedure involved generating 100 bootstrap replicates: for each replicate, a set of Poisson-distributed weights ($\lambda = 1$) was applied to the training samples to create resampled datasets, and a Cubist model was fitted using 10-fold cross-validation. The hyperparameters of the Cubist models, including the number of committees (1, 2, 3, 5, 10, and 20) and nearest neighbors (from 1 to 5), were optimized within a predefined grid. Predictions from each bootstrap model were computed on a hold-out validation set. The following model-level performance metrics were calculated for each replicate, and 95 % confidence intervals were derived across the 100 models, providing an assessment of uncertainty at the model level; Root Mean Square Error (RMSE) (Eq. (4)), normalized Root Mean Square Error (nRMSE) (Eq. (5)), coefficient of determination R^2 (Eq. (6)), and ratio of performance to interquartile distance (RPIQ) (Eq. (7))

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

$$nRMSE = \frac{RMSE}{\bar{y}} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

$$RPIQ = \frac{Q_{75}(y) - Q_{25}(y)}{RMSE} \quad (7)$$

where y_i are the observed values, $\hat{y}_i$ the estimated values, n the number of observations, $\bar{y}$ the mean of the observed values and $Q_{75}(y)$ and $Q_{25}(y)$ the third and first quantiles of the observed value distribution, respectively.

Moreover, the same bootstrapping approach was employed to implement a hybrid prediction model combining ordinary kriging and Cubist regression (hereafter KpR–Cubist).

In this context, KpR stands for *Kriging-prior Regression*, an approach in which the outputs of an ordinary kriging model, specifically, the kriging mean and its associated variance, are used as additional input features for a machine learning algorithm [12].

In this hybrid framework, spatial dependence is first characterized through ordinary kriging, which produces two spatially explicit predictors for both calibration and validation datasets. These kriging-derived features are subsequently incorporated into the Cubist regression model as supplementary covariates, allowing the algorithm to account for spatial autocorrelation patterns without directly performing geostatistical interpolation. Each bootstrap iteration involves independent kriging feature computation and Cubist model fitting, thereby generating an ensemble of spatially informed models. The methodology aligns conceptually with recent developments in spatially enhanced machine learning [12,36–38], where kriging or related spatial predictors are employed to inform data-driven models through feature

engineering rather than residual post-processing.

To assess differences in predictive performance among the different combinations of models and covariate sets, the Kruskal–Wallis test was first applied to the distribution of RPIQ values from the bootstrap replicates. This non-parametric test evaluates whether there is a statistically significant difference in the median RPIQ values across all groups without assuming normality and homoscedasticity (i.e., it is robust to heteroscedasticity). Because bootstrap resampling was conducted independently for each model–covariate combination, RPIQ distributions were treated as independent samples.

When the Kruskal–Wallis test indicated a significant difference ($p < 0.05$), pairwise post-hoc comparisons were performed using the Wilcoxon rank-sum test (Mann–Whitney U test) for all pairs of groups. P-values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure to control the false discovery rate. This approach allowed the identification of specific dataset pairs exhibiting statistically significant differences in predictive performance as measured by RPIQ.

2.4.3. Variable importance and robustness

To assess the relative contribution of each predictor to SOC estimation, variable importance values were extracted from each Cubist model and summarized across bootstrap iterations. In Cubist, variable importance (expressed as a percentage) is derived from both the frequency with which a covariate is used in the model's rule-based regressions and the extent to which it contributes to reducing prediction error within each rule. Each variable importance value is scaled between 0 and 100, but these scores are computed independently for each covariate and therefore do not sum to 100 across variables. Specifically, Cubist constructs a set of decision rules, each followed by a linear regression. For each rule, the importance of a variable reflects how often it appears in the rule conditions and how strongly it influences the predicted outcome, i.e., how much it helps to reduce the residual error. These local contributions are then averaged across all rules and committees, resulting in a mean importance score for each covariate. Variables with higher importance scores are those that consistently help partition the feature space and improve the accuracy of SOC predictions across the model.

The mean importance across bootstrap realizations was used to quantify the stability and consistency of each predictor's influence on SOC predictions. To further assess the robustness of these predictors, the mean importance was divided by the coefficient of variation (CV) of each covariate. Predictors showing high importance and low variability (i.e., high importance-to-CV ratio) were considered more robust, as they provide stable contributions across the dataset. This interpretation aligns with established concepts of model stability and variable reliability discussed in Strobl et al. [39], Saltelli et al. [40], and Gregorutti et al. [41].

3. Results

3.1. Soil datasets

The mean SOC values are almost identical for FRA and TW (1.59 % and 1.63 %, respectively) and slightly higher for ITA (1.72 %) (Table 1).

Table 1
Mean statistics of the Soil Organic Carbon values in the soil datasets.

Dataset code	Dataset	N°	Min %	Max %	Mean %	Median %	σ %
ITA	Entire	597	0.40	6.26	1.72	1.61	0.67
ITA_red	Reduced	529	0.40	6.31	1.72	1.60	0.67
FRA	Entire	325	0.49	4.96	1.63	1.49	0.67
FRA_red	Reduced	240	0.72	4.96	1.61	1.45	0.68
TW	Entire	1180	0.34	5.3	1.59	1.51	0.69
TW_red	Reduced	520	0.34	4.9	1.53	1.46	0.63

The standard deviations are also very similar, ranging from 0.67 to 0.69. The SOC distributions are quite similar across the three datasets, all showing a positively skewed pattern, as illustrated in Fig. 3a and b

The reduced datasets (ITA_red, FRA_red, and TW_red), obtained by selecting only sampling points for which at least three images with bare soil conditions were available, exhibit SOC distributions very similar to their corresponding full datasets. This similarity is almost perfect between ITA and ITA_red (Fig. 3b), where the dataset reduction was less severe, while some minor differences can be observed between TW and TW_red due to the substantial reduction in sample size (from 1180 to 520).

3.2. Bare soil imagery availability

The availability of cloud-free images within the Sentinel-2 collection, i.e., images not masked by the atmospheric filter, was highest in Italy (39 %), although values varied considerably between the two sampling years (Table 2). In contrast, the proportions were lower in France (24 %) and Taiwan (19.8 %).

The proportion of bare-soil images, and consequently the bare-soil frequency (bare-soil images / cloud-free × 100), was also highest in Italy, with an average frequency of 36.5 %. Taiwan showed a consistently lower value (27.8 %) with little variation across the three sampling years, while France had the lowest frequency (24.7 %).

The bare-soil frequency clearly affects the proportion of useful sampling points within the three datasets, as shown in Fig. 4. For all datasets, almost all sampling points remain after applying the atmospheric mask (ATM_mask). However, after applying the bare-soil mask (BS_mask), the percentages decrease to 94.3 % in Italy, 82.5 % in France, and 67.0 % in Taiwan.

The final step, which involved computing the composite median reflectance (R50), led to a further reduction in the number of usable sampling points, resulting in 88.6 % for Italy, 73.8 % for France, and 44.1 % for Taiwan. The FRA and ITA datasets exhibited similar trends in sampling point reduction, although the decrease was more pronounced for FRA, while the TW dataset showed a much steeper decline compared to the other two.

3.3. SOC vs covariates correlation

Analysis of the Pearson's correlation coefficients between SOC values and the selected covariates revealed clear differences among the three datasets. Overall, BS covariate set exhibited stronger negative correlations in France, moderate in Italy, and weaker in Taiwan (Fig. 5). All BS correlations were statistically significant, except for bands B11 and B12 in Taiwan.

Among the Sentinel-2 spectral bands, the highest correlations were observed in the visible range, followed by the NIR (B8) and SWIR (B12) bands, while B11 consistently showed the weakest relationship. Regarding VI covariate set, the Italian dataset showed the strongest and most significant positive correlations, with GNDVI exhibiting the highest values. Correlations in Taiwan were also significant, though generally weaker, whereas those in France were low and not statistically significant.

