https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Earth System Science Data Discussions
Open Access

Soil organic carbon maps and associated uncertainty at 90 m for peninsular Spain

Durante, Pilar*[1]; Requena-Mullor, Juan Miguel*[2,3]; Vargas, Rodrigo[4]; Guevara, Mario[5]; Alcaraz-Segura, Domingo[6]; Oyonarte, Cecilio[3,7]

(1) Agresta Sociedad Cooperativa, 28012 Madrid, Spain
(2) Departamento de Biología Vegetal y Ecología, University of Almería, 04120 Almería, Spain
(3) Centro Andaluz para el Cambio Global - Hermelindo Castro (ENGLOBA), University of Almería, 04120 Almería, Spain
(4) Department of Plant & Soil Sciences, University of Delaware, Newark, DE, USA 19716
(5) National Autonomous University of Mexico, 04510 Ciudad de México, México
(6) Departamento de Botánica, and Instituto Interuniversitario para la Investigación del Sistema Tierra en Andalucía (IISTA), Universidad de Granada, 18006 Granada, Spain
(7) Departamento de Agronomía, University of Almería, 04120 Almería, Spain

Correspondence to: Juan M. Requena-Mullor (juanmir@ual.es), Pilar Durante (pdurante@agresta.org)

Abstract

Human activities have significantly disrupted the global carbon cycle, leading to increased atmospheric CO_2 levels and altering ecosystems' carbon absorption capacities, with soils serving as the largest carbon reservoirs in terrestrial ecosystems. The complexity and variability of soil properties, shaped by long-term transformations, make it crucial to study these properties at various spatial and temporal scales to develop effective climate change mitigation strategies. However, integrating disparate soil databases presents challenges due to the lack of standardized protocols, necessitating collaborative efforts to standardize data collection and processing to improve the reliability of Soil Organic Carbon (SOC) estimates. This issue is particularly relevant in peninsular Spain, where variations in sampling protocols and calculation methods have resulted in significant discrepancies in SOC concentration and stock estimates. This study aimed to improve the understanding of SOC storage and distribution in peninsular Spain by focusing on two specific goals: integrating and standardizing existing soil profile databases, and modeling SOC concentrations (SOCc) and stocks (SOCs) at different depths using an ensemble machine-learning approach. The research produced four high-resolution SOC maps for peninsular Spain, detailing SOCc and SOCs at depths of 0-30 cm, 30-100 cm and the effective soil depth, along with associated uncertainties. These maps provide valuable data for national soil carbon management and contribute to compiling Spain's National Greenhouse Gas Emissions Inventory Report. Additionally, the findings support global initiatives like the Global Soil Organic Carbon Map, aligning with international efforts to improve soil carbon assessments. The soil organic carbon concentration (g/kg) maps for the 0-30 cm and 30-100 cm standard depths, along with the soil organic carbon stock (tC/ha) maps for the 0-30 cm standard depth and the effective soil depth, including their associated uncertainties, —all at a 90-meter pixel resolution— (SOCM90) are freely available at https://doi.org/10.6073/pasta/48edac6904eb1aff4c1223d970c050b4 (Durante et al., 2024).

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

1. Introduction

Human activities have profoundly disrupted the global carbon cycle, leading to a significant increase in atmospheric CO_2 levels and a consequential alteration of ecosystems' inherent carbon absorption capacity (Crowther et al., 2016). Soils emerge as the primary carbon reservoirs within terrestrial ecosystems due to their remarkable capacity to store approximately two to four times more organic carbon in the top meter (i.e., approximately 1500 Pg C; (Scharlemann et al., 2014). Furthermore, the residence time of soil organic carbon (SOC) exceeds that of aboveground biomass, underscoring the crucial role of SOC storage in climate change mitigation strategies (Jobbágy et al., 2000; Saatchi et al., 2007).

Understanding the biogeochemical properties of soils is crucial for assessing and tracking changes in SOC storage capacity. Soils inherently undergo long-term transformations, with processes affecting SOC storage varying significantly over time and across different regions. This complexity and spatiotemporal heterogeneity necessitate studying the biogeochemical properties of soils at various spatial scales, from local fields to entire landscapes, and at different temporal scales, from short-term seasonal changes to long-term geological shifts (Stockmann et al., 2015). Consequently, numerous national and global initiatives have been launched to gather and standardize empirical data on soil properties accumulated over decades and make them readily available (Harden et al., 2018; Shangguan et al., 2014).

Integrating legacy and collaborative regional and global datasets of soil properties can pose challenges due to the absence of standardized protocols. For instance, in Spain, numerous organizations and institutions have collected soil profile data over fifty years, using different methods, laboratory techniques, standards, scales, and georeferencing systems (Llorente et al., 2018). Hence, this valuable information is currently scattered and fragmented, requiring substantial effort to integrate the historical soil databases into a cohesive, harmonized, and geographically well-defined dataset. Moreover, many of these databases lack the necessary information for accurately calculating Soil Organic Carbon stocks (SOCs), which depend on data such as SOC concentration (SOCc), bulk density, and coarse fragments (Calvo de Anta et al., 2020; Poeplau et al., 2017). This deficiency can lead to biased estimates of SOCs across different ecosystems. This challenge is also exacerbated by the considerable costs and operational complexities associated with soil data collection (Smith et al., 2020; Vargas et al., 2017). The absence of readily available databases containing consistent and comprehensive information on soil properties poses a significant challenge to effective soil monitoring across Spain and other regions of the world. The integration process is hindered by discrepancies in data formats, resolution, and quality, leading to potential inaccuracies and gaps in soil information. These challenges underscore the need for systematic methodologies and collaborative efforts to standardize data collection and processing protocols, thereby enhancing the reliability and usability of soil data for assessing SOCs and supporting climate change mitigation strategies.

The dynamics of SOC are determined by both soil physical and chemical properties, along with environmental soil-forming factors (Jenny, 1941). Soil properties exhibit diverse and complex patterns across scales due to the broad spatial and temporal range of soil formation conditions (Allen and Starr, 2019). To account for these complexities, SOC modeling has evolved from simple

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

qualitative approaches to sophisticated quantitative estimations and uncertainty assessment through models such as CLORPT or SCORPAN (Jenny, 1941; McBratney et al., 2003). Rooted in these theoretical models, digital soil mapping (DSM) has introduced a plethora of empirical models that estimate SOCs as a function of concurrent environmental factors (i.e., explanatory variables) (Chen et al., 2022), potentially reducing the number of in situ samples required for accurate spatial predictions (McBratney et al., 2003; Searle et al., 2021). These empirical models usually shows variability in their SOCs estimates, so ensemble methods are used to merge predictions by leveraging the strengths of each modeling technique (Shangguan et al., 2017; Wang et al., 2018b). Thus, ensemble modeling is assumed to provide more accurate and robust spatial predictions than individual models, especially when different models capture distinct aspects of soil dynamics (Padarian and McBratney, 2020). In conjunction with the spatial predictions of SOCs, it is also crucial to report their associated uncertainty, as it conveys valuable information for the proper interpretation of these empirically derived estimates (Poggio et al., 2021).

In Spain, the aforementioned challenges associated with sampling protocols and stock calculation procedures have yielded considerable variation in SOCs estimates. Local-scale estimations of SOCs have been developed for various ecosystems, including agricultural lands (Albaladejo et al., 2009; Álvaro-Fuentes et al., 2008; Muñoz-Rojas et al., 2012), forests, and pastures (Doblas-Miranda et al., 2013). At the national level, SOCs estimates within the upper 30 cm depth have shown a notable range, varying from 2.82 Pg C (Rodríguez Martín et al., 2016) to 3.25 Pg C (Calvo de Anta et al., 2020). While SOCs are typically standardized to the upper 30 cm, the subsoil carbon pool (i.e., >30 cm) may contribute up to 50% of the total stock in Mediterranean soils (Mulder et al., 2016). This discrepancy can lead to an underestimation of a substantial portion of carbon within the effective soil depth (i.e., the soil depth where most SOC storage occurs). Therefore, addressing these multifaceted challenges is pivotal in designing more accurate national-scale assessments of SOCs that support effective management strategies for mitigating climate change.

The overall goal of this study was to enhance our understanding of SOC storage and distribution in peninsular Spain by distinguishing between different carbon variables: SOCc (g/kg), and SOCs (tC/ha). To this aim, we addressed two specific goals: 1) to integrate and standardize disparate soil profile databases developed over the years within peninsular Spain, and 2) to model and map SOCc and SOCs at two soil depths using a machine-learning-based ensemble modeling approach, including their associated uncertainties. Furthermore, we present four 90-meter pixel resolution SOC maps for peninsular Spain (i.e., SOCM90). These maps outline the spatial estimate of SOCc at depths of 0-30 cm and 30-100 cm, along with the SOCs at 0-30 cm and its effective depth, including their associated uncertainties. This information can serve as a reference point for effectively estimating and managing soil carbon sinks at the national level, as well as for compiling the National Greenhouse Gas Emissions Inventory Report (Ministry for Ecological Transition and the Demographic Challenge, 2024). Furthermore, the insights gained from this study contribute to global efforts, including the Global Soil Organic Carbon Map and the GlobalSoilMap (FAO and ITPS, 2018; Arrouays et al., 2014a), as it aligns with the specifications of the Global Soil Organic Carbon Map consortium (Arrouays et al., 2014b).

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

2. Materials and Methods

We followed a three-step methodological framework (Fig. 1). First, we collected soil data from various public sources at different administrative levels and compiled static environmental predictors. Next, we built and assessed ensemble spatial models for SOCc and SOCs using three distinct supervised learning approaches. The final step involved generating spatial predictions and evaluating their accuracy.

