SCAP
SOCIEDADE DE CIÊNCIAS AGRÁRIAS DE PORTUGAL

Analysing the different interrelationships of soil organic carbon using machine learning approaches: Assessing the specific case of Portugal

Utilização de abordagens *machine learning* para analisar as diferentes inter--relações do carbono orgânico do solo: Avaliação do caso específico de Portugal

Vítor J.P.D. Martinho[1,*], Tiago C.B. Ramos[2], Nádia L. Castanheira[3], Carlos Cunha[4], António J.D. Ferreira[5], José L.S. Pereira[1] and María del Carmen Sánchez-Carreira[6]

[1] *School of Agriculture (ESAV) and CERNAS-IPV Research Centre, Polytechnic Institute of Viseu, Portugal*
[2] *MARETEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[3] *National Institute of Agricultural and Veterinary Research, IP (INIAV), Portugal*
[4] *School of Technology and Management (ESTGV), Polytechnic Institute of Viseu, Portugal*
[5] *Polytechnic Institute of Coimbra, School of Agriculture (ESAC) and CERNAS Research Centre, Portugal*
[6] *ICEDE Research Group, Applied Economics Department, Faculty of Economics and Business Sciences, CRETUS, Universidade de Santiago de Compostela, Spain*
(* *E-mail: vdmartinho@esav.ipv.pt)*
https://doi.org/10.19084/rca.40281

Received/recebido: 2025.02.11
Accepted/aceite: 2025.04.01

ABSTRACT

Given the importance of soil organic carbon (SOC) for sustainability, policymakers and researchers are particularly concerned with identifying the conditions that promote carbon storage in the soil. These assessments provide relevant support for the design of policy instruments aimed at increasing soil quality and its carbon sequestration capacity. The new technologies associated with the digital transition can bring relevant added value, namely through artificial intelligence methodologies, where machine learning approaches are important. In this context, this research aims to analyse the several interrelationships of SOC in the specific Portuguese context, with a focus on highlighting its main predictors and providing proposals for stakeholders (including policymakers). To achieve these objectives, statistics from the INFOSOLO database were considered and evaluated using machine learning algorithms to select the most important SOC predictors and identify accurate models. These interrelationships were quantified with cross-sectional regressions and optimisation models. The results obtained provide relevant information for the design of adjusted policy measures that promote sustainable practices and increase soil quality. Generally, Portuguese soils have low organic carbon content due to soil features, climate circumstances and land management. Adjusted management of agroforestry activities is possibly the easiest part to deal with in this context.

Keywords: INFOSOLO Database; Artificial Intelligence; Cross-Sectional Regressions; Optimisation Approaches.

INTRODUCTION

The levels of soil organic carbon (SOC) are fundamental to improve the quality of the soils (Emamgholizadeh *et al*., 2018), sustainability (Reddy & Shwetha, 2024) and mitigating the threats created by anthropogenic activities. In some cases, the SOC contributes to reduce the levels of pollutants in the groundwater (Alam *et al*., 2025). SOC is also considered an important predictor of soil macronutrients (Zolfaghari *et al*., 2020). The new challenges arising from the dynamics associated with climate change require adjusted agricultural practices (Birru *et al*., 2024). The cation exchange capacity (CEC) is another important variable to assess soil quality and also clay, silt, sand, pH, and SOC are, in some cases, used in modelling processes (Mishra *et al*., 2022). The potentially mineralisable nitrogen is also considered an important indicator of soil quality (Pacci *et al*., 2024).

The new technologies associated with the digital transition offer better conditions for the researchers (Pavlovic *et al*., 2024) and may bring relevant contributions to the assessments related to soil characteristics (Alqadhi *et al*., 2023), dynamics (Chen *et al*., 2024a), functions (Ramcharan *et al*., 2017) and monitoring (Vahedi, 2017). Several efforts have been made to improve the methodologies for predicting soil characteristics (Minasny *et al*., 2024) and this provides valuable insights to better support the farmers' decisions (Banger *et al*., 2019). Information and adjusted methods are crucial to promote a more sustainable development (Jiang *et al*., 2019), and effective agricultural planning (Romero *et al*., 2012). Better assessment conditions may support the farmers in the crop selection process, taking into account the specific soil particularities (Bhat *et al*., 2023), and provide relevant insights to related institutions (Samarinas *et al*., 2024), policymakers (Samarinas *et al*., 2023) and managers (Seydi *et al*., 2024). This is important for the design of adjusted policies and to promote eco-friendly strategies (Sanderman *et al*., 2018).

Considering artificial intelligence approaches, for Australian contexts, the most important predictors of SOC are soil depth, pH and geomorphological characteristics (Benke *et al*., 2020). In other frameworks, agricultural management practices, such as the adoption of grassland production systems and conservation agriculture, showed improved levels of SOC (Dal Ferro *et al*., 2018). The same positive effects on the organic carbon (OC) were obtained using no-tillage strategies (Rai *et al*., 2022). For Brazilian soils, the most relevant predictors are soil class, monthly mean temperature, precipitation, slope and vegetation indexes (Gomes *et al*., 2019). In other studies, biotic, hydrological and pedological features appear as relevant explanatory variables for soil carbon (Keskin *et al*., 2019), as well as earth observation factors (Le *et al*., 2021), vegetation, soil water (Xiong *et al*., 2014), and Sentinel-1 and Sentinel-2 information (Zhang *et al*., 2023).

The advantages of artificial intelligence methodologies relative to the traditional approaches are not unanimous and, in some circumstances, is suggested a combination of both methods to overcome the weaknesses of each technique (Bernardini *et al*., 2024). Additionally, the results obtained depend on the assumptions made (Szatmári & Pásztor, 2019). In any case, the added value of artificial intelligence for agricultural management has been highlighted in the scientific literature (Bhatt *et al*., 2024).

The preservation of soil and water quality is a concern for the international organisations and the European Union (EU) institutions. These concerns are present in the agenda of the United Nations, as well as in the policy instruments and measures of the Common Agricultural Policy (CAP) (Bancheri *et al*., 2024). This is particularly important, when the agricultural sector contributes significantly to greenhouse gas (GHG) emissions (Guan *et al*., 2023). Adjusted farming decisions may increase the potential of soil carbon sequestration. Assessing the soil-related dynamics involves multidisciplinary researchers (Tziolas *et al*., 2021) and is fundamental for a better knowledge of the soil potentialities (Tziolas *et al*., 2024).

Considering these current needs, this study aims to provide deeper insights into the assessment of SOC within the Portuguese context. These insights can be considered as support to improve soil policy instruments and measures, particularly those developed by Portuguese and EU institutions.

MACHINE LEARNING APPLICATIONS IN IDENTIFYING SOIL ORGANIC CARBON PREDICTORS: A REVIEW OF PUBLISHED STUDIES

The discussions about the real contributions of each land use and land cover for the GHG emissions are not unanimous, showing that the soil capacity to store carbon depends on decisions made during all productive processes, including in forest lands where the planting phase is decisive to prevent significant soil disruptions (Baggio-Compagnucci *et al*., 2022). Another related issue that has generated discussion is the use of the most adjusted methodologies to predict soil characteristics (Ma *et al*., 2021). These contexts highlight the importance of adjusted assessments about the impacts of the different land use, land cover and soil characteristics on the soil capacity to sequester carbon. The machine learning approaches, for example, may bring relevant added value to these evaluations, specifically supporting the identification of important predictors and accurate models. In any case, the accuracy of the results obtained with these methodologies may be affected by the presence of some factors, such as the existence of iron (Dai *et al*., 2024), the consideration of nutrient and soil structural stability indicators (Delahaie *et al*., 2024), soil water content (Lin *et al*., 2021) and management practices (Karunaratne *et al*., 2024). An understanding of the interrelationships between the different soil characteristics and its dynamics is crucial for effective soil management (Fang *et al*., 2024) and assessment. Soil is the biggest terrestrial pool of OC (Georgiou *et al*., 2022) and the levels of SOC along with the amounts of nitrogen are crucial for the vegetation dynamics (Peng *et al*., 2024). The preservation of soil quality (defined as the capacity of the soil to function (Taghipour *et al*., 2022)) is crucial to avoid soil degradation (Fathizad *et al*., 2020), promote a more sustainable development (Hou *et al*., 2020), mitigate climate change impacts (Hu *et al*., 2024), specifically in critical regions (Liang *et al*., 2024) and vulnerable conditions (Wang *et al*., 2021), and enhance the crop productivity (Sirsat *et al*., 2018).

The following machine learning methods and approaches are some of the methodologies considered by the scientific literature to assess the SOC content: random forest (Gholizadeh *et al*., 2020), extreme gradient boosting, support vector machine (Chen *et al*., 2024b) and artificial neural network (Sun *et al*., 2023). The random forest model has also been considered in other studies along with correlation analysis (Ma *et al*., 2024) and structural equation modelling (Wang *et al*., 2024). Other research has used partial least squares regression, artificial neural networks, support vector regression and random forest (Thabit *et al*., 2024).

The data availability is fundamental to carrying out an adjusted analysis of the SOC. Usually, the following data sources are considered by the related researches: remote sensing technologies, and socio-economic, soil texture and topographic databases (Chen *et al*., 2024b); digital camera and Sentinel-2 remote sensor (Gholizadeh *et al*., 2020); large-scale field observation and published deep permafrost SOC data (Wei *et al*., 2022); Landsat-8, Sentinel-2 and Sentinel-3 (Zhou *et al*., 2021); legacy soil databases such as INFOSOLO (Ramos *et al*., 2017).

