Earth Science Informatics (2025) 18:323
https://doi.org/10.1007/s12145-025-01833-w

RESEARCH

Estimation of soil properties using machine learning techniques to improve hydrological modeling in a semiarid environment: Campo de Cartagena (Spain)

Francisco Alonso-Sarria[1] · Arantzazu Blanco-Bernardeau[2] · Francisco Gomariz-Castillo[1] · Helena Jiménez-Bastida[3] · Asunción Romero-Diaz[3]

Received: 26 November 2024 / Accepted: 28 February 2025

Abstract
Soils are a key element in the hydrological cycle through a number of soil properties that are complex to estimate and exhibit considerable spatial variability. Therefore, several techniques have been proposed for their estimation and mapping from point data along a given study area. In this work, four machine learning methods: Random Forest, Support Vector Machines, XGBoost and Multilayer Perceptrons, are used to predict and map the proportions of organic carbon, clay, silt and sand in the soils of the Campo de Cartagena (SE Spain). These models depend on a number of hyperparameters that need to be optimised to maximise accuracy, although this process can lead to overtraining, which affects the generalisability of the models. In this work it was found that neural networks gave the best results in validation, but on the test data the methods based on decision trees, random forest and xgboost were more accurate, although the differences were generally not significant. Accuracy values, as usual for soil variables, were not high. The RMSE values were 8.040 for SOC, 7.049 for clay, 10.227 for silt and 13.561 for loam. The layers obtained were then used to obtain annual curve number layers whose ability to reproduce runoff hydrographs was compared with the official CN layer. For high flow events, the CN layers obtained in this study gave better results (NSE=0.807, PBIAS=-4.7 and RMSE=0.4) than the official CN layers (NSE=-2.28, PBIAS=135.82 and RMSE=1.8).

Keywords Soil variables · Machine learning · Hyperparameter optimisation · Hydrological modelling

Introduction

Soil is an open system whose properties are determined by a wide variety of physical, chemical and biological processes. This system has important functions in ecosystems, including the storage of water resources, the regulation of water flows that determine the rate of evaporation, the recharge of aquifers and the generation of runoff. Accurate and spatially distributed estimations of their properties are therefore necessary for adequate monitoring of natural resources, water and land management and erosion forecasting (Lagacherie and McBratney 2006; Hartemink 2008; Miller 2012; Schirrmann et al. 2013; Minasny and McBratney 2016; Bobryk et al. 2016; Forkuor et al. 2017; Martínez-Hernández et al. 2017; Rodrigo-Comino et al. 2018; Ramirez-Lopez et al. 2019; Li et al. 2022).

The water behaviour of soils is determined by their physico-chemical properties. The spatial variability of such variables depends on the characteristics of the parent material, topography, climate, vegetation, weather and anthro-

Communicated by: Hassan Babaie

✉ Francisco Alonso-Sarria
alonsarp@um.es

Arantzazu Blanco-Bernardeau
aramucia@googlemail.com

Francisco Gomariz-Castillo
fjgomariz@um.es

Helena Jiménez-Bastida
helena.j.b@um.es

Asunción Romero-Diaz
arodi@um.es

[1] Instituto Universitario del Agua y del Medio Ambiente, Universidad de Murcia, Edificio D. Campus de Espinardo, Murcia 30100, Spain

[2] Kveloce, Plaza de la Reina 19 esc izqda 1B, Valencia 46002, Spain

[3] Departmento de Geografía, Universidad de Murcia, Santo Cristo 1, Murcia 30001, Spain

Published online: 11 March 2025

pogenic activities (Mulder et al. 2011; Umali et al. 2012; Lin et al. 2016) through different soil-landscape processes. These relationships were established by Jenny (1941).

In recent years, digital soil mapping techniques have been used to map soil properties using environmental variables as predictors (Dobos et al. 2006; Grimm and Behrens 2010; Taghizadeh-Mehrjardi et al. 2016; Forkuor et al. 2017; Minasny and Hartemink 2011; McBratney et al. 2003; Grunwald et al. 2011; Zhu et al. 2015; Arrouays et al. 2017; Zeraatpisheh et al. 2017; Meier et al. 2018; Heung et al. 2016; Sharififar et al. 2019; Taghizadeh-Mehrjardi et al. 2019; Zare et al. 2016; Rial et al. 2017).

The amount of organic matter in the soil is one of the most frequently modelled variables, in the form of soil organic carbon (SOC) (Martin et al. 2014; Masri et al. 2015; Were et al. 2015; Khan et al. 2015; Pinheiro et al. 2017; Rahman et al. 2018; Gomes et al. 2019; Emadi et al. 2020) or soil organic matter (SOM) (Byrne and Yang 2016; Khanal et al. 2018; Qi et al. 2018).

Among the physical properties modelled, texture (Sarmadian et al. 2013; Pahlavan-Rad and Akbarimoghaddam 2018), bulk density (Beguin et al. 2017; Bondi et al. 2018) or degree of flocculation (Pinheiro et al. 2017) might be highlighted.

The concentration of different elements has also been studied, e.g. Calcium (Pinheiro et al. 2017; Masri et al. 2015), Phosphorus (Masri et al. 2015; Wilson et al. 2016; Pinheiro et al. 2017; Hengl et al. 2017; Li et al. 2017), Magnesium (Pinheiro et al. 2017; Khanal et al. 2018), and Potassium (Khanal et al. 2018). Soil salinity (Wu et al. 2018), electrical conductivity (Ranjbar and Jalali 2016), cation exchange capacity (Pinheiro et al. 2017; Khanal et al. 2018; Sarmadian et al. 2013) and pH (Pinheiro et al. 2017; Khanal et al. 2018; Zhang et al. 2018; Pahlavan-Rad and Akbarimoghaddam 2018) are the main chemical properties analysed in the literature.

These soil property maps can be made for different objectives such as crop protection, weed detection, plant disease identification and integrated pest management (Behmann et al. 2015), yield and crop suitability forecasting (González-Sánchez et al. 2014; Khanal et al. 2018; Elavarasan et al. 2018), irrigation planning (Goldstein et al. 2018), soil temperature modelling (Bilgili et al. 2010; Kim and Singh 2014), development and evaluation of precision farming techniques (Hengl et al. 2017) or soil dryness assessment (Coopersmith et al. 2014).

The models used in digital soil mapping are trained with data on the soil properties to be estimated, measured at sites where other environmental attributes are available as predictors: topographic variables, climate, parent material, vegetation type and land use (Gessler et al. 1995). The result is a model that predicts the spatial distribution of soil properties (Minasny and Hartemink 2011; Zhu et al. 2006).

Despite concerns that the application of machine learning (ML) techniques may ignore soil science knowledge and produce misleading or erroneous results (Rossiter 2018), ML techniques have provided useful tools for soil mapping (Khaledian and Miller 2020).

Among the most classical models used in soil mapping are spline interpolation (Bilgili et al. 2010), geostatistical methods (Wälder et al. 2008), multiple linear regression (Sumfleth and Duttmann 2008; Bonfatti et al. 2016; da Silva Chagas et al. 2016; Angelini et al. 2017), generalised linear models (Karunaratne et al. 2014; Guisan and Harrell 2000) or generalised additive models (Poggio et al. 2010).

The first ML models used include K-nearest neighbours (KNN) (Mansuy et al. 2014; Taghizadeh-Mehrjardi et al. 2016), regression trees (Taghizadeh-Mehrjardi et al. 2014; Scull et al. 2005; Wiesmeier et al. 2011). Such models are quite simple and are considered nowadays slightly naive. However, ensembles of decision trees have been a common and powerful tool in the last years, e.g. Random Forest (RF) (Grimm et al. 2008; Akpa et al. 2014; Hengl et al. 2015; Forkuor et al. 2017; Pahlavan-Rad and Akbarimoghaddam 2018) and Boosting (Lemercier et al. 2012). Similar to decision trees are the rules based models, such as Cubist (Miller et al. 2015; Akpa et al. 2016; Rudiyanto et al. 2018; Taghizadeh-Mehrjardi et al. 2016). Other usual models are Support Vector Machines (SVM) (Yao et al. 2008; Kovačević et al. 2010; Pradhan 2013; Kavzoglu et al. 2014; Forkuor et al. 2017; Cai et al. 2010; Brungard et al. 2015), and Neural Networks (NN) (Behrens et al. 2005; Elshorbagy and Parasuraman 2008; Kalambukattu et al. 2018).

ML techniques are nowadays commonly used in several environmental applications, such us Climatology (Yang et al. 2024; Ruiz-Álvarez et al. 2019), Biogeography and Ecology (Cutler et al. 2007; Giménez-Casalduero et al. 2020), Hydrogeology (Baudron et al. 2013), Remote Sensing (Lary et al. 2016; Alonso-Sarria et al. 2024), etc.

Padarian et al. (2020) provide a comprehensive review of the application of ML techniques in soil science. They note an increase over time in the number of publications using ML to model various aspects of soils. They attribute this increase to a combination of increased computational power and access to high performance computers, increased data availability (e.g. remote sensing) and growing interest in data science. They also conclude that of the more than 100 different model variants that have been used in soil science, most have been used experimentally in one or two papers, and only a few have been used systematically. There is a general increase in the use of all models, but it is possible to see a proportional decrease in the use of some models such as SVM, splines and decision trees, giving way to more advanced alternatives such as RF. According to Padarian et al. (2020), these more advanced modelling techniques tend to produce better results than simpler, more traditional approaches. In another

extensive comparisons, Sirsat et al. (2018) compared 76 different algorithms, with regression tree ensembles coming out on top in predicting soil fertility indices. Other comparative studies also showed a consistent superior performance of ML methods (NN, SVM, RF) over simpler approaches (principal component regression or partial least squares regression).

Motia and Reddy (2021) conducted a similar meta-analysis, reviewing 57 research papers to analyse the contribution of different ML techniques in soil analysis. SVM, RF and NN and their variants were found to be the most widely used methods in soil analysis and prediction applications.

The selection of the most appropriate algorithm for a given set of resource constraints and desired level of uncertainty depends on the complexity of the algorithm's hyperparameters, the number of soil samples available, and the time and resources available for calibration. Unfortunately, many studies do not provide their rationale for the selection of a particular algorithm (Khaledian and Miller 2020). Another potential problem with many studies that use different machine learning techniques to show which one performs better is that they often do not analyse the statistical significance of the differences in accuracy between different techniques. This is a common problem in the application of ML techniques according to Spiegelhalter (2019).

In general, the number of data points plays a key role in the robustness of ML results. However, the response of model performance to the amount of training data differs between ML algorithms. Cubist and RF are less sensitive to sample size (Morgan et al. 2003), while neural networks are more vulnerable to small sample sizes (Hernández-Lobato and Adams 2015; Tu 1996). On the other hand, the running time for KNN and SVM increases exponentially with sample size (Khaledian and Miller 2020). Small sample sizes can lead to unstable results (Khaledian and Miller 2020).

Some studies suggest that environmental variables may be even more important than the models used for digital soil mapping (McKenzie and Ryan 1999; McBratney et al. 2000). McBratney et al. (2003) proposed the SCORPAN model as the basis for digital soil mapping, where soil (S) or its properties are a function of environmental factors such as climate (C), organisms (O), relief (R), parent material (P), time expressed as soil age (A) and position in space (N). The spatial variability of the soil would be seen as the result of the complex combinations of these factors through the processes of edaphogenesis. The SCORPAN model is the basis of the digital soil mapping (Lagacherie and McBratney 2006; Hartemink 2006).

Environmental variables can have different effects on different soil properties due to the different mechanisms of soil property formation (Shi et al. 2018). Terrain characteristics are among the most important covariates affecting the spatial distribution of soil properties, as they control the process of energy and mass fluxes (Moore et al. 1993). The most commonly used terrain properties are elevation, slope, orientation, curvatures, topographic wetness index (TWI) (Beven and Kirkby 1979; Quinn et al. 1991), multiresolution valley flatness (MrVBF) and ridge flatness (MrRTF) indices (Gallant and Dowling 2003). Many studies have shown that the use of multi-scale terrain attributes increases prediction accuracy (Smith et al. 2006; Behrens et al. 2014; Maynard and Johnson 2014; Miller et al. 2015). The usual way to generate variables with different scales of analysis is to derive them from digital elevation models with different resolutions, or to use variable window sizes to calculate topographic attributes from a single DEM (Shi et al. 2018). Selecting the most informative or relevant predictors before training the model can increase interpretability (Xiong et al. 2014; Prasad et al. 2018; Wang et al. 2018; Keskin et al. 2019).

Soil Organic carbon (SOC) plays an important role in fertility and nutrient cycling by providing a reservoir for other organisms and by allowing the formation of aggregates that increase soil porosity, aeration, infiltration capacity and erosion resistance. The amount and type of SOC depends on biomass inputs and the rate of biomass decomposition, and thus on microbial activity, which is dependent on changes in soil moisture, temperature and acidity (Alexander 1977).

Climatic variables, depending mainly on latitude, are important at large scales (Liu et al. 2012). However, at regional or local scales, geomorphometric variables derived from elevation and land use or vegetation type are the dominant factors (Rezaei and Gilkes 2005). These factors determine the activity of organisms, erosion and sedimentation processes, drainage conditions and the exposure of the soil to climatic conditions. Relationships have been found between topographic factors such as slope or catchment area and SOC content (Buol et al. 1989; Thompson and Kolka 2005; Nadeu Puig-Pey 2013). Land use and land cover, as well as land use changes and land management practices, also linked to topographic factors, also influence the SOC content (Bergstrom et al. 2001; Lal 2002).

Soil texture is an important factor in other soil properties: temperature, structure, aggregate formation, cohesion, moisture holding capacity and hydraulic conductivity. It is therefore an important property in hydrological modelling and a common input to pedotransference functions.

Clay in particular has an important influence on cation exchange capacity, protects organic matter from degradation and facilitates aggregate formation. The proportions of the different particle size fractions depend on the mineralogical composition of the parent material and its resistance to weathering agents, which in turn is linked to climate (especially temperature and moisture). However, in deep soils this

relationship may be less direct due to the presence of several superimposed pedogenetic processes. The relief also influences the processes of erosion, transport and sedimentation.

Several authors (Gessler et al. 2000; Pachepsky et al. 2001; Wilcke et al. 2008) have studied the relationship between morphometric factors and texture, taking them as input variables, and found, for example, a positive correlation between altitude and sand content, while clay content was negatively correlated with altitude and slope. The relationship between terrain curvature and texture varies with scale (Brown et al. 2004). Organisms can influence the distribution of particles and can assist weathering through the physical action of roots or the chemical action of exudates and metabolites, or by retaining sediments in the face of water or wind erosion. However, in low areas, environmental factors such as relief or vegetation are poorer indicators of the spatial variability of particle distribution (McKenzie and Ryan 1999; Zhu et al. 2010).