Concerning the correlation coefficients between SOC and the covariates obtained from the embedding dataset, due to the intrinsic, dimensionless nature of the embeddings, which result from the synthesis of heterogeneous data sources, consistent trends across covariates that could be shared among the three soil datasets cannot be clearly detected. FRA showed the highest maximum and mean correlations, both negative (−0.43) and positive (0.36), compared with the other two datasets (Table 3). The percentage of covariates showing a significant correlation ($p < 0.05$) was 70 % for FRA and 65 % for TW, while it was markedly lower for ITA (38 %).

Longitude showed a positive and significant correlation with SOC ($p < 0.05$) for TW (0.15) and FRA (0.34), whereas for ITA the correlation

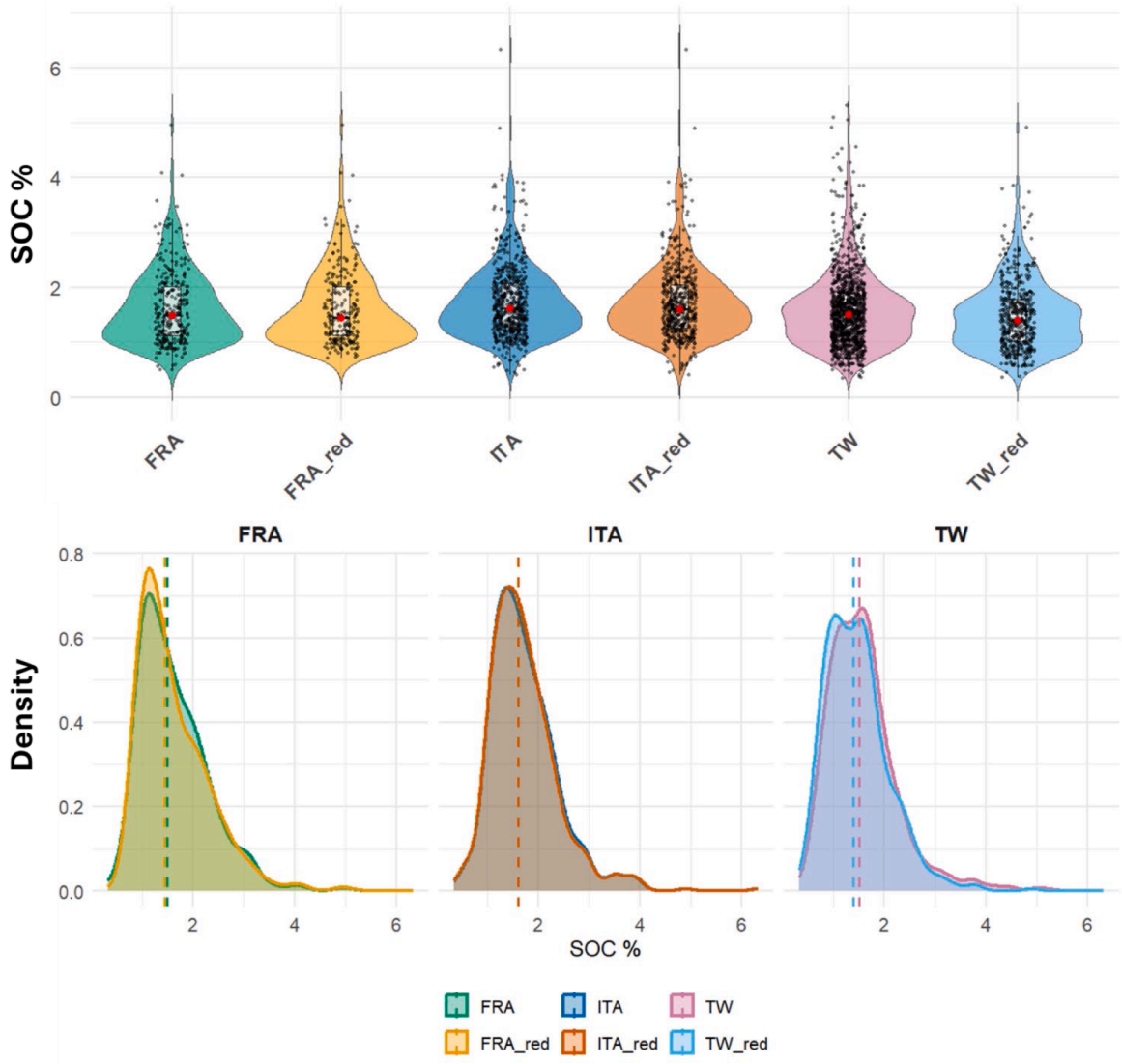

Fig. 3. Violin diagrams and distribution curve relative to soil organin carbon values of the soil datasets.

Table 2
Mean values of the percentage of cloud-free, bare soil (BS) and bare soil frequencies observed in the three regions of interest for each year.

Dataset	Year	Cloud-free %	BS %	BS frequency %
ITA	2020	36.4	10.6	29
	2022	40.7	16.7	41.2
	All	39	14.3	36.5
FRA	2023	24	5.9	24.7
TW	2022	19.1	5	27.1
	2023	19.2	5.1	28.1
	2024	20.6	5.3	28.1
	All	19.8	5.1	27.8

was not significantly different from zero. Latitude exhibited a positive and significant correlation with SOC for ITA (0.26) and TW (0.18), while for FRA it was significant but negative (−0.26).

3.4. SOC estimation models

The Cubist models based solely on satellite-derived covariates (BS, VI, and EMB) showed significant differences in prediction accuracy. Among them, the EMB-derived models consistently achieved higher RPIQ values than those obtained from BS and VI sets across both the entire and reduced datasets in all three countries, although the RPIQ confidence interval was generally broader for EMB (Table 4; Fig. 6). Although the correlation coefficients between SOC and the three VI covariates were comparable across the soil datasets, GNDVI exhibited the highest mean importance and robustness in Italy and Taiwan, whereas NDVI was most influential in France (Table 5). For the BS models, the most relevant and stable predictors were B2, B3, and B8A in Italy; B11 and B12 in France; and B3, B4, B7, B11, and B12 in Taiwan. In contrast, the EMB feature set displayed a more even distribution of variable importance, with mean importance values ranging from 5 to 65.

Including geographic coordinates (LON and LAT) in the Cubist models generally led to significant performance improvements, particularly for BS and EMB sets. Notably, the EMBxy configuration outperformed all other model setups, achieving median RPIQ values of 1.65 in Italy, 2.24 in France, and 1.44 in Taiwan (Table 4; Fig. 6). However, the range between the upper and lower CI RPIQ values is broader for EMBxy compared to BSxy, highlighting a higher estimation uncertainty for the EMBxy models. Variable importance analysis revealed a pronounced influence of spatial coordinates, especially for the VIxy models, where LAT consistently exceeded an importance value of 80 and very Robustness level (Table 5). In the BSxy covariate configuration, LAT was dominant in Italy and France, while in Taiwan, LON ranked among the

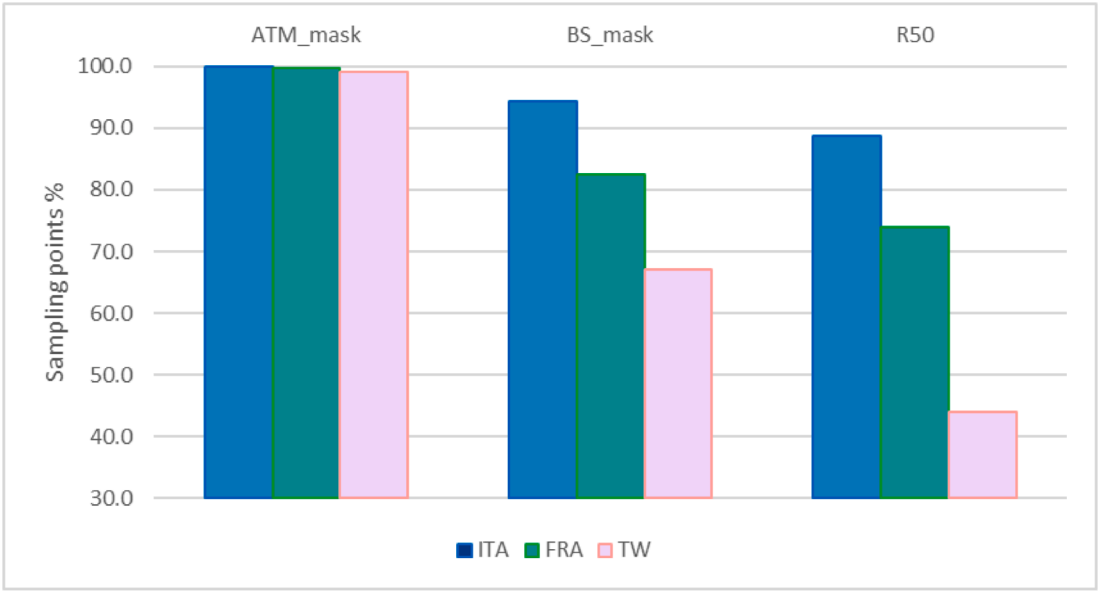

Fig. 4. Percentage of points relative to the total number of sampled points remaining after the application of the atmospheric mask (ATM_mask), the bare soil mask (BS_mask) and the R50 filter.