Between the variables considered to explain the levels of SOC appear the following: elevation and clay (Chen *et al*., 2024b); climate factors, clay, sand and silt (Hosseinpour-Zarnaq *et al*., 2024); soil organic matter (SOM) and climate factors (Kaushal & Baishya, 2024); annual precipitation (Li *et al*., 2024); mean annual temperature (Ma *et al*., 2024); cropping system, climate, soil characteristics and vegetation index (Ou *et al*., 2024); climate variables, pH, coarse fragments and land cover (Rial *et al*., 2017); soil depth, surface temperature and elevation (Sothe *et al*., 2022); vegetation and topography (Wadoux *et al*., 2023); mean annual air temperature and Normalized Difference Vegetation Index (Wu *et al*., 2022). Soil bulk density is another example of a soil variable interrelated with OC (Shi *et al*., 2023). Additionally, the age of the vegetation and mean annual precipitation were identified as important explanatory variables of the SOC in the recovery of soils affected by mining activities (Zhu *et al*., 2024).

The soil's potential to store carbon varies depending on the specifics of each location. In the Chinese context, the mountain areas with forests have higher levels of OC than the coastal and plain regions with crops (Chen *et al*., 2024b). The expected rise of the sea level in Australian regions may contribute to increase the of blue carbon with the migration of some ecosystems for the lands (Duarte de Paula Costa *et al*., 2021). In the Moroccan forest, the level of SOC varies with the soil characteristics and the species considered (El Mderssa *et al*., 2024). The potential of the soil to sequester carbon is region-specific (Padarian *et al*., 2022) and is crucial to maintain the soil quality (Song *et al*., 2020). The levels of SOM and pH are also relevant indicators of soil quality (Suleymanov *et al*., 2023).

MATERIAL AND METHODS

To provide insights into the assessment of soil organic carbon (SOC) within the Portuguese context, the INFOSOLO database was used. INFOSOLO provide a compilation of soil data relevant to Portugal. The database aggregates information derived from soil surveys, research studies, and various projects, offering an extensive dataset encompassing thousands of soil horizons/layers, and soil profiles (Ramos *et al*., 2017). This database has the added value of providing information that can be used as support by researchers, policymakers and organisations related to soil management. However, some limitations are acknowledged, which have to do with the absence of specific variables in certain soil horizons/layers.

The soil data available in this database was analysed through machine learning approaches, specifically to identify the most important predictors of the SOC and the most accurate models, following the IBM SPSS Modeler software (IBM SPSS Modeler, 2025) procedures. In this software, the accuracy levels of each model are evaluated using the relative error. Through cross-section regression methodologies (with robust standard error, for standard error type), the relationships between the most important covariates and the SOC were quantified, following the suggestions presented by the Stata software (Stata, 2025; StataCorp, 2023a, 2023b). To find optimised results, linear programming models were considered and the procedures proposed by the LINGO software (LINGO, 2025) were taken into account. For the data analysis and results interpretation about the soil characteristics and dynamics, the World Reference Base (WRB) for Soil Resources document was also consulted (IUSS Working Group WRB, 2022).

RESULTS AND DISCUSSION

This section is organised into two subsections, one for data analysis and the other for modelling.

INFOSOLO SOC data

Histosols (soils with relevant organic layers), Umbrisols (soils with great accumulation of organic matter in the topsoil), Leptosols (shallow soils with limitations for the vegetation growth and with great quantity of coarse sediments), Anthrosols (with strong anthropogenic impact namely through agricultural practices) and Solonchaks (presence of great quantities of soluble salts) (IUSS Working Group WRB, 2022) are the soil groups with the highest organic soil content (Table 1). This information highlights the influence of soil characteristics on the SOC levels. It should be noted that all the information available in the database for all horizons was used together for the analyses carried out in this study.

The qualifiers (considered for a more detailed explanation of soils within each soil group) with the

Table 1 - Top 20 Reference Soil Groups with the highest percentage of OC content

Reference Soil Groups	Organic carbon content mean (%)	Organic carbon content count (number of observations)
Histosols	8.20	8
Umbrisols	2.67	176
Leptosols	2.65	200
Anthrosols	2.16	1110
Solonchaks	1.95	11
Regosols	1.73	1265
Fluvisols	1.54	677
Cambisols	1.52	1674
Acrisols	1.22	148
Ferralsols	1.17	37
Alisols	0.81	69
Podzols	0.72	90
Gleysols	0.68	58
Calcisols	0.67	281
Vertisols	0.61	320
Luvisols	0.50	1025
Arenosols	0.45	227
Planosols	0.44	77
Solonetz	0.32	39
Plinthosols	0.28	11

highest percentage of OC content are presented in Tables 2, 3 and 4. In general, the top 5 qualifiers are those exhibited following: Anthric (has anthropogenic impacts); Hyperhumic (with more than 5% of SOC); Umbric (surface horizon rich in OC); Epieutric (≥ 50% of base saturation in superficial horizons); Hypersalic (very salty salic horizon); Humic (with more than 1% of SOC); Aric (with a superficial horizon showing disturbance, usually by ploughing); Escalic (terraces with truncated or transported soil); Hyperdystric (soils with too much exchangeable Al); Siltic (silt or silt loam texture); Dystric (low base saturation); and Sodic (having a subsurface horizon high in exchangeable sodium and magnesium) (IUSS Working Group WRB, 2022).

Generally, the forest and grasslands are the land uses with the largest levels of OC content, however, Table 5 reveals that it is not easy to identify a clear pattern for the different types of agroforestry activities (forestry, arable crops, horticulture, permanent crops, grassland, etc). These findings

Table 3 - Top 20 qualifiers 2 with the highest percentage of OC content

Qualifier2	Organic carbon content mean (%)	Organic carbon content count (number of observations)
Hyperhumic	6.12	39
Aric	2.69	2
Humic	2.32	450
Hyperdystric	2.24	1724
Escalic	2.23	717
Episkeletic	1.88	4
Regic	1.76	55
Dystric	1.65	885
Gleyic	1.62	19
Plinthic	1.38	31
Endodystric	1.23	2
Leptic	1.17	2
Endoleptic	1.10	17
Tecnic	1.09	2
Alcalic	1.04	3
Endogleyic	1.03	17
Sodic	0.99	97
Endoarenic	0.86	7
Hyperdystic	0.86	2
Eutric	0.76	320

Table 2 - Top 20 qualifiers 1 with the highest percentage of OC content

Qualifier1	Organic carbon content mean (%)	Organic carbon content count (number of observations)
Anthric	5.62	5
Hyperhumic	5.56	119
Umbric	4.28	50
Epieutric	3.41	9
Hypersalic	2.81	7
Leptic	2.30	128
Lithic	2.22	21
Plaggic	2.20	1060
Humic	2.03	1892
Colluvic	2.02	156
Hyposalic	1.92	31
Epileptic	1.77	113
Escalic	1.52	28
Alumic	1.41	19
Endoleptic	1.30	221
Hortic	1.19	4
Vetic	1.17	37
Sodic	0.98	24
Aric	0.88	50
Epidystric	0.83	68

Table 4 - Top 20 qualifiers 3 with the highest percentage of OC content

Qualifier3	Organic carbon content mean (%)	Organic carbon content count (number of observations)
Hyperdystric	2.52	885
Escalic	2.22	389
Siltic	1.88	138
Dystric	1.84	397
Sodic	1.74	14
Endoarenic	1.70	14
Endoskeletic	1.61	84
Skeletic	1.42	167
Chromic	1.39	756
Geoabruptic	1.34	9
Rhodic	1.11	39
Arenic	0.93	89
Novic	0.83	3
Clayic	0.74	136
Episkeletic	0.73	28
Epiarenic	0.62	4
Eutric	0.60	67
Hypereutric	0.51	171
Ochric	0.51	5
Pellic	0.50	70

Table 5 - Land uses with the highest percentage of OC content

Land Use	Organic carbon content mean (%)	Organic carbon content count (number of observations)
Floriculture and ornamental plants	6.89	1
Rocks and stones	5.41	6
Pine dominated mixed woodland	5.01	19
Mixed woodland	4.49	16
Shrubland without tree cover	4.25	76
Pine dominated coniferous woodland	3.69	66
Coniferous woodland	3.56	47
Shrubland with sparse tree cover	3.21	80
Durum wheat	2.89	2
Other fruit trees and berries	2.88	6
Mediterranean woodland	2.87	172
Permanent industrial crops	2.86	6
Broadleaved woodland	2.80	338
Other coniferous woodland	2.64	3
Apple fruit	2.46	4
Nuts trees	2.40	21
Pasture	2.29	514
Grassland without tree/shrub cover	2.26	132
Grassland with sparse tree/shrub cover	2.20	52
Maize	2.15	27
Other mixed woodland	2.14	1
Eucalypt forest	2.12	45
Non built-up linear features	2.12	5
Spontaneously vegetated surfaces	2.04	91
Irrigated crop	2.01	207
Bare land	1.88	8
Irrigated arable crop	1.84	1145
Horticulture	1.83	352
Other bare soil	1.83	12
Other fresh vegetables	1.83	4
Pine forest	1.78	380
Temporary grasslands	1.76	13
Oranges	1.73	3
Non built-up area features	1.71	3
Fallow	1.67	1141
Golf course	1.61	3
Other leguminous and mixtures for fodder	1.54	2
Mixed cereals for fodder	1.51	9
Olive groves	1.48	119
Rape and turnip rape	1.48	1
Arable crop	1.47	50
Triticale	1.45	1
Vineyard	1.32	368
Rice	1.25	22
Sunflower	1.22	12
Dry pulses	1.14	3
Common wheat	1.11	9
Rainfed crop	1.06	39
Barley	1.00	5
Oats	1.00	19
Potatoes	0.99	9
Vineyards	0.99	43
Fruit trees	0.98	109
Rainfed arable crop	0.96	1188
Chestnut forest	0.95	29
Rye	0.91	7
Olive grove	0.89	583
Pear fruit	0.85	6
Other cereals	0.84	2
Quercus forest	0.82	91
Other root crops	0.80	1
Mixed crops	0.73	16
Almond	0.69	31
Forest	0.69	418
Cherry fruit	0.62	2
Tomatoes	0.51	2
Cotton	0.48	9
Sugarbeet	0.48	47
Cedars	0.41	3
Melon	0.35	12

confirm that the soil capacity to store carbon depends on a multiplicity of factors, with land use being one of the key influences. These outputs suggest, as shown in the literature review, that the variables with influence on the levels of SOC are local-specific, where various factors are complexly interrelated.

