

Satellite-based soil organic carbon mapping on European soils using available datasets and support sampling

Onur Yuzugullu^{a,*¹}, Noura Fajraoui^{a,1}, Axel Don^b, Frank Liebisch^c

^a AgriCircle AG, Bahnhofstrasse 28b, Pfäffikon, Schwyz, Switzerland

^b Thünen Institute of Climate-Smart Agriculture, Braunschweig, Germany

^c Agroscope, Water Protection, and Substance Flows, Reckenholzstrasse 191, Zurich, Switzerland



ARTICLE INFO

Keywords:

Remote sensing
Machine learning
Soil carbon mapping
Precision agriculture
Soil health
Soil carbon sequestration

ABSTRACT

Soil organic carbon (SOC) plays a major role in the global carbon cycle and is an important factor for soil health and fertility. Accurate mapping of SOC and other influencing parameters are crucial to guide the optimization of agricultural land management to maintain and restore soil health, to increase soil fertility, and thus to quantify its potential for sequestering CO₂. Remote sensing and machine learning techniques offer promising approaches for predicting SOC distribution. In this study, we used remote sensing data and machine learning algorithms to map SOC at regional to large scale, which we then combined with temporospatial and spectral signature-based soil sampling to integrate local ground measurements. A rigorous validation approach was performed where several independent unseen datasets with a high number of samples were used, which additionally involved densely sampled fields. We found that our approach could predict SOC with an average percentage error of less than 10 % with an R² of 0.91 using support sampling on croplands located on mineral soils, demonstrating the potential of remote sensing, machine learning, and specific ground measurements for mapping SOC. Our results suggest that this approach could make small carbon differences measurable and inform carbon sequestration efforts and improve our understanding of the impacts of land use and field management practices on soil carbon cycling.

1. Introduction

Soil organic carbon (SOC) determines soil quality and health (Tiessen et al., 1994). It plays a pivotal and multifaceted role in enhancing the fertility and productivity of agricultural soils as a crucial regulator of the global carbon cycle (Loveland and Webb, 2003; Karchegani et al., 2012; Falahatkar et al., 2014). Its presence brings significant benefits to soil structure, particularly through the augmentation of water-holding capacity, as well as soil nutrient content (Rawls et al., 2003; Lal, 2020). A high SOC content signifies soil robustness and its capacity to support optimal plant growth. In addition to its impact on soil quality, SOC assumes a key role in the context of climate change mitigation. Soils can act as a vital carbon sink, facilitating carbon sequestration by effectively storing atmospheric carbon in the soil matrix. This capacity to sequester carbon can contribute to global efforts to reduce greenhouse gas emissions. Therefore, monitoring and mapping SOC emerge as indispensable practices, as they enable assessing and managing SOC levels in

agricultural soils. Such endeavors facilitate the implementation of sustainable agricultural practices, leading to long-term productivity and the preservation of soil health (Nunes et al., 2021; Reeves, 1997; Weil Magdoff et al., 2004).

Traditional field-based sampling approaches pose practical challenges in terms of cost and time, especially when dealing with extensive agricultural areas (Tajik et al., 2020; Poelplau et al., 2022; Zeraatpisheh et al., 2022). Due to the high spatial variability of SOC (Huang et al., 2007; Vasenev et al., 2013), cumulative SOC stock calculations via extrapolation to larger areas give inaccurate results, making it impossible to quantify significant changes in soil carbon stocks over several years.

In this context, remote sensing is a promising and affordable alternative for SOC assessment. In recent years, there has been a growing interest in mapping SOC using satellite technologies, with researchers and companies actively pursuing this endeavor. This interest can be attributed to four primary reasons.

* Corresponding author.

E-mail address: onur@agricircle.com (O. Yuzugullu).

¹ Authors contributed equally.

(i). Satellite data offers the advantage of efficiently and cost-effectively monitoring large areas, enabling rapid assessment of SOC distribution.

(ii). Satellite data can provide information on regions where collecting ground data is challenging or even impossible, thereby overcoming logistical limitations.

(iii). The high temporal resolution of satellite data facilitates the tracking of SOC changes over time, enhancing our understanding of dynamic soil carbon dynamics.

(iv). Methods based solely on satellite data are non-invasive, requiring no direct soil interaction or disturbance, simplifying the mapping process, and, most importantly, have lower costs.

During the 1990s, significant research was dedicated to the investigation of the relationship between satellite data and SOC mapping. A notable study (Wilcox et al., 1994) demonstrated strong linear correlations between Landsat TM bands, having a spatial resolution of 30 m, and surface SOC based on an analysis of 224 SOC measurements. Subsequent research continued utilizing Landsat TM for SOC mapping and formed the foundation of SOC mapping using remote sensing (Merry and Levine, 1995; Batjes, 1996; Frazier et al., 1997; Levine et al., 1998; Cannell et al., 1999). These studies consistently reported promising correlations with coefficients of determination (R^2) exceeding 0.5, which means that only 50 % of the variation could be explained.

In the 2000s, the focus shifted towards validating SOC predictions from remote sensing against higher-resolution data, such as aerial photography, to investigate spatial variations. For example, Basnyat et al. (2004) compared the spatial distribution of SOC with plant growth through the normalized difference vegetation index (NDVI) derived from Landsat TM. Throughout this decade, research on SOC prediction using remote sensing continued, expanding sample sizes and incorporating field spectroscopy and Landsat TM data (Goidts et al., 2009; He et al., 2002; Jaber, 2006; Nyssen et al., 2008; Szakacs et al., 2004).

The 2010s marked a significant shift in SOC research, coinciding with the launch of Sentinel satellites, having a spatial resolution of 10–20 m. In 2018, researchers from the Czech Republic investigated methods for retrieving SOC and soil texture from Sentinel-2 data, showcasing the superiority of satellite-derived SOC maps over those calculated from aerial images regarding the accuracy and the benefits of high temporal resolution of shorter than 2 weeks (Gholizadeh et al., 2018). Another notable study in 2018 by Castaldi et al. (2018) modeled SOC in croplands using APEX data and the extensive LUCAS topsoil database, incorporating the highest number of samples up to that point. In contrast to previous decades focused on mapping and spatial variation, the 2010s witnessed a shift towards quantitative analysis through model development (Castaldi et al., 2016, 2018, 2019, 2023; Ladoni et al., 2010; Ayoubi et al., 2011; Kumar et al., 2012, 2018; Vågen et al., 2013; Noshadi et al., 2014; Peng et al., 2015; Mondal et al., 2017).

By 2020, SOC sequestration had emerged as a prominent and highly researched topic in the field of remote sensing (Castaldi et al., 2023; Dvorakova et al., 2023). Numerous important factors influencing SOC were investigated during this period, furthering our understanding of this complex and highly volatile phenomenon. Notably, studies delved into the effects of crop residues (Dvorakova et al., 2020), terrain (Guo et al., 2021), climate (Dvornikov et al., 2021), soil texture (Padarian et al., 2022), and management practices (Paul et al., 2020; He et al., 2021; Wang et al., 2021a; Zhang et al., 2022). These investigations shed light on the multifaceted nature of SOC dynamics and their relationship on SOC sequestration with key environmental factors.

Simultaneously, researchers integrated optical and radar data to model changes in SOC (Wang et al., 2021b; Nguyen et al., 2022; Zhou et al., 2022). This fusion of data sources provided a comprehensive perspective on SOC variations, capturing both surface and subsurface characteristics. Shafizadeh et al. (Shafizadeh-Moghadam et al., 2022) combined multi-year Sentinel-1, Sentinel-2, and terrain data to assess the SOC on forest and cultivated lands. During this period, there was a notable trend towards developing more sophisticated models,

incorporating deep learning techniques, and handling larger datasets (Zhang et al., 2022; Odebiri et al., 2022a, 2022b; Meng et al., 2022; Zhao et al., 2022). These advancements allowed for more accurate predictions and improved understanding of the complex relationships between remote sensing data and SOC dynamics.

In line with these advancements, a recent analysis of cost and accuracy conducted by Andries et al. (2021) has highlighted the potential of integrating remote sensing technology and in-situ sampling as a reliable and cost-efficient method for monitoring and assessing SOC stocks at a local scale.

In this study, we investigate two different predictive frameworks for SOC estimation using satellite data and machine learning only and machine learning refined with informed ground measurements. In both frameworks machine-learning parts consist of different algorithms, which are evaluated for their individual performances, including a light gradient boosting machine (LGBM), random forest (RF), and multi-layer perception (MLP). To ensure unbiased evaluation, all frameworks are validated on an independent set of samples that were not used during the model development and refinement phase. These measurements are crucial for building accurate machine-learning models and calibrating the relationships between environmental predictors and SOC content.

The objective of this study was to calculate SOC maps at 10 m × 10 m resolution using a novel approach combining a satellite-based machine-learning model and ground measurements. The machine learning model was trained using a Europe-wide dataset collected over agricultural areas. The ground measurements are acquired via the sampling approach called Precision Sampling (Yuzugullu et al., 2020) for more accurate predictions. The Precision Sampling approach considers the time series of Sentinel-1 and Sentinel-2 images and assigns sampling locations according to the spatial distribution of spectral-temporal heterogeneity. The validation analysis has four levels that challenge the developed approach: (i) During model training using a random split of the acquired data, (ii) after model training on the unseen data, (iii) after model training and refinement using ground measurements on the unseen data, (iv) after the model training and refinement using ground measurements on four densely sampled fields located in Europe.

2. Data

This section describes the datasets providing initial ground SOC data used to train the ML model ensemble and subsequently outlines the input features utilized for modeling. It includes information about climate, terrain, and soil texture, as these factors are known to influence soil organic carbon content (VandenBygaart, 2006; Vos et al., 2019).

2.1. Ground data

2.1.1. Training data

For the spatial modeling of SOC levels, we used ground data from various sources offering openly available data and compiled them into a unified database reflecting a large variance in SOC and soil situations for algorithm development and testing. To ensure unit consistency, we converted the units of the collected data to mass percentages SOC(%). The data were collected from different European locations, as depicted in Fig. 1. Specific datasets such as ITACyL (Incio, 2022), LUFA Nord-West, NABODAT (Marianne Stokar, 2022), and OpenAgrar (Poeplau et al., 2020a) represent their respective countries, showing dense sampling in regions like northern Spain, Germany, and Switzerland. The LUCAS (LUCAS - ESDAC, 2020) dataset covers a significant portion of Europe, including countries like France, Great Britain, Italy, Greece, Czechia, Poland, Slovakia, Austria, and Denmark, contributing to a comprehensive coverage of soil SOC measurements.

Table 1 provides a summary of the data used in this study, including information on SOC and soil texture (clay, silt, and sand) at the topsoil (0–10 cm) level. The SOC values exhibit a right-skewed distribution, as shown in Fig. 2, with values ranging from 0.01 to 6.28 %, a mean of 1.32

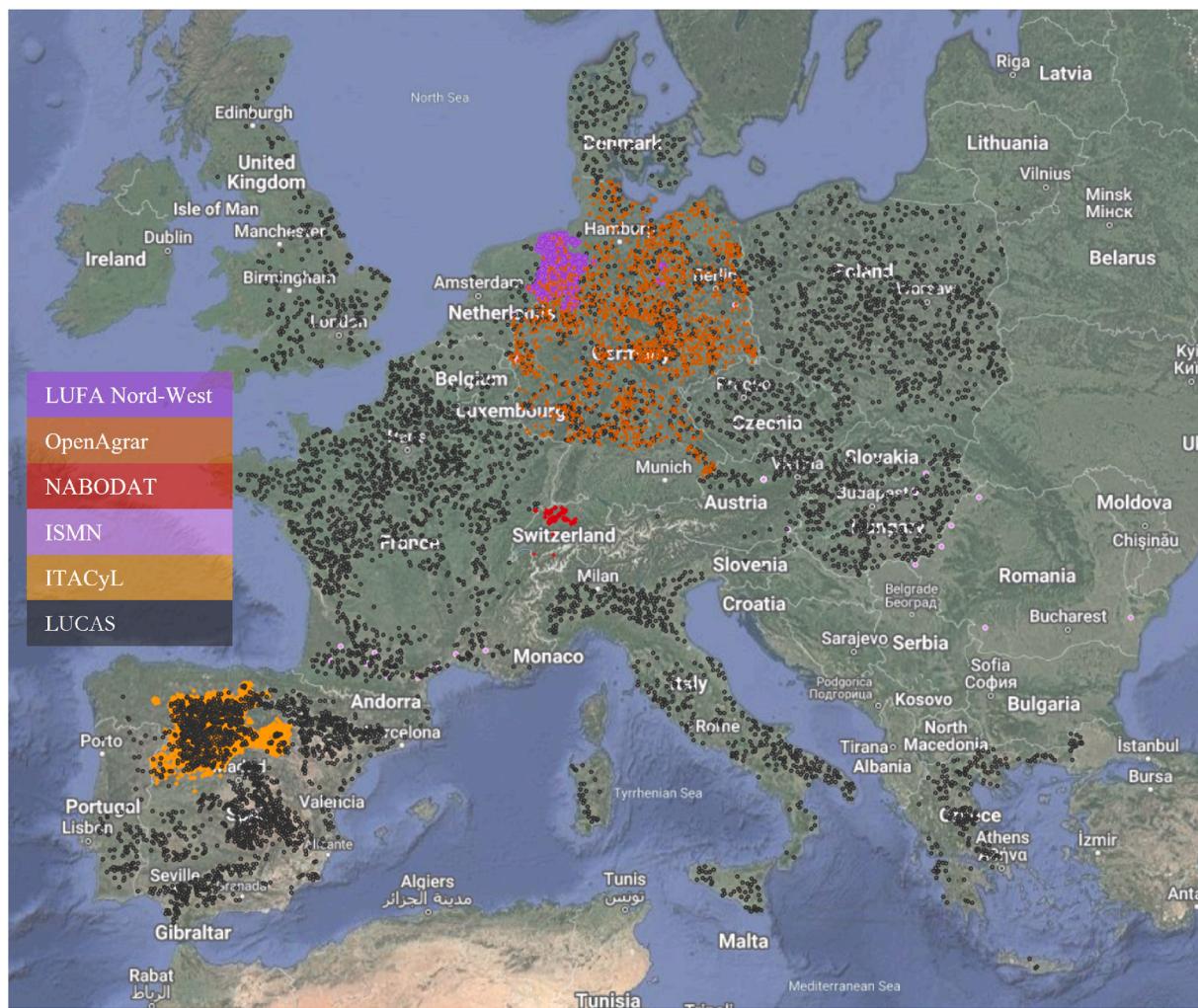


Fig. 1. Available SOC data ($n = 36622$) in this study that are collected from agricultural areas, marked according to their sources.