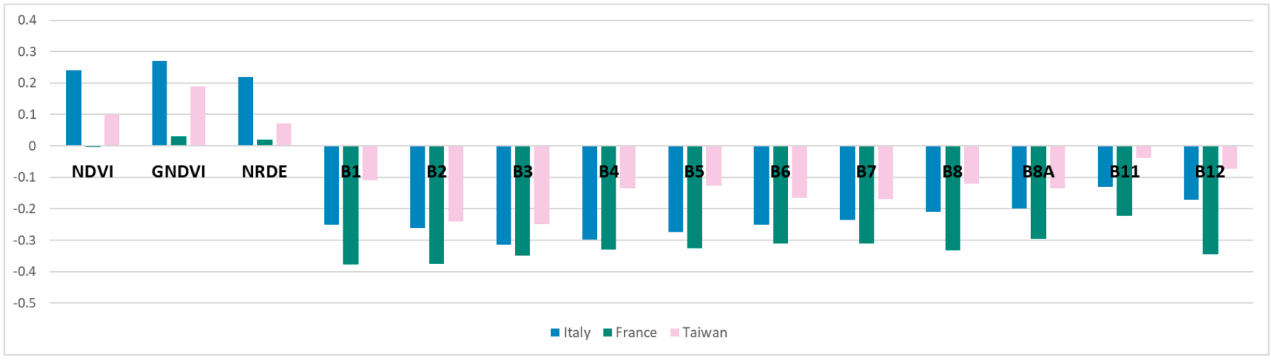

Fig. 5. Pearson's coefficient values for the correlation between SOC and the vegetation indices (VI) and bare soil (BS) covariates.

Table 3
Maximum and mean Pearson's correlation coefficients between SOC and the 64 embedding covariates. The percentage of covariates showing a significant correlation ($p < 0.05$) is also reported.

Dataset	Positive correlation		Negative correlation		% $p < 0.05$
	Max	Mean	Max	Mean	
ITA	0.27	0.08	−0.28	−0.07	38
FRA	0.36	0.18	−0.43	−0.19	70
TW	0.18	0.08	−0.24	−0.09	65

most influential predictors alongside B11, B12, B4, and B7. Due to its lower coefficient of variation, LON also exhibited the highest robustness in this setup. For EMBxy, LAT was the most important predictor in Italy (99.1) and France (93.1), whereas in Taiwan, LON showed the highest importance (88.4), with LAT remaining relatively strong (47.2).

Semi-variogram analysis and Moran's I test confirmed the presence of significant ($p < 0.05$) positive spatial autocorrelation in SOC, which was stronger in FRA (0.77) than in TW (0.42) and ITA (0.33). The strong relevance of the spatial component in Cubist models was further confirmed by the KpR–Cubist approach, where only the kriging-predicted values from Ordinary Kriging were used as input features for the machine learning stage, thus not including any of the other satellite-based covariates. This hybrid model achieved RPIQ values comparable to the EMBxy setup in France and Taiwan, while no improvement was observed for Italy. Although the hybrid framework initially included both kriging-derived and ancillary predictors, the resulting models relied primarily on the kriging estimates, indicating that the spatial component alone captured most of the SOC variability.

4. Discussions

4.1. Constraints and drivers of bare-soil availability for satellite-based SOC assessment

The use of multitemporal satellite imagery increases the likelihood of capturing bare-soil conditions over croplands. However, the actual frequency of bare-soil observations depends on two main factors: the absence of cloud cover and the presence of exposed soil unaffected by vegetation or other disturbing conditions. Major disturbing factors include excessive soil moisture and extreme surface roughness, the latter often observed in the days following tillage.

The most favorable period for linking soil properties with satellite data generally occurs just after seedbed preparation, when soil roughness is minimal [42]. Achieving all these ideal conditions is particularly challenging when using automated approaches across large satellite datasets. Several strategies have been tested to more effectively identify exposed-soil conditions. These include selecting acquisition dates

Table 4
Validation accuracy statistics of the models for predicting organic carbon content. RMSE= root mean square error; RPIQ= ratio to interquartile range; R^2=coefficient of determination; nRMSE: normalized root mean square error.

Dataset	Covariates		N°	N °cal	N°val	RPIQ	RMSE	R^2	nRMSE
ITA	VI	Cubist	597	477	120	1.13	0.69	0.04	0.40
	EMB	Cubist	597	477	120	1.54	0.51	0.46	0.29
	VIxy	Cubist	597	477	120	1.37	0.57	0.31	0.33
	EMBxy	Cubist	597	477	120	1.65	0.48	0.51	0.28
		KpR-Cubist	597	477	120	1.13	0.69	0.01	0.40
ITA_red	VI	Cubist	529	423	106	1.14	0.66	0.03	0.39
	EMB	Cubist	529	423	106	1.62	0.47	0.49	0.28
	BS	Cubist	529	423	106	1.31	0.58	0.22	0.34
	VIxy	Cubist	529	423	106	1.35	0.57	0.31	0.33
	EMBxy	Cubist	529	423	106	1.64	0.47	0.51	0.28
	BSxy	Cubist	529	423	106	1.55	0.49	0.44	0.29
		KpR-Cubist	529	423	106	1.22	0.62	0.04	0.36
FRA	VI	Cubist	325	260	65	1.36	0.62	0.07	0.39
	EMB	Cubist	325	260	65	1.97	0.43	0.56	0.27
	VIxy	Cubist	325	260	65	1.68	0.51	0.41	0.32
	EMBxy	Cubist	325	260	65	2.06	0.41	0.60	0.26
		KpR-Cubist	325	260	65	1.92	0.44	0.54	0.28
FRA_red	VI	Cubist	240	192	48	1.44	0.61	0.11	0.38
	EMB	Cubist	240	192	48	2.16	0.41	0.63	0.26
	BS	Cubist	240	192	48	1.60	0.55	0.28	0.35
	VIxy	Cubist	240	192	48	2.07	0.43	0.56	0.27
	EMBxy	Cubist	240	192	48	2.24	0.40	0.64	0.25
	BSxy	Cubist	240	192	48	2.01	0.44	0.52	0.28
		KpR-Cubist	240	192	48	2.19	0.40	0.60	0.25
TW	VI	Cubist	1172	937	235	1.12	0.69	0.05	0.43
	EMB	Cubist	1180	944	236	1.34	0.57	0.30	0.36
	VIxy	Cubist	1172	937	235	1.29	0.60	0.28	0.38
	EMBxy	Cubist	1180	944	236	1.42	0.54	0.39	0.34
		KpR-Cubist	1180	944	236	1.27	0.60	0.27	0.38
TW_red	VI	Cubist	520	416	104	1.32	0.58	0.09	0.40
	EMB	Cubist	520	416	104	1.39	0.55	0.23	0.38
	BS	Cubist	520	416	104	1.31	0.58	0.10	0.40
	VIxy	Cubist	520	416	104	1.35	0.56	0.16	0.38
	EMBxy	Cubist	520	416	104	1.44	0.54	0.27	0.37
	BSxy	Cubist	520	416	104	1.29	0.59	0.10	0.40
		KpR-Cubist	520	416	104	1.44	0.53	0.31	0.36

corresponding to low surface roughness using NDVI trends [42]; identifying the driest conditions by extracting the 90th percentile of reflectance values (R90) for each band [6] or using the Sentinel-2 Water Index (S2WI) [43]; and computing mean or median reflectance (R50) values [7,44–46] to reduce the influence of extreme values or potential errors.