In any case, when we consider together the soil features and the land use types (Table 6), the soil groups highlighted in Table 1 and the qualifiers identified in Tables 2, 3 and 4, combined with fallow and forest land, pasture and permanent crops emerge as the soil characteristics with the highest OC content.

layer and elevation (Table 7). Nitrogen has an importance of about 80%, assuming, in this way, the most relevant potential to predict the SOC in Portuguese soils. These results emphasise the importance of SOC and nitrogen for soil quality and are in line with the findings identified in the literature review (Pacci *et al*., 2024). In fact, it is known the interdependency of soil organic carbon with nitrogen. When CO_2 increases, it promotes net primary productivity, resulting in a higher C:N ratio and lower mineralisation rates, due to nitrogen fixation in the biomass and nitrogen depletion in the soil (Tashi *et al*., 2016). The importance of a balanced ratio of carbon/nitrogen to guarantee soil fertility and its capacity for carbon storage is also recognised.

Table 6 - Top 20 soil characteristics and land uses with the highest percentage of OC content

Reference Soil Groups	Qualifier1	Qualifier2	Qualifier3	Parent Material	Land Use	Organic carbon content mean (%)	Organic carbon content count (number of observations)
Fluvisols	Hyperhumic	Hyperdystric		Unconsolidated deposits	Fallow	10.45	3
Histosols	Epieutric			Organic materials		9.54	3
Leptosols	Hyperhumic	Hyperdystric		Metamorphic rocks	Mediterranean woodland	9.11	2
Leptosols	Lithic	Hyperhumic	Hyperdystric	Metamorphic rocks	Forest	9.06	1
Histosols	Epidystric			Organic materials		8.84	3
Umbrisols	Epileptic	Hyperhumic	Hyperdystric	Metamorphic rocks	Mediterranean woodland	8.15	2
Cambisols	Hyperhumic	Eutric	Chromic	Metamorphic rocks	Fallow	8.14	2
Leptosols	Umbric	Hyperdystric		Igneous rocks	Fallow	7.95	2
Leptosols	Umbric	Hyperhumic	Hyperdystric	Metamorphic rocks	Pine forest	7.78	1
Umbrisols	Endoleptic	Hyperhumic	Hyperdystric	Igneous rocks	Fallow	7.74	3
Leptosols	Umbric	Hyperdystric		Metamorphic rocks	Pine forest	7.40	2
Regosols	Leptic	Hyperhumic	Hyperdystric	Igneous rocks	Pasture	7.24	1
Regosols	Colluvic	Hyperhumic	Hyperdystric	Unconsolidated deposits	Pasture	7.18	4
Leptosols	Umbric	Hyperhumic	Hyperdystric	Metamorphic rocks	Mediterranean woodland	7.08	1
Fluvisols	Umbric	Hyperdystric		Unconsolidated deposits	Pasture	7.00	3
Leptosols	Hyperhumic	Hyperdystric		Igneous rocks	Mediterranean woodland	6.98	3
Regosols	Aric	Hyperhumic	Hyperdystric	Metamorphic rocks	Eucalypt forest	6.97	2
Regosols	Hyperhumic	Hyperdystric		Igneous rocks	Mediterranean woodland	6.92	8
Leptosols	Lithic	Hyperhumic	Hyperdystric	Metamorphic rocks	Mediterranean woodland	6.88	2
Umbrisols	Hyperhumic	Hyperdystric		Unconsolidated deposits	Olive grove	6.86	4

Modelling

The most important predictors of the OC content in the Portuguese context, identified using IBM SPSS Modeler software (IBM SPSS Modeler, 2025) procedures, are the total nitrogen content, cation exchange capacity, coarse sand, lower limit of soil horizon/layer, pH, upper limit of soil horizon/

The most accurate models are linear regressions, C&R (classification and regression) tree, linear XGBoost (gradient boosting algorithm considering a linear model as the base), CHAID (Chi-squared Automatic Interaction Detection) and LSVM (linear support vector machine) (Table 8).

Table 7 - The most important variables to predict the OC content

Nodes	Importance
N	0.776
CEC	0.073
CS	0.049
Land_Use	0.028
Hor_bot	0.019
pH	0.011
Hor_top	0.010
Qualifier 1	0.009
Qualifier 2	0.006
Z	0.006

N, Total nitrogen content (g/kg); CEC, Cation exchange capacity (cmolc/kg); CS (%), Coarse sand (2.0-0.2 mm); Hor_bot, Lower limit of soil horizon/layer (cm); pH, Soil reaction; Hor_top, Upper limit of soil horizon/layer (cm); Z, Elevation.

Table 8 - The most accurate models to predict the OC content

Model	Build Time (mins)	Correlation	Number Fields Used	Relative Error
Linear	2	0.871	28	0.241
C&R Tree	2	0.872	20	0.239
XGBoost Linear	2	0.880	32	0.227
CHAID	2	0.882	18	0.223
LSVM	2	0.897	32	0.195

The SOC content (%), in the Portuguese context and considering the statistical information from the INFOSOLO database, has a positive and statistically significant correlation with the following variables (Figure 1): Si (Silt (0.020-0.002 mm) – weak correlation); C (Clay (<0.002 mm) – weak correlation); N (Total nitrogen content – strong correlation); P (Extractable phosphorus content – medium correlation); K (Extractable potassium content – medium correlation); pH (Soil reaction – weak correlation); $CaCO_3$ (Total carbonate content – weak correlation); Ca_ex (Exchangeable Ca^{2+} content – weak correlation); K_ex (Exchangeable K+ content – weak correlation); CEC (Cation exchange capacity – weak correlation); Theta_FC (Soil water content at field capacity – weak correlation); and Theta_WP (Soil water content at wilting point – weak correlation). The correlation is negative and statistically significant in the following cases (Figure 1): Hor_top (Upper limit of soil horizon/layer – medium correlation); Hor_bot (Lower limit of soil horizon/layer – medium correlation); CS (Coarse sand (2.0-0.2 mm) – weak correlation); BD (Dry bulk density – medium correlation); and Na_ex (Exchangeable Na+ content – weak correlation). These results confirm the strong correlation between SOC and nitrogen, but also with the phosphorus and potassium.

Considering the results presented before in this section and the objectives proposed for this research, it was considered pertinent to include the variables presented in Table 10 (this table provides information on the number of observations and other statistics related to the variables considered) to carry out a linear regression with the OC content as the dependent variable. Nonetheless, results for relevant statistical tests were obtained and presented in Table 9. The VIF shows the absence of relevant multicollinearity, but the Breusch–Pagan/Cook–Weisberg test reveals the presence of heteroscedasticity. To deal with the heteroscedasticity, it was considered a linear regression with robust

Table 9 - Specification tests for linear regression

Specification tests	Results
Mean VIF (variance inflation factor)	3.970
Breusch–Pagan/Cook–Weisberg test for heteroscedasticity	8513.360 [0.000]

Table 10 - Summary statistics for variables related to soil characteristics and geographical features

Variable	Observations	Mean	Standard Deviation	Min	Max
OC	9361	2	2	0	24
N	7213	1	1	0	13
CEC	9443	14	9	0	65
CS	11342	33	19	0	99
Hor_bot	11342	57	40	2	400
pH	11124	6	1	3	10
Hor_top	11342	31	35	0	270
Z	11342	245	228	0	1880