Table 1

Summary of the data sets used for training the model ensemble and containing data properties related to corg, including according to sources, units, sample counts, resolution, sampling depth, and sampling period.

Name	Unit	Sample Count	Resolution	Sampling Depth (cm)	Sampling Period	Reference
ITACyL	%	21782	Point	0–30	2000–2021	Incio (2022)
LUFA Nord-West	%	16888	Field	0–30	1995–2015	Personal contact
LUCAS	g/kg	8551	Point	0–20	2009–2018	LUCAS - ESDAC (2020)
NABODAT	%	3185	Point	0–30	2000–2017	Marianne Stokar (2022)
OpenAgrar	g/kg	2075	Point	0–10	2011–2018	Poeplau et al. (2020a)
ISRIC	%	1667	Point	0–10	2000–	ISRIC Soil Data Hub (2022)
ISMN	%	110	Point	0–30	2017–	ISMN (2022)
Total		54258				

%, and a standard deviation of around 0.40 %. Considering the distribution of SOC values, it is important to note that the model's predictive capabilities may be limited for SOC values higher than 3.70 % due to the lack of sufficient training data in that range.

2.1.2. Validation data

To validate our approach, we used field samples collected at 0–30 cm depth and followed different levels of validation involving ground measurement. The first level consists of data from [Table 1](#), which are split during the model training. The second and third level includes measurements obtained from fields through the Precision Sampling method ([Yuzugullu et al., 2020](#)). The last level involves densely sampled agricultural fields from Poland, Switzerland, and France, whose details

are given in [Table 2](#). The data from Poland are based on tests of 9032 soil samples in a specific field and made available by Polish companies: Top Farms "Głubczyce" sp. z o. o. and QZ-Solutions sp. z o.o. The tests of these samples were carried out as part of the project of detailed soil research in southern Poland - an experimental field with an area of 127 ha, cultivated by Top Farms "Głubczyce". The data from Switzerland are based on tests of 33 soil samples in a test field included in the EJP Soil STEROPES project. Lastly, the data from France are based on tests of 55 soil samples in two test fields cultivated by SandriCourt Estate.

The Precision Sampling method utilizes a spectral-temporal heterogeneity map of the field, calculated from Sentinel-1 and Sentinel-2 images acquired when the field is either bare soil or at a very low plant biomass coverage. In our validation dataset, we collected data from a

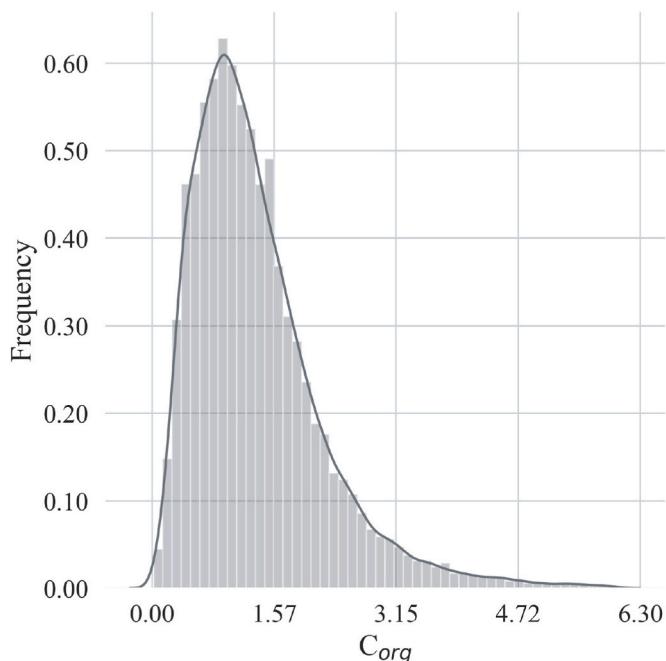


Fig. 2. Distribution of the SOC, used for the development of the prediction models.

Table 2
Information on densely sampled agricultural fields.

Field Code	Country	Area (ha)	SOC Range (%)	Number of Samples	Data Source
(a)	Poland	127	0.5–2.0	9032	Top Farms "Głubczyce" sp. z o.o. and QZ-Solutions sp. z o.o.
(b)	Switzerland	4.2	2.2–3.2	33	Public, EGP Soil STEROPES
(c)	France	7.5	1.7–4.4	32	Sandricourt Estate
(d)	France	14	1.8–3.3	23	Sandricourt Estate

total of 532 fields belonging to AgriCircle AG clients, as shown in Fig. 3. The soil samples were taken at 0–30 cm depth. Among these fields, 225 had independent validation points located at least 30 m away from the sampling locations that are used for the refinement process. The sampled fields were primarily located in Great Britain, Germany, Switzerland, Poland, and Ukraine. Densely sampled fields represent the real spatial variation of SOC, challenging the presented methodology at a finer scale. The fields are located in Poland, Switzerland, and France. For refinement purposes, the closest measurement point to the Precision Sampling location is selected.

By incorporating these field measurements into our analysis, we ensure a robust validation of our approach and enhance the reliability of our results. The diverse geographical distribution of the sampled fields further strengthens the generalizability of our findings across different agricultural areas having different climates, soil textures, terrain, and even management practices, as being tested using the densely sampled fields.

2.2. Satellite data

In this research, we used data from Sentinel-1 (S1) and Sentinel-2 (S2), which are part of the Copernicus program led by the European Space Agency (ESA) in partnership with the European Union (EU). The Copernicus program is a comprehensive Earth observation initiative to

provide accurate and timely information about the Earth's environment. Through a constellation of satellites known as Sentinels, Copernicus collects data on the atmosphere, oceans, and land. The program offers free and open access to its data and services, empowering policymakers, businesses, and citizens to make informed decisions regarding climate change and sustainable development, in particular for agriculture.

S1 carries a C-band Synthetic Aperture Radar (SAR) with a center frequency of 5.405 GHz. The S1 mission consists of two satellites, S1a, and S1b, providing a maximum of 6-day temporal resolution. In our SOC prediction model, we used the Ground Range Detected (GRD) data processed by ESA at a spatial resolution of 10 m. The Sentinel-1 data used in this research includes two polarizations: vertical transmission-vertical received (VV) and vertical transmission-horizontal received (VH). These polarizations and the system's frequency make the data sensitive to the physical and dielectric properties of the surface. According to (Liu et al., 2013), soil particles with higher organic matter content exhibit a greater capacity to adsorb water, resulting in lower dielectric constant values and higher transition water content levels within the soil.

The descriptive statistics (minimum, maximum, mean, standard deviation) of the features derived from the Sentinel-1 data are collected between 2017 and 2023 covering multiple agricultural seasons on the ground measurement locations, and are provided in Table 3. The mean and standard deviation of the VV polarization in time are higher compared to the VH polarization. This difference can be attributed to the fact that VH is lower than VV over bare soil with low roughness, as only a limited portion of the signal is polarized and returns to the sensor.

S2 is equipped with the Multispectral Instrument (MSI), which captures 13 electromagnetic spectrum bands ranging from the visible (VIS) to the short infrared (SWIR) wavelengths. The revisit period of S2 is 6 days, providing frequent and regular observations of the Earth's surface. The 13 bands have different spatial resolutions, with some bands at 10 m resolution and some at 20 m resolution. Additionally, there are three bands at 60 m resolution specifically dedicated to cloud screening and atmospheric corrections. The combination of these bands allows for detailed and comprehensive earth observation.

Table 4 summarizes the descriptive statistics (minimum, maximum, mean, standard deviation) for the features derived from the S2 data, which are collected between 2017 and 2023 covering multiple agricultural seasons on the ground measurement locations. The table includes the temporal mean and standard deviation for each band. It is observed that the temporal mean of reflectances converges between 0.3 and 0.4, indicating a consistent average reflectance across different bands. The standard deviations remain relatively stable within the range of 0.05–0.1, suggesting limited variability in reflectance values over time on agricultural soils.

This spectral behavior aligns with the expected spectral signature of soil (Mohamed et al., 2018). The interaction between incident radiation and soil particles can explain the saturation behavior observed. In particular, near-infrared (NIR) and shortwave infrared (SWIR) wavelengths are absorbed and scattered by soil particles, limiting the additional information that can be extracted about soil composition and properties. As the NIR and SWIR wavelengths are absorbed and scattered, the soil's response becomes less sensitive to further changes in the electromagnetic spectrum. This saturation effect occurs because the soil particles effectively absorb the energy at NIR and SWIR wavelengths, and additional increases in energy do not lead to significant changes or improvements in the measurement of soil characteristics.

2.3. Climate data

The climate data used in this study was obtained from the Meteomatics API (Weather API, 2022). The API call involved providing the latitude and longitude coordinates of the sampling locations, as well as specifying the desired parameters Data was retrieved for daily temperature (T) and precipitation (Prec).

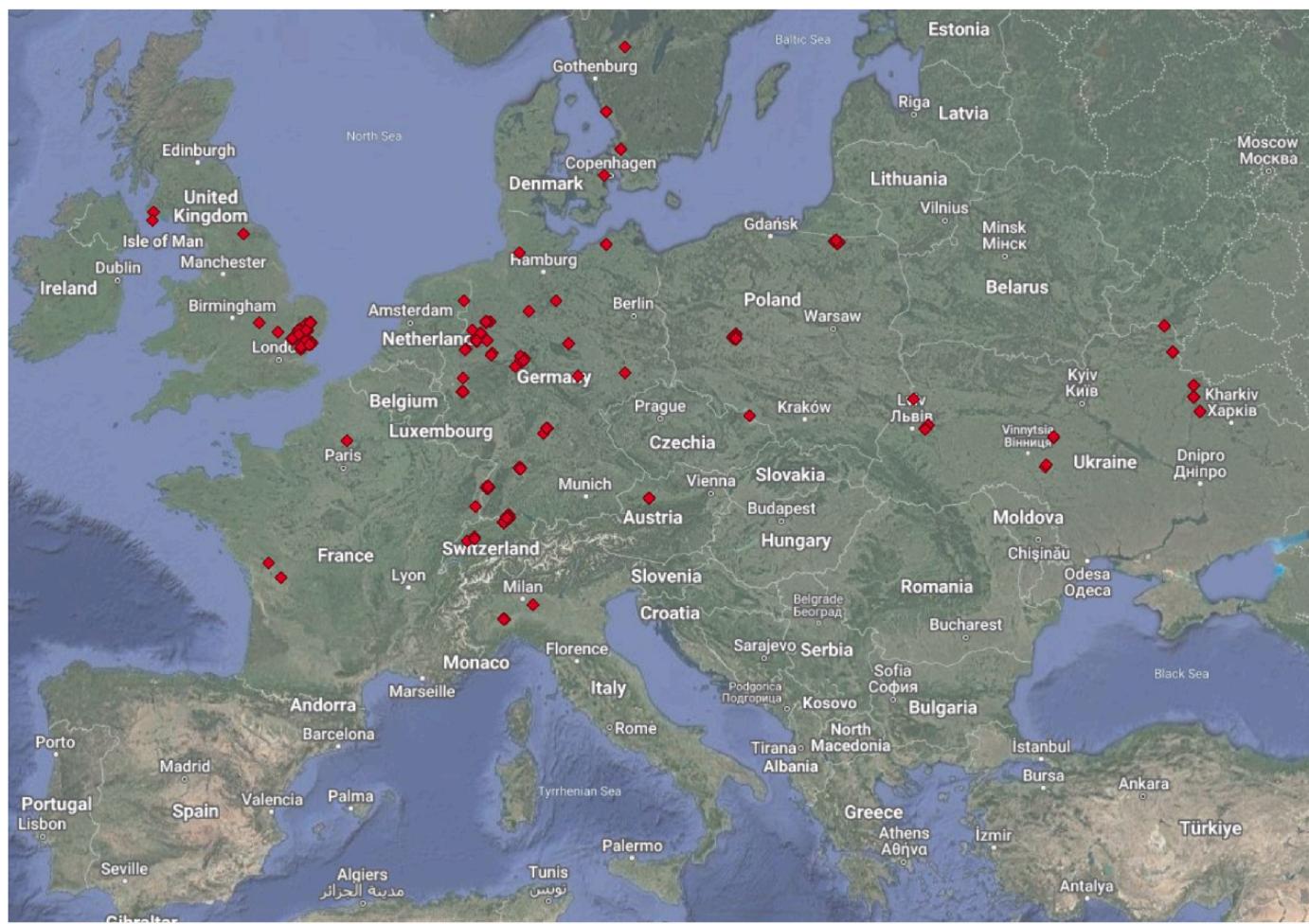


Fig. 3. Validation SOC data ($n = 255$) in this study were collected from agricultural areas with Precision Sampling.