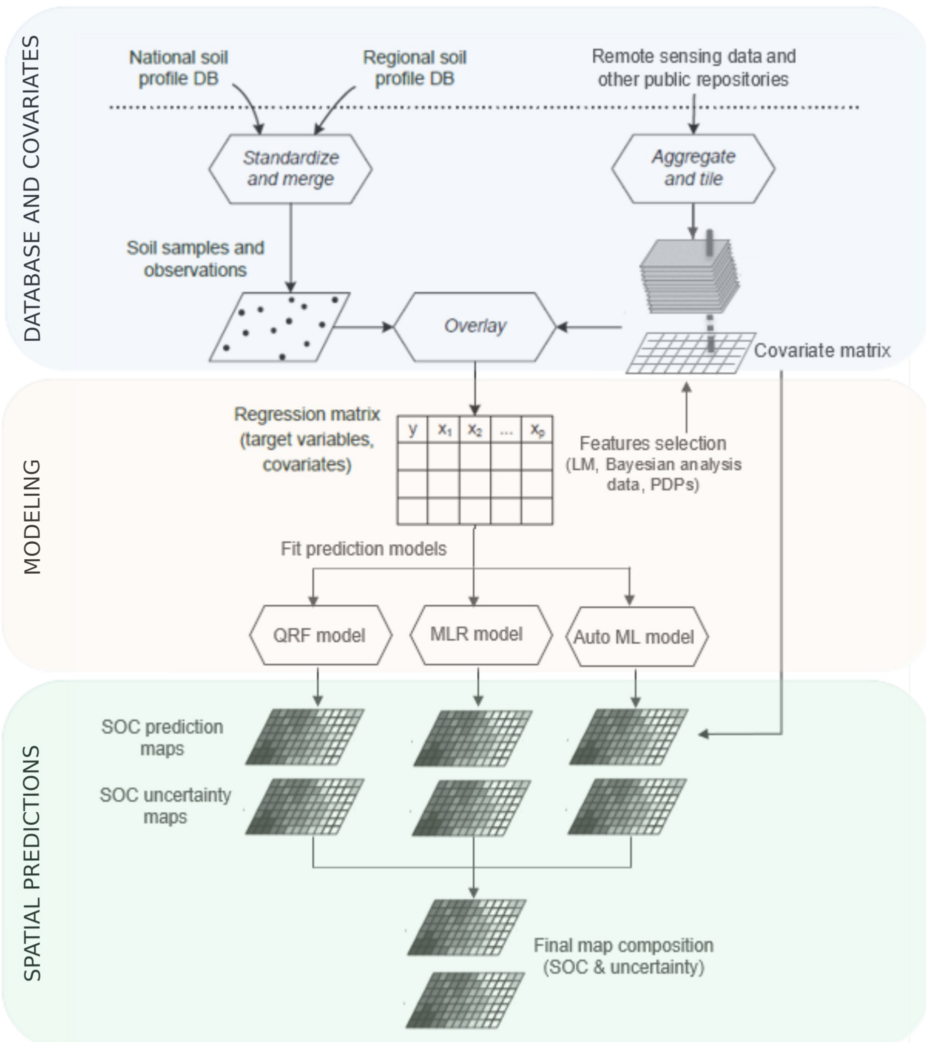


Figure 1. The three-setp methodological framework for this study (adapted from the World Soil Information Service; Hengl et al., 2017).

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Study area

The study area encompassed peninsular Spain, spanning an area of 491,258 km2 (Fig. 2). Peninsular Spain is characterized by an intricate topography dominated by rugged mountain systems, expansive plateaus, and broad watershed depressions. The expansive Central Plateau covers most of the peninsula, with elevations ranging from 600 to 760 masl. The plateau gently slopes towards the west, directing the flow of most watercourses towards the Atlantic Ocean. Surrounding the plateau lie hills and steep mountain ranges, with a maximum altitude of 3,478 masl (Serrano, 2000).

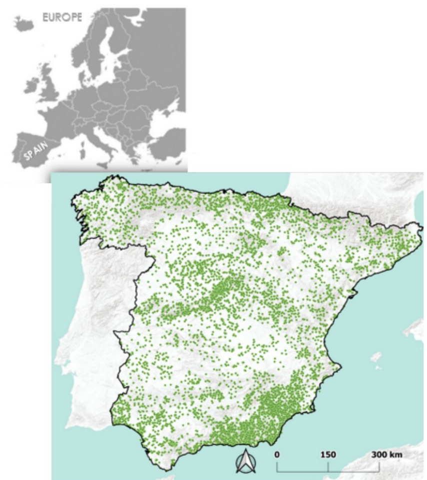


Figure 2. Study area (outlined in black) showing the locations of soil samples collected between 1954 and 2018 (points in green; n=8,361).

Spain is one of the most diverse countries in Europe in terms of climate, ranging from humid to semiarid conditions (AEMET IPMA, 2011). The average annual precipitation ranges from 200 mm (in the southeast) to 2200 mm (in the northern and mountainous regions), while the mean annual temperature spans from below 2.5ºC at higher altitudes to over 17.5ºC in the southern and southeastern. The Mediterranean climate is dominant, extending across the inland plateaus (continental Mediterranean) to the coastal areas (coastal Mediterranean). This climate is characterized by wet, cold to mild winters and dry, hot, or mild summers, with variable temperatures and rainfall during autumn and spring. However, these contrasting climatic conditions are weaker along the coast, transitioning predominantly to arid or semi-desert conditions in the southeast. Conversely, in the north and northwest, the climate tends to be predominantly oceanic, characterized by high humidity and mild temperatures.

The diverse topography and wide-ranging climatic conditions facilitate a mosaic of land covers and uses. Despite a decreasing trend in agricultural areas over the past two decades, agricultural land still occupies approximately 33% of the total land area (MAPA, 2021). The agricultural landscape boasts diverse systems, encompassing herbaceous crops, primarily dryland cereals, and orchard crops such as grapes, olives, almonds, and various fruits. Forested areas cover more than 59% of the peninsula, predominantly comprising natural forests, plantations—mostly

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

found in mountainous regions within humid and subhumid areas—and shrublands. Extensive grasslands and other herbaceous vegetation thrive, particularly in high-altitude regions and the northern part of the country. Wetlands and water surfaces cover 0.9% of peninsular Spain's total area, while artificial surfaces occupy 7.1%.

2.1 Data compilation

2.1.1 Soil database

The database comprised 8,361 georeferenced soil profiles, containing 27,931 pedogenetic soil horizons. We collected soil data from public domain resources or were facilitated by national institutions responsible for the information. Specifically, the Red Carbosol database contributed 78% of the samples, compiled through a collaborative network of Spanish soil experts across multiple research centers and universities, aggregating data from 635 different sources (Llorente et al., 2018). The second major source (18% of profiles) was the Consejería de Sostenibilidad, Medio Ambiente y Economía Azul (Andalusian Government, personal communication). The remaining 4% of the data were extracted from the LUCDEME database, which was compiled by various regional institutions, including Región de Murcia (Alias and Ortiz, 1986), the Agrarian Technological Institute (Junta de Castilla y León), and the University of Castilla La-Mancha (Bravo et al., 2019). Sampling periods spanned from 1954 to 2018, with most samples collected between 1965 and 2000.

To facilitate standardized and reconciled information on soil properties across sampling units, the compiled data structure was modified to create a unified database. The database included information on soil properties related to the computation of the variables, i.e., profile and horizon ID, horizon depth (cm), total carbon content (g/kg), bulk density (g/cm^3), sand, silt, and clay (%), and coarse fragment content (> 2 mm; % of total volume).

Profile inclusion in the final database was contingent upon meeting quality criteria aligned with our research objectives and guided by quantitative pedological methodologies (Beaudette et al., 2013). These criteria encompassed accurate georeferencing, eliminating duplicate information, handling missing data values, verifying information consistency with the horizon, and fixing format inconsistencies. Furthermore, soil properties were explored through basic descriptive statistics, such as minimum, maximum, average, and standard deviation values, to ensure data consistency and evaluate their variability. Inconsistent data were appropriately reclassified as "no data". Further database adjustments included log-transforming the original SOC data. This step was taken to capitalize on the log-normal distribution tendency previously observed in SOC (Yigini et al., 2018). This transformation aimed to improve the correlation between SOC and its predictive factors, ultimately enhancing the accuracy of SOC spatial distribution modeling. To prioritize the organic carbon estimation in mineral soils, profiles or horizons containing more than 20% organic matter were omitted from the database, which is in line with WRB criteria (WRB-IUSS, 2014). As a result, Histosol profiles and horizons primarily composed of organic materials (H, O, L) were excluded from the analysis.

2.1.2 Standard soil depths and estimation of SOC concentration and stock

The soil depth of morphological horizons was standardized to facilitate the integration of the outcomes. The resulting range was established using widely recognized worldwide criteria for SOC estimation (Brus et al., 2017). The information concerning SOCc was obtained through analytical

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

measurements and used directly in the estimation process. However, some data were converted into SOCc from organic matter values, employing a conversion factor of 0.67, based on the assumption of 58% carbon content in organic matter (Rosell et al., 2001). SOCc values were then discretized into two standard depths: 0-30 cm and 30-100 cm. For each standard depth, the final value was determined as the sum of SOCc from morphological horizons, weighted by their original depth. Subsequently, the SOCs within a profile -the total amount of soil carbon per unit area at its effective depth- were computed using Equation [1]. The effective soil depth (ESD) was defined as the solum, encompassing surface and subsurface horizons with root presence and biological activity (Baillie, 2001).

$$SOCs\left(Kg\cdot m^{2}\right)=\sum_{i=1}^{n} SOC\left(g/Kg\right)_{i}\cdot BD\left(Kg\cdot m^{3}\right)_{i}\cdot\left[1-\left(\frac{CRFVOL}{100}\right)\right]_{i}\cdot HSIZE\left(cm_{i}\right) \qquad \text{Eq. [1]}$$

where *i* represents the horizon, *n* denotes the total number of horizons in the profile, *BD* stands for bulk density, *CRFVOL* is the percentage of coarse fragments (i.e., over 2 mm in diameter), and *HSIZE* is the horizon thickness. SOCs were estimated for the standard depth of 0-30 cm before standardizing it to the 0-30 cm range. To prevent the propagation of errors in estimating parameters that may be absent in certain profiles from the existing soil data, the SOCs was computed only for profiles containing available information on bulk density and coarse fragment content. Thus, the final number of profiles used for modeling carbon content as a function of depth was 8,332 profiles for SOCc (g/kg) at a standard depth of 0-30 cm, 6,947 profiles for SOCc (g/kg) at a standard depth of 30-100 cm, 1,475 profiles for SOCs (tC/ha) at a standard depth of 0-30 cm, and 1,499 profiles for SOCs (tC/ha) at its effective depth. Finally, the primary characteristics of SOC in peninsular Spain were determined through basic descriptive statistics of carbon-related variables, including density distribution, mean profiles, and spatial autocorrelation. Data was processed using QGIS and the R packages: *aqp*, *PerformanceAnalytics*, *GSIF*, and *gstat* (R Foundation for Statistical Computing, 2022; Beaudette et al., 2023; Hengl et al., 2020; Pebesma, 2004; Peterson et al., 2014).