	OC	Longitude	Latitude	Z	Hor_top	Hor_bot	Coarse	CS	FS	Si	C	BD	EC	N	P	K	pH	CaCO3	Ca_ex	Mg_ex	K_ex	Na_ex	CEC
OC	1.000																						
	.																						
Longitude	0.075	1.000																					
	(0.220)	.																					
Latitude	-0.088	0.3787*	1.000																				
	(0.154)	(0.000)	.																				
Z	-0.099	0.5860*	0.7509*	1.000																			
	(0.108)	(0.000)	(0.000)	.																			
Hor_top	-0.6303*	0.000	0.031	0.057	1.000																		
	(0.000)	(1.000)	(0.610)	(0.350)	.																		
Hor_bot	-0.5065*	0.011	-0.1490*	-0.081	0.8462*	1.000																	
	(0.000)	(0.863)	(0.015)	(0.188)	(0.000)	.																	
Coarse	0.019	-0.004	0.016	-0.044	-0.047	-0.1705*	1.000																
	(0.759)	(0.948)	(0.801)	(0.472)	(0.442)	(0.005)	.																
CS	-0.2687*	0.091	0.2918*	0.2044*	-0.044	-0.1730*	0.089	1.000															
	(0.000)	(0.137)	(0.000)	(0.001)	(0.471)	(0.005)	(0.145)	.															
FS	-0.102	-0.2651*	0.2576*	0.1740*	-0.2006*	-0.2876*	-0.045	0.1586*	1.000														
	(0.095)	(0.000)	(0.000)	(0.004)	(0.001)	(0.000)	(0.465)	(0.010)	.														
Si	0.3011*	-0.030	-0.2501*	-0.2150*	-0.042	0.020	0.111	-0.7024*	-0.3070*	1.000													
	(0.000)	(0.631)	(0.000)	(0.000)	(0.494)	(0.743)	(0.072)	(0.000)	(0.000)	.													
C	0.2018*	0.1524*	-0.3404*	-0.2252*	0.1914*	0.3514*	-0.1264*	-0.7076*	-0.7237*	0.4728*	1.000												
	(0.001)	(0.013)	(0.000)	(0.000)	(0.002)	(0.000)	(0.039)	(0.000)	(0.000)	(0.000)	.												
BD	-0.4534*	-0.088	0.1390*	0.063	0.3252*	0.2808*	0.1304*	0.3752*	0.2881*	-0.3689*	-0.3663*	1.000											
	(0.000)	(0.152)	(0.023)	(0.306)	(0.000)	(0.000)	(0.033)	(0.000)	(0.000)	(0.000)	(0.000)	.											
EC	0.007	-0.3939*	-0.1827*	-0.2372*	-0.1774*	-0.2491*	0.029	0.075	0.3751*	-0.1503*	-0.2707*	0.1510*	1.000										
	(0.913)	(0.000)	(0.003)	(0.000)	(0.004)	(0.000)	(0.641)	(0.221)	(0.000)	(0.014)	(0.000)	(0.014)	.										
N	0.8297*	0.106	0.038	0.035	-0.6749*	-0.6272*	0.052	-0.1523*	0.050	0.2091*	0.020	-0.4208*	0.088	1.000									
	(0.000)	(0.084)	(0.538)	(0.572)	(0.000)	(0.000)	(0.398)	(0.013)	(0.419)	(0.001)	(0.746)	(0.000)	(0.151)	.									
P	0.5004*	0.066	-0.001	0.056	-0.6267*	-0.5364*	-0.040	0.1331*	0.1355*	-0.097	-0.1855*	-0.1812*	0.2868*	0.5680*	1.000								
	(0.000)	(0.286)	(0.982)	(0.366)	(0.000)	(0.000)	(0.514)	(0.030)	(0.027)	(0.115)	(0.002)	(0.003)	(0.000)	(0.000)	.								
K	0.4787*	0.1696*	0.041	-0.019	-0.3533*	-0.3322*	-0.040	-0.108	-0.1711*	0.050	0.2218*	-0.3155*	0.049	0.5415*	0.4163*	1.000							
	(0.000)	(0.006)	(0.509)	(0.759)	(0.000)	(0.000)	(0.521)	(0.079)	(0.005)	(0.412)	(0.000)	(0.000)	(0.428)	(0.000)	(0.000)	.							
pH	0.1816*	0.069	-0.4372*	-0.2508*	0.070	0.3091*	-0.1494*	-0.5071*	-0.4878*	0.4192*	0.6279*	-0.2186*	-0.1286*	0.042	0.084	0.1222*	1.000						
	(0.003)	(0.265)	(0.000)	(0.000)	(0.251)	(0.000)	(0.015)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.036)	(0.491)	(0.173)	(0.046)	.						
CaCO3	0.1471*	0.1320*	-0.2365*	-0.114	-0.042	0.1274*	0.071	-0.3670*	-0.1513*	0.2785*	0.2895*	-0.1806*	-0.054	0.1507*	0.1318*	0.116	0.4976*	1.000					
	(0.016)	(0.031)	(0.000)	(0.064)	(0.494)	(0.037)	(0.250)	(0.000)	(0.013)	(0.000)	(0.000)	(0.003)	(0.383)	(0.014)	(0.031)	(0.059)	(0.000)	.					
Ca_ex	0.3212*	0.1924*	-0.3251*	-0.1754*	0.054	0.2590*	-0.1895*	-0.6195*	-0.6476*	0.4930*	0.8372*	-0.4242*	-0.2586*	0.1645*	0.055	0.2317*	0.7831*	0.4595*	1.000				
	(0.000)	(0.002)	(0.000)	(0.004)	(0.379)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.007)	(0.367)	(0.000)	(0.000)	(0.000)	.				
Mg_ex	-0.017	0.038	-0.043	-0.100	0.2564*	0.3188*	-0.1944*	-0.3575*	-0.3683*	0.1894*	0.5815*	-0.054	-0.1769*	-0.1584*	-0.2892*	0.035	0.2325*	-0.1777*	0.4620*	1.000			
	(0.777)	(0.532)	(0.483)	(0.102)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	(0.000)	(0.379)	(0.004)	(0.010)	(0.000)	(0.568)	(0.000)	(0.004)	(0.000)	.			
K_ex	0.3921*	0.105	-0.070	-0.1226*	-0.3178*	-0.2170*	-0.1470*	-0.081	-0.1959*	0.042	0.2305*	-0.2345*	0.082	0.4684*	0.4126*	0.7462*	0.2553*	0.1573*	0.2984*	0.113	1.000		
	(0.000)	(0.087)	(0.257)	(0.045)	(0.000)	(0.000)	(0.016)	(0.187)	(0.001)	(0.491)	(0.000)	(0.000)	(0.181)	(0.000)	(0.000)	(0.000)	(0.000)	(0.010)	(0.000)	(0.065)	.		
Na_ex	-0.1967*	-0.3513*	-0.1684*	-0.2554*	0.2609*	0.2136*	-0.025	-0.1404*	-0.015	0.024	0.1633*	0.2106*	0.3054*	-0.2195*	-0.1935*	-0.103	0.085	-0.1849*	0.068	0.3791*	-0.032	1.000	
	(0.001)	(0.000)	(0.006)	(0.000)	(0.000)	(0.001)	(0.686)	(0.022)	(0.811)	(0.696)	(0.008)	(0.001)	(0.000)	(0.000)	(0.002)	(0.092)	(0.165)	(0.002)	(0.271)	(0.000)	(0.601)	.	
CEC	0.2660*	0.1840*	-0.3146*	-0.1705*	0.105	0.3129*	-0.1760*	-0.5834*	-0.6966*	0.4669*	0.8478*	-0.3858*	-0.3491*	0.082	-0.058	0.1656*	0.7057*	0.3219*	0.9244*	0.6209*	0.2588*	0.114	1.0
	(0.000)	(0.003)	(0.000)	(0.005)	(0.087)	(0.000)	(0.004)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.182)	(0.346)	(0.007)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.064)	.
V	0.083	0.106	-0.1487*	-0.094	0.115	0.1732*	-0.1391*	-0.4854*	-0.3016*	0.3664*	0.5267*	-0.1886*	0.046	0.064	0.025	0.2088*	0.4883*	0.4252*	0.6135*	0.2216*	0.1648*	0.1533*	0.38
	(0.176)	(0.085)	(0.015)	(0.128)	(0.060)	(0.005)	(0.023)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.450)	(0.300)	(0.688)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.007)	(0.012)	(0.0
Theta_FC	0.2205*	0.1350*	-0.3438*	-0.2175*	0.1795*	0.3266*	-0.042	-0.7541*	-0.6969*	0.6289*	0.9421*	-0.3873*	-0.2844*	0.045	-0.2183*	0.1558*	0.6236*	0.3238*	0.8186*	0.5310*	0.1527*	0.1506*	0.82
	(0.000)	(0.028)	(0.000)	(0.000)	(0.003)	(0.000)	(0.490)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.461)	(0.000)	(0.011)	(0.000)	(0.000)	(0.000)	(0.000)	(0.013)	(0.014)	(0.0
Theta_WP	0.2094*	0.1476*	-0.3422*	-0.2214*	0.1803*	0.3382*	-0.090	-0.7148*	-0.7168*	0.5318*	0.9765*	-0.3730*	-0.2738*	0.038	-0.1987*	0.1829*	0.6290*	0.3118*	0.8360*	0.5714*	0.1944*	0.1689*	0.84
	(0.001)	(0.016)	(0.000)	(0.000)	(0.003)	(0.000)	(0.143)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.533)	(0.001)	(0.003)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.006)	(0.0

Figure 1 - Spearman's rank correlation matrix between variables related to soil characteristics and geographical features. Note: *, statistically significant at 5%; The meaning of the acronyms is in the text.

standard error (for the standard error type) (Table 11). The findings from the Spearman's rank correlation matrix (Spearman, 1904) and the creation of Tables 9, 10 and 11 were obtained following Stata software procedures (Stata, 2025; StataCorp, 2023a, 2023b).

The coefficients of the independent variables considered reveal that these factors have a positive and statistically significant marginal impact on the SOC content in Portugal, with nitrogen having the strongest marginal effect. The exceptions are the lower limit of soil horizon/layer and pH with negative marginal impacts, and the upper limit of soil horizon/layer without statistical significance (Table 11).