Table 3

Features obtained from S1, namely temporal mean and standard deviation after the incidence angle correction.

Feature	Unit	Min	Max	Mean	stDev	Feature	Unit	Min	Max	Mean	stDev
VV (mean)	dB	-17.58	-1.11	-11.09	1.24	VV (stDev)	dB	0.00	10.23	2.73	0.67
VH (mean)	dB	-23.75	-2.76	-18.54	1.28	VH (stDev)	dB	0.00	12.56	2.64	1.01

Table 4

Reflectance-based features obtained from S2, namely temporal mean and standard deviation after atmospheric correction.

Feature	Unit	Min	Max	Mean	stDev	Feature	Unit	Min	Max	Mean	stDev
B2 (mean)	–	0.00	0.98	0.12	0.05	B2 (stDev)	–	0	0.62	0.08	0.05
B3 (mean)	–	0.00	0.98	0.16	0.05	B3 (stDev)	–	0	0.58	0.08	0.05
B4 (mean)	–	0.00	0.98	0.20	0.05	B4 (stDev)	–	0	0.55	0.08	0.04
B5 (mean)	–	0.00	0.98	0.24	0.05	B5 (stDev)	–	0	0.55	0.09	0.04
B6 (mean)	–	0.00	0.95	0.26	0.06	B6 (stDev)	–	0	0.49	0.09	0.04
B7 (mean)	–	0.00	0.91	0.28	0.06	B7 (stDev)	–	0	0.48	0.09	0.04
B8 (mean)	–	0.00	0.96	0.30	0.06	B8 (stDev)	–	0	0.46	0.10	0.04
B8A (mean)	–	0.00	0.89	0.31	0.06	B8A (stDev)	–	0	0.46	0.10	0.03
B9 (mean)	–	0.00	0.99	0.31	0.06	B9 (stDev)	–	0	0.59	0.10	0.04
B11 (mean)	–	0.00	0.68	0.36	0.07	B11 (stDev)	–	0	0.31	0.09	0.02
B12 (mean)	–	0.00	0.61	0.29	0.05	B12 (stDev)	–	0	0.31	0.08	0.02

The descriptive statistics of the climate features, including the temporal mean, standard deviation of temperature, as well as the temporal mean and sum of precipitation, is shown in Table 5. These statistics offer insights into the variability of temperature and precipitation across the measurement points.

2.4. Terrain data

The terrain data was obtained from the Copernicus Global 30 m Digital Elevation Model, accessed through the Google Earth Engine (GEE) with the identifier COPERNICUS/DEM/GLO30. This dataset has a spatial resolution of 30 m, allowing for detailed information on the

Table 5

Summary of the climate data over last three years.

Feature	Unit	Min	Max	Mean	stDev	Feature	Unit	Min	Max	Mean	stDev
T (mean)	C°	2.13	17.88	10.31	1.40	Prec (mean)	mm	0.01	0.27	0.07	0.03
T (stDev)	C°	2.79	11.74	6.72	0.58	Prec (sum)	mm	19.17	538.37	133.34	65.38

elevation of the study area.

In addition to the elevation data, we derived several terrain-driven features from the digital elevation model (DEM) to further characterize the terrain and hydrological properties of the study area. These features include slope, aspect, hillshade, and topographic wetness index (TWI), which are summarized in [Table 6](#).

2.5. Soil texture

Soil texture plays a crucial role in influencing the variation of SOC levels, as it directly affects the soil's capacity to retain and sequester SOC. Clayey soils, characterized by higher proportions of clay particles, have a greater potential for organic matter retention and typically exhibit higher SOC content. On the other hand, sandy soils, with a higher proportion of sand particles, generally have lower organic matter and SOC levels.

To obtain a broad understanding of soil composition and its implications for carbon storage, low-resolution soil texture maps provided by organizations like ISRIC (International Soil Reference and Information Centre) categorize soil types based on their relative proportions of sand, silt, and clay. Such maps provide valuable insights into the dynamics of SOC. However, it is important to acknowledge the limitations of low-resolution maps, as they may not accurately capture fine-scale spatial variations in soil texture and SOC content. [Table 7](#) summarizes the descriptive statistics (min, max, mean, standard deviation) of essential soil texture features, offering an overview of their distribution and variability within the study area.

3. Methods

The developed SOC mapping approach consists of four building blocks, as illustrated in [Fig. 4](#):

1. Preprocessing of the data: This step involves applying temporal limitations, eliminating outliers, and selecting relevant features from the input datasets. These preprocessing techniques ensure the quality and relevance of the data used for modeling.

2. Training of different model frameworks: Three different models are trained using the preprocessed data, namely Light Gradient Boosting Machine, Random Forest, and Multi-Layer Perceptron. The purpose is to explore different modeling approaches and assess their performance in predicting SOC levels.

3. Selection of the best-performing model framework: Based on evaluation metrics and performance measures, the model framework that demonstrates the highest accuracy and reliability in predicting SOC levels is selected as the best-performing model and used for the next steps.

4. Refinement of generated maps using ground measurements: The selected model framework is then utilized to generate initial SOC maps. These maps are further refined and validated using ground measurements, ensuring the accuracy and precision of the predicted SOC values.

By following these four building blocks, the SOC mapping approach

aims to provide reliable and accurate predictions of SOC levels across the study area.

3.1. Pre-processing

3.1.1. Temporal data selection

SOC is known to display temporal dynamics influenced by various factors. Interventions aimed at carbon sequestration, such as organic fertilizer application and regenerative agriculture practices, can enhance SOC levels. Conversely, activities leading to bare soil exposure or excessive soil disturbance result in a SOC decline.

To ensure the relevance and consistency of the SOC data used in our model development, a rigorous filtering process was implemented. Only measurements obtained after the year 2000 were considered, while older data were excluded. This filtering criterion serves two purposes. Firstly, it ensures that the selected data represents the current SOC status, as older measurements may not accurately reflect the present conditions due to the aforementioned factors. Secondly, it accounts for potential changes in SOC levels resulting from evolving management practices over time. The temporal variation in SOC is assumed to be negligible in this study, as the variation is expected to be within the range of the laboratory error.

By implementing this temporal filtering process, we aimed to enhance the accuracy and applicability of our model by utilizing the most relevant and up-to-date SOC data available.

3.1.2. Outlier elimination

To ensure data quality and reliability of the data input, we conducted outlier elimination during the preprocessing. We analyzed the distribution of each feature and the target parameter (SOC) individually. We identified samples located in the top and bottom 1 % percentile as potential outliers for each feature. The samples that were marked at least in 80 % of the features were identified as an outlier. These outliers were then removed from the dataset to eliminate extreme values that could distort the analysis and modeling results. Specifically for the SOC values, an additional criterion was applied. Values exceeding 3.70 %, which corresponds to the 0.95 percentile of the data, as shown in [Fig. 2](#), were poorly populated and were therefore removed from the dataset. This criterion helps to maintain the dataset's relevance and coherence with real-world conditions.

As a result of the outlier elimination process, the number of samples in the dataset was reduced from 48877 to 36622 (see [Table 8](#)). While this reduction may seem significant, it contributes to enhancing the quality and reliability of the remaining data by eliminating potential outliers that could introduce bias into the analysis. Furthermore, after the outlier elimination, we observed that the distributions of the parameters tended to converge towards a Gaussian distribution with a slight skewness. This helps feature scaling, which is beneficial for the convergence of machine learning (ML) algorithms.

Table 6

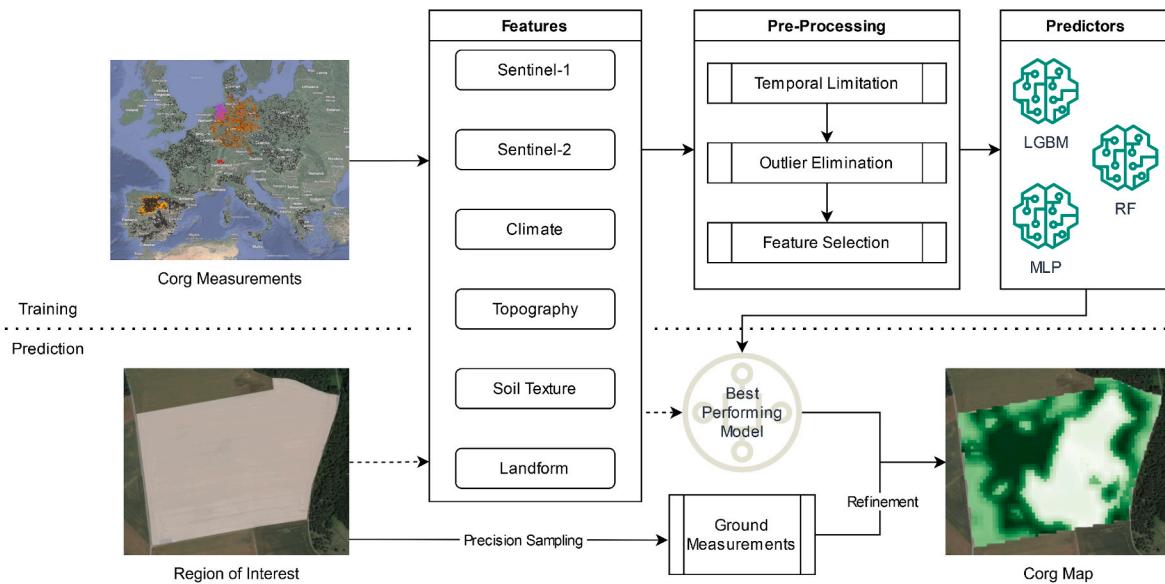
Summary of the terrain data.

Feature	Unit	Min	Max	Mean	stDev	Feature	Unit	Min	Max	Mean	stDev
elevation	m	0	1721.00	531.77	338.83	hillshade	-	107.00	251.00	179.88	7.78
Aspect	-	0	36.00	2.74	2.21	TWI	-	-4.41	18.06	-0.10	2.91
Slope	degree	0	355.00	144.84	106.21						

Table 7

Summary of the low-resolution soil-texture data.

Feature	Unit	Min	Max	Mean	stDev	Feature	Unit	Min	Max	Mean	stDev
Clay	%	3.50	64.00	21.88	7.41	Sand	%	2.00	89.70	44.12	17.40
Silt	%	5.10	80.30	34.00	12.04	Bulk Density	kg/m ³	800	1580	1248	102

**Fig. 4.** Applied process chain for the SOC model training and prediction.**Table 8**

Summary of the ground data after each pre-processing step.

Name	Unit	Original Sample Count	Temporal Limitation	Outlier Elimination
ITACyL	%	21782	20336	20225
LUFA Nord-West	%	16888	4876	4571
LUCAS	g/kg	8551	7814	6962
NABODAT	%	3185	2443	1933
OpenAgrar	g/kg	2075	1859	1624
ISRIC	%	1667	1439	1211
ISMN	%	110	110	96
Total		54258	48877	36622

3.1.3. Feature selection

In our study, feature selection was performed using linear correlation analysis, which allows us to understand the influence of individual parameters and their linear interactions, namely, Pearson correlation coefficient (Benesty et al., 2009). We assessed the correlation among all the features within the dataset. During the feature elimination process, we focused on highly correlated features. Specifically, we identified features with an absolute correlation coefficient greater than 0.9. From this subset of highly correlated features, we retained the one that exhibited a strong correlation with the target parameter (SOC), and removed the others. By considering the correlation between features, we aimed to identify the most relevant and informative features for our modeling purposes. This approach helps to reduce the dimensionality of the dataset and select the features that contribute the most to predicting SOC levels.

3.2. Model development

3.2.1. Algorithms

To map SOC using satellite data, we evaluated multiple algorithms that are used for SOC mapping using remote sensing data, each offering unique characteristics and advantages (Ayoobi et al., 2011; Ye et al., 2021; Wang et al., 2016; Kim and Grunwald, 2016). The following algorithms, that are implemented in *scikit-learn* package (Pedregosa et al., 2011) were tested in our study:

LGBM (Light Gradient Boosting Machine): LGBM is a gradient boosting framework that utilizes tree-based learning algorithms. It is known for its efficient training and prediction capabilities. LGBM is based on histogram-based algorithms and supports parallel computing, making it suitable for handling large-scale data efficiently. It offers a wide range of hyperparameters that can be customized to optimize performance (Ke et al., 2017).