The frequency of bare-soil observations can be quite low, particularly when a combination of cloudy images coincides with the months when bare-soil conditions are more common. This may explain the observed differences in bare-soil frequency between 2020 and 2022 in Italy (Table 2). The sampling points of the Italian dataset were collected from agricultural areas dominated by winter wheat cultivation [47]. This crop typically presents higher bare-soil occurrence at the end of summer and the beginning of autumn, before the onset of the rainy season [48]. As a result, Italy showed both higher bare-soil frequency and a smaller loss of samples in the training dataset compared to France and Taiwan dataset. Although France and Taiwan exhibited similar percentages of bare-soil pixels and bare-soil frequency, an extensive loss of samples was observed only in Taiwan. For nearly half of the Taiwanese sites, fewer than two dates showing bare-soil conditions were available. This is likely due to the very low proportion of cloud-free images (below 20 %), particularly along the eastern part of the island, where cloudiness is more persistent because of frequent summer rainfall and the winter monsoon. It should be noted that most of the soil samples collected in Taiwan are paddy soils, which are flooded for >100 days during the growing season. This practice drastically reduces the occurrence of bare-soil conditions unaffected by excessive water content.

As shown in Fig. 4, although less pronounced than in Taiwan, a noticeable reduction in usable samples is also evident in France compared to Italy. The French dataset includes sampling areas from distinct and geographically distant regions, characterized by different pedoclimatic conditions. However the scarcity of bare-soil observations in France may be attributed to the widespread adoption of soil conservation practices (minimum tillage and cover crops), which maintain ground cover for most of the year.

Therefore, based on Fig. 4 and the considerations discussed above, it can be inferred that farm management practices can strongly influence the occurrence of bare soil. In particular, the growing attention to soil organic matter conservation and carbon storage, driven by the increasing interest in Carbon Farming and the carbon credit market [49, 50], is leading to a progressively wider adoption of soil conservation practices. Such practices, including minimum or no tillage and the use of cover crops between main crops, enhance soil health and fertility. At the same time, they significantly reduce the frequency of bare-soil conditions and may lead to scenarios in which the soil surface is rarely, if ever, exposed.

4.2. Spectral and vegetation-based covariates for SOC estimation: opportunities and limitations

Even when R50 data can be retrieved, the correlation between Sentinel-2 bands and SOC can be weak, as observed for the TW dataset (Fig. 5). A possible explanation for these lower correlation coefficients, especially for B1, B4, B5, B11, and B12 compared to the other two datasets, lies in the distinct pedoclimatic conditions of Taiwan. Acrisols are widespread across the island, particularly in the northern region. These reddish soils can strongly mask the relationship between SOC and spectral response, particularly in the visible and near-infrared regions due to the spectral overlap between organic matter and iron oxide features [51]. As a result, correlations observed for brownish soil types occurring elsewhere in Taiwan may be confounded. Correlation

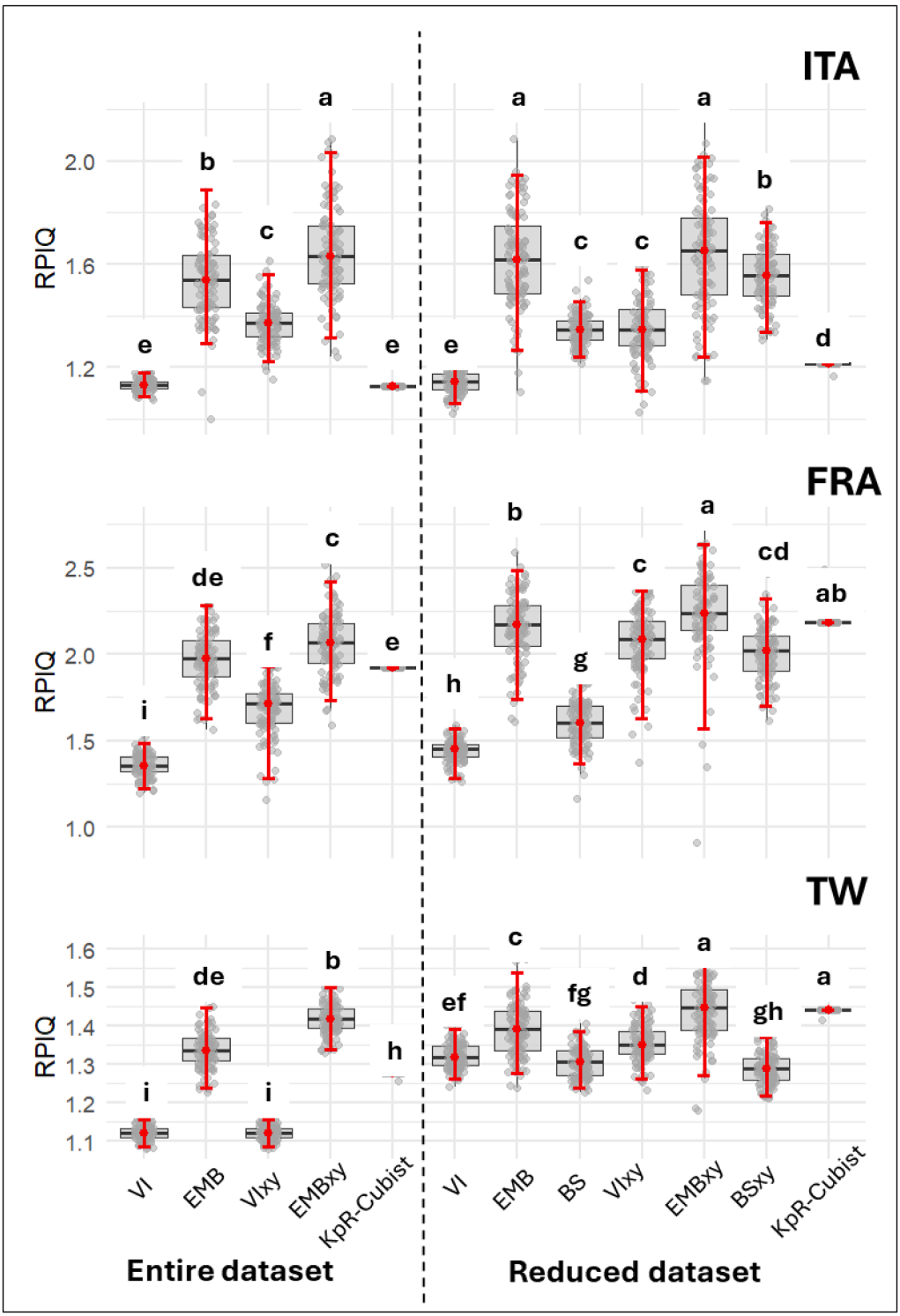

Fig. 6. Boxplots of the validation RPIQ (Ratio to Interquartile Range) values obtained from 100 bootstrap iterations of the soil organic carbon models. Red lines indicate the 95 % confidence intervals. Results are shown for all combinations of covariates across the three datasets. The letters above each boxplot corresponds to the Kruskal–Wallis test followed by Wilcoxon rank-sum post-hoc comparisons carried independently for each region. Different letters indicate significant differences; higher values correspond to earlier letters ($a > b > c$…). ($p < 0.05$).

coefficients are also weak for B1, while for B11 and B12 the relationships are not statistically significant. This is consistent with the high sensitivity of the SWIR region to soil moisture [52,53]. Considering the high annual rainfall in Taiwan and the widespread practice of paddy field flooding, it is reasonable to conclude that the soils are rarely dry.

The development of new approaches and methodologies becomes crucial to implement effective SOC monitoring systems based on satellite data. The three vegetation indices selected in this study have shown good correlations with SOC in previous works [3,54–56]. However, in our analysis, a clear and consistent correlation with SOC was observed only for the Italian site. One possible explanation for the differences in SOC–VIs relationships among the study areas may again lie in variations in farm management. The less diverse crop rotations in Italy and Taiwan may have enhanced the correlation coefficients. The Po River Plain is

Table 5
Mean variable importance and robustness values for all covariates, calculated across 100 bootstrap iterations of the soil organic carbon models.