The aim of this work is to use four machine learning methods, RF, SVM, XGBoost and MLP, to estimate SOC, clay, silt and sand soil content in the Campo de Cartagena. For the four models, their hyperparameters will be optimised to maximise their accuracy, measured as root mean square error (RMSE). Part of the dataset will be reserved as test data to evaluate the models with the optimised parameters and to observe the degree of overfitting that the optimisation has produced. The best model for each variable is used to generate a map of these variables. However, a permutation analysis is performed to determine whether the results are significantly different between each other.

The obtained layers are used to estimate the layers of curve number (Chow et al. 1987; Ferrer-Julià 2003; Al-Ghobari et al. 2020). In addition to the conventional validation in ML, the results obtained in predicting the runoff hydrographs of the Rambla del Albujón in the case of some rainfall events will be compared using this CN layer and the CN layer proposed in Ministerio de Medio Ambiente y Medio Rural y Marino (2011). This way the ability of both datasets to predict flash floods is compared. The curve number is a method developed by the US Soil Conservation Service to calculate the amount of storm rainfall that is not infiltrated into the soil and then converted to storm runoff. Its main parameter is the curve number, which is calculated from soil texture, land use and slope. A full description can be found in Chow et al. (1987) or Al-Ghobari et al. (2020). Despite its age, it is a widely used method in rainfall hydrology, especially in land regional planning of large territories. This is because it does not require several different input variables. However, it does require soil texture data, which is easy to measure in the laboratory for soil samples, but not easy to estimate in a spatially distributed way.

Methodology

Study area

The study area (Fig. 1) is the catchment area of the Mar Menor (1265 km^2). It is located in the south-eastern part of the Segura river basin district. It is a vast area, with few slopes and some small ridges, whose maximum altitude is barely 150 metres above sea level. In turn, it is characterised by tabular reliefs that extend from the foothills of the nearby mountain ranges, such as the Sierra de Carrascoy, to the Mar Menor (Romero Díaz and Belmonte Serrato 2011). It is made up of 14 basins, which in turn contain their respective ephemeral watercourses or wadis. The largest basin is that of the *Rambla del Albujón*, with a surface area of 543 km^2 and a 40 km long main channel, which is the largest contributor of water to the Mar Menor.

This basin was characterised by a natural network of ramblas and gullies with a radial layout, i.e. a system of ramblas that function individually. This is due to the characteristics of the terrain and the semi-arid climate. However the extensive agricultural land use has obliterated most of the network. Temperatures in the area are above 3°C all year round, so there is no risk of frost. The average annual temperature is around 18°C, with average minimum temperatures in winter between 10°C and 11°C and average maximum temperatures in summer around 26°C, with maximum temperatures reaching 38-42°C (Albaladejo-García et al. 2021). One of the most characteristic features of the climate of Campo de Cartagena is its aridity. The annual rainfall average is between 300-350 mm, depending on the specific area of the basin, which causes a rather long dry season of almost 9 months (Conesa García 1990). The irregularity of rainfall is remarkable, and years with less than 200 mm are common. At the same time, rainfall episodes of more than 150 mm in a few hours are not uncommon. Water scarcity issues have led to a high degree of mechanisation in agriculture, with drip irrigation used for about 90% of crops (Alcón et al. 2011).

The aridity that characterises this basin, together with the low gradient, means that the drainage network consists solely of dry channels for most of the year (Romero Díaz and Belmonte Serrato 2011). Due to the transformation that the Campo de Cartagena has undergone in favour of agricultural activities, farmers have been able to compensate for the climatic conditions, and develop irrigation, thanks to the arrival of the Tagus-Segura water transfer thus occupying a large part of this territory with cultivation and irrigation plots (Romero Díaz and Belmonte Serrato 2011).

In the Campo de Cartagena, soil degradation has been due to several factors related to the intensive exploitation of the cultivated area, including the use of agricultural inputs,

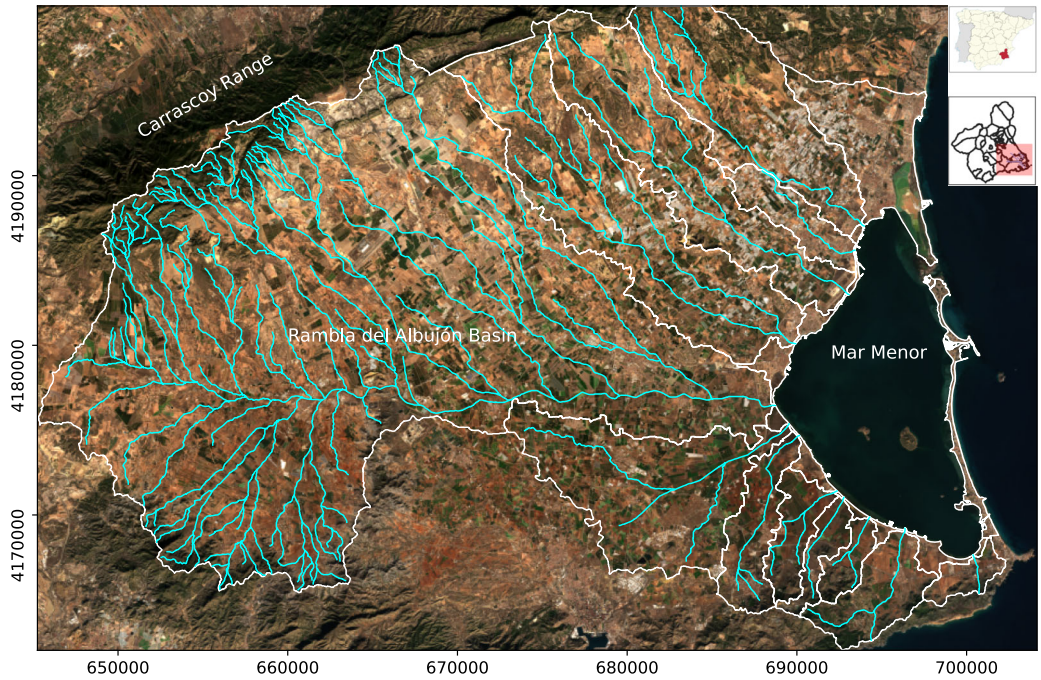

Fig. 1 Study area. Campo de Cartagena and Rambla del Albujón basin

reduction of biodiversity precipitated by the expansion of monocultures, cultivation in greenhouses, and toxic waste filtering caused by poor farm management (Rodríguez-Calles 2022).

In summary, the study area is characterised by intensive agricultural use combined with water scarcity, flash floods and erosion problems. In addition, the coastal area has both an intensive use as a tourist attraction pole and an environmental relevance as one of the most interesting coastal lagoons in the Mediterranean. In order to reconcile the different human activities with the protection of the environment and the mitigation of environmental risks, it is necessary to have the best possible knowledge of all the environmental and socio-economic systems integrated in the area. To this end, it is important to have better maps of different and relevant environmental variables. In particular, distributed soil texture and soil carbon information could improve agricultural productivity, flash flood forecasting and estimation of sediment flows into the lagoon.

Information sources and predictors

In order to obtain predictors to calibrate the models, different sources of information have been used:

- The soil data come from the LUCDEME soil database (Alias and Ortiz 2004), extended with data from Blanco Bernardeau (2015). The LUCDEME project was part of the activities linked to the United Nations Plan of Action to Combat Desertification (DESCON), due to the seriousness of the desertification processes in the southeastern peninsula. It generated a large amount of information, including 132 sheets of 1:100.000 soil maps covering the entire south-east of the peninsula. Although originally produced on paper, some of these were later digitised. The maps are accompanied by the corresponding data files, which contain the description of the soils according to the FAO-UNESCO (1974) classification system and include a data file with the data of 547 profile samples and 1922 samples of the arable layer in the first 30 centimetres of the soil according to a 3x3 kilometre grid. From these data, the Campo de Cartagena database used in this work was extracted, including 274 points. Table 1 shows the variables analysed and the determination methods used. Table 2 shows basic statistics of both the dependent and the independent variables, and Table 3 shows the correlation coefficients between variables. These coefficients

Table 1 Analytical data included in the LUCDEME database and analysed in this work

Variable	Units	Description
Organic carbon (SOC)	g kg^{-1}	Anne (1945)
Clay (0-2 μ)	%	Robinson pipette
Silt (2-20 y 20-50 μ)	%	Robinson pipette
Sand (50 μ - 2 mm)	%	Sieving

Further details on these methods in Pansu and Gautheyrou (2006)

Table 2 Basic statistics of dependent and independet variables

Variable	mean	stddev
Soil Organic carbon	11.62	6.27
Clay	21.99	9.61
Silt	42.81	10.94
Sand	35.19	15.02
Height	206.92	181.62
Slope	5.29	6.62
aspect	208.88	115.75
Curvature	0.0	0.0
TPI	0.68	1.65
TWI	7.96	2.8
LS	500.91	1901.42
MRVBF	1.97	1.92
X	670158.5	14650.14
Y	4180639.06	11381.76

are quite low; however they detect mainly linear relations whereas the models we are going to test are able to detect non linear relations. Figure 2 shows the distribution of organic carbon values and texture classes of the analysed data, and Fig. 3 shows the spatial distribution of the original samples. This is a database collected before the availability of GNSS systems, so the uncertainty in the positions is greater than that of current sampling campaigns. This is indeed a challenge for any attempt to interpolate the data; however, we believe that it is an interesting database and that it is worth trying to use it to obtain layers of soil variables at a higher resolution (25 m) than those currently available (500 m).

- Digital Elevation Model (DEM) with 25 m resolution from LiDAR data. It was obtained from the CNIG website (National Geographic Institute of Spain, 2013). The following predictors are obtained from this DEM:
  - Height in meters above mean sea level
  - Slope in degrees
  - Aspect in degrees counterclockwise from North
  - Profile curvature is defined as the rate of change of the slope in the direction of the maximum gradient, and therefore depends on the second order derivatives of the height. It reflects the acceleration or deceleration of material flow along a slope, so that a negative profile curvature is concave and the flow undergoes a relative deceleration, while a positive profile curvature reflects a convex slope and implies a relative acceleration. Profile curvature has a significant relationship with soil moisture, indicating the tendency of a cell to accumulate water or not.
  - Topographic Position Index (TPI), developed by Jennes (2005), is a focal operator that compares the height of a cell with the average height of the window. Positive values of the TPI represent cells with an elevation higher than the average of the window under consideration, representing ridges, while negative values represent cells with a lower elevation (valleys).m.o.s.l. Values close to zero can represent cells with a slope of zero or with a constant slope.
  - Terrain roughness index (TRI) is an index of the heterogeneity of the terrain, defined by Riley et al. (1999) as:

$$TRI = \sqrt{\sum_{i=-1,j=-1}^{i=1,j=1} (X_{i,j} - X_{0,0})^2} \quad (1)$$

  - Topographic Wetness Index (TWI) is based on the idea that topography controls water movement on a slope, indicating the spatial distribution of soil moisture and soil surface saturation. This index is part of the distributed hydrological model TOPMODEL (Beven and Kirkby 1979; Quinn et al. 1991), which is used to model topography-related hydrological processes at the slope and catchment scale. This index is calculated as

$$TWI = ln\frac{a}{tan\beta} \quad (2)$$

    where a is the specific flow accumulation area, reflecting the tendency of that cell to accumulate water, and β is the slope. The calculation of a requires knowledge of the flows occurring on the slope to determine the total accumulation area A flowing through the cell under consideration, along with the effective contour length L orthogonal to the flow, such that $a = A/L$, where L is weighted by multiplying by the local slope angle, thus β, which represents the hydraulic gradient, i.e. the tendency of gravitational forces to move water down the slope.
  - The USLE (Universal Soil Loss Equation) is an empirical equation used for agricultural purposes, where the LS factor represents the effect of slope on soil loss, which increases with slope steepness and slope length (Moore and Burch 1986). It is defined as:

$$LS = (\frac{\lambda}{22})^m(\frac{0.43 + 0.3S + 0.043S^2}{6.574}) \quad (3)$$

    where λ is the hillslope length (m), S i the slope (%), and m receives different values depending on slope:

Table 3 Pearson correlation coefficients between dependent and independet variables

	SOC	Clay	Silt	Sand	Height	Slope	Aspect	Curvature	TPI	TWI	LS	MRVBF	X	Y
SOC	1.0	−0.329	−0.2	0.356	0.206	0.101	0.176	0.033	0.003	−0.03	0.307	−0.305	−0.134	0.035
Clay	−0.329	1.0	0.064	−0.687	−0.294	−0.41	−0.161	0.159	−0.126	0.249	−0.096	0.405	0.32	0.091
Silt	−0.2	0.064	1.0	−0.769	−0.322	−0.465	0.212	−0.361	−0.461	0.477	0.104	0.426	0.172	0.103
Sand	0.356	−0.687	−0.769	1.0	0.422	0.601	−0.051	0.161	0.417	−0.507	−0.014	−0.57	−0.33	−0.133
Height	0.206	−0.294	−0.322	0.422	1.0	0.584	0.107	0.275	0.371	−0.439	0.057	−0.565	−0.57	0.33
Slope	0.101	−0.41	−0.465	0.601	0.584	1.0	−0.037	0.457	0.737	−0.721	−0.13	−0.647	−0.093	0.041
Aspect	0.176	−0.161	0.212	−0.051	0.107	−0.037	1.0	−0.088	−0.044	0.195	−0.011	0.141	−0.083	0.232
Curv.	0.033	0.159	−0.361	0.161	0.275	0.457	−0.088	1.0	0.772	−0.532	−0.093	−0.334	−0.098	0.103
TPI	0.003	−0.126	−0.461	0.417	0.371	0.737	−0.044	0.772	1.0	−0.664	−0.139	−0.466	−0.053	0.199
TWI	−0.03	0.249	0.477	−0.507	−0.439	−0.721	0.195	−0.532	−0.664	1.0	0.377	0.67	0.156	−0.172
LS	0.307	−0.096	0.104	−0.014	0.057	−0.13	−0.011	−0.093	−0.139	0.377	1.0	0.045	−0.247	−0.026
MRVBF	−0.305	0.405	0.426	−0.57	−0.565	−0.647	0.141	−0.334	−0.466	0.67	0.045	1.0	0.335	−0.055
X	−0.134	0.32	0.172	−0.33	−0.57	−0.093	−0.083	−0.098	−0.053	0.156	−0.247	0.335	1.0	0.061
Y	0.035	0.091	0.103	−0.133	0.33	0.041	0.232	0.103	0.199	−0.172	−0.026	−0.055	0.061	1.0