2.1.3 Representativeness of the soil database

The spatial representativeness of the compiled soil database was evaluated using a probability distribution approach based on the maximum entropy method (Maxent), which has extensively been applied in modeling spatial point patterns (Phillips et al., 2006) and representativeness of environmental monitoring networks (Villarreal et al., 2018). We used the Maxent approach to estimate the relationship between the number of soil samples and soil-forming factors. Thus, the Maxent model was used to identify distinct representative areas based on pertinent environmental covariates chosen through covariate selection methods (see “Feature selection” subsection below). The model's predictive accuracy was evaluated using the Area Under the Curve (AUC) metric for the training data (see section 2.3 below). This metric was compared against the AUC expected for a random model. Thus, AUC values lower than 0.5 would indicate that the model's predictive performance was worse than random estimation (Fielding and Bell, 1997). We built a model-based predictive map depicting the similarity of environmental predictor variables, thereby minimizing relative entropy between them and locations containing sampled soil data (Elith et al., 2011). This method effectively conveyed information on the spatial representativeness of soil samples across different environmental factors in peninsular Spain.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Open Access Earth System Science Data Discussions

2.1.4 Environmental Covariates

We identified the environmental variabcorporated into the SOC model following the SCORPAN conceptual spatial inference model. This conceptual model categorizes soil property predictions based on seven forming factors, encompassing soil properties (S), climatic variables (C), biota (O), relief (R), parent material (P), time, age (A), and spatial location (N). These covariates were grouped by static and dynamic variables as follows:

Static variables

Relief Factor. Geomorphometric variables depicting terrain characteristics were assessed. Topographic relief was evaluated through geomorphometry and feature extraction derived from the Geomorpho90m global dataset at a resolution of 90 meters under the WGS84 geodetic datum. This dataset comprised 26 fully standardized geomorphometric variables derived from the MERIT-Digital Elevation Model (DEM), encompassing layers that depict (i) the rate of change across the elevation gradient using first and second derivatives, (ii) ruggedness, and (iii) geomorphological forms (Amatulli et al., 2020). Details regarding the source and resolution of these products are outlined in Table 1.

Human Factor. Land cover and land use information (IGN, 2012) were reclassified into 13 classes, i.e., non-irrigated arable land, permanently irrigated land, heterogeneous agricultural areas, agro-forestry areas, broad-leaved forest, coniferous forest, mixed forest, sclerophyllous vegetation, pastures, moors and heathland, sparsely vegetated areas, and transitional woodland/shrub.

Parent Material Factor. Lithological classes were derived from the lithological map of Spain 1M, which comprises 22 hierarchical levels (IGME, 1995).

Soil Properties. Soil information was obtained from the Digital District Soil Atlas (USDA, 1987). The soil map was digitized based on the 1:2,000,000 map in the Atlas Nacional de España (Soil Science) published by the CSIC/IRNAS (De la Rosa et al., 2001).

Dynamic variables

Climate Factor. Precipitation and temperature climatic variables were obtained from Ninyerola et al. (2005).

Biota Factor. We computed a set of ecosystem functioning attributes derived from remotely sensed indices. These attributes are associated with the carbon cycle, water cycle, and energy or heat balance. Functional attributes relate to each index's quantity, seasonality, and phenology. The satellite products spanned a sufficient time interval to ensure their stability over time (2000-2019). The complete remotely sensed set comprised 172 ecosystem functioning attributes as candidate predictors for SOC. Details regarding the sources and resolution of the satellite products are provided in Table 1. The Google Earth Engine platform (Gorelick et al., 2017) was employed to derive these attributes, including the annual mean (serving as a proxy for annual total amount), annual maximum and minimum (indicating annual extremes), seasonal standard deviation (describing seasonality), and sine and cosine of the dates of maximum and minimum (indicating phenology) (Alcaraz-Segura et al., 2017).

In summary, 254 environmental covariates, capturing the diverse forming factors across peninsular Spain, were computed as an initial step preceding the spatial modeling of SOC.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Open Access Earth System Science Data Discussions

Table 1. Description of predictors for spatial modeling of SOC.

Category	Variables	Source	Scale/ Resolution
Topographic			
	(i) Slope, Aspect, Aspect cosine, Aspect sine, Eastness, Northness, Convergence, Compound topographic index, Stream power index, East-West first order partial derivative, North-South first order partial derivative (ii) Profile curvature, Tangential curvature, East-West second order partial derivative, North-South second order partial derivative, Second order partial derivative (iii) Elevation standard deviation, Terrain ruggedness index, Roughness, Vector ruggedness measure, Topographic position index, Maximum multiscale deviation (dev-magnitude), Scale of the maximum multiscale deviation, Maximum multiscale roughness, Scale of the maximum multiscale roughness (iv) Geomorphon	Geomorpho90m (Amatulli et al., 2020)	90m
Climate			
	Mean annual precipitation (mm). Mean, minimum, and maximum annual temperature (ºC). Radiation (kW/m^2). Period 1951-1999.	University of Barcelona (Ninyerola et al., 2005)	20 m
Land features			
	Soil types	Proyecto SEIS.net (MIMAM- CSIC)	1:100000
	Lithology	SGE-IGME (Spain)	1:200000
	Land use/cover	IGN-Corine Land Cover	1:100000
Remotely sensed indices			
Carbon cycle	Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI	MOD13Q1	250 m
Water cycle	Precipitation, Normalized Difference Water Index (NDWI) Evapotranspiration (ET)	CHIRPS MCD43A4	1km 500 m
Radiative balance	Albedo	MCD43B3	500 m
Sensible heat	Land Surface Temperature (LST)	MOD11A2	1 km
Ecosystem Functional attributes (inter-annual and monthly mean): **Amount**: mean, maximum, minimum **Seasonality**: *standard deviation, coefficient of variation, range, relative range* **Phenology**: *sine and cosine of the dates of maximum and minimum*		(Alcaraz-Segura et al., 2017)	

(i) First order derivative, (ii) Second order derivative, (iii) Ruggedness, (iv) Geomorphological forms.

Finally, the environmental covariates were compiled into a covariate matrix with spatially explicit information. A method alternative to reprojection and rescaling was employed to accommodate the diverse formats of coordinate reference systems (CRS) and spatial resolutions of the covariates. This approach aimed to reduce geometric distortion and mitigate computational limitations stemming from the substantial volume of data (Bauer-Marschallinger et al., 2014). Firstly, a reference matrix was created based on the pixel center locations (x, y) of the most detailed resolution layer, i.e., MERIT-DEM (90 m), with WGS84 (EPSG 4326) as the CRS. Subsequently,

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

geoprocessing techniques extracted the covariate values for each location in the reference matrix. When applicable, the reference matrix coordinates were then reprojected to match the CRS of each covariate. The resulting matrix comprised the value of each covariate (columns) extracted for every point in the study area at intervals of 90 meters. This matrix was organized into a tiling system to enhance computing processing time during geoprocessing analyses. Categorical variables were rasterized. To do that, only categories with a sufficiently representative number of soil samples (i.e., over 100 data points) were considered. These categories were subsequently transformed into a binary variable, indicating the presence or absence of the specific category (Yigini et al., 2018). The covariance matrix was generated using the R packages *sp*, *rgdal*, and *raster* (Hijmans, 2024; Keitt et al., 2012; Pebesma and Bivand, 2005).

2.2 Modeling

The data matrix comprised the log transformations of SOCc and SOCs and the covariate point data extracted at the same locations as the soil profiles. To mitigate geometric distortions, the coordinates of the profiles were reprojected to match the CRS of each covariate, where applicable. This organized structure facilitated analyzing the relationship between carbon and the covariates, ensuring accurate spatial alignment for meaningful interpretation.

2.2.1 Covariate selection

We analyzed covariate importance (CVI) to discern which covariates had the most significant impact on soil carbon models, thus reducing the substantial number of covariates (254) and mitigating the risk of potential overfitting (Gregorutti et al., 2017). CVI was evaluated using three selection methods: multiple linear regression, Bayesian analysis, and projection pursuit regression models (coupled with partial dependence plots).

Multiple linear regression is an easily interpretable method where the dependent variable (here SOCc and SOCs) is represented as a linear combination of regression coefficients, predictor variables, and a random error term. This error term accounts for the variation in the dependent variable that cannot be explained by the linear relationship with the predictors and their coefficients. Two fundamental assumptions are considered: a) the existence of a linear relationship between the response variable and the predictors (environmental covariates), and b) the absence of multicollinearity among the predictors (Yan and Su, 2009). To evaluate the importance of each covariate, we utilized the absolute value of the t-statistic. The t-statistic was computed by dividing the regression coefficient associated with each predictor by its standard error, serving as a measure of predictor accuracy (Greenwell and Boehmke, 2020). This analysis was conducted using the *vip* R package (Greenwell and Boehmke, 2020).

Bayesian analysis is an inferential approach rooted in the probability distribution of parameters derived from observed data and additional available information. Unlike traditional multilinear models, Bayesian probability models treat parameters as random variables and integrate data and prior information on the parameter distribution through a likelihood function to form a new posterior distribution (Gelman et al., 2013). The resulting parameter estimates are conditional on the observed data, which, following the rules of probability theory, ensures a consistent posterior distribution interpretation (McElreath, 2018). The Bayesian models were fitted using the Markov Chain Monte Carlo (MCMC) sampling methodology through an iterative selection of the most significant covariates. The final covariate selection was determined by evaluating the models based

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

on their Watanabe–Akaike information criteria (WAIC) scores, with preference given to the model exhibiting the lowest WAIC value. This criterion serves as a measure of model performance, balancing goodness of fit with model complexity. We implemented the Bayesian approach using the *rethinking* package in R (McElreath, 2020).

Projection Pursuit Regression (PPR) involves forming linear combinations of non-parametric functions of the predictor variables, enabling the exploration of nonlinear relationships within the data (Friedman and Stuetzle, 1981). Thus, PPR captures complex relationships between predictor variables and the dependent variable. Partial dependence plots (PDPs) were constructed for each covariate of interest, illustrating how changes in each covariate affect the predicted outcome while keeping other covariates fixed. The flatness of each PDP was assessed to determine the strength of the association between the covariate and the predicted outcome. CVI scores were computed based on the variability in the partial dependence values, supporting the identification of the most influential covariates. These scores captured the variability in the partial dependence values for each main effect by calculating the standard deviation of the y-axis values for each PDP. Covariates were ranked based on their CVI scores and the shape of their PDPs, guiding the selection of covariates for inclusion in the final model. The PPR models and PDPs were constructed using the *stats* and *pdp* packages in R (Greenwell, 2017).