Considering these results, the model obtained and presented in Table 11 was analysed through models of optimisation, following LINGO software (LINGO, 2025) suggestions. The results are those exhibited in Table 12. These findings reveal that in a hypothetical scenario, with the maximum values for the variables with positive marginal impacts and the minimum for those with negative effects, it could be possible to obtain 20% of OC content. This means that in the real world contexts, we should generally expect, values strongly below than this optimised result.

Figure 2, obtained through the IBM SPSS Modeler software, reveals that a random sample has an 84% probability of belonging to node 1 (for lower values of total nitrogen content) with a predicted OC content of 1.016%. The terminal node 6 has the highest predicted OC content (8.482%) and a random sample has a 2% probability of belonging to this node. The samples belonging to this node have higher values of total nitrogen content. This modelling process is considered among the most accurate models to predict the OC content (Table 8).

Table 11 - Linear regression results with robust standard error (for standard error type) to deal with heteroscedasticity

OC	Coefficient	Robust Standard Error	t	P>t
N	1.355	0.024	56.530	0.000
CEC	0.023	0.002	12.480	0.000
CS	0.006	0.001	10.690	0.000
Hor_bot	-0.002	0.001	-2.490	0.013
pH	-0.067	0.012	-5.760	0.000
Hor_top	0.001	0.001	1.300	0.195
Z	0.000	0.000	5.140	0.000
_constant	-0.249	0.087	-2.850	0.004

Table 12 - Results obtained through linear programming

Variable	Value	Reduced Cost
N	13	0
CEC	65	0
CS	99	0
Hor_bot	2	0
pH	3	0
Z	1880	0
Row	**Slack or Surplus**	**Dual Price**
1	20	1
2	0	1
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0

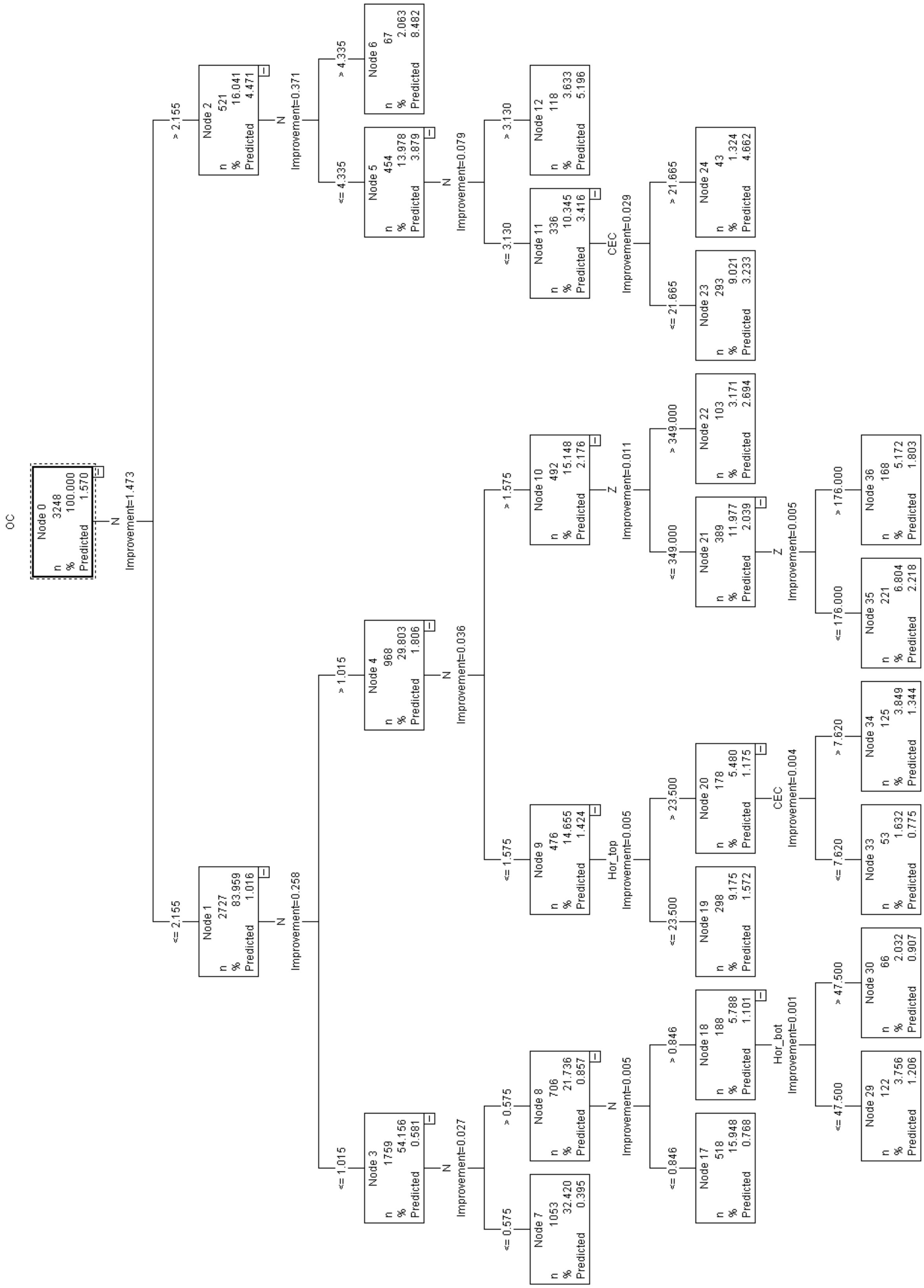

Figure 2 - Classification and regression (C&R) tree results with the OC content as the target.

CONCLUSIONS

The literature survey highlighted the importance of the SOC to preserve and improve soil quality, mitigating the impacts of human activities and addressing the challenges created by climate change. Other variables are also considered important indicators for assessing soil quality, such as cation exchange capacity, clay, silt, sand and pH. The new technologies associated with the digital Era open interesting potentialities to evaluate and manage the soil frameworks, nonetheless, improvements in the datasets available and in the methodologies adopted are still needed. The literature emphasises accurate models, relevant sources of information and the most important predictors of SOC. Depth, pH, agricultural practices, soil type, temperature, precipitation, slope and vegetation indexes are some of the relevant SOC predictors. The agricultural sector contributes significantly to GHG emissions, namely nitrous oxide (N_2O) and methane (CH_4), but multidisciplinary approaches and adjusted farming management may improve the soil's capacity to store carbon.

The data analysis reveals that the soil characteristics (identified by soil groups and qualifies) and the land use impact the levels of SOC in the Portuguese context. However, when all these indicators are considered together it is difficult to identify a general pattern, suggesting that soil capacity for the carbon sequestration is site specific. In any case, some soil groups (Histosols, Umbrisols, Leptosols, Anthrosols and Solonchaks) appear to be more prone to storage carbon. However, the location seems to be a decisive element as most of these soils are located in central and northern Portugal, where temperatures are lower, precipitation is higher, and in areas with florest and shrubland.

The results obtained with machine learning approaches and econometric methodologies show that the total nitrogen content is the most important predictor of the SOC in Portuguese soils. A balanced relationship between carbon and nitrogen is required to maintain equilibrium among the soil fertility and its capacity to sequester carbon. The process of mineralisation of the SOM is needed to improve the soil fertility, but reduce the levels of SOC stored (Veloso *et al*., 2022).

In terms of practical implications, referring that, in general, the soils in Portugal have a low level of OC content because of the soil characteristics, climate conditions and the agroforestry practices. The management of agroforestry land is perhaps the easiest part of this problem to control. In terms of policy recommendation, it is suggested to reinforce the national, European and international policy instruments that promote more sustainable farming practices, namely those that mitigate soil disruptions. It would also be relevant to improve the datasets available, enhance the national soil monitoring systems and involve the stakeholders in the process of carbon farming. For future research, it would be important to explore the INFOSOLO database through other approaches to benchmark with the results identified here.

ACKNOWLEDGMENTS

This work is funded by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project Refª UIDB/00681 (https://doi.org/10.54499/UIDP/00681/2020). Furthermore we would like to thank the CERNAS Research Centre and the Polytechnic Institute of Viseu for their support. This work was developed under the Science4Policy 2023 (S4P-23): annual science for policy project call, an initiative by PlanAPP - Competence Centre for Planning, Policy and Foresight in Public Administration in partnership with the Foundation for Science and Technology, financed by Portugal's Recovery and Resilience Plan.