RF (Random Forest): RF is an ensemble learning method that combines multiple decision trees to make predictions. By aggregating the outcomes of several models, RF improves accuracy and robustness. It also provides the ability to assess the importance of features in the data, facilitating the feature selection process (Breiman, 2001).

MLP (Multi-Layer Perceptron): MLP is a type of artificial neural network that consists of multiple layers of interconnected neurons. It can learn complex patterns in the data and make predictions through feed-forward propagation. MLP models can be trained with or without back-propagation to optimize their performance (Gardner and Dorling, 1998).

These algorithms provide diverse modeling approaches, each with its own strengths in capturing the intricate relationships between several features and SOC levels. Through a thorough assessment of these algorithms, our objective is to identify the model that can accurately explain the variations observed in the features with the highest level of accuracy and use that model for the next steps. This assessment involves rigorous evaluation and comparison of the algorithms' performance using

appropriate metrics and validation techniques.

3.2.2. Data split

The data split step in ML model development is a critical stage for assessing the model's performance and its ability to generalize to new, unseen data. This step involves dividing the dataset into training and validation sets. The training set typically accounts for 80 % of the dataset and is used to train the model. During this process, the model learns the underlying relationship between the input features and the target SOC by minimizing the prediction errors, consequently improving its ability to make accurate predictions.

On the other hand, the validation set plays a crucial role in independently evaluating the model's performance. It consists of the remaining 20 % of the data and should be completely independent of the training sets. This ensures that the model has not been exposed to the validation points during the training process, allowing for an unbiased assessment of its predictive capabilities. By evaluating the model's performance on unseen data, we can gauge its ability to generalize and make reliable predictions under unknown conditions.

3.2.3. Hyperparameter optimization

Hyperparameter optimization plays a crucial role in ML model development as it involves finding the best combination of hyperparameters that maximize the model's performance for a given dataset. Hyperparameters are configuration settings that are not learned from the data but must be set manually.

In this study, hyperparameter optimization is performed using the Hyperopt package (Bergstra et al., 2013). It is a widely used Python library that offers a flexible and efficient framework for hyperparameter tuning. It incorporates Bayesian optimization algorithms, which iteratively explore the hyperparameter space to find the optimal configuration by specifying the objective function to be optimized and controlling the optimization process. It automates the search for the best hyperparameters by intelligently selecting new configurations to evaluate based on the results of previous evaluations. This iterative process continues until an optimal configuration is found that maximizes the model's performance on the validation data.

In this study, we focus on the optimization of the hyperparameters, including learning rate, maximum depth, and bin size, to minimize the validation error between the measured and the estimated values and maximize the coefficient of determination.

3.3. Selection of the best-performing model

To ensure high predictive accuracy, three previously described ML algorithms were evaluated and compared in this study, namely LGBM, RF, and MLP. The selection of the final model was based on the comparison of their validation scores. To assess the accuracy of the models, two metrics were employed: Mean Absolute Percent Error (MAPE) and the coefficient of determination (R^2).

MAPE measures the average percentage difference between the predicted and observed values of SOC. It is calculated using the following equation (1), where n is the number of samples.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{Predicted_i - Measured_i}{Measured_i} \right| \quad (1)$$

The coefficient of determination, R^2 , measures how well the model explains the variance in the SOC values. It ranges from 0 to 1, with a higher value indicating a better model fit to the observed data.

By comparing the MAPE scores of the different models, the algorithm that achieved the highest validation R^2 and lowest MAPE was selected as the final model. A lower MAPE indicates a better fit between the predicted and actual values, indicating higher predictive accuracy and better model performance in capturing the variations in SOC levels.

3.4. Output SOC map refinement and post-processing

Different potential approaches can be considered to reduce the error and enhance the accuracy of the predictions: incorporating additional ground measurements for model training and applying a refinement process using ground measurements in a specific region of interest.

The first option involves retraining the model with a substantial number of additional ground measurements to reduce the error. Given the complex nature of the SOC system in agricultural areas, this option can be costly and time-consuming. In addition, it would require financial resources to conduct field measurements and obtain accurate SOC values across various locations and conditions.

The second option involves a refinement process using ground measurements in a specific region of interest. Instead of retraining the entire model, ground measurements can be selectively collected in the target region, and the predicted SOC map can be adjusted based on these measurements. This approach provides a more straightforward, more localized solution for achieving highly accurate SOC maps while empowering the first option of extending the dataset in the long term.

3.4.1. Precision sampling

In this study, we choose the second option by following the Precision Sampling approach (Yuzugullu et al., 2020) to refine the predicted SOC map, which was previously mentioned in Section 2.1.2. The process involves sampling each region of interest at multiple points based on the spectral-temporal heterogeneity of the soil observed in Sentinel-1 and Sentinel-2 images. This sampling strategy allows for a more targeted and representative collection of ground measurements.

3.4.2. Post-processing

A 2D interpolation technique with thin-plate spline (Bookstein, 1989) was applied to the sampled SOC measurements to conduct the refinement process. Thin-plate spline interpolation is commonly used in geostatistics and spatial analysis to estimate values at unmeasured locations based on the values observed at sampled locations. It provides a smooth surface that captures the spatial variation of the feature of interest. After the interpolation, a 3x3 smoothing window was applied to the interpolated SOC map. This smoothing process helps reduce noise and further enhances the spatial variation representation over the area of interest. The 3x3 window means that each pixel in the map is smoothed using the values of its neighboring pixels within a 3x3 grid, which corresponds to a 30mx30 m area for the Sentinel satellites.

The predicted SOC map can be adjusted and refined based on the variation of SOC in the field and error observed at the Precision Sampling locations by applying the refinement process using interpolation and smoothing techniques. As we apply a 2D interpolation for the refinement, the quality of the output depends on the location of the Precision Sampling locations, the quality of the ground measurements, and the prediction accuracy of the model. Combining these three factors, this approach aims to improve the spatial accuracy and representation of SOC content across the study area, providing a more reliable and detailed SOC map.

4. Results and discussion

This section initially presents the data insights where the correlations between the SOC and related features that are remotely sensed or obtained from available datasets are investigated. This provides a better understanding of the relationships between the features and their potential influence on SOC and the prediction of it. Next, the model results are presented for different validation cases with and without refinement via ground measurements. Additionally, case studies from Poland, Switzerland, and France, which have dense sampling, are provided to illustrate how the model's results can be applied in different geographical contexts. Lastly, the practical implications of the findings are discussed. The insights gained from the model can have real-world

applications, particularly in the context of soil management and carbon sequestration efforts.

4.1. Data insight

The linear relationships between the target parameter, SOC, and the selected features are represented in Fig. 5. Upon examining the correlations, shown in Fig. 5, it is evident that no single feature exhibits a strong (≥ 0.5) negative or positive correlation with SOC. For example, the Sentinel-1 features, which capture the physical and dielectric properties of the soil, do not show a significant correlation with SOC, as shown in a case study conducted in Iran (Shafizadeh-Moghadam et al., 2022). On the other hand, in our study, the Sentinel-2 features, particularly the temporal mean values, showed a negative correlation ranging from 0.2 to 0.4, consistent with findings from previous investigations (Gholizadeh et al., 2018; Castaldi et al., 2023). In contrast, the temporal standard deviation derived from Sentinel-2 features showed only a weak positive correlation or no correlation at all with SOC.

Turning to the terrain-driven features, none of the parameters displayed high correlations (≥ 0.5), with SOC except for elevation. In the collected data, the elevation exhibited a relatively strong negative correlation with SOC. However, this finding contradicts the literature, where researchers have found a positive correlation between elevation and SOC (Dieleman et al., 2013; Tsui et al., 2013). On the other hand,

our data's negative correlation, which was previously also observed in (Aksoy et al., 2016), explained considering the SOC enhancing effect of shallow groundwater, which is mainly found in the lowlands with low elevation (Poeplau et al., 2020b) and likely to the much larger extend of the dataset compared to the previous studies (Dieleman et al., 2013; Tsui et al., 2013) in a different climate introducing a bias towards lower regions because of their larger spatial extent in Europe. Another reason can be climate conditions, where higher altitudes have lower temperatures and precipitation, leading to lower soil respiration and crop residues, resulting in more SOC at lowlands (Jiang et al., 2013; Badraghi et al., 2021).

Regarding soil structure, there is a relatively strong negative correlation between bulk density and SOC. This finding can be attributed to the fact that SOC is reducing the bulk density of soils by improving their structure (De Vos et al., 2005; Zeraatpisheh et al., 2021). The ability of soil to infiltrate water is influenced by bulk density, and plants rely on the soil solution to uptake nutrients. Therefore, areas with high bulk density may have limited water availability, impacting plant nutrient uptake, including carbon. Consequently, soils with high bulk density are expected to have lower SOC content (Sakin, 2012; Gupta and Larson, 1979; Adams, 1973).

When examining the inter-correlations between features, several noteworthy observations emerge. Firstly, there is a strong positive correlation between the Sentinel-1 and the Sentinel-2 features, depending

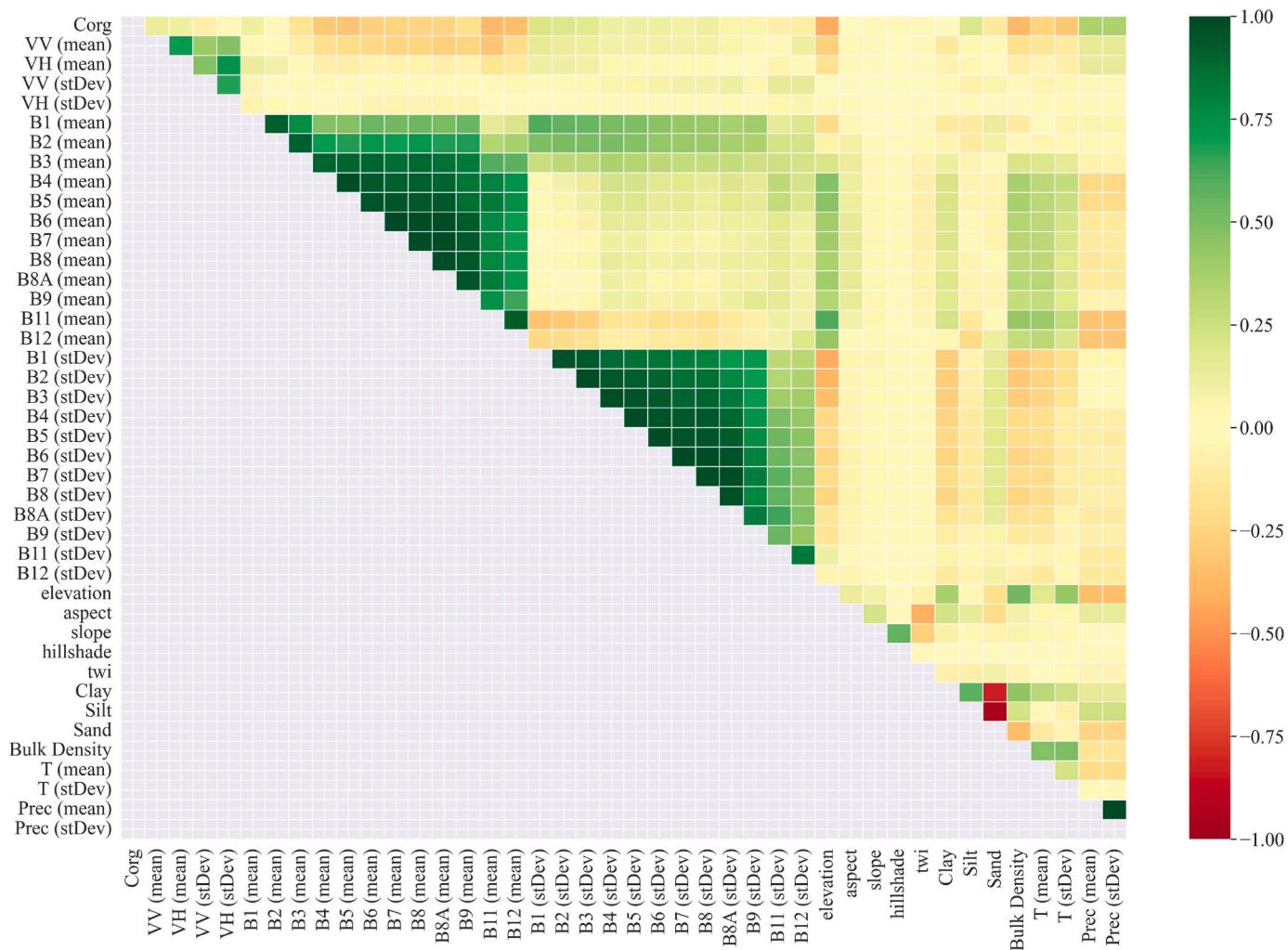


Fig. 5. Correlation between SOC and the features, $n = 36622$. In the color-coded heatmap, red indicates a strong negative correlation, green indicates a strong positive correlation, and yellow represents a weak or no correlation between the features. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

on the satellite. This high correlation leads to eliminating multiple Sentinel-1 and Sentinel-2 features during the feature reduction stage of the pre-processing before the model training stage. Moving on to the terrain-driven features, there is a strong positive correlation between slope and hillshade, which is expected since the hillshade is calculated based on the slope and aspect of the terrain. In terms of soil texture, all three features (clay, silt, and sand) exhibit high correlations due to their sum totaling 100 %. Lastly, the climate features demonstrate a correlation greater than 0.95 between the temporal mean and standard deviation of precipitation, indicating a close relationship between the mean and variability of precipitation values.