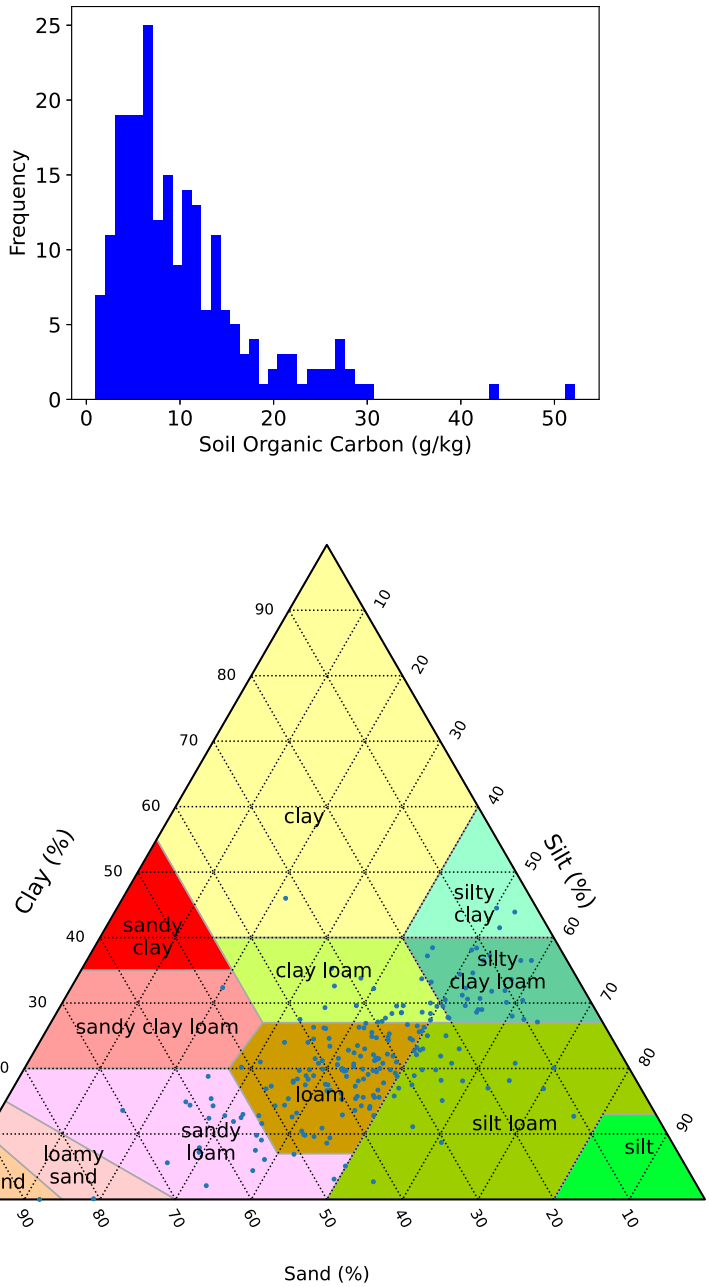

Fig. 2 Histogram of soil organic carbon concentration values (top) and triangle of textures (bottom) showing the distribution of values in the samples used

if S $\leq$ 1, m = 0.2; if 1 $<$ S $\leq$ 3.5, m = 0.3; if 3.5 $<$ S $\leq$ 4.5, m = 0.4 and if S $>$ 4.5, m = 0.5.

– The Multiresolution Index of Valley Bottom Flatness (MRVBF) , purposed by Gallant and Dowling (2003), is an index that allows the description of valley bottom morphology over a wide range of scales by combining the results into a single multiresolution index. MRVBF values less than 0.5 are not considered to be valley bottoms. Values between 0.5 and 1.5 would correspond to smaller, steeper valleys, while wider, flatter valleys would be represented by values higher than 1.5.

- The CORINE Land Cover land use maps for 1990, 2000, 2006, 2012 and 2018 (European Environment Agency 1995; Bossard et al. 2000). As these maps are available for several years, estimates can be made for the same years.
- Soil maps from the LUCDEME project maps digitised by the Autonomous Community of the Region of Murcia (Ramírez-Santiagosa et al. 1999).
- Lithostratigraphic, permeability and hydrogeological map of Spain at 1:200,000 scale from the Spanish Geological and mineralogical Institute (IGME). It shows the geological units according to lithostratigraphic and

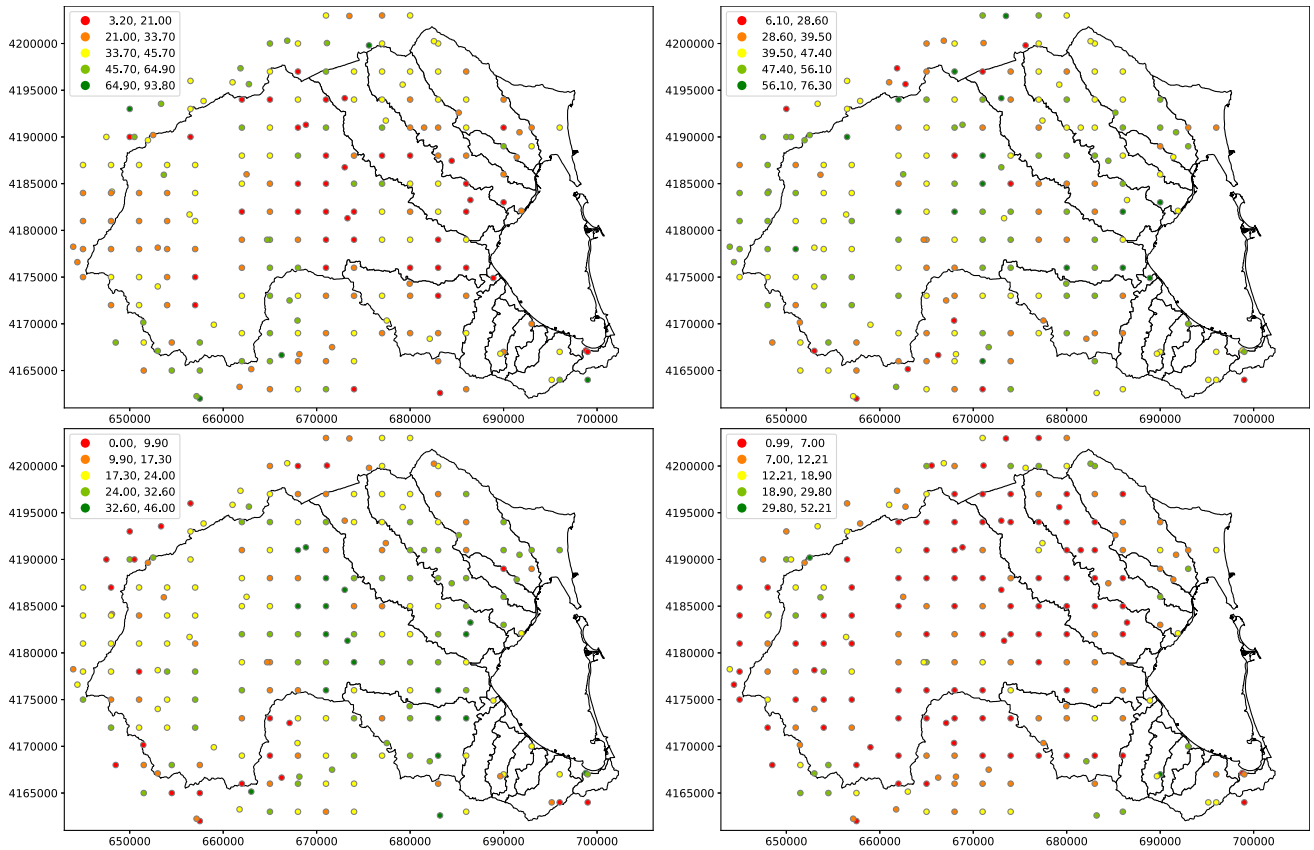

Fig. 3 Spatial distribution of Sand percentage (top left), Silt percentage (top right), Clay percentage (bottom left) and Soil Organic Carbon (bottom right)

hydrogeological criteria. A map of lithological units and a semi-quantitative permeability map with five levels were obtained from these sources.

- X and Y UTM coordinates are added to these predictions.

Algorithms

The large distance between observation points makes difficult the use of local interpolation methods such as geostatistical techniques (Burgess and Webster 1980a, b; Webster and Burgess 1980; Goovaerts 1997) or others. Therefore, only global interpolation methods, based on four algorithms belonging to three of the most common types of machine learning algorithms, were used.

RF (Breiman 2001) consists of an ensemble of decision trees (on the order of hundreds or thousands). Each of them is trained with a resampled subsample obtained by bootstrapping the original data set; the excluded data are then used to obtain an error estimate (Stum 2010). On the other hand, only a subset of the predictor variables is considered at each decision node of each tree. The size of this subset (*mtry*) and the number of trees (*ntree*) are the parameters to be optimised in this model. The default values for regression are $ntree = 500$ and $mtry = int(p/3)$ where p is the number of predictors (Liaw and Wiener 2002). Once the trees have been trained, the prediction for new cases will be the average of the predictions of all the trees. RF has been used in digital soil mapping (Grimm et al. 2008; Wiesmeier et al. 2011; Ließ et al. 2012).

Boosting (Wade 2020) is another type of decision tree ensemble with the aim of improving the predictive behaviour of the trees. In this case, instead of training all the trees in parallel, they are trained one at a time, so that as each tree is trained, the cases predicted with more error by the previous tree are given more weight. At the same time, the maximum depth that the trees can reach is limited, so that each individual tree has less predictive power than an isolated decision tree. Its main hyperparameter is the learning rate; a high learning rate reduces the weight of the first trees in the total.

The goal of SVM is to obtain a linear model that maximises the number of cases that are less than ϵ away from the straight

line defined by that model. If the relationships between the predictors and the response variable are non-linear, it is possible to try to make them linear by transforming the space of variables using a kernel function. The most common kernel functions are the gaussian kernel and the polynomial kernel. The former is controlled by the parameter γ and the latter by γ and the degree of the polynomial (g). SVM has been used in the field of soil science, both for predicting properties and for classifying soil types or detecting cases of soil salinisation or contamination (Bhattacharya and Solomatine 2006; Ballabio 2009; Kovačević et al. 2010; Cai et al. 2010; Brungard et al. 2015). In this paper we test the polynomial kernel optimising ϵ, γ and g and the Gaussian kernel optimising ϵ and γ.

Artificial Neural Networks (ANN) are non-parametric machine learning methods that allow the detection of non-linear relationships and have been used both in modelling soil properties when large databases with a large number of variables are availables (Elshorbagy and Parasuraman 2008), and in the elaboration of edaphotransfer functions. These networks have a system of many interconnected nodes organised in layers: an input layer, an output layer and one or more hidden layers that extract useful information from the input layer and use it to predict the results. Neural networks have generally been applied to the prediction of edaphototransfer functions, but also to the determination of other soil properties (Behrens et al. 2005; Anagu et al. 2009; Hattab et al. 2013).

A multilayer perceptron is a simple type of neural network that is widely used in regression. It consists of multiple layers of interconnected neurons. The regularisation strategy is to drop a proportion of the neurons (drop rate). Other hyperparameters of the model are the batch size (number of cases analysed before updating the values of the neuron weights) and the learning rate (magnitude of the changes in the values of the neuron weights at each update).

Software

The terrain features extracted from the DEM were computed using GRASS (GRASS Development Team 2023) and SAGA (Conrad et al. 2015). The prediction models were developed in Python using the libraries scikit-learn (Pedregosa et al. 2011) for RF and SVM, xgboost (Chen and Guestrin 2016) for the XGBOOST model and tensorflow (Abadi et al. 2016) for MLP. A detailed explanation of the algorithms used and their implementation in Python can be found in Géron (2019) and James et al. (2023).In addition, a Python library was created with functions to draw the texture triangle and calculate texture classes, hydrological groups and curve number from Ferrer-Julià (2003) and USLE K-factor from Corral-Pazos-de-Provens et al. (2023).

Procedure

The variables analysed are SOC, clay, silt and sand. Five models are tested: RF, Polynomial Kernel Support Vector Machines, Gaussian Kernel Support Vector Machines, XGBoost and Multilayer Perceptron.

In order to avoid overfitting of the models to the training data, we used 20 % of the samples to test and 80 % to train. In order to avoid overfitting in the hyperparameter optimization process, we used 4-fold cross validation with the training data. The data set is randomly divided into five parts. The first part will be used as test data and the other four parts will be used for a four-folds cross-validation. This cross-validation is used to optimise the hyperparameters of the different algorithms by minimising the mean square error. Leave-one-out cross validation (LOO-CV), that is k-fold CV with k=n, would give a less biased error estimation than k-fold cross validation; however, test errors resulting from LOO-CV tend to have higher variance than test error derived from k-fold CV (James et al. 2017). This authors suggest using k=5 or k=10. We decided to use k=4 to reduce computational burden (models have to be calibrated only 4 times) inside every hyperparameter optimization loop.

The optimisation of the hyperparameters is done by a systematic search in the hyperparameter space. After obtaining the results, it is determined which is the minimum and whether it is necessary to do a second search around this minimum. This is determined by checking whether the results show a clear trend towards a decrease in the RMSE at the minimum or whether, on the contrary, a random result appears.

After obtaining the set of hyperparameters that minimises RMSE, the corresponding models are calibrated with the 4 training sets and used to predict the corresponding variable in the test data. To evaluate these results, the RMSE and the coefficient of determination R^2 are calculated. Finally, the statistical significance of the differences in the RMSE values is determined by taking 100000 resamples of the test points to obtain the statistical distribution of these differences. From these differences, the p-value of the observed difference can be obtained. For a difference in the RMSE to be considered significantly different from zero, the p-value should be sufficiently low, in principle less than 0.05, but considering that multiple comparisons are made, this threshold should be lower.

Finally, the best model for each variable is used to obtain a final map from models calibrated on the entire dataset. With these maps, the layers of texture classes, hydrological groups and curve number are finally obtained. As the original data

correspond to 1990, the validation of the models is based on the estimates for this year.

Hydrological model

In order to evaluate the effect of the estimated CN values in an applied case, a simple hydrological HEC-HMS based model was used. HEC-HMS uses three sub-models: 1) The Curve Number Method (Soil Conservation Service 1972) to separate precipitation into infiltrated water and effective precipitation, 2) the instantaneous unit hydrograph method for converting effective precipitation into channel outflow (Chow et al. 1987), and 3) Muskingum's method (McCarthy 1938) to convey, in a given catchment, the flow from tributary catchments to the outflow. As the aim of this study is to determine the differences between the two CN estimation strategies, it was not considered appropriate to carry out a prior calibration of the model.