REFERENCES

Alam, S.M.K.; Li, P.; Rahman, M.; Fida, M. & Elumalai, V. (2025) - Key factors affecting groundwater nitrate levels in the Yinchuan Region, Northwest China: Research using the eXtreme Gradient Boosting (XGBoost) model with the SHapley Additive exPlanations (SHAP) method. *Environmental Pollution*, vol. 364, n. 1, art. 125336. https://doi.org/10.1016/j.envpol.2024.125336

Alqadhi, S.; Mallick, J.; Talukdar, S. & Alkahtani, M. (2023) - An artificial intelligence-based assessment of soil erosion probability indices and contributing factors in the Abha-Khamis watershed, Saudi Arabia. *Frontiers in Ecology and Evolution*, vol. 11, art. 1189184. https://doi.org/10.3389/fevo.2023.1189184

Baggio-Compagnucci, A.; Ovando, P.; Hewitt, R.J.; Canullo, R. & Gimona, A. (2022) - Barking up the wrong tree? Can forest expansion help meet climate goals? *Environmental Science and Policy*, vol. 136, p. 237–249. https://doi.org/10.1016/j.envsci.2022.05.011

Bancheri, M.; Basile, A.; Terribile, F.; Langella, G.; Botta, M.; Lezzi, D.; Cavaliere, F.; Colandrea, M.; Marotta, L.; De Mascellis, R.; Manna, P.; Agrillo, A.; Mileti, F.A.; Acutis, M. & Perego, A. (2024) - A web-based operational tool for the identification of best practices in European agricultural systems. *Land Degradation and Development*, vol. 35, n. 13, p. 3965–3980. https://doi.org/10.1002/ldr.5114

Banger, K.; Nafziger, E.D.; Wang, J. & Pittelkow, C.M. (2019) - Modeling Inorganic Soil Nitrogen Status in Maize Agroecosystems. *Soil Science Society of America Journal*, vol. 83, n. 5, p. 1564–1574. https://doi.org/10.2136/sssaj2019.05.0140

Benke, K.K.; Norng, S.; Robinson, N.J.; Chia, K.; Rees, D.B. & Hopley, J. (2020) - Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. *Geoderma*, vol. 366, art. 114210. s://doi.org/10.1016/j.geoderma.2020.114210

Bernardini, L.G.; Rosinger, C.; Bodner, G.; Keiblinger, K.M.; Izquierdo-Verdiguier, E.; Spiegel, H.; Retzlaff, C.O. & Holzinger, A. (2024) - Learning vs. Understanding: When does artificial intelligence outperform process-based modeling in soil organic carbon prediction? *New Biotechnology*, vol. 81, p. 20–31. https://doi.org/10.1016/j.nbt.2024.03.001

Bhat, S.A.; Hussain, I. & Huang, N.-F. (2023) - Soil suitability classification for crop selection in precision agriculture using GBRT-based hybrid DNN surrogate models. *Ecological Informatics*, vol. 75, art. 102109. https://doi.org/10.1016/j.ecoinf.2023.102109

Bhatt, R.; Hossain, A.; Majumder, D.; Chandra, M.S.; Ghimire, R.; Faisal Shahzad, M.; Verma, K.K.; Riar, A.S.; Rajput, V.D.; Oliveira, M.W.; Nisi, A.; Almalki, R.S.; Bárek, V.; Brestic, M. & Maitra, S. (2024) - Prospects of artificial intelligence for the sustainability of sugarcane production in the modern era of climate change: An overview of related global findings. *Journal of Agriculture and Food Research*, vol. 18, art. 101519. https://doi.org/10.1016/j.jafr.2024.101519

Birru, G.; Shiferaw, A.; Tadesse, T.; Wardlow, B.; Jin, V.L.; Schmer, M.R.; Awada, T.; Kharel, T. & Iqbal, J. (2024) - Cover crop performance under a changing climate in continuous corn system over Nebraska. *Journal of Environmental Quality*, vol. 53, n. 1, p. 66–77. https://doi.org/10.1002/jeq2.20526

Chen, C.; Li, S.-L.; Chen, Q.-L.; Delgado-Baquerizo, M.; Guo, Z.-F.; Wang, F.; Xu, Y.-Y. & Zhu, Y.-G. (2024a) - Fertilization regulates global thresholds in soil bacteria. *Global Change Biology*, vol. 30, n. 8, art. e17466. https://doi.org/10.1111/gcb.17466

Chen, Q.; Wang, Y. & Zhu, X. (2024b) - Soil organic carbon estimation using remote sensing data-driven machine learning. *PeerJ*, vol. 12, n. 8, art. e17836. https://doi.org/10.7717/peerj.17836

Dai, Z.; Liu, X. & Ding, Y. (2024) - Iron-removal learning machine for multicolor determination of soil organic carbon. *Journal of Soils and Sediments*, vol. 24, n. 5, p. 2058–2067. https://doi.org/10.1007/s11368-024-03770-5

Dal Ferro, N.; Quinn, C. & Morari, F. (2018) - A Bayesian belief network framework to predict SOC dynamics of alternative management scenarios. *Soil and Tillage Research*, vol. 179, p. 114–124. https://doi.org/10.1016/j.still.2018.01.002

Delahaie, A.A.; Cécillon, L.; Stojanova, M.; Abiven, S.; Arbelet, P.; Arrouays, D.; Baudin, F.; Bispo, A.; Boulonne, L.; Chenu, C.; Heinonsalo, J.; Jolivet, C.; Karhu, K.; Martin, M.; Pacini, L.; Poeplau, C.; Ratié, C.; Roudier, P.; Saby, N.P.A.; Savignac, F. & Barré, P. (2024) - Investigating the complementarity of thermal and physical soil organic carbon fractions. *Soil*, vol. 10, n. 2, p. 795–812. https://doi.org/10.5194/soil-10-795-2024

Duarte de Paula Costa, M.; Lovelock, C.E.; Waltham, N. J.; Young, M.; Adame, M.F.; Bryant, C.V.; Butler, D.; Green, D.; Rasheed, M.A.; Salinas, C.; Serrano, O.; York, P.H.; Whitt, A.A. & Macreadie, P.I. (2021) - Current and future carbon stocks in coastal wetlands within the Great Barrier Reef catchments. *Global Change Biology*, vol. 27, n. 14, p. 3257–3271. https://doi.org/10.1111/gcb.15642

El Mderssa, M.; Elmalki, M.; Whalen, J.K.; Ikraoun, H.; Aliyat, F.Z.; Dallahi, Y.; Abbas, Y.; Nassiri, L. & Ibijbijen, J. (2024) - Forest stand and soil types determine soil organic carbon storage in the Middle Atlas region of Morocco using machine learning models. *All Earth*, vol. 36, n. 1, p. 1–10. https://doi.org/10.1080/27669645.2024.2400432

Emamgholizadeh, S.; Esmaeilbeiki, F.; Babak, M.; Zarehaghi, D.; Maroufpoor, E. & Rezaei, H. (2018) - Estimation of the organic carbon content by the pattern recognition method. *Communications in Soil Science and Plant Analysis*, vol. 49, n. 17, p. 2143–2154. https://doi.org/10.1080/00103624.2018.1499750

Fang, W.; Zhu, Y.; Liang, C.; Shao, S.; Chen, J.; Qing, H. & Xu, Q. (2024) - Deciphering differences in microbial community characteristics and main factors between healthy and root rot-infected *Carya cathayensis* rhizosphere soils. *Frontiers in Microbiology*, vol. 15, art. 1448675. https://doi.org/10.3389/fmicb.2024.1448675

Fathizad, H.; Ardakani, M.A.H.; Heung, B.; Sodaiezadeh, H.; Rahmani, A.; Fathabadi, A.; Scholten, T. & Taghizadeh-Mehrjardi, R. (2020) - Spatio-temporal dynamic of soil quality in the central Iranian desert modeled with machine learning and digital soil assessment techniques. *Ecological Indicators*, vol. 118, art. 106736. https://doi.org/10.1016/j.ecolind.2020.106736

Georgiou, K.; Jackson, R.B.; Vindušková, O.; Abramoff, R.Z.; Ahlström, A.; Feng, W.; Harden, J.W.; Pellegrini, A.F.A.; Polley, H.W.; Soong, J.L.; Riley, W.J. & Torn, M.S. (2022) - Global stocks and capacity of mineral--associated soil organic carbon. *Nature Communications*, vol. 13, n. 1, art. 3797. https://doi.org/10.1038/s41467-022-31540-9

Gholizadeh, A.; Saberioon, M.; Viscarra Rossel, R.A.; Boruvka, L. & Klement, A. (2020) - Spectroscopic measurements and imaging of soil colour for field scale estimation of soil organic carbon. *Geoderma*, vol. 357, art. 113972. https://doi.org/10.1016/j.geoderma.2019.113972

Gomes, L.C.; Faria, R.M.; de Souza, E.; Veloso, G.V.; Schaefer, C.E.G.R. & Filho, E.I.F. (2019) - Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma*, vol. 340, p. 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

Guan, K.; Jin, Z.; Peng, B.; Tang, J.; DeLucia, E.H.; West, P.C.; Jiang, C.; Wang, S.; Kim, T.; Zhou, W.; Griffis, T.; Liu, L.; Yang, W.H.; Qin, Z.; Yang, Q.; Margenot, A.; Stuchiner, E.R.; Kumar, V.; Bernacchi, C.; Coppess, J.; Novick, K.A.; Gerber, J.; Jahn, M.; Khanna, M.; Lee, D.; Chen, Z. & Yang, S.-J. (2023) - A scalable framework for quantifying field-level agricultural carbon outcomes. *Earth-Science Reviews*, vol. 243, art. 104462. https://doi.org/10.1016/j.earscirev.2023.104462

Hosseinpour-Zarnaq, M.; Moshiri, F.; Jamshidi, M.; Taghizadeh-Mehrjardi, R.; Tehrani, M.M. & Ebrahimi Meymand, F. (2024) - Monitoring changes in soil organic carbon using satellite-based variables and machine learning algorithms in arid and semi-arid regions. *Environmental Earth Sciences*, vol. 83, n. 20, art. 582. https://doi.org/10.1007/s12665-024-11876-9

Hou, D.; Bolan, N.S.; Tsang, D.C.W.; Kirkham, M.B. & O'Connor, D. (2020) - Sustainable soil use and management: An interdisciplinary and systematic approach. *Science of the Total Environment*, vol. 729, art. 138961. https://doi.org/10.1016/j.scitotenv.2020.138961

Hu, H.; Qian, C.; Xue, K.; Jörgensen, R.G.; Keiluweit, M.; Liang, C.; Zhu, X.; Chen, J.; Sun, Y.; Ni, H.; Ding, J.; Huang, W.; Mao, J.; Tan, R.-X.; Zhou, J.; Crowther, T.W.; Zhou, Z.-H.; Zhang, J. & Liang, Y. (2024) - Reducing the uncertainty in estimating soil microbial-derived carbon storage. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 121, n. 35, art. e2401916121. https://doi.org/10.1073/pnas.2401916121

IBM SPSS Modeler (2025) - *Software*. https://www.ibm.com/products/spss-modeler

IUSS Working Group WRB (2022) - *World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps*. 4th edition. International Union of Soil Sciences (IUSS), Vienna, Austria.