Finally, covariates exhibiting the highest CVI scores, consistently identified across all three selection techniques, were integrated into the final dataset. Finally, expert criteria guided the decision-making process for the final selection of variables. This selection procedure was conducted individually for each dependent variable (i.e., SOCc and SOCs) at both the 0-30 cm and 30-100 cm standard depths (SOCc), and at both the 0-30 cm and the effective depth (SOCs).

2.2.2 Predictive models

There are multiple algorithms for predicting SOC using DSM and arguably there is not a single ideal one for predicting SOC across large geographical areas (Guevara et al., 2018). Given the intricate and often nonlinear relationship between SOC and environmental variables, the integration of multi-model ensemble methods with machine learning (ML) algorithms has been adopted to predict SOC spatial variability along with associated uncertainties (Gray et al., 2015; Shangguan et al., 2017; Wang et al., 2018a). Ensemble learning, a branch of ML, combines multiple base ML models, homogeneous (e.g., a combination of multiple decision trees) or heterogeneous (e.g., a combination of decision trees with support vector machines), to enhance predictive performance by mitigating errors between observed and predicted data (Zhang and Ma, 2012). Here, we employed three ensemble modeling approaches for predicting SOC.

Quantile regression forest (QRF), unlike traditional regression methods, QRF can handle sparse legacy data effectively without the need for kriging interpolation of residuals (Meinshausen, 2006). The QRF algorithm can predict SOC values at various quantiles without relying on specific assumptions about the; therefore, it extends the capabilities of the random forest by providing accurate estimates across the entire distribution of the response variable. We leverage the *quantregForest* package in R (Meinshausen, 2006) to implement the QRF method. Validation statistics for QRF models were computed using out-of-bag error estimation. This method evaluates the model's performance by measuring the prediction error on data points not included in the samples used to train each decision tree in the ensemble. Thus, out-of-bag error estimation provides

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

an unbiased estimate of the model's predictive accuracy, helping to measure its performance in generalizing to new data.

Ensemble Machine Learning (MLR) combines linear model regression predictors with non-parametric models using bagging and boosting algorithms (Dietterich, 2000). Bagging and boosting are ensemble learning techniques used to improve the performance of ML models by combining multiple base models. Overall, bagging involves creating multiple subsets of the training data through bootstrapping (random sampling with replacement). Then, a base model (e.g., a decision tree) is trained on each subset independently. Finally, predictions from all base models are combined, typically by averaging, to make the final prediction. Conversely, Boosting works sequentially by training a series of weak models (i.e., models that perform slightly better than random models) and giving more weight to mispredicted data in subsequent iterations. Each new model focuses on the observations mispredicted by the previous models, thus gradually improving the model's performance. The final prediction is typically a weighted sum of predictions from all weak models. Both bagging and boosting aim to reduce overfitting and improve the overall performance and robustness of the model. We constructed a stacked model by integrating predictions from five modeling techniques: linear model regression, random forest, deep learning, cubist, and weighted k-nearest neighbor classifier. The predictions were used as features for a stacked model trained to compress the predictions from the base models. Each base model underwent independent assessment through ten-fold cross-validation. To incorporate spatial information, the dataset was split using spatial partitioning by k-means clustering, considering two classification layers: the XY locations of soil data and the Köppen climate classification (Kottek et al., 2006). The *MLR* R package was used to implement this ensemble approach (Bischl et al., 2016).

Auto-machine learning (AutoML) automates building ML models by efficiently selecting algorithms, tuning hyperparameters, and optimizing computational resources. With the vast array of ML algorithms available and the complexities of hyperparameter tuning, manual selection, and optimization can be challenged. We used the *H2O* package in R (Fryda T; Erin LeDell, 2024) to address this issue, which employs various Gradient Boosting Machine (GBM) algorithms such as generalized linear models, distributed random forests, deep neural networks, XGBoost, and gradient boosting machines. This diversity of models allows stacked ensembles to produce robust final predictions. The models were evaluated using 10-fold cross-validation and ranked based on their root-mean-square error (RMSE).

2.3 Spatial Prediction

We built model-based predictive maps for SOCc at depths of 0-30 cm and 30-100 cm, and for SOCs at depths of 0-30 cm and the effective depth, with a pixel resolution of 90 meters. To do that, we first generated three prediction maps using each ensemble modeling approach (i.e., QRF, MLR, and AutoML) and calculated the standard deviation for each pixel in these maps. Then, we assigned the predicted value from the most accurate map -determined by the lowest standard deviation value (i.e., the highest agreement among models)- to each pixel (Fig. 3). This approach ensures that the final predictive maps reflect the most reliable estimates of SOC content at each pixel, incorporating the collective insights from multiple modeling techniques and accounting for the associated uncertainty. This approach has been successfully applied in digital soil mapping

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

(Varón-Ramírez et al., 2022; Arroyo-Cruz et al., 2017) and in reducing uncertainty when modeling ecosystem-related variables (Gavilán-Acuña et al., 2021).

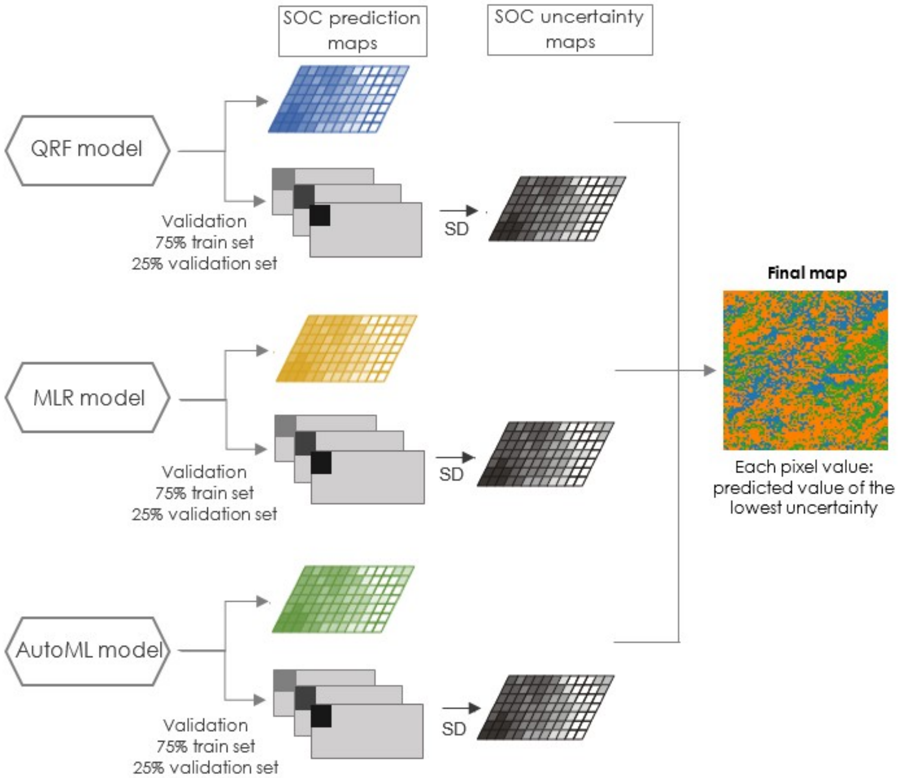


Figure 3. Scheme illustrating the generation of predictive maps of soil organic carbon using an ensemble modeling approach to reduce model uncertainty. SD (standard deviation).

The metrics and information criteria used to evaluate the models' performance generated by the three ensemble modeling approaches included coefficient of determination (R^2), concordance correlation coefficient (CC), which measures the concordance level between predicted and observed values, root mean square error (RMSE), and mean absolute error (MAE). The dataset was randomly split into calibration (75%) and validation (25%) data for each modeling approach. Conditional quantile plots were generated to further assess model performance across the entire distribution of observed SOC values (Wilks, 2019). These plots evenly divide the predicted values and identify corresponding observation values, including the median, 25th/75th, and 10th/90th percentiles. By doing so, they offer insights into the alignment between predictions and observations across the entire range of values. The conditional quantile plots were generated using the *openair* package in R (Carslaw and Ropkins, 2012).

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Open Access Earth System Science Data Discussions

3. Soil database overview

The harmonized soil database comprised 8,332 soil profiles with 25,370 morphological horizons distributed across peninsular Spain. Only profiles with complete information were used for SOC content calculation to prevent error propagation. This resulted in 1,499 profiles, slightly exceeding the 1,475 used for SOC estimation at the standard depth of 0-30 cm. SOCc exhibited high variability across soil horizons, ranging from 0.02 g/kg to 296.9 g/kg, with a mean of 16.53 g/kg (Table 2). The mean SOCc at the standard depth of 0-30 cm was 20.7 g/kg; at the 30-100 cm depth, it was 5.8 g/kg, representing 35% of the soil profile mean. Similarly, SOCs varied from 0.006 kg/m² to 87.1478 kg/m², with a mean of 2.965 kg/m². SOCs at the standard depth of 0-30 cm accounted for 65% of the soil profile mean.

Table 2. Statistical summary of SOCc and SOCs at different soil depths.

Variable	Depth	Number of profiles	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
SOCc (g/kg)	0-30	8,332	0.017	7.148	14.008	20.691	27.098	257.95
	30-100	6,947	0.017	1.700	3.371	5.833	6.814	185.743
SOCs (kg/m^2)	0-30	1,475	0.119	2.066	3.738	5.31	6.95	39.967
	ESD[1]	1,499	0.119	3.000	5.300	8.198	10.260	93.892

(1) Effective soil depth. Soil organic carbon concentration (SOCc). Soil organic carbon stock (SOCs).

Both SOCc and SOCs followed a normal distribution with a right-skew after transforming the original values to a natural log (Fig. S1 in Supplementary material). The highest SOCc values were concentrated at the upper layers (0-30 cm), decreasing rapidly with depth (Fig. 4). Mean profile values ranged from 23 g/kg in the upper horizon (0-5 cm) to 3 g/kg in the deepest horizon (>200 m). In general, peninsular Spain's soils showed shallow depths, with only 35% of horizons extending over 100 cm and decreasing to 3% at depths exceeding 150 cm.