In this study, the schematisation of the model (Fig. 4) was carried out by discretizing the Albujón basin into 15 sub-basins, 11 sections of riverbed, one outlet (near the mouth) and 16 connection nodes between the elements. With regard to the meteorological model, precipitation hyetograms from 25 meteorological stations (Fig. 4) belonging to the Murcia Region Agrarian Information System (SIAM) and the Spanish Automatic Hydrological Information System (SAIH) networks have been used, with a time step of 1 hour. Spatialisation of rainfall data performed by interpolation using an inverse of the squared distance weighted average.

This model was used to simulate the hydrological response of the Rambla del Albujón to real rainfall events, using the curve number layers obtained in this work and those used by the Spanish Ministry of the Environment and Rural and Marine Affairs for the development of the national flood zone mapping system (Ministerio de Medio Ambiente y Medio Rural y Marino 2011). This last study uses a layer of the P0 parameter, complementary to CN, calculated for the whole of peninsular Spain by CEDEX in collaboration with the University of León (ULE 2009). This layer has a resolution of only 500x500 m. The variables from which P0 has been calculated in this work are the hydrological soil group obtained from the method proposed by Ferrer-Julià (2003), the land use from the data of the CORINE LAND COVER project of 2000, the slope of the terrain obtained from the DTM with a spatial resolution of 500x500 m, distinguishing between slopes greater and less than 3 %.

As different goodness-of-fit statistics measure different aspects of model performance (Bennett et al. 2013), three different statistics were used to assess the error of the run-off model: Root Mean Square Error (RMSE), Modified Nash-Sutcliffe Efficiency (NSE), which measures the relative size of the residual variance compared to the measured data variance and is less sensitive to extreme values than R^2 (Legates and McCabe 1999), and Percent Bias (PBIAS), which measures the average tendency of the estimated values to be greater or less than the observations and is not as sensitive to extreme values or to the magnitude of the variables as RMSE. These three statistics were calculated for each estimated hydrograph compared to the observed data.

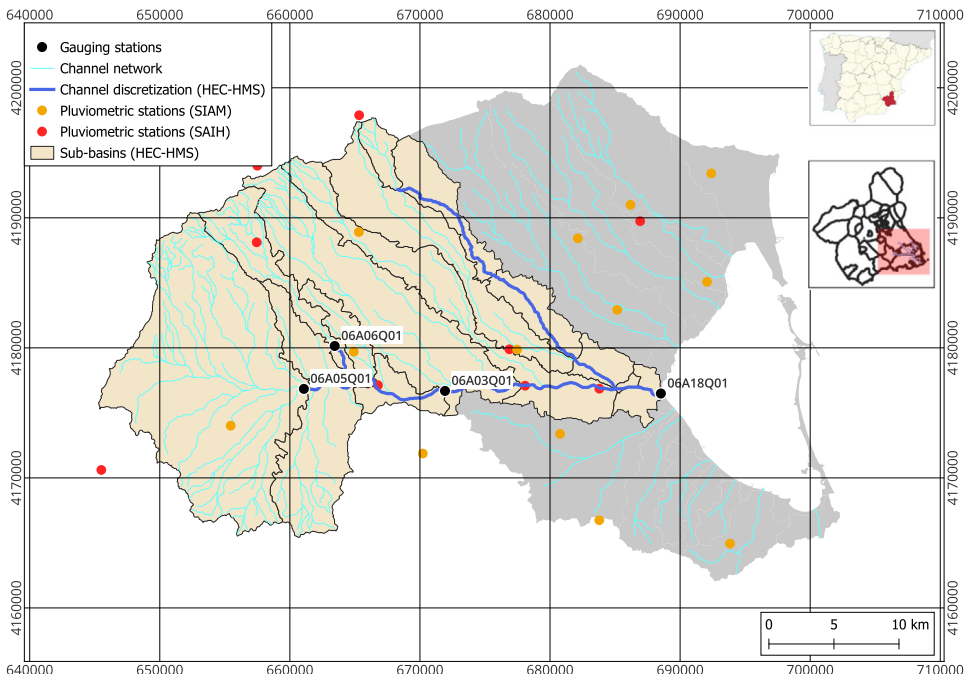

Fig. 4 Study area. Campo de Cartagena and Rambla del Albujón basin

In this study, 4 gauging stations belonging to the SAIH network (Fig. 4) have been used, although after a previous analysis of the behaviour of the observed hydrographs, the results of station 06A03Q01 are presented.

Results and discussion

Table 4 shows the optimal parameters and RMSE values obtained by cross-validation for each variable and model. In general, the best values are obtained with MLP. Only in the case of sand percentage, the result of the SVM with Gaussian kernel is preferable to that of the MLP. It is obvious that the higher the number of parameters of a model, the better the fit can be and the more combinations are made when optimising the hyperparameters, so it is reasonable to think that a model with more parameters can obtain a higher accuracy, at least with the training data.

Table 4 Optimal hyperparameters

RF					
	ntree	mtry	RMSE		
SOC	100	2	6.389		
Clay	400	13	7.646		
Silt	100	2	9.579		
Sand	300	2	13.56		
SVM (Polinomial kernel)					
	degree	gamma	epsilon	RMSE	
SOC	1	0.025	7	6.607	
Clay	1	0.075	0.01	7.635	
Silt	1	0.03	6.3	9.765	
Sand	1	0.5	7.5	14.513	
SVM (gaussian kernel)					
	gamma	epsilon	RMSE		
SOC	0.02	0.25	6.613		
Clay	0.02	0.05	7.75		
Silt	0.0235	4.4	9.597		
Sand	0.03	0.02	13.598		
XGB					
	ntrees	maxDepth	learning rate	RMSE	
SOC	80	1	0.08	6.407	
Clay	44	3	0.075	7.678	
Silt	400	1	0.02	9.595	
Sand	500	5	0.008	14.149	
Multilayer Perceptron					
	nlayer	dropRate	learning rate	batch Size	RMSE
SOC	4	0.15	0.15	16	5.840
Clay	5	0.3	0.02	8	7.008
Silt	6	0.2	0.025	8	0.046
Sand	5	0.35	0.01	4	14.115

For this reason, it is necessary to have test data with which to evaluate the accuracy of the best model obtained. Table 5 shows the results obtained with the different models. Both RMSE and R^2 are shown and the values indicating the highest accuracy for each variable are highlighted in bold, in this case the best results for SOC, silt percentage and sand percentage are obtained with RF, while the best results for clay percentage are obtained with XGBoost. It can be seen that, in general, it is the results of the models with fewer hyperparameters and therefore fewer optimisation trials that ultimately give the best results on the test data, as they are not overtrained. In the case of clay and silt, although RF and XGBoost have the lowest RMSE values, MLP has the highest R^2 values. This fact and the small differences in the RMSE values already seem to indicate that the observed differences are insignificant.

Table 6 shows the results of the permutation tests to determine whether the difference in accuracy of two models are statistically significant. P-values lower than 0.05 appear only when comparing RF with both SVM models for SOC, and it must be taken into account that we are making 40 comparisons, which clearly poses the problem of multiple comparisons. Furthermore, the 2 p-values obtained that could be considered significant are not much lower than 0.05, so we could conclude that there are no significant differences between the different methods. We think, as does (Spiegelhalter 2019), that this might be the case for a great deal of ML comparison results when comparing different state-of-the-art machine learning algorithms. However, it is not common to check the significance of accuracy differences in similar papers.

Comparing SOC results with other studies in terms of RMSE is complex because different studies express these measurements in different units, the conversion of which depends on bulk density and is not always taken into account in the work. On the other hand, the organic matter content will be very different in different areas. It is therefore preferable to make comparisons using R^2.

Martin et al. (2014) use Boosted Regression Trees (BRT) to estimate organic carbon in France using different sets of predictors. The values of R^2 obtained range from 0.17 to 0.35, so the values obtained in this paper would be within this range. In contrast, Were et al. (2015) compare SVM, RF and ANN to estimate organic carbon in a small area of Kenya and the reported values of R^2 are 0.64 for SVR, 0.61 for ANN and 0.53 for RF. These values are significantly higher than those obtained in this study.

As for the clay, silt and sand contents, they are easier to compare with RMSE, Reza Pahlavan-Rad and Akbarimoghaddam (2018) using RF and obtain RMSE values of 21.4 for sand, 17.45 for silt and 6.02 for clay. The values for

Table 5 Error statistics with test data

	RF		SVM polinomic		SVM gaussian		XGB		MLP1	
	RMSE	r2	RMSE	r2	RMSE	r2	RMSE	r2	RMSE	r2
SOC	**8.040**	**0.283**	8.688	0.173	8.935	0.18	8.45	0.18	8.775	0.166
Clay	7.183	0.278	7.345	0.234	7.701	0.151	**7.049**	0.288	7.327	**0.288**
Silt	**10.227**	0.041	10.263	0.063	10.705	0.012	10.705	0.012	14.32	**0.167**
Sand	**13.561**	**0.214**	14.608	0.117	14.328	0.126	14.328	0.126	14.768	0.077

clay are similar to ours, but those for silt and sand are significantly worse than ours. Bashir et al. (2024) use different model distributions to improve the results of the SVM and RF machine learning models. They report RMSEs between 1.1 and 9.7 for sand, between 3 and 18.1 for silt and between 1.8 and 18.1 for clay. Our results are therefore slightly worse, but within the range of values. Martinelli and Gasser (2022) obtain the best results with RF after comparing this method with KNN, NN, XGBOOST and linear regression, the latter being the worst performer. The RMSE values obtained by these authors with RF are 12.5 for sand, 10.2 for clay and 7.16 for silt. Our results are better for clay, slightly worse for sand and significantly worse for silt.

Taking into account the uncertainty in the positioning of the points, the results can be considered quite adequate compared to those reported in the literature. Another relevant issue in validating the clay, sand and silt estimates is to check that the pattern obtained in the texture triangle correctly reproduces the patterns observed with the real data. Figure 5 (left) shows the heat maps of the estimates in the study area on the texture triangle. The distribution does not correctly reproduce the expected distribution, that of the LUCDEME project data (Fig. 2). The deviation is clear, and due to a bias of the model. Ensemble models tend to reduce variability in the modelled variable, at the end of the day, it works by calculating means. We decided then to expand the distribution of the predicted results to fit the real distribution of the variables. In order to obtain a more correct estimation of the texture classes, hydrological groups and number of curves, a correction coefficient was applied to the values estimated

Table 6 P-values of permutation tests of prediction models for organic carbon content, clay content, sand content and silt content

SOC	RF	SVM (Poly)	SVM (RBF)	XGB	MLP1
RF		**0.0344**	**0.0423**	0.2317	0.6205
SVM (Poly)	**0.0344**		0.1667	0.0572	0.7103
SVM (RBF)	**0.0423**	0.1667		0.1341	0.8166
XGB	0.2316	0.0571	0.1341		0.8382
MLP1	0.6205	0.7103	0.8166	0.8382	
Clay	RF	SVM (Poly)	SVM (RBF)	XGB	MLP1
RF		0.3136	0.2604	0.6569	0.6532
SVM (Poly)	0.3136		0.632	0.1834	0.7904
SVM (RBF)	0.2605	0.632		0.2236	0.3495
XGB	0.6569	0.1834	0.2236		0.5497
MLP1	0.6532	0.7904	0.3495	0.5497	
Silt	RF	SVM (Poly)	SVM (RBF)	XGB	MLP1
RF		0.8403	0.5627	0.2171	0.1050
SVM (Poly)	0.8403		0.3299	0.4054	0.086
SVM (RBF)	0.5627	0.3299		0.2411	0.0697
XGB	0.217	0.4053	0.2411		0.1446
MLP1	0.105	0.086	0.0697	0.1446	
Sand	RF	SVM (Poly)	SVM (RBF)	XGB	MLP1
RF		0.2414	0.8119	0.2235	0.2092
SVM (Poly)	0.2414		0.2535	0.8318	0.5483
SVM (RBF)	0.8119	0.2535		0.3386	0.3108
XGB	0.2235	0.8318	0.3386		0.615
MLP1	0.2092	0.5483	0.3108	0.615	

P-values lower than 0.05 are highlighted

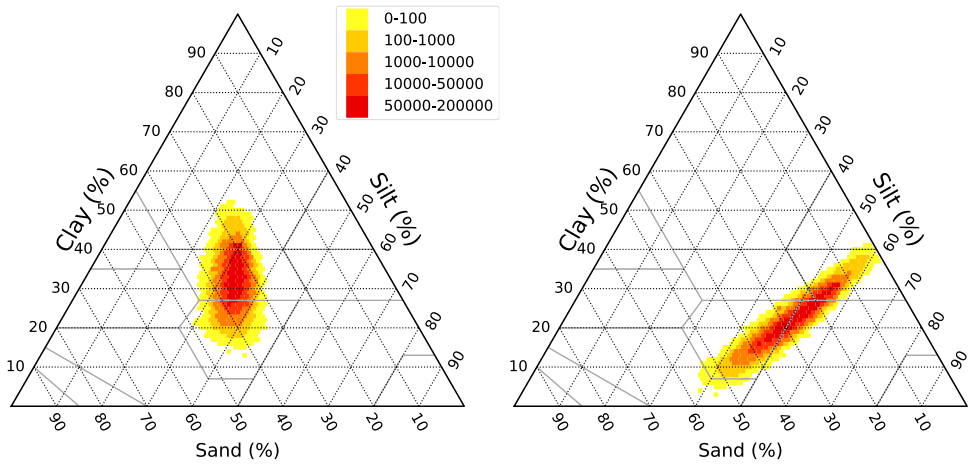

Fig. 5 Texture triangle showing the distribution of predicted values in 1990 uncorrected (left) and corrected (right)

with the RF models with the objective of reproducing the observed distribution. This is a simple linear transformation consisting of:

- clay = $clay_0$ -10
- sand = $sand_0$ - $clay_0$ + 25
- silt = $silt_0$ + $clay_0$ - 25

where $clay_0$, $sand_0$ and $silt_0$ stands for the values produced by the RF model. A posterior normalization ensures that the three fractions add up to 100 %

This changes the texture distribution from that on the left of Fig. 5 to that of the right. This transformation, which applies to all years, is necessary if the resulting layers are to