Jiang, R.; He, W.; Zhou, W.; Hou, Y.; Yang, J.Y. & He, P. (2019) - Exploring management strategies to improve maize yield and nitrogen use efficiency in northeast China using the DNDC and DSSAT models. *Computers and Electronics in Agriculture*, vol. 166, art. 104988. https://doi.org/10.1016/j.compag.2019.104988

Karunaratne, S.; Asanopoulos, C.; Jin, H.; Baldock, J.; Searle, R.; Macdonald, B. & Macdonald, L.M. (2024) - Estimating the attainable soil organic carbon deficit in the soil fine fraction to inform feasible storage targets and de-risk carbon farming decisions. *Soil Research*, vol. 62, n. 2, art. SR23096. https://doi.org/10.1071/SR23096

Kaushal, S. & Baishya, R. (2024) - Inclusive Indian Central Himalayan soil carbon estimates underscores significant inorganic carbon contribution and temporal dynamics: Implications for carbon sequestration. *Journal of Environmental Management*, vol. 372, art. 123312. https://doi.org/10.1016/j.jenvman.2024.123312

Keskin, H.; Grunwald, S. & Harris, W.G. (2019) - Digital mapping of soil carbon fractions with machine learning. *Geoderma*, vol. 339, p. 40–58. https://doi.org/10.1016/j.geoderma.2018.12.037

Le, N.N.; Pham; T.D.; Yokoya, N.; Ha, N.T.; Nguyen, T.T.T.; Tran, T.D.T. & Pham, T.D. (2021) - Learning from multimodal and multisensor earth observation dataset for improving estimates of mangrove soil organic carbon in Vietnam. *International Journal of Remote Sensing*, vol. 42, n. 18, p. 6866–6890. https://doi.org/10.1080/01431161.2021.1945158

Li, Q.; Chen, B.; Yuan, H.; Li, H. & Zhuang, S. (2024) - Characterization of controlling factors for soil organic carbon stocks in one Karst region of Southwest China. *PLoS One*, vol. 19, art. e0296711. https://doi.org/10.1371/journal.pone.0296711

Liang, B.; Wei, J.; Zhao, H.; Wu, S.; Hou, Y. & Zhang, S. (2024) - Mechanisms driving spatial and temporal changes in soil organic carbon stocks in saline soils in a typical county of the western Songnen Plain, northeast China. *Soil Research*, vol. 62, n. 1, art. SR23198. https://doi.org/10.1071/SR23198

Lin, L.; Wang, Y.; Liu, X. & Zhang, X. (2021) - Water-absorption-trough dewatering machine for estimation of organic carbon in moist soil. *Environmental Pollution*, vol. 284, art. 117445. https://doi.org/10.1016/j.envpol.2021.117445

LINGO (2025) - *Software*. https://www.lingo.com/

Ma, H.; Peng, M.; Yang, Z.; Yang, K.; Zhao, C.; Li, K.; Guo, F.; Yang, Z. & Cheng, H. (2024) - Spatial distribution and driving factors of soil organic carbon in the Northeast China Plain: Insights from latest monitoring data. *Science of the Total Environment*, vol. 911, art. 168602. https://doi.org/10.1016/j.scitotenv.2023.168602

Ma, Y.; Minasny, B.; McBratney, A.; Poggio, L. & Fajardo, M. (2021) - Predicting soil properties in 3D: Should depth be a covariate? *Geoderma*, vol. 383, art. 114794. https://doi.org/10.1016/j.geoderma.2020.114794

Minasny, B.; Bandai, T.; Ghezzehei, T.A.; Huang, Y.-C.; Ma, Y.; McBratney, A.B.; Ng, W.; Norouzi, S.; Padarian, J.; Sharififar, A.; Styc, Q. & Widyastuti, M. (2024) - Soil Science-Informed Machine Learning. *Geoderma*, vol. 452, art. 117094. https://doi.org/10.1016/j.geoderma.2024.117094

Mishra, G.; Sulieman, M.M.; Kaya, F.; Francaviglia, R.; Keshavarzi, A.; Bakhshandeh, E.; Loum, M.; Jangir, A.; Ahmed, I.; Elmobarak, A.; Basher, A. & Rawat, D. (2022) - Machine learning for cation exchange capacity prediction in different land uses. *Catena*, vol. 216, part A, art. 106404. https://doi.org/10.1016/j.catena.2022.106404

Ou, J.; Wu, Z.; Yan, Q.; Feng, X. & Zhao, Z. (2024) - Improving soil organic carbon mapping in farmlands using machine learning models and complex cropping system information. *Environmental Sciences Europe*, vol. 36, n. 1, art. 80. https://doi.org/10.1186/s12302-024-00912-x

Pacci, S.; Dengiz, O.; Alaboz, P. & Saygın, F. (2024) - Artificial neural networks in soil quality prediction: Significance for sustainable tea cultivation. *Science of the Total Environment*, vol. 947, art. 174447. https://doi.org/10.1016/j.scitotenv.2024.174447

Padarian, J.; Minasny, B.; McBratney, A. & Smith, P. (2022) - Soil carbon sequestration potential in global croplands. *PeerJ*, 10, art. e13740. https://doi.org/10.7717/peerj.13740

Pavlovic, M.; Ilic, S.; Ralevic, N.; Antonic, N.; Raffa, D. W.; Bandecchi, M. & Culibrk, D. (2024) - A Deep Learning Approach to Estimate Soil Organic Carbon from Remote Sensing. *Remote Sensing*, vol. 16, n. 4, art. 655. https://doi.org/10.3390/rs16040655

Peng, S.; Bao, N.; Wang, S.; Gholizadeh, A.; Saberioon, M. & Peng, Y. (2024) - Mapping vertical distribution of SOC and TN in reclaimed mine soils using point and imaging spectroscopy. *Ecological Indicators*, vol. 158, art. 111437. https://doi.org/10.1016/j.ecolind.2023.111437

Rai, T.; Kumar, S.; Nleya, T.; Sexton, P. & Hoogenboom, G. (2022) - Simulation of maize and soybean yield using DSSAT under long-term conventional and no-till systems. *Soil Research*, vol. 60, n. 6, p. 520–533. https://doi.org/10.1071/SR21042

Ramcharan, A.; Hengl, T.; Beaudette, D. & Wills, S. (2017) - A soil bulk density pedotransfer function based on machine learning: A case study with the NCSS soil characterization database. *Soil Science Society of America Journal*, vol. 81, n. 6, p. 1279–1287. https://doi.org/10.2136/sssaj2016.12.0421

Ramos, T.B.; Horta, A.; Gonçalves, M.C.; Pires, F.P.; Duffy, D. & Martins, J.C. (2017) - The INFOSOLO database as a first step towards the development of a soil information system in Portugal. *Catena*, vol. 158, p. 390–412. https://doi.org/10.1016/j.catena.2017.07.020

Reddy, B.S. & Shwetha, H.R. (2024) - Integrating Soil Spectral Library and PRISMA Data to Estimate Soil Organic Carbon in Crop Lands. *IEEE Geoscience and Remote Sensing Letters*, vol. 21, art. 2502005. https://doi.org/10.1109/LGRS.2024.3374824

Rial, M.; Martínez Cortizas, A. & Rodríguez-Lado, L. (2017) - Understanding the spatial distribution of factors controlling topsoil organic carbon content in European soils. *Science of the Total Environment*, vol. 609, p. 1411–1422. https://doi.org/10.1016/j.scitotenv.2017.08.012

Romero, C.C.; Hoogenboom, G.; Baigorria, G.A.; Koo, J.; Gijsman, A.J. & Wood, S. (2012) - Reanalysis of a global soil database for crop and environmental modeling. *Environmental Modelling and Software*, vol. 35, p. 163–170. https://doi.org/10.1016/j.envsoft.2012.02.018

Samarinas, N.; Tsakiridis, N.L.; Kalopesa, E. & Zalidis, G.C. (2024) - Soil Loss Estimation by Water Erosion in Agricultural Areas Introducing Artificial Intelligence Geospatial Layers into the RUSLE Model. *Land*, vol. 13, n. 2, art. 174. https://doi.org/10.3390/land13020174

Samarinas, N.; Tsakiridis, N.L.; Kokkas, S.; Kalopesa, E. & Zalidis, G.C. (2023) - Soil Data Cube and Artificial Intelligence Techniques for Generating National-Scale Topsoil Thematic Maps: A Case Study in Lithuanian Croplands. *Remote Sensing*, vol. 15, n. 22, art. 5304. https://doi.org/10.3390/rs15225304