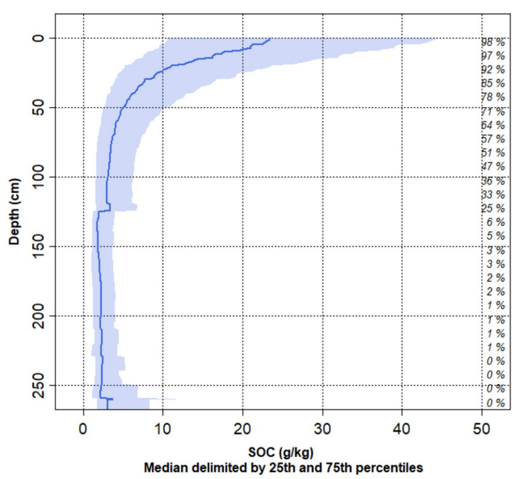


Figure 4. Average Soil Organic Carbon concentration (SOCc; g/kg) at different depths.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Nugget-to-sill ratio (NSR) analysis indicated weak spatial autocorrelation of SOC with an NSR >75% for the log-transformed SOCc. The large nugget effect (i.e., 1.3; Fig. S2 in Supplementary material) suggested that the data did not capture a significant portion of fine-scale SOCc variation. Despite the weak autocorrelation evidence, SOCc kriging predictions aligned with the expected distribution and demonstrated relatively low standard errors (Fig. S2 in Supplementary material). To delve deeper into the spatial dependence of SOCc with soil depth, autocorrelation analysis was conducted at six different depths, ranging from 5 cm to depths exceeding 100 cm (Table S1 in Supplementary material). A moderate dependence was observed, with the best spatial correlation (NSR = 35%) at the shallowest horizon (0-5 cm), gradually decreasing with depth to an NSR of 72% at 2 m.

4. Representativeness of the soil database

The representativeness of a soil type in the database (i.e., the probability of a soil type being sampled) as a function of soil-forming factors is illustrated in Fig. 5. The representativeness values ranged from 0 (low) to 1 (high). The Maxent model yielded an AUC value of 0.611, indicating a predictive capacity greater than that yielded by a random model (i.e., 0.5). The predictive map showed variations in representativeness across different ecosystems. Overall, mountainous regions such as the Central Mountain System in central Spain, Sierra Morena and the Betic System in the south, as well as the Sierra de la Demanda, Cantabrian Range, and the Pyrenees in the north, showed high representativeness, as depicted by the blue areas in Fig. 5. In contrast, the Central Plateau, Ebro Depression, and Tajo and Guadiana Basins (i.e., interior regions in both the south and the north) were among the least represented areas (areas in green).

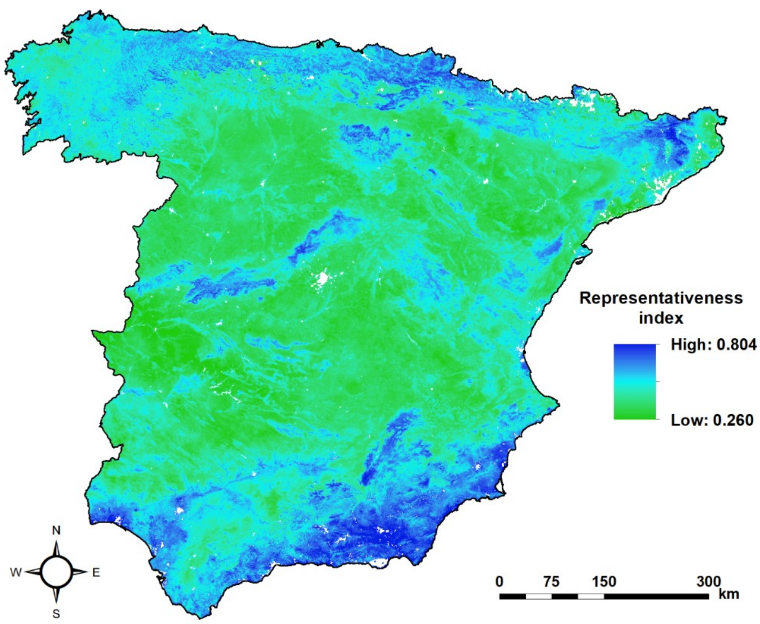


Figure 5. Maxent model-based representativeness of soil types sampled across peninsular Spain.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

5. SOC modeling and prediction

The CVI analysis resulted in a highly reduced covariate space across selection methods for SOCs and SOCc. Table 3 shows the selected covariates at different depths. Seventeen and nineteen covariates were selected for modeling SOCc at the 0-30 cm and 30-100 cm depth, respectively. For modeling SOCs, thirteen covariates were selected for the 0-30 cm depth and the effective depth. Overall, annual precipitation, and the mean and minimum spring temperatures, were identified as the most influential climatic factors. Similarly, among the remotely sensed variables related to the carbon cycle, the annual mean and maximum NDVI and the monthly mean EVI (in March) were found relevant. Moreover, annual and monthly indices linked to the water cycle, such as ET and NDWI, exhibited notable importance. Regarding topographic covariates, indices derived from terrain roughness, including maximum rough-magnitude, dev-scale related to topographic position, and slope, were identified as particularly relevant. Additionally, soil covariates, such as lithology, soil type, and land use/cover, were included in all SOC models due to their fundamental role in modeling soil properties (Jenny, 1941).

Table 3. Covariates selection for modeling soil organic carbon concentration (SOCc) and soil organic carbon stock (SOCs) at different standard depths.

Frequency	SOCc (0-30 cm)	SOCc (30-100 cm)	SOCs (0-30 cm/ESD) [(1)]
	Climate variables [(2)]		
Annual	- Min Temp	- Mean Pp	- Mean Pp - Mean Temp
Monthly	- Mean Pp for May - Mean Temp for May - Min Temp for May	- Max Temp for Feb - Mean Temp for May - Min Temp for March	- Max Temp for April - Min Temp for May
	Remotely sensed indices [(2)]		
Annual	- Max Albedo - Mean NDVI - Max NDWI	- Mean ET - Max LST - Max NDVI - Max NDWI	- Max ET - Mean NDVI
Monthly	- Mean Albedo for August - Mean ET for March - Mean LST for March - Mean NDVI for June	- Mean EVI for March - Mean ET for May - Mean LST for July - Mean NDWI for July	- Mean EVI for March
	Topographic variable [(2)]		
Static	- dev-magnitude - dev-scale - rough-magnitude	- dev-magnitude - dev-scale - rough-magnitude - slope	- elev-stdev - rough-magnitude - slope
	Land features		
Static	- Lithology	- Soil types	- Land use/cover
Total number	17	19	13

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

(1) ESD: effective soil depth. (2) Max: maximum; Min: minimum; dev-magnitude: Max multiscale deviation; dev-scale: Scale of the Max multiscale; rough-magnitude: Max multiscale roughness; elev-stdev: Elevation standard deviation; Pp: precipitation; Temp: temperature; NDVI: Normalized Difference Vegetation Index; NDWI: Normalized Difference Water Index; ET: Evapotranspiration, LST: Land Surface Temperature.

Analysis of residuals revealed R^2 values of 0.68 for SOCc and 0.54 for SOCs in the upper 30 cm, with lower values in deeper horizons. Table 4 summarizes SOC models' performance at different depths. Model validation included computing the CC for both calibration (CCcal) and validation (CCval) data and the normalized root mean square error (nRMSE) and normalized mean absolute error (nMAE) for validation data. The inconsistency between CCcal and CCval was minimal, except in the AutoML model, where the GBM family of algorithms notably contributed to a higher CCcal. Based on CCval and nMAE metrics, a decline in accuracy with depth was observed for both SOC variables. For instance, at the 0-30 cm depth, CCval was 0.583, and nMAE was 0.441, while at the 30-100 cm depth, CCval was 0.351, and nMAE was 0.668 for SOCc. Comparing the same depth intervals, SOCc demonstrated greater accuracy than SOCs across all models in the upper 30 cm, whereas it exhibited the lowest accuracy at 30-100 cm. Although the disparities in nMAE among the three predictive ensemble approaches were minimal, there were noticeable differences in CC values, indicating a decreasing trend of higher performance in the AutoML, QRF, and MLR models.

Table 4. Average model performance for soil organic carbon concentration (SOCc) and stock (SOCs) at different depths.

Parameter	CC_{cal}	CC_{val}	nRMSE	nMAE	CC_{cal}	CC_{val}	nRMSE	nMAE
SOC concentration								
Predictive model	**0-30 cm**				**30-100 cm**			
AutoML	0.825	0.583	0.684	0.433	0.629	0.351	1.296	0.668
QRF	0.485	0.472	0.733	0.458	0.354	0.307	0.680	0.610
MLR	0.434	0.350	0.775	0.474	0.278	0.227	0.737	0.578
SOC stock								
	0-30 cm				**Effective soil depth**			
AutoML	0.672	0.417	0.548	0.441	0.563	0.378	0.651	0.516
QRF	0.445	0.381	0.545	0.504	0.358	0.295	0.646	0.552
MLR	0.401	0.270	0.628	0.535	0.323	0.232	0.711	0.570

Normalized root mean square error (nRMSE) and Normalized mean absolute error (nMAE) in parts per unit.

The conditional quantile plots revealed that the predicted SOC values did not encompass the entire range of observed values (Fig. S3 in Supplementary material). The highest predicted SOCc at 0-30 cm reached a maximum of 136.1 g/kg (compared to the maximum observed value of 257.95 g/kg), while the highest predicted SOCs value at the effective soil depth (ESD) reached 38.33 kg/m2 (compared to the observed value of 93.892 kg/m2) (Fig. S3 in Supplementary material).

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Model-based predictive maps of SOCc and SOCs in peninsular Spain are shown in Figs. 6 and 7, respectively. Overall, SOCc and SOCs exhibited similar and consistent spatial patterns across depths (Fig, 6a and 7a). Higher SOC predicted values correlated with both climatic and topographic features. For example, the northwest and north regions, characterized by a humid climate, exhibited the highest SOC content, gradually decreasing toward the east. Similarly, major mountainous areas, such as the Central Mountain System, Iberian System, and Subbetic System, also achieved noticeable SOC predictions. In contrast, low SOC predictions were obtained in extensive agricultural regions, including the Central Plateau, southwest Guadiana alluvial plain, and the Guadalquivir depression, along with arid zones in the southeast. The uncertainty around SOC predictions revealed distinct spatial patterns for SOC predictions, with notable disparities between the northern and southern regions of peninsular Spain. These disparities positively correlated with SOC content. North areas with higher SOC predicted values also exhibited greater uncertainty (Figs. 6b and 7b). Although the spatial distribution of SOC uncertainty remained consistent between concentration and stock, SOCc exhibited larger areas with higher uncertainty. Regarding standard depths, regions with higher uncertainty were more prevalent in upper horizons (0-30 cm) than in deeper horizons.