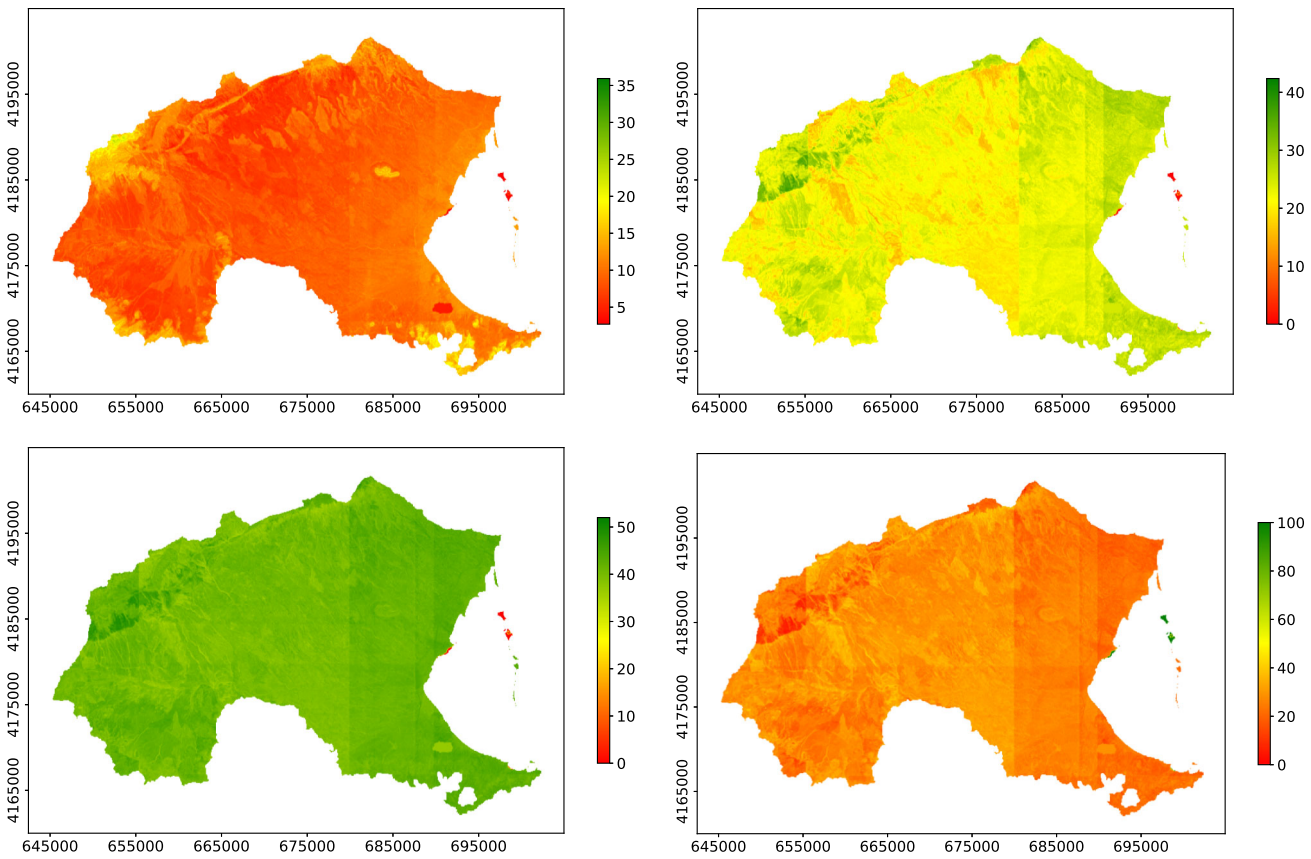

Fig. 6 Estimated SOC (top left), clay (top right), silt (bottom left) and sand (bottom right) in 1990

be used as input data for any type of hydrological, agronomic or erosion model. This type of transformation is clearly not ideal, and further work is needed to avoid such biases in soil texture models. The use of other algorithms may give better results. Figure 6 shows the final maps.

We obtained uncertainty maps, which are shown in Fig. 7. As the best models in all cases were ensembles of trees (XGBoost in one case and Random Forest in the others), it is possible to take the individual prediction from each of the trees and calculate the standard deviation of the prediction. However, it is important to note that this is a measure of how confident the model is about the prediction, not how accurate the prediction is.

Runoff model

Due to the high RMSE values and the need to transform the values obtained to correctly reproduce the pattern observed in the texture triangle, we decided, as a second form of validation, to use the produced layers to obtain layers of the CN parameters of the SCS abstractions model and compare the results of simulating the response to a series of precipitation events with those obtained with the official layers from Ministerio de Medio Ambiente y Medio Rural y Marino (2011). If better results are obtained, we can assume that the layers obtained are useful from an applied point of view, despite the issues in their estimation.

Figure 8 shows the CN values of Ministerio de Medio Ambiente y Medio Rural y Marino (2011) and those obtained in this work aggregated by basin and Fig. 9 shows the scstterplot of such values. The map of CN values from Ministerio de Medio Ambiente y Medio Rural y Marino (2011) has a spatial resolution of 500 m, whereas the maps produced in this work have a spatial resolution of 25 m, therefore, we call the first map CN500 and the second CN25.

The CN500 values range from 60 to 82, while the CN25 values range from 71 to 82. The correlation coefficient between the two is 0.25. It is evident that the values obtained in this work are systematically higher, although there is some correlation between the values of the two sets. It is therefore expected that the ability to reproduce the observed hydrograph will be different in both cases.

Table 7 shows the main results of the simulation of three events (09/12-14/2019, 11/18-20/2018 and 03/06-08/2021) using both the 500 m CN layer and the 25 m CN layer. For the small runoff events (2018 and 2021 events) there are no

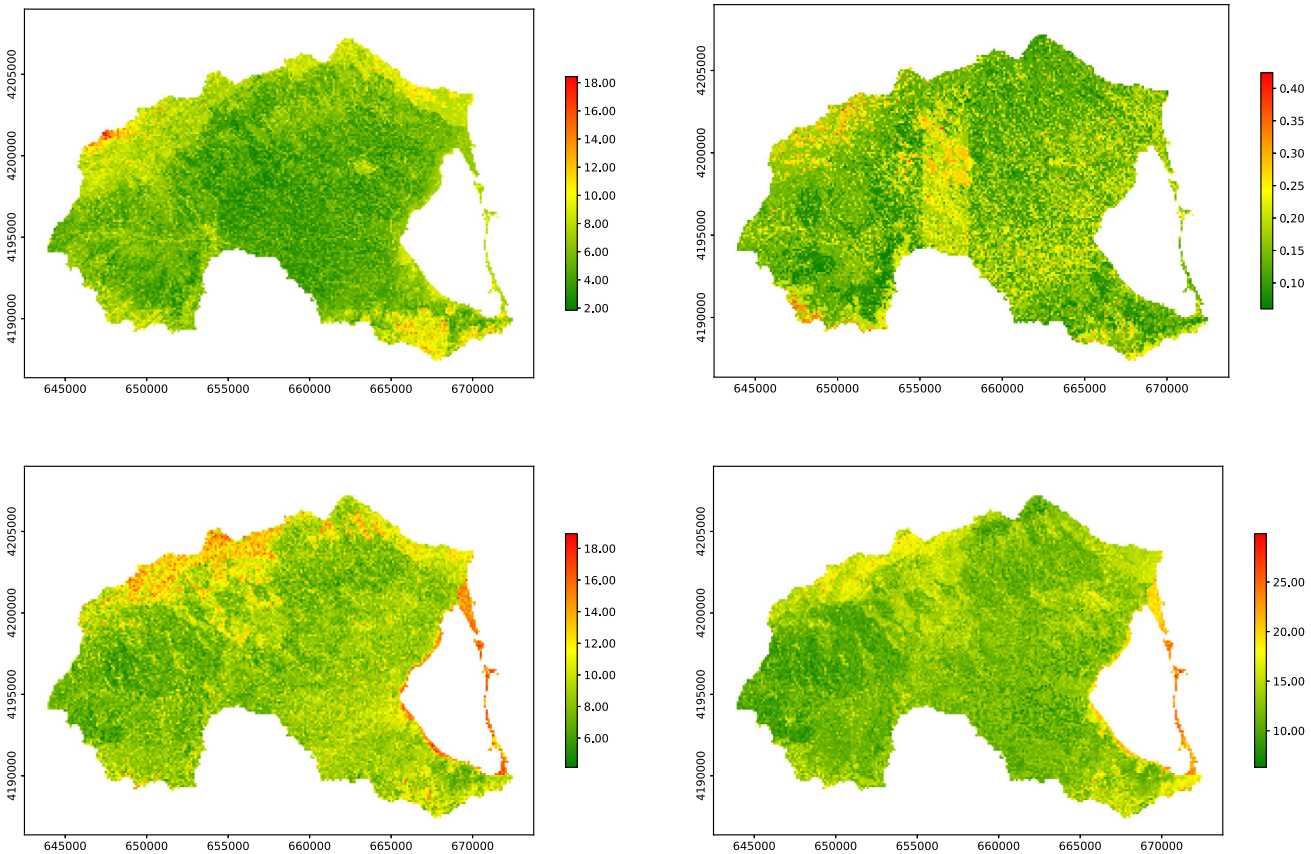

Fig. 7 Uncertainty (standard deviation of estimations) in SOC (top left), clay (top right), silt (bottom left) and sand (bottom right) in 1990

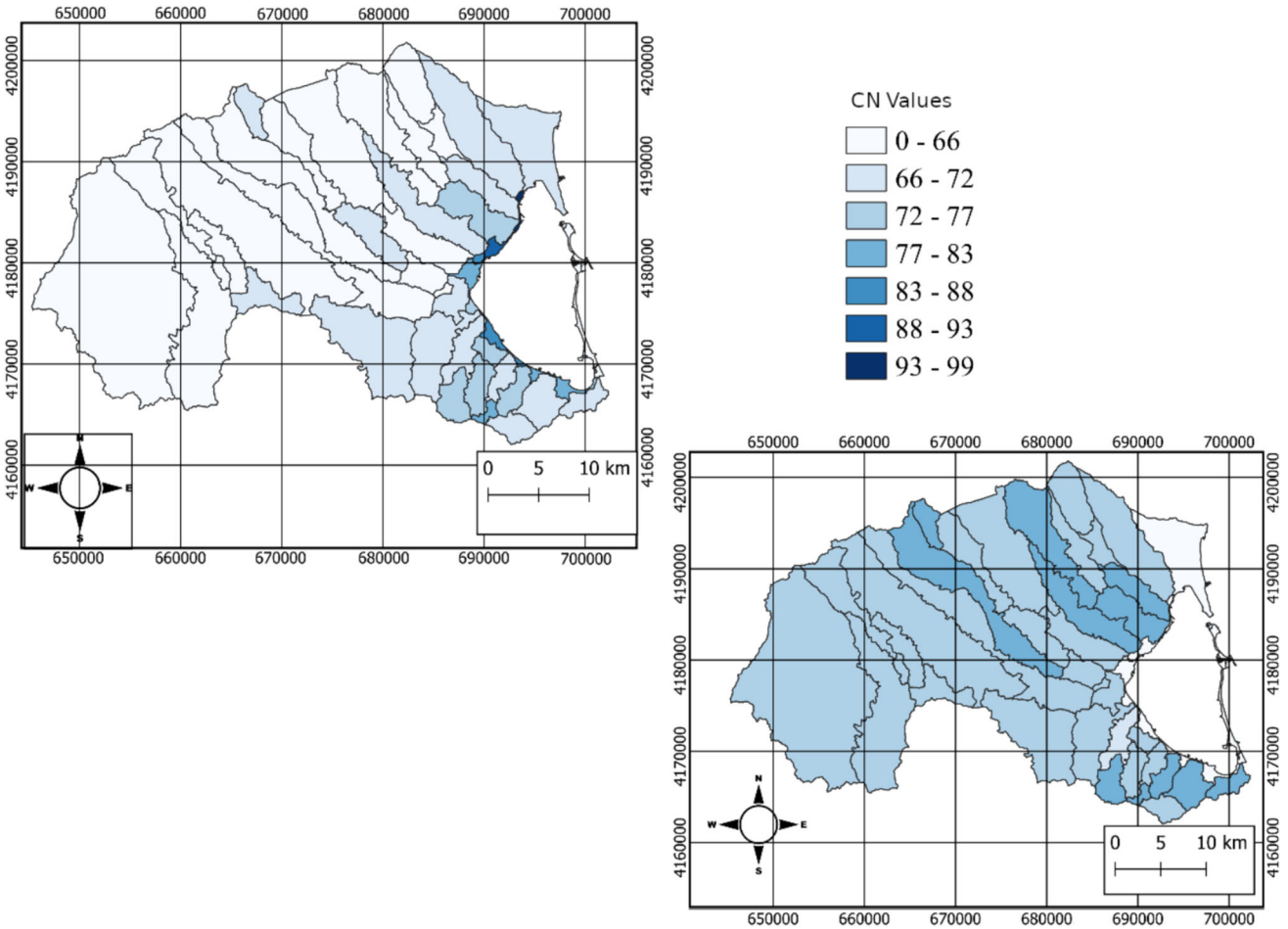

Fig. 8 Aggregation by sub-basin of CN values at 500 m from Ministerio de Medio Ambiente y Medio Rural y Marino (2011) (top left) and CN values at 25 m obtained in this work (bottom right)

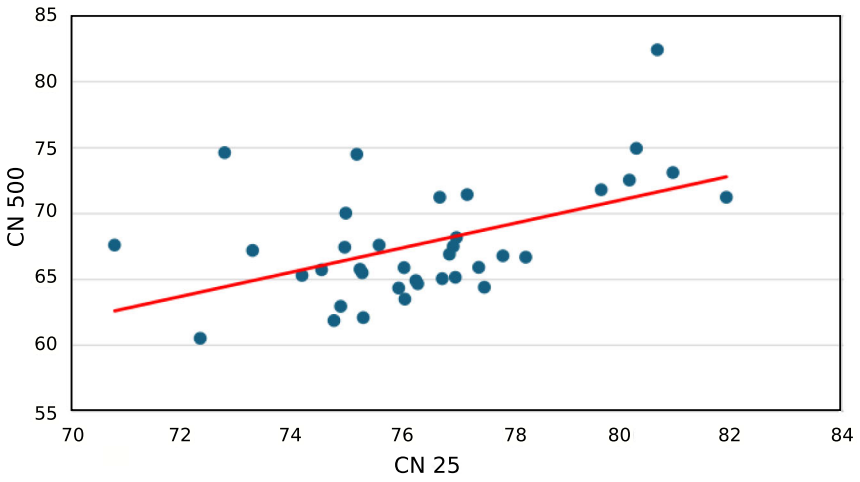

Fig. 9 Comparison of CN values at 500 metres from Ministerio de Medio Ambiente y Medio Rural y Marino (2011) and CN values at 25 metres obtained in this work. The red line shows the regression line between the values of both models

Table 7 Goodness-of-fit statistics of the estimated hydrographs using both the 500 m CN estimation (CN500) and the 25 m CN estimation (CN25)

Start	End	CN layer	ObVol	EsVol	ObPeak	EsPeak	NSE	PBIAS	RMSE
09/12/2019 12	09/14/2019 04	CN500	7.01	13.61	100.50	146.90	−2.28	135.82	1.80
09/12/2019 12	09/14/2019 04	CN25	7.01	5.52	100.50	61.30	0.81	−4.70	0.40
18/11/2018 00	20/11/2018 23	CN500	3.00	3.14	53.60	33.50	−0.06	0.22	1.00
18/11/2018 00	20/11/2018 23	CN25	3.00	3.14	53.60	33.50	−0.06	0.22	1.00
03/06/2021 00	03/08/2021 23	CN500	0.47	1.74	14.70	18.60	−1.31	51.65	1.50
03/06/2021 00	03/08/2021 23	CN25	0.47	1.74	14.70	18.60	−1.31	51.65	1.50

Other hydrograph variables shown are observed volume in Hm^3 (ObVol), estimated volume in Hm^3 (EsVol), observed peak (obPeak) in m^3/s and estimated Peak (EsPeak) m^3/s

differences using different CN layers. However, for the large runoff event, the differences are very significant. CN25 gives more accurate results, with a very good shape and position of the hydrograph, while the hydrograph generated with CN500 is delayed by about 5 hours. The runoff volume simulated by CN500 is almost double that of the observed data, while the runoff volume simulated by CN25 is 78 % of the observed. Interestingly, the peak in CN500 is overestimated, as is the runoff volume in general, but the peak in CN25 is underestimated. Figure 10 shows the observed hydrographs and those simulated with the values of CN500 and CN25.