		ITA		FRA		TW	
Covariate set up	Covariate	Mean Imp.	Robustness	Mean Imp.	Robustness	Mean Imp.	Robustness
VI	GNDVI	100	369.8	26.8	128.4	97.1	359.9
	NRDE	0	0	70.4	180.3	95.5	277.7
	NDVI	4.66	9.5	98.1	232.1	0	0
BS	B1	16.7	64.6	37.0	157.4	34.8	184.5
	B2	81.7	326.5	51.1	219.8	28.4	170.6
	B3	83.1	342.4	32.0	149.1	97.6	584.4
	B4	10.9	44.2	16.1	79.7	84.4	460.3
	B5	8.1	33.0	38.2	190.5	33.5	190.6
	B6	54.2	228.0	23.0	120.6	49.5	278.8
	B7	62.7	266.2	25.3	134.4	75.3	413.4
	B8	67.4	305.4	36.9	209.1	23.6	123.7
	B8A	76.5	323.1	33.8	182.3	21.6	113.8
	B11	23.1	105.8	84.4	523.9	85.0	398.8
	B12	22.1	105.7	81.8	478.4	85.4	407.4
VIxy	LAT	99.8	17,720.9	86.2	2874.7	91.8	3343.6
	LON	44.9	315.2	85.3	176.6	86.0	23,431.9
	GNDVI	26.3	91.3	6.4	30.5	53.6	169.5
	NRDE	14.2	27.7	11.0	28.2	37.9	84.4
	NDVI	9.8	19.1	19.4	46.0	1.0	2.1
BSxy	B1	5.3	20.4	12.1	51.6	34.1	181.2
	B2	63.7	254.7	22.0	94.8	23.1	138.7
	B3	64.2	264.4	30.0	139.7	85.1	509.2
	B4	35.8	145.5	23.9	118.1	70.6	385.0
	B5	19.5	79.0	48.0	239.4	28.5	162.3
	B6	48.3	203.2	26.6	140.0	47.7	268.6
	B7	40.7	172.6	25.2	133.7	65.5	359.7
	B8	41.7	189.1	30.1	170.6	23.1	120.8
	B8A	50.5	213.4	26.2	141.3	25.2	132.4
	B11	34.9	160.1	46.7	289.9	72.9	342.3
	B12	34.0	162.3	31.2	182.6	68.9	328.7
	LAT	94.0	17,247.2	95.5	3183.6	10.6	435.7
	LON	44.7	311.3	66.5	137.8	60.8	25,439.9

characterized by intensive agriculture with short rotations, while the Taiwanese agricultural landscape is dominated by paddy soils, where spring–summer rice is often followed by an autumn rice crop, a dry-season crop, or a fallow period [57]. Conversely, in France, where most soil samples were collected from farms adopting soil-conservation-oriented management, crop rotations tend to be longer and include greater diversity, often using cover crops. This kind of management likely caused a flattening of the correlation between annual mean vegetation index values and SOC content. The selected vegetation indices generally provided the poorest predictive performance when used as covariates in the present study. Although vegetation indices have been successfully employed in combination with other covariates for SOC prediction in previous studies [55,56], we tested their effectiveness in a no-bare-soil scenario. Under these conditions, they proved unsuitable for SOC estimation across all regions of interest. However, the variability in the importance of VI covariates across the three datasets suggests that alternative vegetation indices could be explored, and that their selection should be guided by the predominant farm management practices in the area.

4.3. Integrating satellite embeddings and spatial information in SOC modelling

The Satellite Embedding dataset generated by the AlphaEarth Foundations model can potentially be applied to a wide range of purposes, from the production of thematic maps to the estimation of biophysical variables or change detection [20]. However, these applications have not yet been explored in the literature. Consequently, in this study, we tested for the first time the capability of this dataset (EMB covariates) for SOC prediction. Models trained using EMB covariates consistently outperformed the other approaches (Fig. 6), demonstrating their potential for SOC estimation and monitoring. Different behaviors were observed among the three soil datasets with respect to the correlation between EMB covariates and SOC. This variability may limit the generalization capability of models trained using EMB data. The FRA and TW datasets showed significant correlations for >65 % of the covariates; however, many of these covariates were also highly correlated with each other. This redundancy led to a reduction in the mean importance values within the Cubist models, particularly for FRA, where the average mean importance was 17.4 and its distribution was strongly skewed toward lower values.

When latitude (LAT) and longitude (LON) were included in the VI, BS, or EMB covariate sets, average RPIQ values generally increased. To further verify and explicitly quantify the contribution of spatial covariates, we performed an additional Wilcoxon signed-rank test on paired RPIQ values obtained from identical bootstrap iterations, comparing model configurations with and without geographical coordinates. This supplementary analysis confirmed that the inclusion of LAT and LON leads to a statistically significant improvement in predictive performance ($p < 0.05$), thereby supporting the relevance of spatial information in SOC estimation. However, this improvement in predictive performance was accompanied by greater uncertainty, as evidenced by the wider confidence intervals shown in Fig. 6. The increased uncertainty likely reflects the need to estimate additional spatial effects, as well as potential collinearity between geographic coordinates and other covariates. This suggests that while including spatial coordinates can help capture geographic variability in the data, it may also introduce additional variability in the estimates, highlighting the need to balance predictive gain with the reliability of predictions. Nevertheless, the Cubist models appeared to effectively cope with multicollinearity, as their predictive performance and robustness were not noticeably affected. However, approaches aimed at reducing the number of covariates, such as PCA or minimum noise fraction feature selection [58], could help mitigate multicollinearity in high-dimensional dataset such as EMB. Alternatively, models that are less sensitive to multicollinearity, such as Random Forest (RF) or Support Vector Machines (SVM), could

be explored. Actually, RF models were also tested using the EMBxy covariate sets. Although the average RPIQ values were generally slightly lower than those obtained with Cubist, the standard deviation was also smaller and the range narrower, thereby reducing the uncertainty of the predictions. The best Cubist models, obtained by combining the FRA_red dataset with the EMBxy covariates, yielded RPIQ values between 0.91 and 2.76, with a standard deviation of 0.26 (mean value 2.24) (Fig. 6). Using Random Forest, the corresponding maximum and minimum RPIQ values were 2.49 and 1.77 (mean 2.21), with a standard deviation of 0.14. Similarly, for ITA_red, RF was able to reduce the uncertainty; however, in this case, the mean RPIQ decreased from 1.63 for the Cubist models to 1.51 using RF.

The consistently high mean importance values of the two spatial covariates confirm the presence of spatial autocorrelation in SOC. For the FRA dataset, LAT and LON show similar importance (Table 5), whereas for ITA, LAT is generally more influential than LON, and the opposite trend was observed for TW.

In Taiwan, the higher importance of LON likely reflects the strong pedoclimatic and geological gradient extending from the coastal plains to the central mountainous area. Compared with upland fields, paddy soils are mostly subjected to alternating oxidation–reduction conditions during the rice-growing period, leading to differences in soil organic matter decomposition rates between the two land uses [59]. In addition, Xie et al. [60] also indicated that the SOC content in paddy soils was higher than that in upland soils. Conversely, in the Italian region, the topographic and geological gradient runs from the Po riverbed toward the surrounding hilly areas to the south and north; given that the Po River flows west to east, this could explain the stronger contribution of LAT in the models as compared to LON.

Traditional geostatistical approaches such as ordinary kriging can effectively map soil properties when significant spatial autocorrelation is present. However, these models are inherently static: they predict target values at unsampled locations under the assumption that spatial relationships remain stable over time. In highly dynamic and anthropized environments such as croplands, the spatial structure of SOC may change due to land use modifications and variations in farming practices.

In this regard, hybrid approaches combining geostatistical models with satellite-based covariates can better integrate spatial structure with temporally dynamic EO information [5,37], such as regression kriging [61], linear mixed-effects models [6], or more recently the KpR approach [12].