4.2. Model results

This section provides an overview of the modeling results of different machine learning frameworks and identifies the best-performing model. The results are categorized into four subsections: internal model assessment (4.2.1), validation without refinement (4.2.2), validation with refinement (4.2.3), and validation with refinement on densely sampled fields (4.2.4).

By assessing the model's performance on both the validation data from the data sources and the separate validation data collected according to Precision Sampling and grid sampling, we can accurately gauge the model's effectiveness in predicting SOC content. This comprehensive evaluation approach ensures that the model's performance is thoroughly assessed and provides valuable insights into its predictive capabilities.

4.2.1. Internal model assessment

We conduct a comparative analysis of three different modeling frameworks discussed in Section 3.2.1, whose performances are summarized in Fig. 6. The scatter plots in the matrix structure allow for a visual comparison of the performance of the different modeling frameworks across various evaluation metrics. They also provide a clear representation of the relationships between the predicted values and the actual SOC values for each model and data set.

We selected the model for further work based on the R^2 and MAPE values of the validation dataset. Upon examining the training column of Fig. 6, we can observe that the LGBM model achieves the highest R^2 value of 0.73, indicating that it explains 73 % of the variation in SOC using the given features. The RF and MLP models follow with slightly lower R^2 values. A similar ranking is observed when considering the MAPE values, where the LGBM model outperforms the other models.

However, it is essential to note that an R^2 value of 0.73 indicates that there is still a substantial portion of the variation in SOC that the model does not explain. This could be due to factors such as unaccounted features, measurement errors, or inherent variability in SOC content that cannot be captured solely by the selected features. Additionally, the scatter plots reveal a tendency of the models to overestimate SOC below 1.5 % and underestimate it above 2 % content. These biases in estimation could be attributed to the complexities of the relationship between the input features and SOC, as well as the limitations of the modeling frameworks that tend to predict towards the mean.

Comparing the performance of LGBM between the training ($R^2 = 0.73$) and validation sets ($R^2 = 0.54$), we observe a difference in R^2 values of 0.19. This difference suggests a slight occurrence of overfitting, where the model may have fit the training data too closely and may not

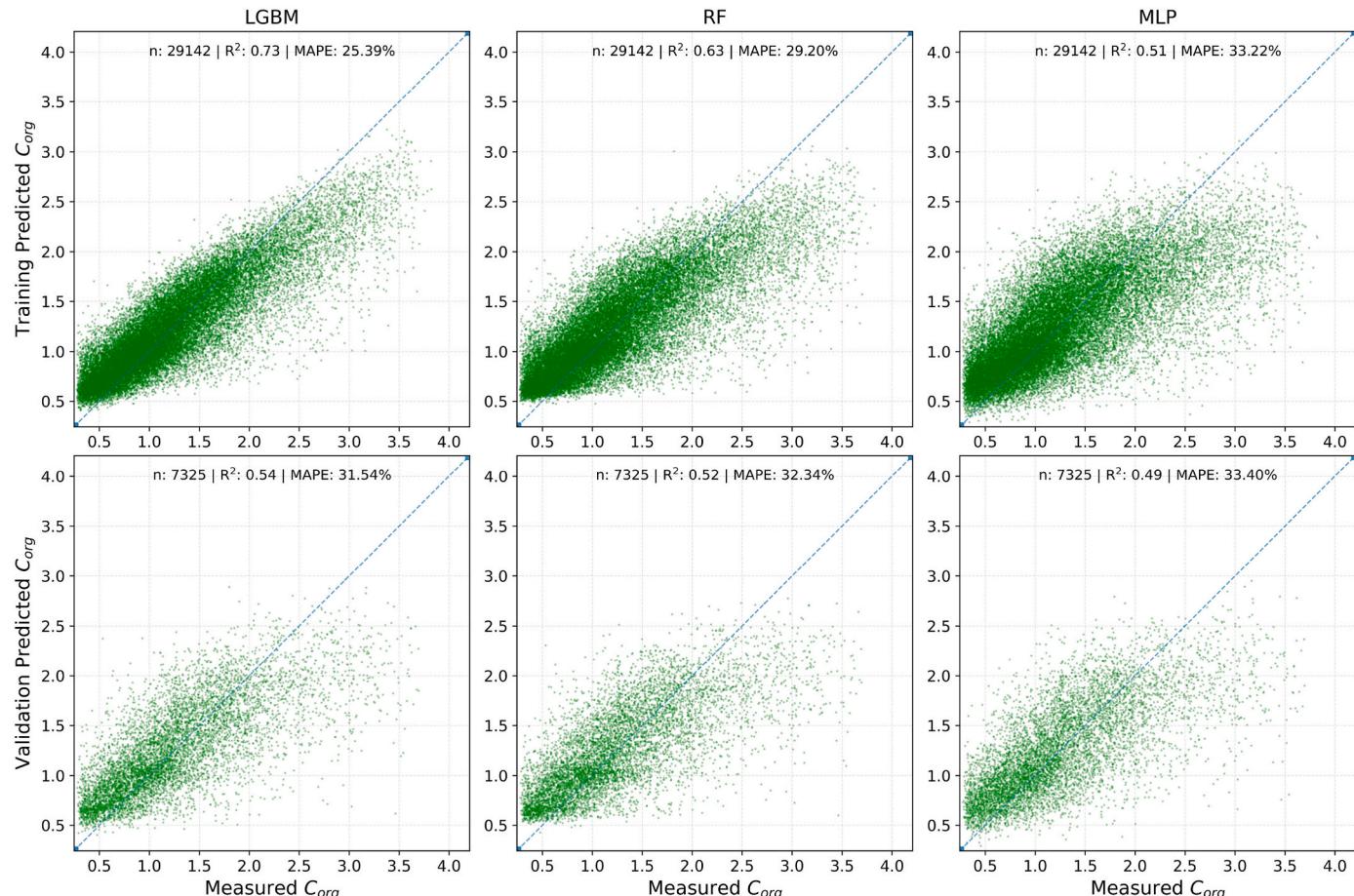


Fig. 6. Scatter plots between measured and predicted SOC values using different frameworks for their training, testing, and validation data sets. Each column in the figure represents a different modeling framework, while each row represents the evaluation results for training and validation data sets.

generalize well to unseen data. However, considering the difference in sample sizes between the training and validation sets, as well as the size of the complete dataset, this difference in performance can be deemed acceptable.

Agricultural management practices, such as crop rotations, cover crops, tillage practices, and nutrient management, play a significant role in the spatial and temporal variations of SOC (Wiesmeier et al., 2013; Hoyle et al., 2016), making it complex to be captured by satellite data alone. At the same time, the spectral information is obtained for the surface of the soil, but soil samples comprise down to 30 cm depth at maximum, causing a potential mismatch of features and the target. Moreover, SOC shows a higher variability also at a small scale that might be possible to resolve using satellite data (Poeplau et al., 2022). The temporal variability of SOC and the timing misalignment between ground measurements and satellite data can introduce additional uncertainties in the predictions (Rochette et al., 2011). Measurement uncertainties arising from different laboratories, equipment, sampling devices and sampling protocols can also impact the accuracy of the model predictions (Jandl et al., 2014).

Given these complexities and limitations, it is crucial to interpret the model's performance in the context of the specific challenges associated with SOC prediction in agricultural systems. Continued research and refinement of modeling approaches, as well as the integration of additional contextual information, can further improve the accuracy and robustness of SOC predictions.

In the validation plot of the LGBM model, we observe an R^2 value of 0.54 and a MAPE value of 32 percent. These results indicate that the LGBM model, trained using the given features, can predict the SOC content at the pixel level with moderate success. However, it is important to consider the spatial range of the SOC measurements, which encompass different climates, soil textures, spectral properties, and cultivation practices in Europe. These factors introduce variability and complexity into the prediction task.

The MAPE value of 32 % indicates a relatively high error rate when it comes to estimating changes in SOC stocks, for example, as part of carbon sequestration schemes. Carbon stock change calculations typically require multiple measurements over time, and the mean SOC values should have minimal overlap with the associated errors to ensure a reliable assessment. In this context, incorporating ground measurements can be beneficial or even necessary to refine the predictions at the field level. Ground measurements provide direct and more accurate information about SOC content, and integrating them into the prediction process can improve their accuracy and reliability.

4.2.2. Validation without precision sampling

In this research section, the previously trained LGBM model was tested on a separate set of points where SOC measurements were collected for Precision Sampling calibration from different fields and validation without performing any prior refinement. This section serves as independent validation, similar to the one conducted in Section 4.2.1, but using recent data collected between 2019 and 2023. The results in Fig. 7 demonstrate similar findings to the validation conducted within internal model evaluations, with a dataset size of 1725 ground measurements.

By evaluating the performance of the LGBM model on this separate validation dataset, we can assess its generalization ability and robustness in predicting SOC content in recent and unseen locations. The scatter plots in Fig. 7 provide insights into the relationship between the predicted SOC values and the actual measurements at these validation points.

The R^2 value of 0.49 obtained from the validation data set indicates that the LGBM model explains approximately 49 % of the variation in SOC using the given features. While this value is lower than the R^2 value obtained from the internal model assessment, it still indicates a moderate performance level in predicting SOC content at the pixel level. However, it is essential to consider the spatial and temporal range of the

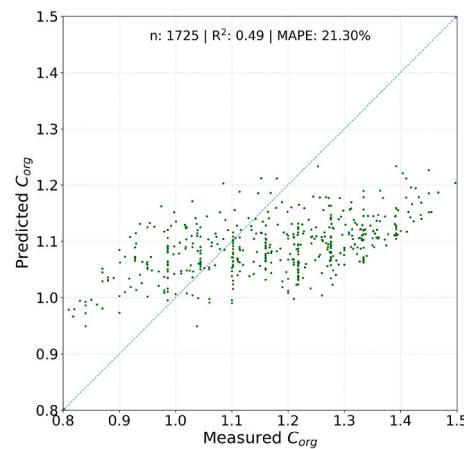


Fig. 7. Correlation between measured and predicted SOC values on the validation data without refinement.

validation data set, which encompasses various environmental conditions, soil types, and agricultural practices. The MAPE value of 21 % suggests that there is still a relatively high error level in the predictions. Similar to the previous validation, the under/over-estimation trend can also be observed in this dataset, with a central tendency around 1.1 % SOC content.

These results further support the model's ability to predict SOC content at the pixel level but also highlight the need for refinement. The MAPE value of 21 percent indicates that there is still room for improvement in the accuracy of the predictions. Incorporating additional ground measurements to retrain the model and refining the prediction map can help mitigate the under/over-estimation tendencies and enhance the accuracy of the SOC predictions for SOC content.

4.2.3. Validation with precision sampling

The validation conducted in Sections 4.2.1 and 4.2.2 indicates that there is room for improvement in achieving accurate SOC maps.

Fig. 8 presents the calibration and validation point results from the dataset described in Section 2.1.2, in which the Precision Sampling measurements of SOC were used to adjust the predicted SOC map via thin-plate spline. It is important to note that these 225 validation points in the agricultural fields are located at least 30 m from the calibration points, ensuring their independence. By examining the scatter plots in Fig. 8, we can observe the performance of the refined SOC map at the calibration (left) and validation (right) points.

The refined SOC map achieves a high R^2 value of 0.93 at the calibration points, indicating a strong correlation between the predicted SOC values and the ground truth measurements. The low MAPE value of 6.92 % suggests a relatively small average difference between the predicted and actual SOC values at the calibration points. These results demonstrate the effectiveness of the refinement process in accurately estimating the SOC content using the interpolation and smoothing techniques.

Moving to the validation points, the refined SOC map maintains a high level of performance with an R^2 value of 0.91. This indicates a good correlation between the predicted SOC values and the ground truth measurements at the validation points, demonstrating the generalizability of the refined map. The slightly lower MAPE value of 6.5 % at the validation points than the calibration points suggests (6.9 %) a slightly larger average prediction error. However, this MAPE value is still significantly lower than the initial MAPE of 21 % observed without incorporating ground measurements.