The spatial distribution of the ensemble modeling approach used for predicting SOC content varied depending on the combination of SOCc and SOCs and depth (Figs 6c and 7c). Regarding SOCc at 0-30 cm, all three models exhibited a uniform spatial distribution, except for regions with higher SOCc in the north and mountainous areas, where the QRF model was dominant. In contrast, predictions for SOCc at 30-100 cm showed a reduced contribution from the AutoML approach, with MLR and QRF models dominating distinct areas. Models' capacity to predict SOCs was spatially consistent at different depths, with no discernible spatial pattern.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Open Access Earth System Science Data Discussions

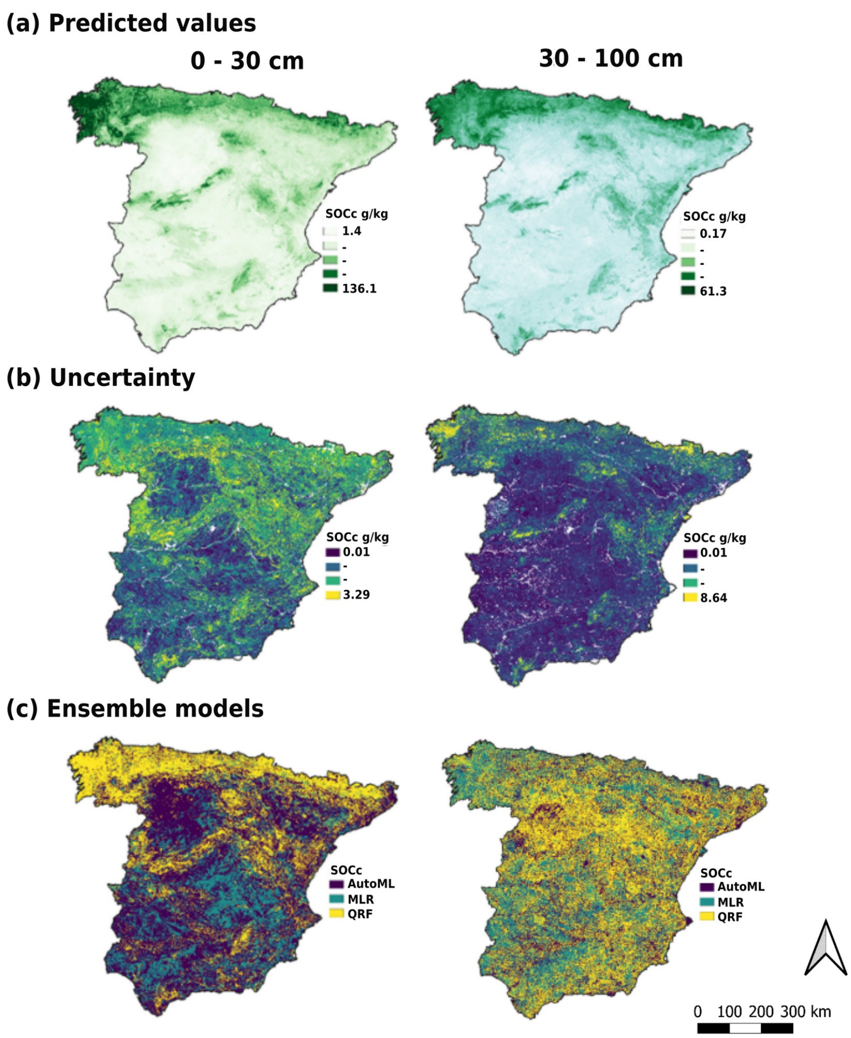
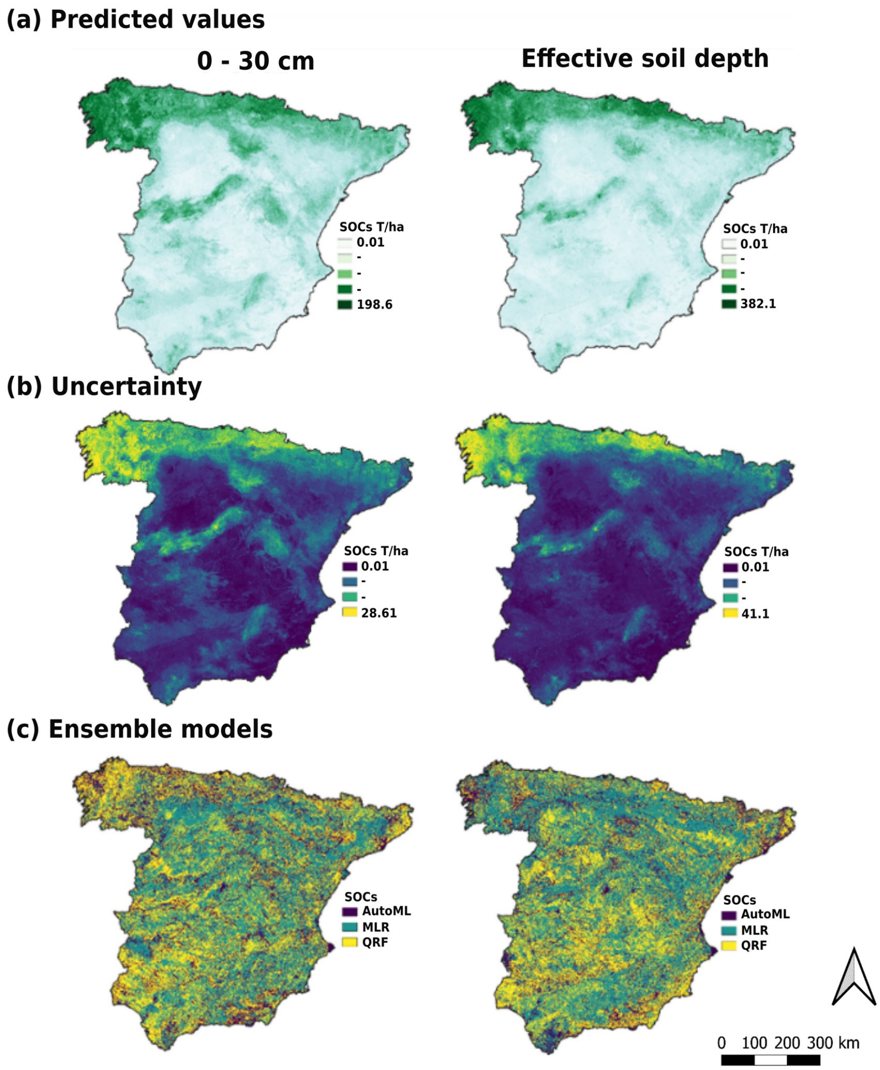


Figure 6. (a) Soil organic carbon concentration (SOCc) maps in peninsular Spain at a 90-meter pixel resolution. (b) Uncertainty maps are based on the standard deviation of predictions obtained through the three ensemble modeling approaches. (c) Ensemble algorithm employed for pixel-level SOC predictions. The visualization used the cumulative pixel count cut method, with a default range from 2% to 98%.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024



538 **Figure 7**. (a) Soil organic carbon stock (SOCs) maps in peninsular Spain at a 90-meter pixel resolution. (b) Uncertainty maps are based on the standard deviation of predictions obtained through the three ensemble modeling approaches. (c) Ensemble algorithm employed for pixel-level SOC predictions. The visualization used the cumulative pixel count cut method, with a default range from 2% to 98%.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

6. Data availability

The soil organic carbon concentration (g/kg) maps for the 0-30 cm and 30-100 cm standard depths, along with the soil organic carbon stock (tC/ha) maps for the 0-30 cm standard depth and the effective soil depth, including their associated uncertainties, —all at a 90-meter pixel resolution— (SOCM90) are freely available at https://doi.org/10.6073/pasta/48edac6904eb1aff4c1223d970c050b4 (Durante et al., 2024).

7. Further considetations

Our study seeks to enhance the availability and reliability of soil information essential for informed decision-making regarding SOC management and climate change mitigation strategies. By integrating disparate soil profile databases and employing advanced ensemble modeling techniques, we aimed to provide comprehensive and standardized SOC maps for peninsular Spain, facilitating access to critical soil data at the national scale. As part of the broader effort to enhance soil data accessibility and usability, our methodology demonstrates the transformation of previously inaccessible soil information into actionable insights for spatial variability studies and carbon stock assessments.

By establishing a systematic approach to organizing national soil information, we mitigate potential errors and discrepancies in future data generation processes, ensuring the reliability and consistency of soil carbon estimates. Furthermore, the enhanced spatial modeling approach of soil information in peninsular Spain supports ongoing global soil information initiatives, including the Global Soil Organic Carbon Map, a project of FAO, the Global Soil Partnership, and the GlobalSoilMap.Net project. It enables informed decision-making regarding land use planning, agricultural practices, and environmental conservation efforts.

The SOCM90 integrated information on more than eight thousand profiles for peninsular Spain soils. Despite these advancements, it is essential to acknowledge the existence of data gaps in certain areas and incentivize future soil survey programs to increase sampling efforts in underrepresented regions. By expanding soil monitoring networks and improving spatial coverage, the SOCM90 can contribute to more comprehensive assessments of SOC content and inform targeted soil management strategies.

8. Author contribution

Conceptualization: all; Data curation: PD and CO; Formal analysis: PD; Funding acquisition: PD and CO; Methodology: PD, RV, MG, DA, and CO; Supervision; RV, MG, DA, and CO; Validation: PD, RV, MG, and CO; Writing – original draft preparation: JMRM, PD, CO; Writing – review & editing: JMRM, PD, RV, MG, and CO.

9. Competing interests

The authors declare that they have no conflict of interest.

10. Acknowledgements

PD acknowledges support from the pre-doctoral grant [DI-15-08093] awarded by the ‘National Programme for the Promotion of Talent and Its Employability’ of the Ministry of Economy, Industry, and Competitiveness, which are partially funded by the European Social Fund (ESF) from the European Commission. JMRM was funded by the University of Almería through the Spanish Ministry of Universities (María Zambrano Program) [grant number RR_C_2021_09]; the University of Almeria’s programme for research and knowledge transfer [grant number P_FORT_GRUPOS_2023/26]. RV was supported by the NASA Carbon Monitoring System grant 80NSSC21K0964. MG was funded by UNESCO-IGCP (grant no. 765) and Conahcyt (grant no. CF-2023-I-1846). DA acknowledges support from the project “Plan Complementario de I+D+i en el área de Biodiversidad (PCBIO)” funded by the European Union within the framework of the Recovery, Transformation and Resilience Plan - NextGenerationEU and by the Regional Government of Andalucia, by the EarthCul project (PID2020-118041GB-I00 Spanish National Research and Innovation Plan 2020). CO acknowledges support from the projects INTEGRATYON3 (PID2020-117825GB-C21 and C22), both funded by MCIN/AEI/10.13039/501100011033.