The model gives the same results for both low rainfall events. This is probably because the rainfall intensity in such events was too low to overcome infiltration with both CN25 and CN500 values. Figure 11 shows a sensitivity analysis to evaluate the model response of the Albujón catchment in gauge 06A03Q01 (see Fig. 4). This sensitivity analysis was performed by sistematically changing the area-weighted mean CN of the seven sub-basins, but maintaining the differences between them to preserve the spatial variability, and using the real yetographs of the three events as input. Runoff volume (hm^3) and peak runoff (m^3/s) are calculated for each CN value and hyetograph. The lines in the figure represent the results of the sensitivity analysis and the points represent the results obtained with the CN25 values estimated in this work (mean=63.4, sd=2.3) and the official CN500 values (mean=74.06, sd=1.21) in the basins.

Figure 11 shows that both runoff volume and peak are the same for both CN values in the 2018 and 2021 events. Both hydrograph parameters are insensitive to CN values up to a threshold, and this threshold depends on the characteristics of the event, mainly the rainfall amount and intensity. For the high rainfall event of 2019, the threshold is around 60, thus the two CN values give different results; however, for the low rainfall events, the threshold is slightly higher than 80 and the two CN values give the same results.

Conclusions

Four models were tested for predicting soil organic carbon content and percentages of clay, silt and sand. The accuracy of the models is rather low, but within the ranges observed in previous work using similar models. It is well known that soil variables are particularly difficult to model due to their high spatial variability over small distances. Furthermore, this is a 1990 database with a much higher positional uncertainty than those obtained in later studies using GNSS. We also assume that the changes in the geomorphometric properties we use as dependent variables are negligible, but agricultural practices and run-off processes may have altered them. For these reasons reasons, we consider the results to be quite adequate under the circumstances and at the same time as an encouragement to obtain better data. The resulting NC layer still has higher accuracy in HMS than the official NC layer. We believe that these results encourage new systematic sampling campaigns to improve soil information in order to have a better modelling capability.

The differences between the different models used are very small and not statistically significant. It is common in papers using ML techniques not to carry out this check and end up finding that one model performs better than others with minimal differences in accuracy that may not be significant. The statistical significance of differences in accuracy statistics between different models should therefore be analysed, especially when these differences are small. Another interesting conclusion that can be drawn from these results is the need for better data, in addition to trying to find better algorithms.

Using the curve number layers derived from the estimates obtained in this work, the accuracy of a simple hydrological model, whose objective is to reproduce the runoff hydro-

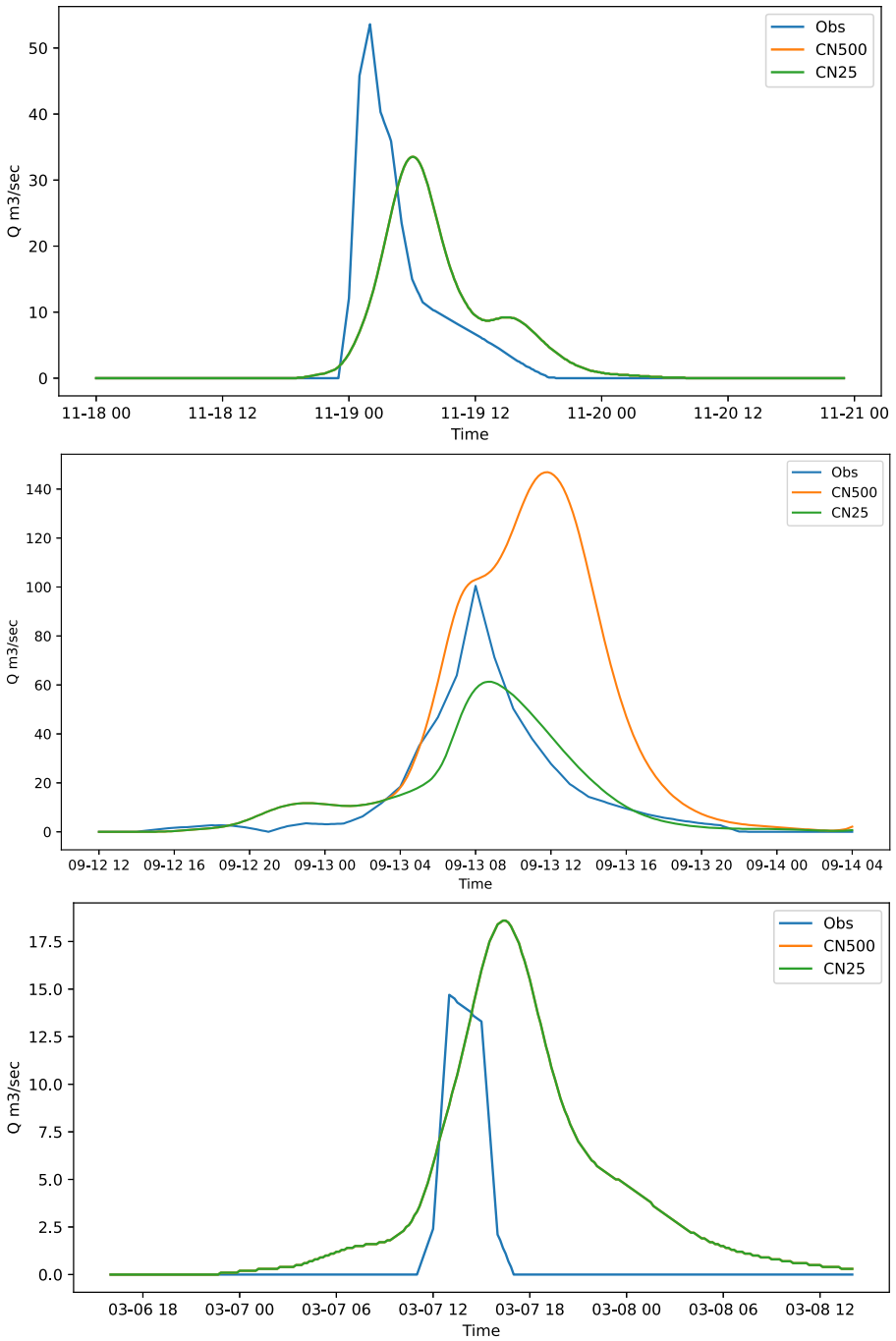

Fig. 10 Hydrographs of the 11/18-20/2018 (top), 09/12-14/2019 (middle) and 03/06-08/2021 (bottom) precipitation events. Both CN25 and CN500 produce the same hydrograph for the 2018 and 2021 events

graph of the Rambla del Albujón for the rainfall events of 09/12-14/2019, 11/18-20/2018 and 03/06-08/2021. The results show that the CN layer obtained in this work provides better hydrograph estimates than the official CN layer when the runoff volume is high, specially the time to peak, which is relevant for nowcasting, and the total runoff. Although these results show that the official data overestimate runoff, this does not mean that it would be the same in other study areas. The missrepresentation of soil properties in the model may have different effects in different places and for different events. Regardless of the accuracy of the models, we believe that this result reflects the interest of improving the parameters of environmental models in general and hydrological models in particular using ML techniques.

The importance of flooding as an environmental hazard makes it advisable to develop more work to improve the

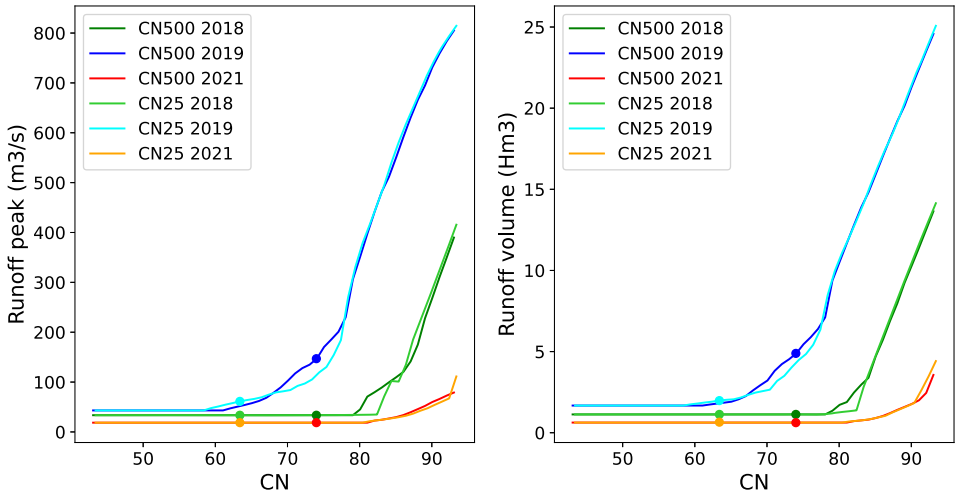

Fig. 11 Curve number sensitivity analysis. En el eje de X se presenta el CN medio ponderado por superficie de cada una de las siete subcuencas aguas arriba del punto de aforo 06A03Q01. Los cuadros representan el caudal pico simulado utilizando CN25; los puntos representan el caudal pico simulado utilizando CN500

knowledge of soil properties in areas prone to such events. In a broader context, several hydrological, erosion or agricultural models have been proposed and used in spatial planning to reconcile agricultural activity, tourism, environmental protection and environmental risk management in semi-arid areas. To achieve these objectives, it is clear that further work on soil sampling is needed to provide the scientific community with more densely sampled databases to obtain more accurate estimates of the different soil variables. Faster sampling and analysis techniques are needed to overcome the trade-off between higher sampling density and larger areas covered. Remote sensing data could also help. Recently, NASA installed a hyperspectral sensor (EMIT) on the International Space Station. The aim of this sensor is to identify minerals in the soil surface in arid and semi-arid areas. The data is only available since 2023, but it could be used as an additional predictor for models that estimate soil properties. Another interesting line of future work is the use of ensembles of different models to try to increase the resulting accuracy.

Author Contributions FAS did the machine learning interpolation and wrote the manuscript paper, ABB prepared the soil database, FGC and HJB did the hydrological modeling, ARD wrote the manuscript paper. All authors reviewed the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Grant TED2021-131131B-I00 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

Data Availability The database is not public.

Declarations

Competing Interests The authors declare no competing interests.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al (2016) Tensorflow: a system for large-scale machine learning. In: 12th {*USENIX*} Symposium on operating systems design and implementation ({*OSDI*} 16), pp 265–283

Akpa SI, Odeh IO, Bishop TF, Hartemink AE (2014) Digital mapping of soil particle-size fractions for Nigeria. Soil Sci Soc Am J 78(6):1953–1966. https://doi.org/10.2136/sssaj2014.05.0202

Akpa SI, Odeh IO, Bishop TF, Hartemink AE, Amapu IY (2016) Total soil organic carbon and carbon sequestration potential in Nigeria. Geoderma 271:202–215. https://doi.org/10.1016/j.geoderma.2016.02.021