The significant reduction in MAPE from 21 % (without refinement) to 6.5 % (with refinement) highlights the value of incorporating ground measurements, particularly with the Precision Sampling approach, in improving the accuracy of the SOC map. Moreover, achieving these

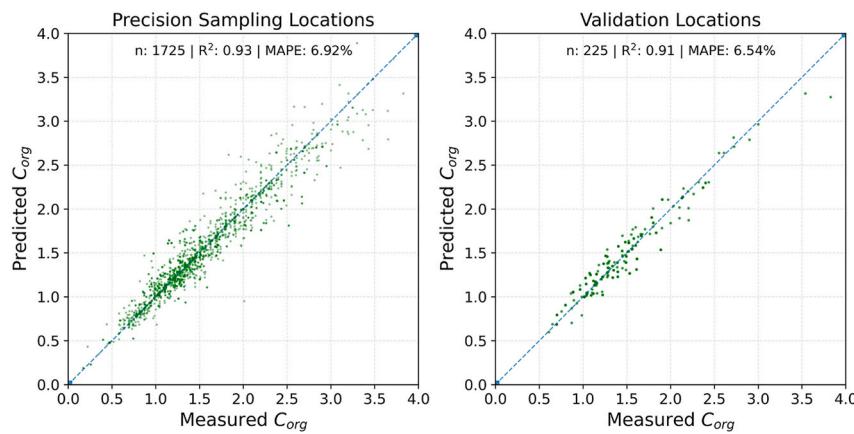


Fig. 8. Correlation between measured and predicted SOC values on the validation data with refinement.

results at a spatial resolution of 10 m further enhances the utility of the refined SOC map for detailed carbon sequestration assessments and land management decisions while having the field sampling cost as a financial drawback of the proposed approach.

4.2.4. Validation with precision sampling on densely sampled fields

This validation section evaluated the SOC prediction model on four densely sampled fields presented in Fig. 9, one in Poland (a), one in Switzerland (b), and two in France (c&d). This evaluation allows for a higher-resolution assessment of the model's performance and provides insights into the spatial variability of SOC within the fields.

Fig. 9(a) presents the results for the field located in Poland, which is cultivated by Top Farms "Głubczyce", having a total area of 127 ha. The data was made available by Polish companies: Top Farms "Głubczyce" sp. z o. o. and QZ-Solutions sp. z o.o. The left sub-figure shows the grid sampling results, where the field was sampled at a grid size of 25 m. The grid sampling results are visualized as a map rasterized to a resolution of 10x10 m. It is important to note a gap in the sampling coverage and a misreported single-sampled value in the southern part of the field. The middle sub-figure displays the model prediction refined with the Precision Sampling approach and the corresponding sampling locations. The Precision Sampling locations are spread across the field, which provides representative ground measurements for the refinement process. The right sub-figure presents a scatter plot at the pixel level, comparing the ground measurements with the corresponding predictions from the refined SOC map. Each point in the scatter plot represents a pixel in the field where both the ground measurement and the predicted SOC value are available. By examining the scatter plot, we can assess the accuracy of the refined SOC map in capturing the variability of the ground measurements.

The scatter plot indicates an R^2 value of 0.52 and a MAPE of 10.6 % for the refined SOC map in capturing the variability of the ground measurements. The R^2 value of 0.52 suggests a moderate correlation between the predicted SOC values and the ground truth measurements. The MAPE value of 10.6 % indicates the average difference between the predicted and actual SOC values at the pixel level. It is worth noting that the calculated MAPE in this densely sampled field is slightly higher than the one reported in Section 4.2.3 by a difference of 4 %. This discrepancy may arise due to the field's specific characteristics, spatial variability, and the density and distribution of the ground measurements in the sampling process.

We can observe discrete ground measurements in the scatter plot shown in the right sub-plot of Fig. 9(a). The discrete nature of the ground measurements is because the SOC values provided by the laboratory reports are rounded to one decimal point. Unfortunately, without having access to the actual measurements with two decimal points, which corresponds to a change up to 10 % of the reported value, it is impossible

to determine whether the accuracy of the model would have been higher or lower in this specific validation case. The discrete nature of the ground measurements limits our ability to assess the model's performance at a finer scale and evaluate its agreement with the predicted values.

The second field is located in Switzerland and was sampled within the STEROPES project, which Agroscope leads. Fig. 9(b) shows the results for the corresponding field, having a total area of 4.2 ha. The left sub-figure shows the sampling locations with a color coding for the measured SOC value. The middle sub-figure displays the model prediction refined with the Precision Sampling approach and the corresponding sampling locations. The right sub-figure presents a scatter plot at the pixel level, comparing the ground measurements with the related predictions from the refined SOC map. Each point in the scatter plot represents a pixel in the field where both the ground measurement and the predicted SOC value are available. By examining the scatter plot, we can assess the accuracy of the refined SOC map in capturing the variability of the ground measurements.

The scatter plot indicates an R^2 value of 0.76 and a MAPE of 2.9 % for the refined SOC map in capturing the variability of the ground measurements. The R^2 value of 0.76 suggests a high correlation between the predicted SOC values and the ground truth measurements. The MAPE value of 2.9 % indicates the average difference between the predicted and actual SOC values at the pixel level. This finding of 2.9 % is higher than the one calculated in validation with adjustment (6.5 %) and the field located in Poland (10.6 %), possibly due to its smaller size.

The third and fourth fields are located north of Paris, France. The data was made available by Sandricourt Estate. Fig. 9(c) and (d) show the results for the corresponding fields, having an area of 7.5 and 14 ha, respectively. The scatter plot indicates R^2 values of 0.95 and 0.85 and MAPE values of 7.4 % and 6.4 % for the refined SOC maps in capturing the variability of the ground measurements. The R^2 values greater than 0.85 suggest a strong correlation between the predicted SOC values and the ground truth measurements. These MAPE findings of around 7 % align with the MAPE values calculated in validation with adjustment (6.5 %) and less than the field located in Poland (10.6 %), possibly due to their smaller sizes.

4.3. Practical implications

The assessment of carbon sequestration is essential for the carbon market, where farmers can be rewarded for the SOC that they sequester. Monitoring requires calculating SOC stock changes over time by considering the accrual or loss of carbon from the soil. Two main methodologies exist for this purpose, namely measure-remesure and carbon-balance models. The measure-remesure approach requires multiple measurements in time and has the highest assessment accuracy

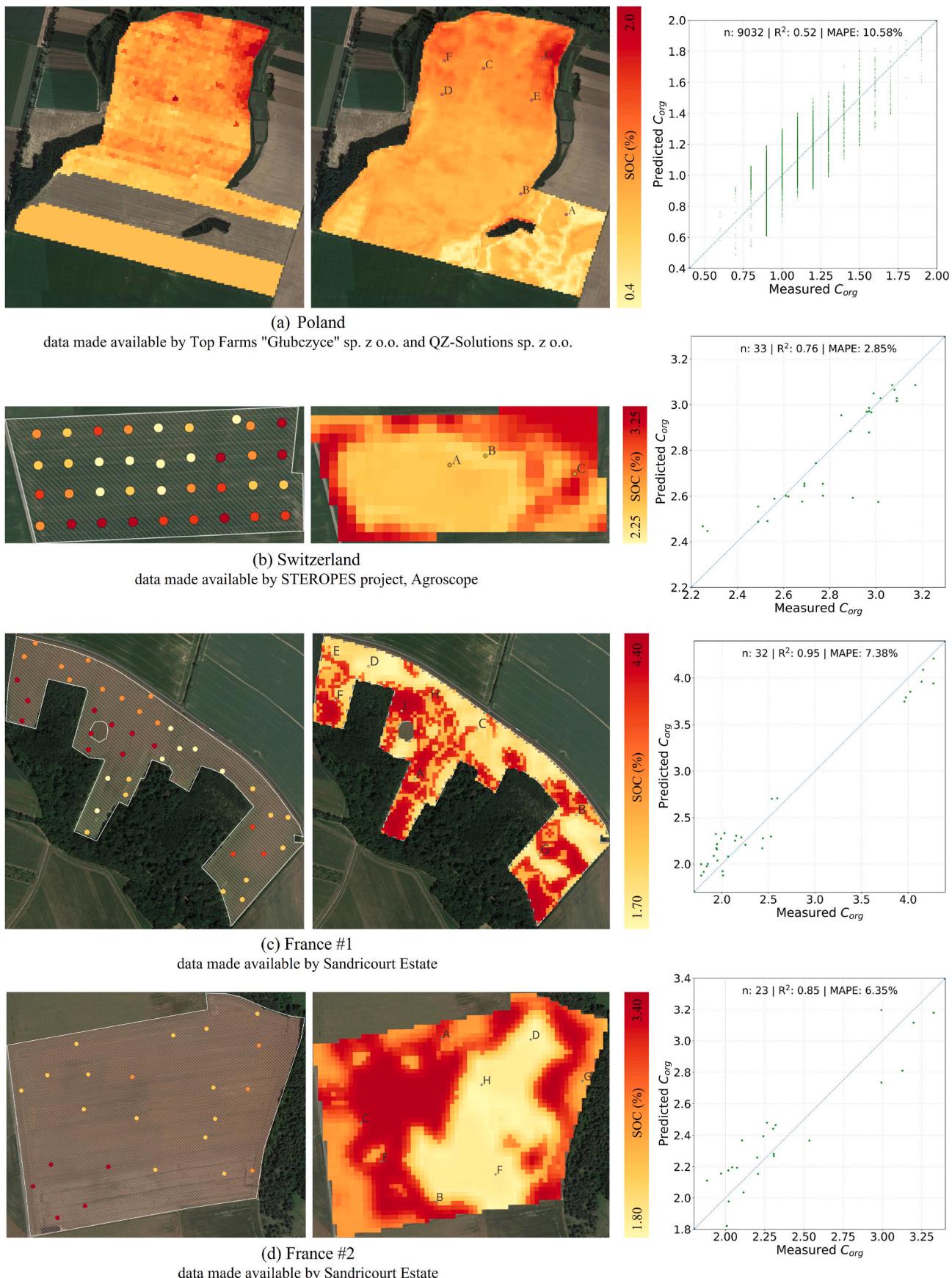


Fig. 9. Validation SOC data collected from densely sampled fields located in Poland, Switzerland, and France. In each sub-figure, the order of the visuals from left to right are measurement map, predicted map, and validation plot, respectively.

as it relies on the ground measurements directly. On the other hand, carbon-balance models require management practices and some inputs such as soil type, climate, and crop information. At higher levels, they use temporal trends of indices such as leaf area index for the forecast of SOC stock. The models, such as LPJ-Guess (Smith, 2001) and RothC (Coleman and Jenkinson, 1996), are proven to make reliable estimations by simulating the factors in an ecosystem within given boundaries if calibrated correctly. The downside of such models is that they require many inputs to forecast the SOC stock, which leads to a high degree of uncertainty. At the same time, it is costly to evaluate them at a fine spatial resolution.

For an accurate SOC sequestration assessment, the method that provides the lowest uncertainty is the measure-remesure approach, which aligns perfectly with this study's ground measurement supported SOC mapping approach. The high-resolution SOC maps generated through the presented framework can serve as a baseline and status map needed for a measure-remesure method. SOC stock map calculation requires SOC content, bulk density, and stone content, where SOC can be calculated by the presented approach. A similar model can be developed for the bulk density, and generated maps can be improved using ground measurements at the Precision Sampling locations. Of course, fast and largely automatized measurement of soil density along with the soil sampling is a prerequisite for this. Subsequently, the SOC content and soil density maps can be combined to derive SOC stocks and related maps. To implement the measure-remesure approach, the initial step would be to create a baseline SOC stock map using the model and refinement via Precision Sampling. This map represents the spatial distribution of SOC stock within the agricultural field at a specific point in time. After a certain time interval, a second measurement of SOC stock content is taken across the same field at the exact locations. This measurement captures the current SOC stock status of the soil. Using the same set of features and the trained model, a new SOC stock map represents the current spatial distribution of SOC stock. By comparing the baseline and status SOC stock maps, the changes in SOC stock can be calculated for different areas within the agricultural field. This information is crucial for certification and monitoring purposes. Integrating the proposed SOC prediction framework into the measure-remesure method offers several advantages. It provides a cost-effective and efficient way to estimate the spatial variation of SOC at a high resolution, allowing for more accurate assessments of SOC stock changes.

The validation of the measure-remesure method relies on accurate quantification of the uncertainty ensuring conservative accounting and minimizing the risk of overestimation. The uncertainty information accounts for various sources of uncertainty, including measurement errors, sampling variability, and model error. These uncertainties are crucial for calculating the discount factor in carbon crediting schemes. The discount factor is typically calculated using statistical methods to quantify the confidence intervals (CI), standard errors, or other statistical measures of uncertainty associated with the SOC stock measurements. The discount factor is then applied to adjust the credited SOC stock, reducing it by a certain percentage to reflect the level of uncertainty. **Table 9** shows the summary of validation datasets used in this research with 50 %, 70 %, 80 %, and 90 % CI.

The agreement between the cases with and without refinement using ground measurements via Precision Sampling, as shown in **Table 9**, is indeed an important finding. It indicates that including ground measurements through Precision Sampling significantly reduces the error and uncertainty associated with the SOC stock estimation. Comparing cases (b) and (c) for the 90 % confidence interval (CI), which is commonly used in carbon crediting programs, we observe a substantial reduction in error by 35.3 % when ground measurements are incorporated. This highlights the value of integrating ground measurements into the modeling process, leading to more accurate and reliable estimates of SOC.