11. References

AEMET IPMA: Atlas climático ibérico/Iberian climate atlas, Agencia Estatal de Meteorología (España) ; Instituto de Meteorología (Portugal), 2011.

Albaladejo, J., Martinez-Mena, M., Almagro, M., Ruiz-Navarro, A., Ortiz, R., and Albaladejo Montoro, Juan; Martínez-Mena García, María; Almagro Costa, Mercedes; Ruiz-Navarro, Ana; Ortiz Silla, R.: Factores de control en la dinámica del Carbono Orgánico de los suelos de la Región de Murcia, Murcia, 155–158 pp., 2009.

Alcaraz-Segura, D., Lomba, A., Sousa-Silva, R., Nieto-Lugilde, D., Alves, P., Georges, D., Vicente, J. R., and Honrado, J. P.: Potential of satellite-derived ecosystem functional attributes to anticipate species range shifts, International Journal of Applied Earth Observation and Geoinformation, 57, 86–92, https://doi.org/10.1016/j.jag.2016.12.009, 2017.

Alias, L. and Ortiz, R.: Memorias y mapas de suelos de las hojas del MTN a escala 1:100.000, 1986.

Allen, T. F. H. and Starr, T. B.: Hierarchy: perspectives for ecological complexity, University of Chicago Press, 2019.

Álvaro-Fuentes, J., López, M. V., Cantero-Martinez, C., and Arrúe, J. L.: Tillage Effects on Soil Organic Carbon Fractions in Mediterranean Dryland Agroecosystems, Soil Science Society of America Journal, 72, 541–547, https://doi.org/10.2136/sssaj2007.0164, 2008.

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., and Domisch, S.: Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers, Sci Data, 7, 162, https://doi.org/10.1038/s41597-020-0479-6, 2020.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B. M., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., Mendonca-Santos, M. d. L.,

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Minasny, B., Montanarella, L., Odeh, I. O. A., Sanchez, P. A., Thompson, J. A., and Zhang, G.-L.: GlobalSoilMap, 93–134, https://doi.org/10.1016/B978-0-12-800137-0.00003-0, 2014a.

Arrouays D., McBratney A.B., Minasny B., Hempel J.W., Heuvelink G.B.M., MacMillan R.A., Hartemink A.E., Lagacherie P., and McKenzie N.J.: The GlobalSoilMap project specifications. In: Arrouays D., McKenzie N., Hempel J., de Forges A.R., McBratney A.B. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System, CRC Press, London, 2014b.

C.E. Arroyo-Cruz, Larson, J., Guevara, M.: A machine learning approach for mapping soil properties in Mexico using legacy data, climate and terrain covariates at a coarse scale 23–27. In: Arrouays, D., Savin, I., Leenaars, J., McBratney, A.B. (Eds.) GlobalSoilMap - Digital Soil Mapping from Country to Globe: Proceedings of the Global Soil Map 2017 Conference, July 4-6, 2017, Moscow, Russia. CRC Press, London. https://doi.org/10.1201/9781351239707, 2017.

Baillie, I. C.: Soil survey staff 1999, soil taxonomy: a basic system of soil classification for making and interpreting soil surveys, agricultural handbook 436, Natural Resources Conservation Service, USDA, Washington DC, USA, pp. 869, 2001.

Beaudette, D. E., Roudier, P., and O'Geen, A. T.: Algorithms for quantitative pedology: A toolkit for soil scientists, Comput Geosci, 52, 258–268, https://doi.org/10.1016/J.CAGEO.2012.10.020, 2013.

Beaudette, D. E., Skovlin, J., Brown, A. G., Roudier, P., and Roecker, S. M.: Algorithms for Quantitative Pedology, in: Geopedology: An Integration of Geomorphology and Pedology for Soil and Landscape Studies, Springer, 201–222, 2023.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M.: mlr: Machine Learning in R, Journal of Machine Learning Research, 1–5 pp., 2016.

Bravo, S., García-Ordiales, E., García-Navarro, F. J., Amorós, J. Á., Pérez-de-los-Reyes, C., Jiménez-Ballesta, R., Esbrí, J. M., García-Noguero, E. M., and Higueras, P.: Geochemical distribution of major and trace elements in agricultural soils of Castilla-La Mancha (central Spain): finding criteria for baselines and delimiting regional anomalies, Environmental Science and Pollution Research, 26, 3100–3114, https://doi.org/10.1007/s11356-017-0010-6, 2019.

Brus, D., Hengl, T., Heuvelink, G., Kempen, B., Mulder, T. V. L., Olmedo, G. F., Poggio, L., Ribeiro, E., Thine Omuto, C., Yigini, Y., and others: Soil organic carbon mapping: GSOC map cookbook manual, FAO, 2017.

Calvo de Anta, R., Luís, E., Febrero-Bande, M., Galiñanes, J., Macías, F., Ortíz, R., and Casás, F.: Soil organic carbon in peninsular Spain: Influence of environmental factors and spatial distribution, Geoderma, 370, 114365, https://doi.org/10.1016/j.geoderma.2020.114365, 2020.

Carslaw, D. C. and Ropkins, K.: Openair - An R package for air quality data analysis, Environmental Modelling & Software, 27–28, 52–61, https://doi.org/10.1016/j.envsoft.2011.09.008, 2012.

Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., and Walter, C.:

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Digital mapping of GlobalSoilMap soil properties at a broad scale: A review, Geoderma, 409, 115567, https://doi.org/10.1016/j.geoderma.2021.115567, 2022.

Crowther, T. W., Todd-Brown, K. E. O., Rowe, C. W., Wieder, W. R., Carey, J. C., Machmuller, M. B., Snoek, B. L., Fang, S., Zhou, G., Allison, S. D., Blair, J. M., Bridgham, S. D., Burton, A. J., Carrillo, Y., Reich, P. B., Clark, J. S., Classen, A. T., Dijkstra, F. A., Elberling, B., Emmett, B. A., Estiarte, M., Frey, S. D., Guo, J., Harte, J., Jiang, L., Johnson, B. R., Kröel-Dulay, G., Larsen, K. S., Laudon, H., Lavallee, J. M., Luo, Y., Lupascu, M., Ma, L. N., Marhan, S., Michelsen, A., Mohan, J., Niu, S., Pendall, E., Peñuelas, J., Pfeifer-Meister, L., Poll, C., Reinsch, S., Reynolds, L. L., Schmidt, I. K., Sistla, S., Sokol, N. W., Templer, P. H., Treseder, K. K., Welker, J. M., and Bradford, M. A.: Quantifying global soil carbon losses in response to warming, Nature, 540, 104–108, https://doi.org/10.1038/nature20150, 2016.

Dietterich, T. G.: Ensemble Methods in Machine Learning, 1–15, https://doi.org/10.1007/3-540-45014-9_1, 2000.

Doblas-Miranda, E., Rovira, P., Brotons, L., Martínez-Vilalta, J., Retana, J., Pla, M., and Vayreda, J.: Soil carbon stocks and their variability across the forests, shrublands and grasslands of peninsular Spain, Biogeosciences, 10, 8353–8361, https://doi.org/10.5194/bg-10-8353-2013, 2013.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J.: A statistical explanation of MaxEnt for ecologists, Divers Distrib, 17, 43–57, https://doi.org/10.1111/j.1472-4642.2010.00725.x, 2011.

FAO and ITPS: Global Soil Organic Carbon Map (GSOCmap), 2018.

Fielding, A. H. and Bell, J. F.: A review of methods for the assessment of prediction errors in conservation presence/absence models, Environ Conserv, 24, 38–49, 1997.

Friedman, J. H. and Stuetzle, W.: Projection Pursuit Regression, J Am Stat Assoc, 76, 817, https://doi.org/10.2307/2287576, 1981.

Fryda T; Erin LeDell, N. G. S. A. A. F. A. C. C. C. T. K. T. N. P. A. M. K. M. M.: Package “h2o” Title R Interface for the “H2O” Scalable Machine Learning Platform, https://doi.org/10.32614/CRAN.package.h2o, 2024.

Gavilán-Acuña, G., Olmedo, G.F., Mena-Quijada, P., Guevara, M., Barría-Knopf, B., Watt, M.S.: Reducing the Uncertainty of Radiata Pine Site Index Maps Using an Spatial Ensemble of Machine Learning Models. Forests 12, 77. https://doi.org/10.3390/f12010077, 2021.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B.: Bayesian Data Analysis, Chapman and Hall/CRC, https://doi.org/10.1201/b16018, 2013.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R.: Google Earth Engine: Planetary-scale geospatial analysis for everyone, Remote Sens Environ, 202, 18–27, https://doi.org/10.1016/j.rse.2017.06.031, 2017.

Gray, J. M., Bishop, T. F. A., and Yang, X.: Pragmatic models for the prediction and digital mapping of soil properties in eastern Australia, Soil Research, 53, 24, https://doi.org/10.1071/SR13306, 2015.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Greenwell, B. M. and Boehmke, B. C.: Variable Importance Plots - An Introduction to the vip Package, R J, 12, 343–366, https://doi.org/10.32614/RJ-2020-013, 2020.

Gregorutti, B., Michel, B., and Saint-Pierre, P.: Correlation and variable importance in random forests, Stat Comput, 27, 659–678, https://doi.org/10.1007/s11222-016-9646-1, 2017.

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C. O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J. A., Ibelles Navarro, A. R., Loayza, V., Manueles, A. M., Mendoza Jara, F., Olivera, C., Osorio Hermosilla, R., Pereira, G., Prieto, P., Ramos, I. A., Rey Brina, J. C., Rivera, R., Rodríguez-Rodríguez, J., Roopnarine, R., Rosales Ibarra, A., Rosales Riveiro, K. A., Schulz, G. A., Spence, A., Vasques, G. M., Vargas, R. R., and Vargas, R.: No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America, Soil, 4, 173–193, https://doi.org/10.5194/soil-4-173-2018, 2018.