In all KpR–Cubist model configurations tested here, the kriging-predicted values entirely dominated the deterministic component of the Cubist regression. Consequently, they accounted for 100 % of the variable importance, irrespective of the satellite-based covariates used (EMB, VI, or BS). Consequently, all satellite-derived covariates had zero importance and did not contribute to the model predictions, resulting in identical outcomes across all covariate combinations and a very small variance of RPIQ values across the bootstrap iterations (Fig. 6). Although this narrow confidence interval could be interpreted as indicating low predictive uncertainty, the failure to exploit satellite-derived information suggests that the model effectively behaved as a purely spatial model rather than a truly hybrid one. As a result, it exhibits a static behavior similar to that of purely geostatistical approaches.

4.4. Modelling strategies and implications for operational SOC monitoring

In this study, we employed global models, i.e., models trained on the entire dataset, although data were split into training and testing subsets. However, for large datasets encompassing different pedoclimatic regions, localized modeling approaches could enhance prediction performance by subsetting the data according to spectral similarity, geographical proximity, soil type, or a combination of these factors [46,62,63]. For instance, Tziolas et al. [46] observed an increase from 1.55 to 2.05 when adopting a localized approach that integrated spatial proximity and spectral similarity.

The RPIQ values reported in the present work were derived from a bootstrapping procedure and represent the expected level of prediction accuracy. For the FRA dataset, the average validation RPIQ using EMB covariates consistently exceeded 2, whereas for ITA and TW the best models achieved RPIQ values of 1.64 and 1.44, respectively. A possible explanation for the lower accuracy obtained for the TW dataset could be the sampling strategy, which was not designed to link ground truth data with satellite pixels, but rather to characterize the average SOC value within the small sampled fields.

Although validation accuracy for ITA and TW was not fully satisfactory, it is important to emphasize that the primary goal of this study was to compare satellite-based approaches for SOC estimation, identifying the most promising covariates within the framework of a SOC monitoring system. Considering the high heterogeneity of the datasets analyzed, localized modeling strategies, combined with models less sensitive to multicollinearity and capable of accounting for spatial structure, could further improve the estimates derived from EMB covariates.

5. Conclusions

Our results show that bare-soil availability is strongly limited by climatic and management factors, especially cloud cover, vegetation persistence, and the widespread adoption of cover cropping and no-tillage. These constraints highlight the need for new EO frameworks that integrate multi-temporal and multi-source data to overcome the reduced occurrence of exposed soil.

In this context, the study demonstrates that reliable SOC estimation is possible even under limited bare-soil conditions when advanced remote sensing and modeling techniques are applied. Among the tested approaches, the deep-learning Satellite Embeddings, tested here for the first time for this purpose, achieved the best performance across all three regions of interest, effectively capturing spectral–spatial patterns related to SOC even under vegetation cover. The hybrid KpR–Cubist model achieved comparable accuracy, confirming the benefit of incorporating spatial dependence into predictive frameworks, however a too strong dependence on spatial covariates could affect the model generalization.

Overall, these findings provide clear evidence that SOC assessment in conservation-oriented agroecosystems is feasible without relying exclusively on bare-soil imagery. Combining AI-based image representations with spatially informed models offers a robust pathway to track SOC dynamics across modern croplands.

For operational applications, it is crucial to ensure that soil sampling was congruent with the spatial resolution of the satellite data used. Proper alignment between the sampling area and pixel size enhances model reliability and prevents spurious predictions. Furthermore, when spatial autocorrelation of SOC is detected, incorporating spatially informed models can improve predictive accuracy and generalization, complementing the information provided by remote sensing covariates.

Future research should prioritize temporal monitoring of SOC changes, evaluate the transferability of these approaches across additional agricultural contexts, and explore their integration into operational frameworks for soil carbon accounting and climate-smart land management.

Ethical statement

The tests and procedures described in this manuscript did not involve human participants or animals, and therefore did not require ethical approval. No ethical issues are associated with the methods used, and no specific permits or authorizations were necessary for the activities described.

CRediT authorship contribution statement

Fabio Castaldi: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Chien-Hui Syu:** Writing – original draft, Investigation, Formal analysis, Data curation. **Miguel Conrado Valdez:** Writing – original draft, Investigation, Formal analysis, Data curation. **Chun-Chien Yen:** Writing – original draft, Resources, Formal analysis, Data curation. **Chi-Farn Chen:** Writing – original draft, Formal analysis, Data curation. **Flavio Bertinaria:** Supervision, Project administration, Funding acquisition, Data curation. **Piero Toscano:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Investigation, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Piero Toscano reports financial support was provided by Barilla G. e R. Brothers. Flavio Bertinaria reports a relationship with Barilla G. e R. Brothers that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was partially supported by the SOC research project DBA.AD001.503 funded by Barilla G. e R. Fratelli S.p.A., by the Ministry of Agriculture, Taiwan (grant No 114AS-13.1.1-CI-01), and by the National Research Council of Italy (grant No 0091275/2022).

Data availability

The authors do not have permission to share data.

References

[1] T. Broeg, A. Don, T. Scholten, S. Erasmi, Reducing bias in cropland soil organic carbon and clay predictions using Sentinel-2 composites and data balancing, Remote Sens. Env. 333 (2026) 115109, https://doi.org/10.1016/J.RSE.2025.115109.

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

[3] A. Gholizadeh, D. Žižala, M. Saberioon, L. Borůvka, Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging, Remote Sens. Env. 218 (2018) 89–103, https://doi.org/10.1016/J.RSE.2018.09.015.

[4] B. van Wesemael, A. Abdelbaki, E. Ben-Dor, S. Chabrillat, P. d'Angelo, J.A. M. Demattê, G. Genova, A. Gholizadeh, U. Heiden, P. Karlshoefer, R. Milewski, L. Poggio, M. Sabetizade, A. Sanz, P. Schwind, N. Tsakiridis, N. Tziolas, J. Yagüe, D. Žížala, A European soil organic carbon monitoring system leveraging Sentinel 2 imagery and the LUCAS soil data base, Geoderma 452 (2024) 117113, https://doi.org/10.1016/J.GEODERMA.2024.117113.

[5] E. Vaudour, A. Gholizadeh, F. Castaldi, M. Saberioon, L. Borůvka, D. Urbina-Salazar, Y. Fouad, D. Arrouays, A.C. Richer-de-Forges, J. Biney, J. Wetterlind, B. Van Wesemael, Satellite imagery to map topsoil organic carbon content over cultivated areas: an overview, Remote Sens. (Basel) 14 (2022) 2917, https://doi.org/10.3390/rs14122917.

[6] F. Castaldi, M. Halil Koparan, J. Wetterlind, R. Žydelis, I. Vinci, A. Özge Savaş, C. Kıvrak, T. Tunçay, J. Volungevičius, S. Obber, F. Ragazzi, D. Malo, E. Vaudour, Assessing the capability of Sentinel-2 time-series to estimate soil organic carbon and clay content at local scale in croplands, ISPRS J. Photogramm. Remote Sens. 199 (2023) 40–60, https://doi.org/10.1016/j.isprsjprs.2023.03.016.

[7] K. Dvorakova, U. Heiden, K. Pepers, G. Staats, G. van Os, B. van Wesemael, Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties, Geoderma 429 (2023) 116128, https://doi.org/10.1016/J.GEODERMA.2022.116128.

[8] K. Khosravi Aqdam, S. Rezapour, F. Asadzadeh, A. Nouri, An integrated approach for estimating soil health: incorporating digital elevation models and remote sensing of vegetation, Comput. Electron. Agric. 210 (2023) 107922, https://doi.org/10.1016/J.COMPAG.2023.107922.

[9] B. Basso, J.T. Ritchie, Assessing the impact of management strategies on water use efficiency using soil-plant-atmosphere models, Vadose Zone J. 11 (2012), https://doi.org/10.2136/VZJ2011.0173;WGROUP:STRING:PUBLICATION vzj2011.0173.

[10] Y. Zhang, L. Guo, Y. Chen, T. Shi, M. Luo, Q.L. Ju, H. Zhang, S. Wang, Prediction of soil organic carbon based on landsat 8 monthly NDVI data for the Jianghan Plain in Hubei Province, China, Remote Sens. (Basel) 11 (2019) 1683, https://doi.org/10.3390/RS11141683, 2019Page 1683 11.