Furthermore, the strong similarity between the 90 % CI values in cases (c) and (d), where Precision Sampling was applied, provides

Table 9

MAPE values of different CIs based on the validation datasets presented in this research.

Dataset	Sample Size	50 % CI	70 % CI	80 % CI	90 % CI
(a) Internal Validation	7325	31.5	38.5	49.4	71.2
(b) Validation without Precision Sampling	1725	21.0	40.3	48.0	56.0
(c) Validation with Precision Sampling	225	6.0	11.5	16.4	20.7
(d) Validation Dense Sampling: Poland	9032	10.6	14.1	17.6	21.7
(e) Validation Dense Sampling: Switzerland	33	2.6	3.1	3.9	7.8
(f) Validation Dense Sampling: France #1	32	7.4	8.7	10.9	13.6
(g) Validation Dense Sampling: France #2	23	6.4	8.2	9.3	10.1

statistical support for the validity of the proposed model and the effectiveness of Precision Sampling. This suggests that the model and refinement process are robust and consistent, as the results remain stable even with different datasets. This has significant implications for carbon crediting programs, as it ensures more accurate accounting and better quantifying carbon sequestration efforts.

5. Conclusion

This research focuses on developing and implementing a novel algorithm for predicting soil organic carbon (SOC) content in European soils to assess field scale SOC changes. The algorithm utilizes satellite, climate, terrain, and soil texture data to generate accurate predictions of SOC levels. Satellite data, such as multispectral and radar imagery, provide information about soils' spectral properties and surface characteristics. This data is combined with climate data, including temperature and precipitation features, and terrain data, which includes elevation, slope, aspect, and hillshade information. Additionally, soil texture data, which describes the composition of the soil in terms of its sand, silt, and clay content, is incorporated into the algorithm. By integrating these diverse data sets, the algorithm can capture and analyze the complex relationships between environmental features and SOC content. The Light Gradient Boosting Machine algorithm is trained using a data set of measured SOC values from various European locations.

The outcomes of this study include a prediction model that can generate accurate estimates of SOC content at the European scale average percentage error below 10 %. In addition, the algorithm has been validated using independent data sets and has demonstrated reliable performance in predicting SOC levels across different soil types and geographic regions in Europe with and without the refinement using the ground measurements.

Ultimately, the developed framework allows the estimation of SOC using a model with a relatively low number of features and ground measurements gathered via precision sampling. Including ground measurements by support sampling enhanced the overall accuracy significantly. This provides predictions with low error, which had significant implications for soil management, land use planning, and environmental research. Accurate predictions of SOC content enable better decision-making regarding soil health, carbon sequestration, and sustainable land management practices. By understanding the spatial distribution of SOC across agricultural fields, stakeholders can make informed choices to optimize agricultural practices, enhance carbon sequestration efforts, and effectively mitigate climate change impacts.

In the subsequent phase of this research, our focus will shift towards extending the predictive capabilities to soil texture mapping, employing a comparable and robust approach. Predicting soil texture at large scale has got lower errors than for predicting SOC content (Gebauer et al.,

2022). This expansion holds the potential to support agricultural management strategies and enhance the precision of soil moisture mapping. By seamlessly integrating these additional dimensions, our ongoing efforts aim to provide a comprehensive toolkit for stakeholders in agriculture and environmental factoring, facilitating better and more nuanced decision-making and contributing to the sustainable and efficient management of agricultural landscapes. URL <https://www.sciencedirect.com/science/article/pii/S0016706112002133>.

CRediT authorship contribution statement

Onur Yuzugullu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Noura Fajraoui:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Axel Don:** Data curation, Resources, Supervision, Writing – original draft, Writing – review & editing. **Frank Liebisch:** Conceptualization, Data curation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgment

The authors would like to thank Rodrigo Principe, Witra Tahjudil, and Utku Berkalp Unalan for setting up the Google Earth Engine framework to download the satellite data and Rene Dechow for his help downloading model data for the OpenAgrar measurements and Trupthi Narayan and Peter Fröhlich for their valuable comments. Frank Liebisch acknowledges funding support from the EJP Soil Project STEROPES – Stimulating Novel Technologies from Earth Remote Observation to Map European Soil Organic Carbon. We also acknowledge Sandricourt Estate for the fields in France and Top Farms “Głubczyce” sp. z o. o. and QZ-Solutions sp. z o.o. for providing permission to use the data from the project of detailed soil research in southern Poland - an experimental field cultivated by Top Farms “Głubczyce”.

References

- Adams, W., 1973. The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils. *J. Soil Sci.* 24 (1), 10–17.
- Aksøy, E., Yigini, Y., Montanarella, L., 2016. Combining soil databases for topsoil organic carbon mapping in Europe. *PLoS One* 11 (3), e0152098. <https://doi.org/10.1371/journal.pone.0152098>.
- Andries, A., Morse, S., Murphy, R.J., Lynch, J., Mota, B., Woolliams, E.R., 2021. Can current earth observation technologies provide useful information on soil organic carbon stocks for environmental land management policy? *Sustainability* 13 (21), 12074.
- Ayoubi, S., Shahri, A.P., Karchegani, P.M., Sahrawat, K.L., 2011. Application of Artificial Neural Network (Ann) to Predict Soil Organic Matter Using Remote Sensing Data in Two Ecosystems. *Biomass and remote sensing of biomass*, pp. 181–196.
- Badraghi, A., Ventura, M., Polo, A., Borruso, L., Giannamarchi, F., Montagnani, L., 2021. Soil respiration variation along an altitudinal gradient in the Italian Alps: disentangling forest structure and temperature effects. *PLoS One* 16 (8), e0247893. <https://doi.org/10.1371/journal.pone.0247893>.
- Basnyat, P., McConkey, B., Meinert, B., Gatkze, C., Noble, G., 2004. Agriculture field characterization using aerial photograph and satellite imagery. *Geosci. Rem. Sens. Lett. IEEE* 1 (1), 7–10.
- Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. *Eur. J. Soil Sci.* 47 (2), 151–163.
- Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. In: *Noise Reduction in Speech Processing*. Springer, Berlin, Germany, pp. 1–4. https://doi.org/10.1007/978-3-642-00296-0_5.
- Bergstra, J., Yamins, D., Cox, D., 2013. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: Dasgupta, S., McAllester, D. (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, Vol. 28 of *Proceedings of Machine Learning Research*. PMLR, Atlanta, Georgia, USA, pp. 115–123. In: https://proceedings.mlr.press/v28/bergstr_a13.html.
- Bookstein, F.L., 1989. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* 11 (6), 567–585.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32.
- Cannell, M., Milne, R., Hargreaves, K., Brown, T., Cruickshank, M., Bradley, R., Spencer, T., Hope, D., Billett, M., Adger, W., et al., 1999. National inventories of terrestrial carbon sources and sinks: the UK experience. *Climatic Change* 42 (3), 505–530.
- Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S., Casa, R., 2016. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. *Rem. Sens. Environ.* 179, 54–65.
- Castaldi, F., Chabriat, S., Jones, A., Vreys, K., Boman, B., Van Wesemael, B., 2018. Soil organic carbon estimation in croplands by hyperspectral remote apex data using the lucas topsoil database. *Rem. Sens.* 10 (2), 153.
- Castaldi, F., Hueni, A., Chabriat, S., Ward, K., Buttafuoco, G., Boman, B., Vreys, K., Brell, M., van Wesemael, B., 2019. Evaluating the capability of the sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS J. Photogrammetry Remote Sens.* 147, 267–282.
- Castaldi, F., Koparan, M.H., Wetterlind, J., Žydelis, R., Vinci, I., Savaş, A.Ö., Kivrak, C., Tunçay, T., Volungevičius, J., Obber, S., et al., 2023. Assessing the capability of sentinel-2 time-series to estimate soil organic carbon and clay content at local scale in croplands. *ISPRS J. Photogrammetry Remote Sens.* 199, 40–60.
- Coleman, K., Jenkinson, D., 1996. Rothc-26.3-a model for the turnover of carbon in soil. In: *Evaluation of Soil Organic Matter Models: Using Existing Long-Term Datasets*. Springer, pp. 237–246.
- De Vos, B., Van Meirvenne, M., Quaerat, P., Deckers, J., Muys, B., 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. *Soil Sci. Soc. Am. J.* 69 (2), 500–510.
- Dieleman, W.I.J., Venter, M., Ramachandra, A., Krockenberger, A.K., Bird, M.I., 2013. Soil carbon stocks vary predictably with altitude in tropical forests: implications for soil carbon storage. *Geoderma* 204–205, 59–67. <https://doi.org/10.1016/j.geoderma.2013.04.005>.
- Dvorakova, K., Shi, P., Limbourg, Q., van Wesemael, B., 2020. Soil organic carbon mapping from remote sensing: the effect of crop residues. *Rem. Sens.* 12 (12), 1913.
- Dvorakova, K., Heiden, U., Peper, K., Staats, G., van Os, G., van Wesemael, B., 2023. Improving soil organic carbon predictions from a sentinel-2 soil composite by assessing surface conditions and uncertainties. *Geoderma* 429, 116128.
- Dvorakova, K., Vasenev, V., Romzaykina, O., Grigorieva, V., Litvinov, Y., Gorbov, S., Dolgikh, A., Korneykova, M., Gosse, D., 2021. Projecting the urbanization effect on soil organic carbon stocks in polar and steppe areas of European Russia by remote sensing. *Geoderma* 399, 115039.
- Falahatkar, S., Hosseini, S.M., Salman Mahiny, A., Ayoubi, S., Wang, S.-q., 2014. Soil organic carbon stock as affected by land use/cover changes in the humid region of northern Iran. *J. Mt. Sci.* 11 (2), 507–518. <https://doi.org/10.1007/s11629-013-2645-1>.
- Frazier, B., Walters, C., Perry, E., 1997. Role of Remote Sensing in Site-specific Management, the State of Site Specific Management for Agriculture, pp. 149–160.
- Gardner, M.W., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* 32 (14–15), 2627–2636.
- Gebauer, A., Sakhaei, A., Don, A., Poggio, M., Ließ, M., 2022. Topsoil texture regionalization for agricultural soils in Germany—an iterative approach to advance model interpretation. *Frontiers in Soil Science* 1, 770326. <https://doi.org/10.3389/fsoil.2021.770326>.
- Gholizadeh, A., Zizala, D., Saberioon, M., Boruvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and sentinel-2 spectral imaging. *Rem. Sens. Environ.* 218, 89–103.
- Goids, E., Van Wesemael, B., Crucifix, M., 2009. Magnitude and sources of uncertainties in soil organic carbon (soc) stock assessments at various scales. *Eur. J. Soil Sci.* 60 (5), 723–739.
- Guo, L., Sun, X., Fu, P., Shi, T., Dang, L., Chen, Y., Linderman, M., Zhang, G., Zhang, Y., Jiang, Q., et al., 2021. Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas. *Geoderma* 398, 115118.
- Gupta, S., Larson, W., 1979. Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density. *Water Resour. Res.* 15 (6), 1633–1635.
- He, C., Wang, S., Xu, J., Zhou, C., 2002. Using remote sensing to estimate the change of carbon storage: a case study in the estuary of yellow river delta. *Int. J. Rem. Sens.* 23 (8), 1565–1580.
- He, X., Yang, L., Li, A., Zhang, L., Shen, F., Cai, Y., Zhou, C., 2021. Soil organic carbon prediction using phenological parameters and remote sensing variables generated from sentinel-2 images. *Catena* 205, 105442.
- Hoyle, F.C., O’Leary, R.A., Murphy, D.V., 2016. Spatially governed climate factors dominate management in determining the quantity and distribution of soil organic carbon in dryland agricultural systems. *Sci. Rep.* 6 (31468), 1–12. <https://doi.org/10.1038/srep31468>.