Albaladejo-García JA, Alcon F, Martínez-Paz JM (2021) Irrigation and its influence on the local regulation of sur-

face temperatures in mediterranean citriculture. Revista de Geografía Norte Grande 79:123–137. https://doi.org/10.4067/S0718-34022021000200123
Alcón F, Miguel MD, Burton M (2011) Duration analysis of adoption of drip irrigation technology in southeastern spain. Technol Forecast Soc 78(6):991–1001. https://doi.org/10.1016/j.techfore.2011.02.001
Alexander M (1977) Introduction to soil microbiology. Wiley, New York
Al-Ghobari H, Dewidar A, Alataway A (2020) Estimation of surface water runoff for a semi-arid area using rs and gis-based scs-cn method. Water 12(7):1924. https://doi.org/10.3390/w12071924
Alias L, Ortiz R (2004) Memorias y mapas de suelos de las hojas del mtn a escala 1:100.000. proyecto lucdeme. Technical report, Ministerio de Medio Ambiente-ICONA
Alonso-Sarria F, Valdivieso-Ros C, Gomariz-Castillo F (2024) Analysis of the hyperparameter optimisation of four machine learning satellite imagery classification methods. Comput Geosci 28:551–571. https://doi.org/10.1007/s10596-024-10285-y
Anagu I, Ingwersen J, Utermann J, Streck T (2009) Estimation of heavy metal sorption in german soils using artificial neural networks. Geoderma 152:104–112
Angelini ME, Heuvelink GBM, Kempen B (2017) Multivariate mapping of soil with structural equation modelling. Eur J Soil Sci 68(5):575–591. https://doi.org/10.1111/ejss.12446
Arrouays D, Lagacherie P, Hartemink AE (2017) Digital soil mapping across the globe. Geoderma Reg 9:1–4. https://doi.org/10.1016/j.geodrs.2017.03.002
Ballabio C (2009) Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151(3–4):338–350
Bashir O, Bangroo SA, Shafai SS, Shah TI, Kader S, Jaufer L, Senesi N, Kuriqi A, Omidvar N, Kumar SN, Arunachalam A, Michael R, Ksibi M, Spalevic V, Sestras P, Marković SB, Billi P, Ercişli S, Hysa A (2024) Mathematical vs. machine learning models for particle size distribution in fragile soils of north-western himalayas. J Soils Sediments 24:2294–2308. https://doi.org/10.1007/s11368-024-03820-y
Baudron P, Alonso-Sarria F, García-Aróstegui JL, Cánovas-García F, Martínez-Vicente D, Moreno-Brotóns J (2013) Identifying the origin of groundwater samples in a multi-layer aquifer system with random forest classification. J Hydrol 499:303–315. https://doi.org/10.1016/j.jhydrol.2013.07.009
Beguin J, Fuglstad GA, Mansuy N, Paré D (2017) Predicting soil properties in the Canadian boreal forest with limited data: comparison of spatial and non-spatial statistical approaches. Geoderma 306:195–205
Behmann J, Mahlein AK, Rumpf T, Römer C, Plümer L (2015) A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. Precis Agric 16(3):239–260
Behrens T, Förster H, Scholten T, Steinrüken U, Spies E, Goldschmitt M (2005) Digital soil mapping using artificial neural networks. J Plant Nutr Soil Sci 168:21–33
Behrens T, Schmidt K, Ramirez-Lopez L, Gallant J, Zhu AX, Scholten T (2014) Hyper-scale digital soil mapping and soil formation analysis. Geoderma 213:578–588. https://doi.org/10.1016/j.geoderma.2013.07.031
Bennett ND, Croke BFW, Guariso G, Guillaume JHA, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LTH, Norton JP, Perrin C, Pierce SA, Robson B, Seppelt R, Voinov AA, Fath BD, Andreassian V (2013) Characterisig performance of environmental models. Environ Model Softw 40:1–20. https://doi.org/10.1016/j.envsoft.2012.09.011
Bergstrom DW, Monreal CM, Jacques ES (2001) Influence of tillage practice on carbon sequestration is scale-dependent. Can J Soil Sci 81:63–70
Beven KJ, Kirkby MJ (1979) Physically based, variable contributing area model of basin hydrology. Hydrolol Sci Bull 24:43–69
Bhattacharya B, Solomatine DP (2006) Machine learning in soil classification. Neural Netw 19(2):186–195
Bilgili AV, Es HM, Akbas F, Durak A, Hively WD (2010) Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of turkey. J Arid Environ 74:229–238
Blanco Bernardeau A (2015) Estudio de la distribución espacial y cartografía digital de algunas propiedades físicas, químicas e hidrodinámicas del suelo de la cuenca del segura. PhD thesis, Universidad de Murcia
Bobryk CW, Myers DB, Kitchen NR, Shanahan JF, Sudduth KA, Drummond ST, Gunzenhauser B, Gomez-Raboteaux NN (2016) Validating a digital soil map with corn yield data for precision agriculture decision support. Agron J 108(3):957–965. https://doi.org/10.2134/agronj2015.0381
Bondi G, Creamer R, Ferrari A, Fenton O, Wall D (2018) Using machine learning to predict soil bulk density on the basis of visual parameters: tools for in-field and post-field evaluation. Geoderma 318:137–147
Bonfatti BR, Hartemink AE, Giasson E, Tornquist CG, Adhikari K (2016) Digital mapping of soil carbon in a viticultural region of southern Brazil. Geoderma 261:204–221. https://doi.org/10.1016/j.geoderma.2015.07.016
Bossard M, Feranec J, Otahel J (2000) Corine land cover technical guide - addendum 2000, technical report no 40. Technical report, European Environment Agency, Kongens Nytorv 6, DK-1050 Copenhagen K, Denmark
Breiman L (2001) Random forests. Mach Learn 45(1):5–32
Brown DJ, Clayton MK, McSweeney K (2004) Potential terrain controls on soil color, texture contrast and grain-size deposition for the original catena landscape in uganda. Geoderma 122:51–72
Brungard CW, Boettinger JL, Duniway MC, Wills SA, Edwards TC Jr (2015) Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240:68–83
Buol SW, Hole FD, McCracken RJ (1989) Soil genesis and classification. Iowa State University Press, Ames
Burgess M, Webster R (1980) Optimal interpolation and isarithmic mapping of soil properties, i. The semi-variogramand punctual kriging. Eur J Soil Sci 31:315–331
Burgess M, Webster R (1980) Optimal interpolation and isarithmic mapping of soil properties, ii. Block kriging. Eur J Soil Sci 31:333–341
Byrne JM, Yang M (2016) Spatial variability of soil magnetic susceptibility, organic carbon and total nitrogen from farmland in northern China. Catena 145:92–98
Cai S, Zhang R, Liu L, Zhou D (2010) A method of salt-affected soil information extraction based on a support vector machine with texture features. Math Comput Model 51(11–12):1319–1325
Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 785–794. ACM, New York, NY, USA. https://doi.org/10.1145/2939672.2939785http://doi.acm.org/10.1145/2939672.2939785
Chow VT, Maidment DR, Mays LW (1987) Applied hydrology. McGraw Hill, New York, p 584
Conesa García C (1990) El Campo de Cartagena. Clima e Hidrología de Un Medio Semiárido. Universidad de Murcia. Secretariado de publicaciones: Vol.II., Murcia. https://www.chsegura.es/export/sites/chs/.galleries/descargas_libros-El_campo_de_Cartagena_Clima_e_hidrologia_de_un_medio_semiarido.pdf
Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Böhner J (2015) System for automated geoscientific analyses (saga) v 2.1.4. Geosci Model Dev 8:1991–2007. https://doi.org/10.5194/gmd-8-1991-2015

Coopersmith EJ, Minsker BS, Wenzel CE, Gilmore BJ (2014) Machine learning assessments of soil drying for agricultural planning. Comput Electron Agric 104:93–104
Corral-Pazos-de-Provens E, Rapp-Arrarás I, Domingo-Santos JM (2023) The usle soil erodibility nomograph revisited. Int Soil Water Conserv Res 11:1–13
Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forest for classification in ecology. Ecology 88(11):2783–2792. https://doi.org/10.1890/07-0539.1
Dobos E, Carré F, Hengl T, Reuter HI, Tóth G (2006) Digital soil mapping as a support to production of functional maps. Office for Official Publications of the European Communities, Luxemburg
Elavarasan D, Vincent DR, Sharma V, Zomaya AY, Srinivasan K (2018) Forecasting yield by integrating agrarian factors and machine learning models: a survey. Comput Electron Agric 155:257–282
Elshorbagy A, Parasuraman K (2008) On the relevance of using artificial neural networks for estimating soil moisture content. J Hydrol 362(1–2):1–18
Emadi M, Taghizadeh-Mehrjardi R, Cherati A, Danesh M, Mosavi A, Scholten T (2020) Predicting and mapping of soil organic carbon using machine learning algorithms in northern Iran. Remote Sens 12(14):2234
European Environment Agency (1995) Corine land cover. Technical report, Commision of the European Communities, Roma
FAO-UNESCO (1974) Soil Map of the World. UNESCO, Paris
Ferrer-Julià M (2003) Análisis de Nuevas Fuentes de Datos Para la Estimación del Parámetro Número de Curva Perfiles de Suelos Y Teledetección. CEDEX, Madrid
Forkuor G, Hounkpatin OK, Welp G, Thiel M (2017) High resolution mapping of soil properties using remote sensing variables in south-western burkina faso: a comparison of machine learning and multiple linear regression models. PLoS One 12:0170478
Gallant JC, Dowling TD (2003) A multi-resolution index of valley bottom flatness for mapping depositional areas. Water Resour Res 39:1347–1360
Géron A (2019) Hands-on machine learning with scikit-learn, keras, and tensorflow. O'Reilly Media Inc, Beijing Boston Farnham Sebastopol Tokyo
Gessler PE, Chadwick QA, Chamran F, Althouse LD, Holmes KW (1995) Soil landscape modelling and spatial prediction of soil attributes. Soil Sci Soc Am J 64:2046–2056
Gessler PE, Chadwick OA, Chamron F, Holmes K, Althouse L (2000) Modeling soil-landscape and ecosystem properties using terrain attributes. Soil Sci Soc Am J 64:2046–2056
Giménez-Casalduero F, Gomariz-Castillo F, Alonso-Sarría F, Cortés E, Izquierdo-Muñoz A, Ramos-Esplá AA (2020) Pinna nobilis in the mar menor coastal lagoon: a story of colonization and uncertainty. Mar Ecol Prog Ser 652:77–94. https://doi.org/10.3354/meps13468
Goldstein A, Fink L, Meitin A, Bohadana S, Lutenberg O, Ravid G (2018) Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist's tacit knowledge. Prec Agric 19(3):421–444
Gomes LC, Faria RM, Souza E, Veloso GV, Schaefer CEG, Fernandes Filho EI (2019) Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340:337–350
González-Sánchez A, Frausto Solís J, Ojeda Bustamante W (2014) Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res 12(2):313–328
Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford Univ. Press, Oxford
GRASS Development Team (2023) Geographic Resources Analysis Support System (GRASS GIS) Software, Version 8.3. Open Source Geospatial Foundation, USA. https://doi.org/10.5281/zenodo.5176030https://grass.osgeo.org. Open Source Geospatial Foundation
Grimm R, Behrens T (2010) Uncertainty analysis of sample locations within digital soil mapping approaches. Geoderma 155:154–163
Grimm R, Behrens T, Märker M, Eisenbeer H (2008) Soil organic carbon concentrations and stocks on barro colorado island - digital soil mapping using random forests analysis. Geoderma 146:102–113
Grunwald S, Thompson JA, Boettinger JL (2011) Digital soil mapping and modeling at continental scales: finding solutions for global issues. Soil Sci Soc Am J 75:1201–1213. https://doi.org/10.2136/sssaj2011.0025
Guisan A, Harrell FE (2000) Ordinal response regression models in ecology. J Veg Sci 11:617–626
Hartemink AE (2006) The future of soil science. Wageningen, IUSS
Hartemink AE, McBratney A (2008) A soil science renaissance. Geoderma 148:123–129
Hattab N, Hambli R, Motelica Heino M, Bourrat X, Mench M (2013) Application of neural network model for the prediction of chromium concentration in phytoremediated contaminated soils. J Geochem Explor 128:25–34
Hengl T, Heuvelink GB, Kempen B, Leenaars JG, Walsh MG, Shepherd KD, Sila A, MacMillan RA, Jesus JM, Tamene L (2015) Mapping soil properties of africa at 250 m resolution: random forests significantly improve current predictions. PLoS One 10:0125814
Hengl T, de Jesus Mendes J, Heuvelink GB, Ruiperez-González M, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, Guevara MA (2017) Soilgrids250m: Global gridded soil information based on machine learning. PLoS One 16;12(2):0169748
Hernández-Lobato JM, Adams R (2015) Probabilistic backpropagation for scalable learning of bayesian neural networks. Int Conf Mach Learn 1861–1869
Heung B, Ho HC, Zhang J, Knudby A, Bulmer CE, Schmidt MG (2016) An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265:62–77
James G, Witten D, Hastie T, Tibshirani R (2023) An introduction to statistical learning with applications in python. Springer, New York Heidelberg Dordrecht London
Jennes J (2005) Topographic position index (tpi jen.avx) extension for arcview 3.x. Technical report, Jenness Enterprises. http://www.jennessent.com
Jenny H (1941) Factors of Soil Formation. McGraw-Hill, New York London
Kalambukattu JG, Kumar S, Raj RA (2018) Digital soil mapping in a himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. Environ Earth Sci 77(5):203. https://doi.org/10.1007/s12665-018-7367-9
Karunaratne S, Bishop T, Baldock J, Odeh I (2014) Catchment scale mapping of measureable soil organic carbon fractions. Geoderma 219:14–23
Kavzoglu T, Sahin EK, Colkesen I (2014) Landslide susceptibility mapping using gis-based multi-criteria decision analysis, support vector machines, and logistic regression. Landslides 11(3):425–439. https://doi.org/10.1007/s10346-013-0391-7
Keskin H, Grunwald S, Harris WG (2019) Digital mapping of soil carbon fractions with machine learning. Geoderma 339:40–58
Khaledian Y, Miller BA (2020) Selecting appropriate machine learning methods for digital soil mapping. Appl Math Model 81:401–418. https://doi.org/10.1016/j.apm.2019.12.016
Khan SA, Satyanarayana V, Venugopal B (2015) An approach to predict soil nutrients and efficient irrigation for agriculture with spatial data mining. Int J Sci Res Dev 3(9):476–478
Khanal S, Fulton J, Klopfenstein A, Douridas N, Shearer S (2018) Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. Comput Electron Agric 153:213–225

Kim S, Singh VP (2014) Modeling daily soil temperature using data-driven models and spatial distribution. Theor Appl Clim 118(3):465–479

Kovačević M, Bajat B, Gajić B (2010) Soil type classification and estimation of soil properties using support vector machines. Geoderma 154(3–4):340–347. https://doi.org/10.1016/j.geoderma.2009.11.005

Lagacherie P, McBratney AB (2006) Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. In: McBratney AB, Lagacherie P, Voltz M (eds) Digital Soil Mapping - An Introductory Perspective, vol 31. Elsevier, Amsterdam Oxford, pp 3–22

Lal R (2002) Soil carbon dynamics in cropland and rangeland. Environ Pollut 116:353–362

Lary DJ, Alavi AH, Gandomi AH, Walker AL (2016) Machine learning in geosciences and remote sensing. Geosci Front 7(1):3–10. https://doi.org/10.1016/j.gsf.2015.07.003

Legates DR, McCabe GJ (1999) Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydro-climatic model validation. Water Resourc Res 35(1):233–241. https://doi.org/10.1029/1998WR900018

Lemercier B, Lacoste M, Loum M, Walter C (2012) Extrapolation at regional scale of local soil knowledge using boosted classification trees: a two-step approach. Geoderma 171–172:75–84

Li Y, Liang S, Zhao Y, Li W, Wang Y (2017) Machine learning for the prediction of l. Chinensis carbon, nitrogen and phosphorus contents and understanding of mechanisms underlying grassland degradation. J Environ Manag 192:116–123

Li Y, Rahardjo H, Satyanaga A, Rangarajan S, Lee DT-T (2022) Soil database development with the application of machine learning methods in soil properties prediction. Eng Geol 306:106769. https://doi.org/10.1016/j.enggeo.2022.106769

Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22

Ließ M, Glaser B, Huwe B (2012) Uncertainty in the spatial prediction of soil texture: comparison of regression tree and random forest models. Geoderma 170(5):70–79

Lin Y, Prentice SE III, Tran T, Bingham NL, King JY, Chadwick OA (2016) Modeling deep soil properties on california grassland hillslopes using lidar digital elevation models. Geoderma Reg 7(1):67–75