Harden, J. W., Hugelius, G., Ahlström, A., Blankinship, J. C., Bond-Lamberty, B., Lawrence, C. R., Loisel, J., Malhotra, A., Jackson, R. B., Ogle, S., Phillips, C., Ryals, R., Todd-Brown, K., Vargas, R., Vergara, S. E., Cotrufo, M. F., Keiluweit, M., Heckman, K. A., Crow, S. E., Silver, W. L., DeLonge, M., Nave, L. E., Bond-Lamberty, B., Lawrence, C. R., Loisel, J., Malhotra, A., Jackson, R. B., Ogle, S., Phillips, C., Ryals, R., Todd-Brown, K., Vargas, R., Vergara, S. E., Cotrufo, M. F., Keiluweit, M., Heckman, K. A., Crow, S. E., Silver, W. L., DeLonge, M., and Nave, L. E.: Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter, Glob Chang Biol, 24, e705–e718, https://doi.org/10.1111/gcb.13896, 2018.

Hengl, T., Jesus, J.M. de, Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE 12, e0169748. https://doi.org/10.1371/journal.pone.0169748, 2017.

Hengl, T., Kempen, B., Heuvelink, G., and Malone, B.: Package “GSIF” - Global Soil Information Facilities, 2020.

Hijmans, R. J.: Geographic Data Analysis and Modeling [R package raster version 3.6-28], 2024.

IGME, 1995: Mapa Geológico de España a escala 1:1.000.000 http://info.igme.es/cartografiadigital/geologica/Geologicos1MMapa.aspx?Id=Geologico1000_(1994).

IGN, 2012: Mapa de suelos de España: Escala 1:1.000.000 http://info.igme.es/cartografiadigital/geologica/Geologicos1MMapa.aspx?Id=Geologico1000_(1994).

Jenny, H.: Factors of soil formation; a sytem of quantitative pedology, 1941.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Jobbágy, E. G., Jackson, R. B., Processes, B., and Change, G.: The vertical distribution of soil organic carbon and its relation to climate and vegetation, Ecological Applications, 10, 423–436, https://doi.org/https://doi.org/10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2, 2000.

Keitt, T., Bivand, R., Pebesma, E., Rowlingson, B., and ‘Rgdal,’ P.: Bindings for the geospatial data abstraction library, 2012.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F.: World Map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift, 15, 259–263, https://doi.org/10.1127/0941-2948/2006/0130, 2006.

De la Rosa, D., Mayol, F., Moreno, D., Rosales, A., Castillo, V., Moreno, F., Cabrera, F., Colomer, J. C., Antoine, J., Masui, S., Brinkman, R., Horn, R., and Prange, N.: SEIS.NET: Sistema Español de Información de Suelos en Internet, Edafología, 8, 45–56, 2001.

Llorente, M., Rovira, P., Merino, A., Rubio, A., Turrión, M., Bad\'\ia, D., Romanya, J., and González, J. C. J. A.: The CARBOSOL Database: a georeferenced soil profile analytical database for Spain, https://doi.org/10.1594/PANGAEA.884517, 2018.

MAPA: Anuario dde Estadística de España 2020, Ministerio de Agricultura, Pesca y Alimentación, 2021.

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

McElreath, R.: Statistical Rethinking, Chapman and Hall/CRC, https://doi.org/10.1201/9781315372495, 2018.

McElreath, R.: rethinking: Statistical Rethinking book package, 2016, R package version, 1, 2020.

Meinshausen, N.: Quantile Regression Forests, Journal of Machine Learning Research, 7, 983–999, 2006.

Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., Martin, M. P., and Arrouays, D.: National versus global modelling the 3D distribution of soil organic carbon in mainland France, Geoderma, 263, 16–34, https://doi.org/10.1016/j.geoderma.2015.08.035, 2016.

Muñoz-Rojas, M., Jordán, A., Zavala, L. M., De la Rosa, D., Abd-Elmabod, S. K., and Anaya-Romero, M.: Organic carbon stocks in Mediterranean soil types under different land uses (Southern Spain), Solid Earth, 3, 375–386, https://doi.org/10.5194/se-3-375-2012, 2012.

Ninyerola, M., Pons, X., and Roure, J.: Atlas climático digital de la Península Ibérica: metodología y aplicaciones en bioclimatología y geobotánica, edited by: Universitat Autònoma de Barcelona, Barcelona, 2005.

Padarian, J. and McBratney, A. B.: A new model for intra- and inter-institutional soil data sharing, SOIL, 6, 89–94, https://doi.org/10.5194/soil-6-89-2020, 2020.

Pebesma, E. and Bivand, R. S.: Classes and Methods for Spatial Data: the sp Package, 2005.

Pebesma, E. J.: Multivariable geostatistics in S: the gstat package, Comput Geosci, 30, 683–691, 2004.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Zivot, E., Lestel, M., Balkissoon, K., and Wuertz, D.: PerformanceAnalytics: Econometric tools for performance and risk analysis, R package version, 1, 2014.

Phillips, S. J., Anderson, R. P., and Schapire, R. E.: Maximum entropy modeling of species geographic distributions, Ecol Modell, 190, 231–259, https://doi.org/10.1016/j.ecolmodel.2005.03.026, 2006.

Poeplau, C., Vos, C., and Don, A.: Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content, SOIL, 3, 61–66, https://doi.org/10.5194/soil-3-61-2017, 2017.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

R Foundation for Statistical Computing: R: A Language and Environment for Statistical Computing., 2022.

Rodríguez Martín, J. A., Álvaro-Fuentes, J., Gonzalo, J., Gil, C., Ramos-Miras, J. J., Grau Corbí, J. M., and Boluda, R.: Assessment of the soil organic carbon stock in Spain, Geoderma, 264, 117–125, https://doi.org/10.1016/j.geoderma.2015.10.010, 2016.

Rosell, R. A., Gasparoni, J. C., and Galantini, J. A.: Soil organic matter evaluation, in: Assessment methods for soil carbon, Lewis Publishers Boca Raton, 311–322, 2001.

Saatchi, S. S., Houghton, R. A., Dos Santos Alvalá, R. C., Soares, J. V., and Yu, Y.: Distribution of aboveground live biomass in the Amazon basin, Glob Chang Biol, 13, 816–837, https://doi.org/10.1111/j.1365-2486.2007.01323.x, 2007.

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

Searle, R., McBratney, A., Grundy, M., Kidd, D., Malone, B., Arrouays, D., Stockman, U., Zund, P., Wilson, P., Wilford, J., Van Gool, D., Triantafilis, J., Thomas, M., Stower, L., Slater, B., Robinson, N., Ringrose-Voase, A., Padarian, J., Payne, J., Orton, T., Odgers, N., O'Brien, L., Minasny, B., Bennett, J. M., Liddicoat, C., Jones, E., Holmes, K., Harms, B., Gray, J., Bui, E., and Andrews, K.: Digital soil mapping and assessment for Australia and beyond: A propitious future, Geoderma Regional, 24, e00359, https://doi.org/10.1016/j.geodrs.2021.e00359, 2021.

Serrano, F. S.: Deformaciones Recientes en el Centro de la Península Ibérica, 2000.

Shangguan, W., Dai, Y., Duan, Q., Liu, B., and Yuan, H.: A global soil data set for earth system modeling, J Adv Model Earth Syst, 6, 249–263, https://doi.org/10.1002/2013MS000293, 2014.

Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., and Dai, Y.: Mapping the global depth to bedrock for land surface modeling, J Adv Model Earth Syst, 9, 65–88, https://doi.org/10.1002/2016MS000686, 2017.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Smith, P., Soussana, J. F., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., Batjes, N. H., Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., Klumpp, K., van Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., and Klumpp, K.: How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal, Glob Chang Biol, 26, 219–241, https://doi.org/10.1111/gcb.14815, 2020.

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., Hong, S. Y., Rawlins, B. G., and Field, D. J.: Global soil organic carbon assessment, Glob Food Sec, 6, 9–16, https://doi.org/10.1016/j.gfs.2015.07.001, 2015.

USDA: Module 3. USDA Textural Classification Study Guide, in: Soil mechanics level 1, National Employee Development Staff, Soil Conservation Service, U.S. Department of Agriculture., Washington DC, 1987.

Vargas, R., Alcaraz-Segura, D., Birdsey, R., Brunsell, N. A., Cruz-Gaistardo, C. O., de Jong, B., Etchevers, J., Guevara, M., Hayes, D. J., Johnson, K., Loescher, H. W., Paz, F., Ryu, Y., Sanchez-Mejia, Z., and Toledo-Gutierrez, K. P.: Enhancing interoperability to facilitate implementation of REDD+: case study of Mexico, Carbon Manag, 8, 57–65, https://doi.org/10.1080/17583004.2017.1285177, 2017.

Varón-Ramírez, V.M., Araujo-Carrillo, G.A., Guevara Santamaría, M.A.: Colombian soil texture: building a spatial ensemble model. Earth System Science Data 14, 4719–4741, https://doi.org/10.5194/essd-14-4719-2022, 2022.

Villarreal, S., Guevara, M., Alcaraz-Segura, D., Brunsell, N. A., Hayes, D., Loescher, H. W., and Vargas, R.: Ecosystem functional diversity and the representativeness of environmental networks across the conterminous United States, Agric For Meteorol, 262, 423–433, https://doi.org/10.1016/j.agrformet.2018.07.016, 2018.

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li Liu, D., Simpson, M., McGowen, I., and Sides, T.: Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia, Ecol Indic, 88, 425–438, https://doi.org/10.1016/j.ecolind.2018.01.049, 2018a.

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., and Liu, D. L.: High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia, Science of the Total Environment, 630, 367–378, https://doi.org/10.1016/j.scitotenv.2018.02.204, 2018b.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences. 4th Edition. (International Geophysics), Elsevier, https://doi.org/10.1016/C2017-0-03921-6, 2019.

WRB-IUSS: World Reference Base for Soil Resources. 2014, FAO, Rome, Italy, 1–191 pp., 2014.

Yan, X. and Su, X.: Linear Regression Analysis: Theory and Computing, World Scientific, 2009.

https://doi.org/10.5194/essd-2024-431
Preprint. Discussion started: 12 November 2024

Yigini, Y., Olmedo, G. F., Reiter, S., Baritz, R., Viatkin, K., and Vargas, R.: Soil organic carbon mapping: Cookbook 2nd Edition, FAO, Rome, 220 pp., 2018.

Zhang, C. and Ma, Y.: Ensemble Machine Learning: Methods and applications, edited by: Zhang, C. and Ma, Y., Springer US, Boston, MA, VIII, 332 pp., https://doi.org/10.1007/978-1-4419-9326-7, 2012.