[11] O.D. Adeniyi, H. Bature, M. Mearker, A systematic review on digital soil mapping approaches in lowland areas, Land (Basel) 13 (2024) 379, https://doi.org/10.3390/LAND13030379/S1.

[12] J. Schmidinger, V. Barkov, S. Vogel, M. Atzmueller, G.B.M. Heuvelink, Kriging prior regression: a case for kriging-based spatial features with TabPFN in soil mapping, Comput. Electron. Agric. 243 (2026) 111352, https://doi.org/10.1016/J.COMPAG.2025.111352.

[13] Q. Chen, W. Zhou, W. Shi, Estimation of soil organic carbon density on the Qinghai–Tibet plateau using a machine learning model driven by Multisource remote sensing, Remote Sens. (Basel) 16 (2024) 3006, https://doi.org/10.3390/RS16163006, 2024Page 3006 16.

[14] D. Urbina-Salazar, E. Vaudour, A.C. Richer-de-Forges, S. Chen, G. Martelet, N. Baghdadi, D. Arrouays, Sentinel-2 and Sentinel-1 bare soil temporal mosaics of 6-year periods for soil organic carbon content mapping in Central France, Remote Sens. (Basel) 15 (2023) 2410, https://doi.org/10.3390/RS15092410/S1.

[15] O. Yuzugullu, N. Fajraoui, A. Don, F. Liebisch, Satellite-based soil organic carbon mapping on European soils using available datasets and support sampling, Sci. Remote Sens. 9 (2024) 100118, https://doi.org/10.1016/J.SRS.2024.100118.

[16] Y. Zhou, M.S. Ferdinand, J. van Wesemael, K. Dvorakova, P.V. Baret, K. Van Oost, B. van Wesemael, A framework for mapping conservation agricultural fields using optical and radar time series imagery, Remote Sens. Env. 328 (2025) 114858, https://doi.org/10.1016/J.RSE.2025.114858.

[17] Y. Bao, X. Meng, H. Liu, M. Xu, M. Wang, A novel method for soil organic carbon prediction using integrated 'ground-air-space' multimodal remote sensing data, Geoderma 460 (2025) 117453, https://doi.org/10.1016/J.GEODERMA.2025.117453.

[18] Y. Hong, Y. Chen, S. Chen, Y. Wang, W. Hu, S. Ye, X. Song, F. Liu, Y. Zhao, J.A. M. Demattê, L. Shi, H. Shen, Z. Shi, G. Zhang, Y. Liu, Bridging the gap between laboratory VNIR-SWIR spectra and landsat-8 bare soil composite image for soil organic carbon prediction, Remote Sens. Env. 328 (2025) 114874, https://doi.org/10.1016/J.RSE.2025.114874.

[19] X. Hu, C. Yang, M. Zhang, F. Wu, B. Wu, Y. Tian, Superpixel-refined deep learning framework with texture enhancement and multi-layer attention fusion for automatic crop detection in VGI cropland imagery, Comput. Electron. Agric. 238 (2025) 110746, https://doi.org/10.1016/J.COMPAG.2025.110746.

[20] Brown, C.F., Kazmierski, M.R., Pasquarella, V.J., Rucklidge, W.J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Lydia Zhang, L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., Kohli, P., contributions, E., DeepMind, G., 2025. AlphaEarth foundations: an embedding field model for accurate and efficient global mapping from sparse label data, 2025–2032.

[21] A. Perego, A. Rocca, V. Cattivelli, V. Tabaglio, A. Fiorini, S. Barbieri, C. Schillaci, M.E. Chiodini, S. Brenna, M. Acutis, Agro-environmental aspects of conservation agriculture compared to conventional systems: a 3-year experience on 20 farms in the Po valley (Northern Italy), Agric. Syst. 168 (2019) 73–87, https://doi.org/10.1016/J.AGSY.2018.10.008.

[22] M.C. Peel, B.L. Finlayson, T.A. McMahon, Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci. 11 (2007) 1633–1644, https://doi.org/10.5194/HESS-11-1633-2007.

[23] N. Broothaerts, P. Panagos, C. Arias Navarro, C. Ballabio, D. Belitrandi, T. Breure, D. De Medici, D. De Rosa, A. Fendrich, C. Havenga, J. Koeninger, J. Kreiselmeier, M. Labouyrie, L. Liakos, A. Maréchal, J. Martin Jimenez, F. Matthews, V. Michailidis, L. Montanarella, A. Muntwyler, A. Orgiazzi, S. Scarpa, C. Schillaci, D. Simoes Vieira, E. Van Eynde, M. Van Liedekerke, P. Wojda, F. Yunta Mezquita, A. Jones, EUSO annual bulletin 2023, Publications Office of the European Union, Luxembourg, 2024. https://doi.org/10.2760/46142.

[24] IUSS Working Group WRB, International soil classification system for naming soils and creating legends for soil maps. World Reference Base for Soil Resources 2014, in: World Soil Resources Report No. 106, FAO, Rome, 2015. Update 2015.

[25] P. Panagos, L. Montanarella, M. Barbero, A. Schneegans, L. Aguglia, A. Jones, Soil priorities in the European Union, Geoderma. Reg. 29 (2022) e00510, https://doi.org/10.1016/J.GEODRS.2022.E00510.

[26] J.C. Young, S. Calla, L. Lécuyer, Just and sustainable transformed agricultural landscapes: an analysis based on local food actors' ideal visions of agriculture, Agric. Ecosyst. Env. 342 (2023) 108236, https://doi.org/10.1016/J.AGEE.2022.108236.

[27] L'agriculture en Auvergne-Rhône-Alpes d'après le Recensement agricole 2020. 2023.https://opera-connaissances.chambres-agriculture.fr/doc_num.php?explnum_id=196896. (accessed 19 November 2025).

[28] ATLAS Agricole de Normandie. 2018. https://www.prefectures-regions.gouv.fr/normandie/content/download/56045/369480/file/20190128-DRAFF-Atlas-version-light.pdf. (accessed 19 November 2025).

[29] S.H. Jien, B. Minasny, B.J. Yang, Y.T. Liu, C.C. Yen, M.A. Ocba, Y.T. Zhang, C. H. Syu, Enhancing soil carbon Storage: developing high-resolution maps of topsoil organic carbon sequestration potential in Taiwan, Geoderma 459 (2025) 117369, https://doi.org/10.1016/J.GEODERMA.2025.117369.