Sanderman, J.; Hengl, T.; Fiske, G.; Solvik, K.; Adame, M.F.; Benson, L.; Bukoski, J.J.; Carnell, P.; Cifuentes--Jara, M.; Donato, D.; Duncan, C.; Eid, E. M.; Ermgassen, P.Z.; Lewis, C.J.E.; Macreadie, P.I.; Glass, L.; Gress, S.; Jardine, S.L.; Jones, T.G.; Nsombo, E.N.; Rahman, M.M.; Sanders, C.J.; Spalding, M. & Landis, E. (2018) - A global map of mangrove forest soil carbon at 30 m spatial resolution. *Environmental Research Letters*, vol. 13, n. 5, art. 055002. https://doi.org/10.1088/1748-9326/aabe1c

Seydi, S.T.; Abatzoglou, J.T.; AghaKouchak, A.; Pourmohamad, Y.; Mishra, A. & Sadegh, M. (2024) - Predictive Understanding of Links Between Vegetation and Soil Burn Severities Using Physics-Informed Machine Learning. *Earth's Future*, vol. 12, n. 8, art. e2024EF004873. https://doi.org/10.1029/2024EF004873

Shi, L.; O'Rourke, S.; de Santana, F.B. & Daly, K. (2023) - Prediction of soil bulk density in agricultural soils using mid-infrared spectroscopy. *Geoderma*, vol. 434, art. 116487. https://doi.org/10.1016/j.geoderma.2023.116487

Sirsat, M.S.; Cernadas, E.; Fernández-Delgado, M. & Barro, S. (2018) - Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods. *Computers and Electronics in Agriculture*, vol. 154, p. 120–133. https://doi.org/10.1016/j.compag.2018.08.003

Song, X.-D.; Wu, H.-Y.; Ju, B.; Liu, F.; Yang, F.; Li, D.-C.; Zhao, Y.-G.; Yang, J.-L. & Zhang, G.-L. (2020) - Pedoclimatic zone-based three-dimensional soil organic carbon mapping in China. *Geoderma*, vol. 363, art. 114145. https://doi.org/10.1016/j.geoderma.2019.114145

Sothe, C.; Gonsamo, A.; Arabian, J. & Snider, J. (2022) - Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. *Geoderma*, vol. 405, art. 115402. https://doi.org/10.1016/j.geoderma.2021.115402

Spearman, C. (1904) - The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, vol. 15, n. 1, p. 72–101. https://doi.org/10.2307/1412159

Stata (2025) - *Statistical software for data science*. https://www.stata.com/

StataCorp (2023a) - *Stata 18 Base Reference Manual [Computer software]*. Stata Press.

StataCorp (2023b) - *Stata Statistical Software: Release 18 [Computer software]*. StataCorp LLC.
Suleymanov, A.; Tuktarova, I.; Belan, L.; Suleymanov, R.; Gabbasova, I. & Araslanova, L. (2023) - Spatial prediction of soil properties using random forest, k-nearest neighbors and cubist approaches in the foothills of the Ural Mountains, Russia. *Modeling Earth Systems and Environment*, vol. 9, n. 3, p. 3461–3471. https://doi.org/10.1007/s40808-023-01723-4
Sun, Y.; Ma, J.; Zhao, W.; Qu, Y.; Gou, Z.; Chen, H.; Tian, Y. & Wu, F. (2023) - Digital mapping of soil organic carbon density in China using an ensemble model. *Environmental Research*, vol. 231, art. 116131. https://doi.org/10.1016/j.envres.2023.116131
Szatmári, G. & Pásztor, L. (2019) - Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. *Geoderma*, vol. 337, p. 1329–1340. https://doi.org/10.1016/j.geoderma.2018.09.008
Taghipour, K.; Heydari, M.; Kooch, Y.; Fathizad, H.; Heung, B. & Taghizadeh-Mehrjardi, R. (2022) - Assessing changes in soil quality between protected and degraded forests using digital soil mapping for semiarid oak forests, Iran. *Catena*, vol. 213, art. 106204. https://doi.org/10.1016/j.catena.2022.106204
Tashi, S.; Singh, B.; Keitel, C. & Adams, M. (2016) - Soil carbon and nitrogen stocks in forests along an altitudinal gradient in the eastern Himalayas and a meta-analysis of global data. *Global Change Biology*, vol. 22, n. 6, p. 2255–2268. https://doi.org/10.1111/gcb.13234
Thabit, F.N.; Negim, O.I.A.; AbdelRahman, M.A.E.; Scopa, A. & Moursy, A.R.A. (2024) - Using Various Models for Predicting Soil Organic Carbon Based on DRIFT-FTIR and Chemical Analysis. *Soil Systems*, vol. 8, n. 1, art. 22. https://doi.org/10.3390/soilsystems8010022
Tziolas, N.; Tsakiridis, N.; Chabrillat, S.; Demattê, J.A.M.; Ben-Dor, E.; Gholizadeh, A.; Zalidis, G. & van Wesemael, B. (2021) - Earth observation data-driven cropland soil monitoring: A review. *Remote Sensing*, vol. 13, n. 21, art. 4439. https://doi.org/10.3390/rs13214439
Tziolas, N.; Tsakiridis, N.; Heiden, U. & van Wesemael, B. (2024) - Soil organic carbon mapping utilizing convolutional neural networks and Earth observation data, a case study in Bavaria state Germany. *Geoderma*, vol. 444, art. 116867. https://doi.org/10.1016/j.geoderma.2024.116867
Vahedi, A.A. (2017) - Monitoring soil carbon pool in the Hyrcanian coastal plain forest of Iran: Artificial neural network application in comparison with developing traditional models. *Catena*, vol. 152, p. 182–189. https://doi.org/10.1016/j.catena.2017.01.022
Veloso, A.; Sempiterno, C.; Calouro, F.; Rebelo, F.; Pedra, F.; Castro, I.V.; Gonçalves, M. da C.; Marcelo, M. da E.; Pereira, P.; Fareleira, P.; Jordão, P.; Mano, R. & Fernandes, R. (2022) - *Manual de fertilização das culturas*. 3a Edição. Instituto Nacional de Investigação Agrária e Veterinária, I.P. - INIAV.
Wadoux, A.M.J.-C.; Saby, N.P.A. & Martin, M. P. (2023) - Shapley values reveal the drivers of soil organic carbon stock prediction. *Soil*, vol. 9, n. 1, p. 21–38. https://doi.org/10.5194/soil-9-21-2023
Wang, D.; Wu, T.; Zhao, L.; Mu, C.; Li, R.; Wei, X.; Hu, G.; Zou, D.; Zhu, X.; Chen, J.; Hao, J.; Ni, J.; Li, X.; Ma, W.; Wen, A.; Shang, C.; La, Y.; Ma, X. & Wu, X. (2021) - A 1km resolution soil organic carbon dataset for frozen ground in the Third Pole. *Earth System Science Data*, vol. 13, n. 7, p. 3453–3465. https://doi.org/10.5194/essd-13-3453-2021
Wang, Z.; Wang, G.; Li, Y. & Zhang, Z. (2024) - Determinants of carbon sequestration in thinned forests. *Science of the Total Environment*, vol. 951, art. 175540. https://doi.org/10.1016/j.scitotenv.2024.175540
Wei, Z.; Du, Z.; Wang, L.; Zhong, W.; Lin, J.; Xu, Q. & Xiao, C. (2022) - Sedimentary organic carbon storage of thermokarst lakes and ponds across Tibetan permafrost region. *Science of the Total Environment*, vol. 831, art. 154761. https://doi.org/10.1016/j.scitotenv.2022.154761
Wu, T.; Wang, D.; Mu, C.; Zhang, W.; Zhu, X.; Zhao, L.; Li, R.; Hu, G.; Zou, D.; Chen, J.; Wei, X.; Wen, A.; Shang, C.; La, Y.; Lou, P.; Ma, X. & Wu, X. (2022) - Storage, patterns, and environmental controls of soil organic carbon stocks in the permafrost regions of the Northern Hemisphere. *Science of the Total Environment*, vol. 828, art. 154464. https://doi.org/10.1016/j.scitotenv.2022.154464
Xiong, X.; Grunwald, S.; Myers, D.B.; Kim, J.; Harris, W.G. & Comerford, N.B. (2014) - Holistic environmental soil-landscape modeling of soil organic carbon. *Environmental Modelling and Software*, vol. 57, p. 202–215. https://doi.org/10.1016/j.envsoft.2014.03.004

Zhang, H.; Wan, L. & Li, Y. (2023) - Prediction of Soil Organic Carbon Content Using Sentinel-1/2 and Machine Learning Algorithms in Swamp Wetlands in Northeast China. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, p. 5219–5230. https://doi.org/10.1109/JSTARS.2023.3281732

Zhou, T.; Geng, Y.; Ji, C.; Xu, X.; Wang, H.; Pan, J.; Bumberger, J.; Haase, D. & Lausch, A. (2021) - Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. *Science of the Total Environment*, vol. 755, art. 142661. https://doi.org/10.1016/j.scitotenv.2020.142661

Zhu, Y.; Wang, L.; Ma, J.; Hua, Z.; Yang, Y. & Chen, F. (2024) - Assessment of carbon sequestration potential of mining areas under ecological restoration in China. *Science of the Total Environment*, vol. 921, art. 171179. https://doi.org/10.1016/j.scitotenv.2024.171179

Zolfaghari, A.A.; Abolkheiryan, M.; Soltani-Toularoud, A.A.; Taghizadeh-Mehrjardi, R. & Weldeyohannes, A.O. (2020) - Prediction of soil macronutrients using fractal parameters and artificial intelligence methods. *Spanish Journal of Agricultural Research*, vol. 18, n. 2, art. e1104. https://doi.org/10.5424/sjar/2020182-15460