- Huang, B., Sun, W., Zhao, Y., Zhi, J., Yang, R., Zou, Z., Ding, F., Su, J., 2007. Temporal and spatial variability of soil organic matter and total nitrogen in an agricultural ecosystem as affected by farming practices. *Geoderma* 139 (3–4), 336–345.
- Incio, 2022. ITACyL portal web. <https://www.itacyl.es>. (Accessed 22 December 2022).
- ISMN, 2022. <https://ismn.earth/en>. (Accessed 22 December 2022).
- ISRIC soil data hub. <https://www.isric.org/explore/isric-soil-data-hub>, 2022–. (Accessed 12 December 2022).
- Jaber, S.M., 2006. Monitoring Spatial Variations in Soil Organic Carbon Using Remote Sensing and Geographic Information Systems. Southern Illinois University at Carbondale.
- Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M.F., Bampa, F., van Wesemael, B., Harrison, R.B., Guerrini, I.A., Richter, D.D., Rustad, L., Lorenz, K., Chabbi, A., Miglietta, F., 2014. Current status, uncertainty and future needs in soil organic carbon monitoring. *Sci. Total Environ.* 468–469, 376–383. <https://doi.org/10.1016/j.scitotenv.2013.08.026>.
- Jiang, J., Zong, N., Song, M., Shi, P., Ma, W., Fu, G., Shen, Z., Zhang, X., Ouyang, H., 2013. Responses of ecosystem respiration and its components to fertilization in an alpine meadow on the Tibetan Plateau. *Eur. J. Soil Biol.* 56, 101–106. <https://doi.org/10.1016/j.ejsobi.2013.03.001>.
- Karchegani, P.M., Ayoubi, S., Mosaddeghi, M.R., Honarjoo, N., 2012. Soil organic carbon pools in particle-size fractions as affected by slope gradient and land use change in hilly regions, western Iran. *J. Mt. Sci.* 9 (1), 87–95. <https://doi.org/10.1007/s11629-012-2211-2>.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* 30.
- Kim, J., Grunwald, S., 2016. Assessment of carbon stocks in the topsoil using random forest and remote sensing images. *J. Environ. Qual.* 45 (6), 1910–1918.
- Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. *Geoderma* 189–190, 627–634. <https://doi.org/10.1016/j.geoderma.2012.05.022>.
- Kumar, N., Velmurugan, A., Hamm, N.A., Dadhwal, V.K., 2018. Geospatial mapping of soil organic carbon using regression kriging and remote sensing. *Journal of the Indian Society of Remote Sensing* 46 (5), 705–716.
- Ladoni, M., Bahrami, H.A., Alavipanah, S.K., Norouzi, A.A., 2010. Estimating soil organic carbon from soil reflectance: a review. *Precis. Agric.* 11 (1), 82–99.
- Lal, R., 2020. Soil organic matter and water retention. *Agron. J.* 112 (5), 3265–3277.
- Levine, E.R., Kurz, L., Smid, J., Smid, M., Wolf, P., 1998. Algorithms and analysis tools for carbon content modeling in soil based on satellite data. In: *Remote Sensing for Agriculture, Ecosystems, and Hydrology*, vol. 3499. SPIE, pp. 315–322.
- Liu, J., Zhao, S., Jiang, L., Chai, L., Wu, F., 2013. The influence of organic matter on soil dielectric constant at microwave frequencies (0.5–40 ghz). In: *2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS*. IEEE, pp. 13–16.
- Loveland, P., Webb, J., 2003. Is there a critical level of organic matter in the agricultural soils of temperate regions: a review. *Soil Tillage Res.* 70 (1), 1–18. [https://doi.org/10.1016/S0167-1987\(02\)00139-3](https://doi.org/10.1016/S0167-1987(02)00139-3).
- LUCAS - ESDAC, 2020. European commission. <https://esdac.jrc.ec.europa.eu/projects/lucas>. (Accessed 12 December 2022).
- Marianne Stokar, N., 2022. Startseite - NABODAT. <https://www.nabodat.ch/index.php/de>. (Accessed 12 December 2022).
- Meng, X., Bao, Y., Wang, Y., Zhang, X., Liu, H., 2022. An advanced soil organic carbon content prediction model via fused temporal-spatial-spectral (tss) information based on machine learning and deep learning algorithms. *Rem. Sens. Environ.* 280, 113166.
- Merry, C.J., Levine, E.R., 1995. Methods to assess soil carbon using remote sensing techniques. *Soils and Global Change* 5, 250–265.
- Mohamed, E., Saleh, A., Belal, A., Gad, A., 2018. Application of near-infrared reflectance for quantitative assessment of soil properties. *The Egyptian Journal of Remote Sensing and Space Science* 21 (1), 1–14.
- Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., Mukhopadhyay, A., 2017. Spatial soil organic carbon (soc) prediction by regression kriging using remote sensing data. *The Egyptian Journal of Remote Sensing and Space Science* 20 (1), 61–70.
- Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang, N. B., Guo, W., Ngo, H.H., 2022. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and sar data fusion. *Sci. Total Environ.* 804, 150187.
- Noshadi, E., Bahrami, H., Alavipanah, S., 2014. Prediction of soil organic carbon content using field measurements and remotely sensing data. *Casp. J. Appl. Sci. Res.* 3 (7).
- Nunes, M.R., Veum, K.S., Parker, P.A., Holan, S.H., Karlen, D.L., Amsili, J.P., van Es, H. M., Wills, S.A., Seybold, C.A., Moorman, T.B., 2021. The soil health assessment protocol and evaluation applied to soil organic carbon. *Soil Sci. Soc. Am. J.* 85 (4), 1196–1213.
- Nyssen, J., Temesgen, H., Lemenih, M., Zenebe, A., Haregeweyn, N., Haile, M., 2008. Spatial and temporal variation of soil organic carbon stocks in a lake retreat area of the ethiopian rift valley. *Geoderma* 146 (1–2), 261–268.
- Odebiri, O., Mutanga, O., Odindi, J., Naicker, R., 2022a. Modelling soil organic carbon stock distribution across different land-uses in South Africa: a remote sensing and deep learning approach. *ISPRS J. Photogrammetry Remote Sens.* 188, 351–362.
- Odebiri, O., Mutanga, O., Odindi, J., Naicker, R., 2022b. Modelling soil organic carbon stock distribution across different land-uses in South Africa: a remote sensing and deep learning approach. *ISPRS J. Photogrammetry Remote Sens.* 188, 351–362.
- Padarian, J., Stockmann, U., Minasny, B., McBratney, A., 2022. Monitoring changes in global soil organic carbon stocks from space. *Rem. Sens. Environ.* 281, 113260.
- Paul, S., Coops, N., Johnson, M., Krzic, M., Chandna, A., Smukler, S., 2020. Mapping soil organic carbon and clay using remote sensing to predict soil workability for enhanced climate change adaptation. *Geoderma* 363, 114177.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., Greve, M.H., 2015. Modeling soil organic carbon at regional scale by combining multi-spectral images with laboratory spectra. *PLoS One* 10 (11), e0142295.
- Poeplau, C., Don, A., Flessa, H., Heidkamp, A., Jacobs, A., Prietz, R., 2020a. Erste Bodenzustandserhebung Landwirtschaft – Kerndatensatz. <https://doi.org/10.3220/DATA20200203151139>. (Accessed 12 December 2022).
- Poeplau, C., Jacobs, A., Don, A., Vos, C., Schneider, F., Wittnebel, M., Tiemeyer, B., Heidkamp, A., Prietz, R., Flessa, H., 2020b. Stocks of organic carbon in German agricultural soils—key results of the first comprehensive inventory. *J. Plant Nutr. Soil Sci.* 183 (6), 665–681. <https://doi.org/10.1002/jpln.2020000113>.
- Poeplau, C., Prietz, R., Don, A., 2022. Plot-scale variability of organic carbon in temperate agricultural soils—implications for soil monitoring. *J. Plant Nutr. Soil Sci.* 185 (3), 403–416.
- Rawls, W., Pachepsky, Y.A., Ritchie, J., Sobecki, T., Bloodworth, H., 2003. Effect of soil organic carbon on soil water retention. *Geoderma* 116 (1–2), 61–76.
- Reeves, D., 1997. The role of soil organic matter in maintaining soil quality in continuous cropping systems. *Soil Tillage Res.* 43 (1–2), 131–167.
- Rochette, P., Desjardins, R.L., Pattey, E., 2011. Spatial and temporal variability of soil respiration in agricultural fields. *Can. J. Soil Sci. (Mar.)*. <https://doi.org/10.4141/cjss91-018>.
- Sakin, E., 2012. Organic carbon organic matter and bulk density relationships in arid-semi arid soils in southeast anatolia region. *Afr. J. Biotechnol.* 11 (6), 1373–1377.
- Shafizadeh-Moghadam, H., Minaei, F., Talebi-khiavi, H., Xu, T., Homae, M., 2022. Synergetic use of multi-temporal sentinel-1, sentinel-2, ndvi, and topographic factors for estimating soil organic carbon. *Catena* 212, 106077.
- Smith, B., 2001. Lpj-guess-an Ecosystem Modelling Framework, Department of Physical Geography and Ecosystems Analysis, INES, Sövlgatan, vol. 12, 22362.
- Szakács, G., Eschenbrenner, V., Cerri, C., Bernoux, M., 2004. Soil carbon stocks under pastures in the brazilian cerrado region their assessment by orbital remote sensing. *Proc. ISPRS Congr.*, XXth, Istanbul, Turkey XXXV, 12–23.
- Tajik, S., Ayoubi, S., Zeraatpisheh, M., 2020. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. *Geoderma Regional* 20, e00256. <https://doi.org/10.1016/j.geodrs.2020.e00256>.
- Tiessen, H., Cuevas, E., Chacon, P., 1994. The role of soil organic matter in sustaining soil fertility. *Nature* 371, 783–785.
- Tsui, C.-C., Tsai, C.-C., Chen, Z.-S., 2013. Soil organic carbon stocks in relation to elevation gradients in volcanic ash soils of Taiwan. *Geoderma* 209–210, 119–127. <https://doi.org/10.1016/j.geoderma.2013.06.013>.
- Vägen, T.-G., Winowiecki, L.A., Abegaz, A., Hadgu, K.M., 2013. Landsat-based approaches for mapping of land degradation prevalence and soil functional properties in Ethiopia. *Rem. Sens. Environ.* 134, 266–275.
- VandenBygaart, A., 2006. Monitoring soil organic carbon stock changes in agricultural landscapes: issues and a proposed approach. *Can. J. Soil Sci.* 86 (3), 451–463.
- Vasenev, V., Stoer vogel, J., Vasenev, I., 2013. Urban soil organic carbon and its spatial heterogeneity in comparison with natural and agricultural areas in the moscow region. *Catena* 107, 96–102.
- Vos, C., Don, A., Hobley, E.U., Prietz, R., Heidkamp, A., Freibauer, A., 2019. Factors controlling the variation in organic carbon stocks in agricultural soils of Germany. *Eur. J. Soil Sci.* 70 (3), 550–564.
- Wang, Y., Qi, Y., Chen, Y., Xie, F., et al., 2016. Prediction of soil organic matter based on multi-resolution remote sensing data and random forest algorithm. *Acta Pedol. Sin.* 53 (2), 342–354.
- Wang, L., Wang, X., Wang, D., Qi, B., Zheng, S., Liu, H., Luo, C., Li, H., Meng, L., Meng, X., et al., 2021a. Spatiotemporal changes and driving factors of cultivated soil organic carbon in northern China's typical agro-pastoral ecotone in the last 30 years. *Rem. Sens.* 13 (18), 3607.
- Wang, H., Zhang, X., Wu, W., Liu, H., 2021b. Prediction of soil organic carbon under different land use types using sentinel-1/-2 data in a small watershed. *Rem. Sens.* 13 (7), 1229.
- Weather API, 2022. Meteomatics. <https://www.meteomatics.com/en/weather-api>. (Accessed 13 December 2022).
- Weil, R.R., Magdoff, F., et al., 2004. Significance of soil organic matter to soil quality and health. *Soil organic matter in sustainable agriculture* 1–43.
- Wiesmeier, M., Hübner, R., Barthold, F., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., Schilling, B., von Lützow, M., Kögel-Knabner, I., 2013. Amount, distribution and driving factors of soil organic carbon and nitrogen in cropland and grassland soils of southeast Germany (Bavaria). *Agric. Ecosyst. Environ.* 176, 39–52. <https://doi.org/10.1016/j.agee.2013.05.012>.
- Wilcox, C.H., Frazier, B.E., Ball, S.T., 1994. Relationship between soil organic carbon and landsat tm data in eastern Washington. *Photogramm. Eng. Rem. Sens.* 60 (6), 777–781.
- Ye, Z., Sheng, Z., Liu, X., Ma, Y., Wang, R., Ding, S., Liu, M., Li, Z., Wang, Q., 2021. Using machine learning algorithms based on gf-6 and google earth engine to predict and map the spatial distribution of soil organic matter content. *Sustainability* 13 (24), 14055.
- Yuzugullu, O., Lorenz, F., Fröhlich, P., Liebisch, F., 2020. Understanding fields by remote sensing: soil zoning and property mapping. *Rem. Sens.* 12 (7), 1116.
- Zeraatpisheh, M., Ayoubi, S., Mirbagheri, Z., Mosaddeghi, M.R., Xu, M., 2021. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions

- using machine learning algorithms and environmental variables. *Geoderma Regional* 27, e00440. <https://doi.org/10.1016/j.geodrs.2021.e00440>.
- Zeraatpisheh, M., Garosi, Y., Reza Owliaie, H., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., Xu, M., 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: a comparison of a group of environmental covariates. *Catena* 208, 105723. <https://doi.org/10.1016/j.catena.2021.105723>.
- Zhang, L., Cai, Y., Huang, H., Li, A., Yang, L., Zhou, C., 2022. A cnn-lstm model for soil organic carbon content prediction with long time series of modis-based phenological variables. *Rem. Sens.* 14 (18), 4441.
- Zhao, W., Wu, Z., Yin, Z., Li, D., 2022. Attention-based cnn ensemble for soil organic carbon content estimation with spectral data. *Geosci. Rem. Sens. Lett. IEEE* 19, 1–5.
- Zhou, Y., Zhao, X., Guo, X., Li, Y., 2022. Mapping of soil organic carbon using machine learning models: combination of optical and radar remote sensing data. *Soil Sci. Soc. Am. J.* 86 (2), 293–310.