[30] R Core Team, R A language and environment for statistical computing, R Foundation for Statistical Computing, 2025. - References - Scientific Research Publishing, https://www.scirp.org/reference/referencespapers?referenceid=3967248 (accessed 9.3.25).
[31] C. Aybar, Q. Wu, L. Bautista, R. Yali, A. Barja, rgee: An R package for interacting with Google Earth Engine, J. Open Source Softw. 5 (51) (2020) 2272, https://doi.org/10.21105/joss.02272.
[32] J. Wetterlind, M. Simmler, F. Castaldi, L. Borůvka, J.L. Gabriel, L.C. Gomes, V. Khosravi, C. Kıvrak, M.H. Koparan, A. Lázaro-López, A. Łopatka, F. Liebisch, J. A. Rodriguez, A. Savaş, B. Stenberg, T. Tunçay, I. Vinci, J. Volungevičius, R. Žydelis, E. Vaudour, Influence of soil texture on the estimation of soil organic carbon from sentinel-2 temporal mosaics at 34 European sites, Eur. J. Soil Sci. 76 (2025), https://doi.org/10.1111/EJSS.70054.
[33] F. Castaldi, S. Chabrillat, A. Don, B. van Wesemael, Soil organic carbon mapping using LUCAS topsoil database and Sentinel-2 data: an approach to reduce soil moisture and crop residue effects, Remote Sens. (Basel) 11 (2019), https://doi.org/10.3390/rs11182121.
[34] Quinlan, 1992. Quinlan, J.R. (1992) Learning with continuous classes. Proceedings of Australian Joint Conference on Artificial Intelligence, Hobart 16-18 November 1992, 343-348. - References - Scientific Research Publishing. https://www.scirp.org/reference/referencespapers?referenceid=1865452. (accessed 9.3.25).
[35] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. 28 (2008) 1–26, https://doi.org/10.18637/JSS.V028.I05.
[36] H. Han, J. Suh, Spatial prediction of soil contaminants using a hybrid random forest–ordinary kriging model, Appl. Sci. 14 (2024) 1666, https://doi.org/10.3390/APP14041666, 2024Page 1666 14.
[37] N. Pouladi, A.B. Møller, S. Tabatabai, M.H. Greve, Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging, Geoderma 342 (2019) 85–92, https://doi.org/10.1016/J.GEODERMA.2019.02.019.
[38] C. Zhu, Y. Wei, F. Zhu, W. Lu, Z. Fang, Z. Li, J. Pan, Digital mapping of soil organic carbon based on machine learning and regression kriging, Sens. (Basel) 22 (2022) 8997, https://doi.org/10.3390/S22228997.
[39] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinform. 8 (2007), https://doi.org/10.1186/1471-2105-8-25.
[40] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global sensitivity analysis. The Primer. Global Sensitivity Analysis, 2008, pp. 1–292, https://doi.org/10.1002/9780470725184. The Primer.
[41] B. Gregorutti, B. Michel, P. Saint-Pierre, Correlation and variable importance in random forests, Stat. Comput. 27 (2016) 659–678, https://doi.org/10.1007/S11222-016-9646-1, 2016 27:3.
[42] K. Dvorakova, U. Heiden, B. Van Wesemael, Sentinel-2 exposed soil composite for soil organic carbon prediction, Remote Sens. (Basel) 13 (2021) 1791, https://doi.org/10.3390/RS13091791, 2021Page 1791 13.
[43] E. Vaudour, C. Gomez, P. Lagacherie, T. Loiseau, N. Baghdadi, D. Urbina-Salazar, B. Loubet, D. Arrouays, Temporal mosaicking approaches of Sentinel-2 images for extending topsoil organic carbon content mapping in croplands, Int. J. Appl. Earth Obs. Geoinf. 96 (2021) 102277, https://doi.org/10.1016/J.JAG.2020.102277.
[44] F. Castaldi, Sentinel-2 and landsat-8 multi-temporal series to estimate topsoil properties on croplands, Remote Sens. (Basel) 13 (2021), https://doi.org/10.3390/rs13173345.
[45] J.A.M. Demattê, C.T. Fongaro, R. Rizzo, J.L. Safanelli, Geospatial soil sensing System (GEOS3): a powerful data mining procedure to retrieve soil spectral reflectance from satellite images, Remote Sens Env. 212 (2018) 161–175, https://doi.org/10.1016/J.RSE.2018.04.047.
[46] N. Tziolas, N. Tsakiridis, E. Ben-Dor, J. Theocharis, G. Zalidis, Employing a multi-input deep convolutional neural network to derive soil clay content from a synergy of multi-temporal optical and radar imagery data, Remote Sens. (Basel) (2020) 12, https://doi.org/10.3390/RS12091389.
[47] F. Castaldi, G. Buttafuoco, F. Bertinaria, P. Toscano, A geospatial approach for evaluating impact and potentiality of conservation farming for soil health improvement at regional and farm scale, Soil Tillage Res. 244 (2024) 106212, https://doi.org/10.1016/J.STILL.2024.106212.
[48] N. Mzid, S. Pignatti, W. Huang, R. Casa, An analysis of bare soil occurrence in arable croplands for remote sensing topsoil applications, Remote Sens. (Basel) 13 (2021) 474, https://doi.org/10.3390/RS13030474, 2021Page 474 13.
[49] T. Angelopoulou, A.T. Balafoutis, S. Chabrillat, Earth observation technologies for agricultural carbon credits: a review, Smart Agric. Technol. (2025) 101493, https://doi.org/10.1016/J.ATECH.2025.101493.
[50] European Commission, 2021. EU Soil Strategy for 2030 reaping the benefits of healthy soils for people, food, nature and climate.
[51] D.H. Pearlshtien, E. Ben-Dor, Effect of organic matter content on the spectral signature of iron oxides across the VIS–NIR spectral region in artificial mixtures: an example from a red soil from Israel, Remote Sens. (Basel) 12 (2020) 1960, https://doi.org/10.3390/RS12121960, 2020Page 1960 12.
[52] D. Urbina-Salazar, E. Vaudour, N. Baghdadi, E. Ceschia, A.C. Richer-De-forges, S. Lehmann, D. Arrouays, 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. (Basel) 13 (2021) 5115, https://doi.org/10.3390/RS13245115, 2021Page 5115 13.
[53] J. Yue, J. Tian, Q. Tian, K. Xu, N. Xu, Development of soil moisture indices from differences in water absorption between shortwave-infrared bands, ISPRS J. Photogramm. Remote Sens. 154 (2019) 216–230, https://doi.org/10.1016/J.ISPRSJPRS.2019.06.012.
[54] V.R. Kunkel, T. Wells, G.R. Hancock, Modelling soil organic carbon using vegetation indices across large catchments in eastern Australia, Sci. Total Environ. 817 (2022) 152690, https://doi.org/10.1016/J.SCITOTENV.2021.152690.
[55] L. Wang, H. Liu, X. Wang, X. Xu, L. He, C. Luo, Y. Li, X. Zhang, D. Zang, S. Zheng, X. Mei, Identifying optimal variables to predict soil organic carbon in Sandy, Saline, and black soil regions: remote sensing, terrain, or climate factors? Remote Sens. (Basel) 17 (2025) 237, https://doi.org/10.3390/RS17020237, 2025Page 237 17.
[56] X. Xiao, Q. He, S. Ma, J. Liu, W. Sun, Y. Lin, R. Yi, Environmental variables improve the accuracy of remote sensing estimation of soil organic carbon content, Sci. Rep. 14 (2024) 1–14, https://doi.org/10.1038/s41598-024-68424-5, 20241 14.
[57] A. Hameed, Y.P. Chen, F.T. Shen, S.Y. Lin, H.I. Huang, Y.W. Lin, C.C. Young, Evaluation of a subtropical maize-rice rotation system maintained under long-term fertilizer inputs for sustainable intensification of agriculture, Appl. Soil Ecol. 184 (2023) 104772, https://doi.org/10.1016/J.APSOIL.2022.104772.
[58] F. Castaldi, R. Casa, A. Castrignanò, S. Pascucci, A. Palombo, S. Pignatti, Estimation of soil properties at the field scale from satellite data: a comparison between spatial and non-spatial techniques, Eur. J. Soil Sci. 65 (2014), https://doi.org/10.1111/ejss.12202.
[59] Chen, Z.-S., Hseu, Z.Y., Tsai, C.C., 2015. The soils of Taiwan. World soils book series. https://doi.org/10.1007/978-94-017-9726-9.
[60] Z. Xie, J. Zhu, G. Liu, G. Cadisch, T. Hasegawa, C. Chen, H. Sun, H. Tang, Q. Zeng, Soil organic carbon stocks in China and changes from 1980s to 2000s, Glob. Chang. Biol. 13 (2007) 1989–2007, https://doi.org/10.1111/J.1365-2486.2007.01409.X.
[61] Y. Liu, Y. Chen, Z. Wu, B. Wang, S. Wang, Geographical detector-based stratified regression kriging strategy for mapping soil organic carbon with high spatial heterogeneity, Catena (Amst) 196 (2021) 104953, https://doi.org/10.1016/J.CATENA.2020.104953.
[62] B. Minasny, A.B. McBratney, A conditioned Latin hypercube method for sampling in the presence of ancillary information, Comput. Geosci. 32 (2006) 1378–1388, https://doi.org/10.1016/J.CAGEO.2005.12.009.
[63] K.J. Ward, S. Chabrillat, M. Brell, F. Castaldi, D. Spengler, S. Foerster, Mapping soil organic carbon for airborne and simulated enmap imagery using the lucas soil database and a local plsr, Remote Sens. (Basel) (2020) 12, https://doi.org/10.3390/rs12203451.