Liu F, Geng X, Zhu A, Fraser W, Wadell A (2012) Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from modis. Geoderma 171–172:44–52

Mansuy N, Thiffault E, Paré D, Bernier P, Guindon L, Villemaire P, Poirier V, Beaudoin A (2014) Digital mapping of soil properties in canadian managed forests at 250 m of resolution using the k-nearest neighbor method. Geoderma 235:59–73. https://doi.org/10.1016/j.geoderma.2014.06.032

Martin MP, Orton TG, Lacarce E, Meersmans J, Saby NPA, Paroissien JB, Arrouays D (2014) Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. Geoderma 223:97–107

Martinelli G, Gasser M-O (2022) Machine learning models for predicting soil particle size fractions from routine soil analyses in quebec. Soil Sci Soc Am J 86:1509–1522. https://doi.org/10.1002/saj2.20469

Martínez-Hernández C, Rodrigo-Comino J, Romero-Díaz A (2017) Impact of lithology and soil properties on abandoned dryland terraces during the early stages of soil erosion by water in south-east spain. Hydrol Process 31(17):3095–3109. https://doi.org/10.1002/HYP.11251

Masri D, Woon WL, Aung Z (2015) Soil property prediction: an extreme learning machine approach. In: Arik S, Huang T, Lai WK, Liu Q (eds) International conference on neural information processing. Springer, Cham Heidelberg New York Dordrecht London, pp 18–27

Maynard JJ, Johnson MG (2014) Scale-dependency of lidar derived terrain attributes in quantitative soil-landscape modeling: Effects of grid resolution vs. neighborhood extent. Geoderma 230–231:29–40. https://doi.org/10.1016/j.geoderma.2014.03.021

McBratney AB, Odeh IO, Bishop TF, Dunbar MS, Shatar TM (2000) An overview of pedometric techniques for use in soil survey. Geoderma 97:293–327. https://doi.org/10.1016/S0016-7061(00)00043-4

McBratney AB, Mendonca ML, Minasny B (2003) On digital soil mapping. Geoderma 117(1–2):3–52

McCarthy GT (1938) The unit hydrograph and flow routing. In: Conf. North Atlantic Div., U.S. Corps of Engineers, New London, Connecticut

McKenzie NJ, Ryan PJ (1999) Spatial prediction of soil properties using environmental correlation. Geoderma 89:67–94

Meier M, Souza ED, Francelino MR, Fernandes Filho EI, Schaefer CEGR (2018) Digital soil mapping using machine learning algorithms in a tropical mountainous area. Revista Brasileira de Ciência do Solo 42

Miller BA (2012) The need to continue improving soil survey maps. Soil Horizons 53(3):11–15. https://doi.org/10.2136/sh12-02-0005

Miller BA, Koszinski S, Wehrhan M, Sommer M (2015) Comparison of spatial association approaches for landscape mapping of soil organic carbon stocks. Soil 1:217–233. https://doi.org/10.5194/soil-1-217-2015

Minasny B, Hartemink AE (2011) Predicting soil properties in the tropics. Earth-Sci Rev 106(1–2):52–62

Minasny B, McBratney AB (2016) Digital soil mapping: a brief history and some lessons. Geoderma 264:301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Minasny B, Setiawan BI, Saptomo SK, McBratney AB (2018) Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma 313:25–40. https://doi.org/10.1016/j.geoderma.2017.10.018

de Medio Ministerio, Ambiente y Medio Rural y Marino (2011) Guía metodológica para el desarrollo del sistema nacional de cartografía de zonas inundables. Technical report, Ministerio de Medio Ambiente y Medio Rural y Marino

Moore ID, Burch GJ (1986) Modelling erosion and deposition: topographic effects. Trans Am Soc Agric Eng 29:1624–1630

Moore ID, Gessler PE, Nielsen GA, Peterson GA (1993) Soil attribute prediction using terrain analysis. Soil Sci Soc Am J 57:443–452. https://doi.org/10.2136/sssaj1993.03615995005700020026x

Morgan J, Daugherty R, Hilchie A, Carey B (2003) Sample size and modeling accuracy of decision tree based data mining tools. AIMSJ 6:71–99

Motia S (1950) Reddy S (2021) Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation. J Phys Conf Ser 1:012037. https://doi.org/10.1088/1742-6596/1950/1/012037

Mulder V, De Bruin S, Schaepman M, Mayr T (2011) The use of remote sensing in soil and terrain mapping-a review. Geoderma 162:1–19

Nadeu Puig-Pey E (2013) Soil erosion and organic carbon mobilization at the catchment scale: factors, processes and impact on the carbon balance. PhD thesis, University of Murcia

Pachepsky YA, Timlin DJ, Rawls WJ (2001) Soil water retention as related to topographic variables. Soil Sci Soc Am J 65:1787–1795

Padarian J, Minasny B, McBratney AB (2020) Machine learning and soil sciences: a review aided by machine learning tools. Soil 6(1):35–52. https://doi.org/10.5194/soil-6-35-2020

Pahlavan-Rad MR, Akbarimoghaddam A (2018) Spatial variability of soil texture fractions and ph in a flood plain (case study from eastern iran). Catena 160:275–281

Pansu M, Gautheyrou J (2006) Handbook of soil analysis mineralogical. Organic and Inorganic Methods, Springer, Berlin Heidelberg New York

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Pinheiro EF, Ceddia MB, Clingensmith CM, Grunwald S, Vasques GM (2017) Prediction of soil physical and chemical properties by visible and near-infrared diffuse reflectance spectroscopy in the central amazon. Remote Sens 9(4):293

Poggio L, Gimona A, Brown I, Castellazzi M (2010) Soil available water capacity interpolation and spatial uncertainty modelling at multiple geographical extents. Geoderma 160:175–188

Pradhan B (2013) A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using gis. Comput Geosci 51:350–365. https://doi.org/10.1016/j.cageo.2012.08.023

Prasad R, Deo RC, Li Y, Maraseni T (2018) Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. Soil Till Res 181:63–81

Qi H, Paz-Kagan T, Karnieli A, Jin X, Li S (2018) Evaluating calibration methods for predicting soil available nutrients using hyperspectral vnir data. Soil Tillage Res 175:267–275

Quinn P, Beven K, Chevalier P, Planchon O (1991) The prediction of hillslope flow paths for distributed hydrological modeling using digital terrain models. Hydrol Process 5(1):59–79

Rahman SAZ, Mitra KC, Islam SM (2018) Soil classification using machine learning methods and crop suggestion based on soil series. In: 21st International conference of computer and information technology

Ramirez-Lopez L, Wadoux AC, Franceschini MH, Terra FS, Marques KP, Sayão VM, Demattê JA (2019) Robust soil mapping at the farm scale with vis-nir spectroscopy. Eur J Soil Sci 70(2):378–393. https://doi.org/10.1111/ejss.12752

Ramírez-Santiagosa I, Vicente-Albadalejo M, García-Barceló JA, Vaquero-Gómez A (1999) Mapa digital de suelos de la región de murcia. Technical report, Comunidad Autónoma de la Región de Murcia

Ranjbar F, Jalali M (2016) The combination of geostatistics and geochemical simulation for the site-specific management of soil salinity and sodicity. Comput Electron Agric 121:301–312

Reza Pahlavan-Rad M, Akbarimoghaddam A (2018) Spatial variability of soil texture fractuions and ph in a flood plain (case study from eastern iran. Catena 160:275–281

Rezaei SA, Gilkes RJ (2005) The effects of landscape attributes and plant community on soil chemical properties in rangelands. Geoderma 125:167–176

Rial M, Cortizas AM, Taboada T, Rodríguez-Lado L (2017) Soil organic carbon stocks in santa cruz island, galapagos, under different climate change scenarios. Catena 156:74–81

Riley SJ, De Gloria SD, Elliot R (1999) A terrain ruggedness index that quantifies topographic heterogeneity. Int J Sci 5:123–127

Rodrigo-Comino J, Senciales JM, Cerdà A, Brevik EC (2018) The multidisciplinary origin of soil geography: a review. Earth Sci Rev 177:114–123

Rodríguez-Calles L (2022) Ecological impacts of agribusiness transformation in a spanish mediterranean enclave: impacts on the mar menor coastal lagoon. Eurochoices 21(2):43–49. https://doi.org/10.1111/1746-692X.12362

Romero Díaz A, Belmonte Serrato FB (2011) El campo de cartagena una visión global. In: Bastida JH (ed) Recorridos Por el Campo de Cartagena. Instituto Euromediterráneo del Agua, Murcia, pp 17–48

Rossiter DG (2018) Past, present & future of information technology in pedometrics. Geoderma 324:131–137

Ruiz-Álvarez M, Alonso-Sarria F, Gomariz-Castillo F (2019) Interpolation of instantaneous air temperature using geographical and modis derived variables with machine learning techniques. ISPRS Int J Geo-Inf 8(9):382. https://doi.org/10.3390/ijgi8090382

Sarmadian F, Azimi S, Keshavarzi A, Ahmadi A (2013) Neural computing model for prediction of soil cation exchange capacity: a data mining approach. Intern. J Agron Plant Prod 4(7):1706–1712

Schirrmann M, Gebbers R, Kramer E (2013) Performance of automated near-infrared reflectance spectrometry for continuous in situ mapping of soil fertility at field scale. Vadose Zone J 12(4). https://doi.org/10.2136/vzj2012.0199

Scull P, Franklin J, Chadwick OA (2005) The application of classification tree analysis to soil type prediction in a desert landscape. Ecol Model 181(1):1–15

Sharififar A, Sarmadian F, Malone BP, Minasny B (2019) Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma 350:84–92

Shi J, Yang L, Zhu A-X, Qin C, Liang P, Zeng C, Pei T (2018) Machine-learning variables at different scales vs. knowledge-based variables for mapping multiple soil properties. Soil Sci Soc Am J 82(3):645–656. https://doi.org/10.2136/sssaj2017.11.0392

Silva Chagas C, Carvalho Junior W, Bhering SB, Filho BC (2016) Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena 139:232–240. https://doi.org/10.1016/j.catena.2016.01.001

Sirsat M, 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. Comput Electron Agr 154:120–133

Smith MP, Zhu AX, Burt JE, Stiles C (2006) The effects of dem resolution and neighborhood size on digital soil survey. Geoderma 137:58–69. https://doi.org/10.1016/j.geoderma.2006.07.002

Soil Conservation Service (1972) National engineering handbook. Section 4. Hidrology. Technical report, U.S. Dept. of Agriculture. Whasington D.C

Spiegelhalter D (2019) The art of statistics: learning from data. Pelican, UK, p 424

Stum AK (2010) Random forests applied as a soil spatial predictive model in arid utah. PhD thesis, Utah State University

Sumfleth K, Duttmann R (2008) Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators. Ecol Indic 8(5):485–501

Taghizadeh-Mehrjardi R, Minasny B, Sarmadian F, Malone B (2014) Digital mapping of soil salinity in ardakan region, central Iran. Geoderma 213:15–28

Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R (2016) Digital mapping of soil organic carbon at multiple depths using different data mining techniques in baneh region, iran. Geoderma 266:98–110

Taghizadeh-Mehrjardi R, Minasny B, Toomanian N, Zeraatpisheh M, Amirian-Chakan A, Triantafilis J (2019) Digital mapping of soil classes using ensemble of models in isfahan region, iran. Soil Syst 3(2):37

Thompson JA, Kolka RK (2005) Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. Soil Sci Soc Am J 69:1086–1093

Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 49(11):1225–1231

ULE (2009) Obtención del parámetro del umbral de escorrentía para la españa peninsular a partir de nuevas fuentes de datos. informe técnico para el centro de estudios hidrográficos. Technical report, Universidad de León

Umali BP, Oliver DP, Forrester S, Chittleborough DJ, Hutson JL, Kookana RS, Ostendorf B (2012) The effect of terrain and man-

agement on the spatial variability of soil properties in an apple orchard. Catena 93:38–48
Wade C (2020) Hands-On Gradient Boosting with XGBoost and Scikit-learn: perform accessible machine learning and extreme gradient boosting with python. Packt
Wälder K, Wälder O, Rinklebe J, Menz J (2008) Estimation of soil properties with geostatiscal methods in floodplains. Arch Agron Soil Sci 54(3):275–295
Wang B, Waters C, Orgill S, Cowie A, Clark A, Li Liu D, Simpson M, McGowen I, Sides T (2018) Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern australia. Ecol Indic 88:425–438
Webster R, Burgess TM (1980) Optimal interpolation and isarithmic mapping, iii. Changing drift and universal kriging. J Soil Sci 31:505–524
Were K, Bui DT, Dick OB, Singh BP (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an afromontane landscape. Ecol Indic 52:394–403. https://doi.org/10.1016/j.ecolind.2014.12.028
Wiesmeier M, Berthold F, Blank B, Kögel-Knabner I (2011) Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. Plant Soil 340:7–24
Wilcke W, Yasin S, Schmitt A, Valarezo C, Zech W (2008) Gradients in a Tropical Mountain Ecosystem of Ecuador. 9. Soils Along the Altitudinal Transect and in Catchments. Springer
Wilson HF, Satchithanantham S, Moulin AP, Glenn AJ (2016) Soil phosphorus spatial variability due to landform, tillage, and input management: a case study of small watersheds in southwestern manitoba. Geoderma 280:14–21
Wu W, Zucca C, Muhaimeed AS, Al-Shafie WM, Fadhil Al-Quraishi AM, Nangia V, Liu G (2018) Soil salinity prediction and mapping by machine learning regression in central mesopotamia. Iraq. Land Degrad Dev 29(11):4005–4014
Xiong X, Grunwald S, Myers DB, Kim J, Harris WG, Comerford NB (2014) Holistic environmental soil-landscape modeling of soil organic carbon. Environx Modell Softw 57:202–215
Yang R, Hu J, Li Z, Mu J, Yu T, Xia J, Li X, Dasgupta A, Xiong H (2024) Interpretable machine learning for weather and climate prediction: a review. Atmos Environ 338:120797. https://doi.org/10.1016/j.atmosenv.2024.120797
Yao X, Tham LG, Dai FC (2008) Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, china. Geomorphology 101(4):572–582. https://doi.org/10.1016/j.geomorph.2008.02.011
Zare E, Huang J, Triantafilis J (2016) Identifying soil landscape units at the district scale by numerically clustering remote and proximal sensed data. Comput Electron Agric 127:510–520
Zeraatpisheh M, Ayoubi S, Jafari A, Finke P (2017) Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in iran. Geomorphology 285:186–204. https://doi.org/10.1016/j.geomorph.2017.02.015
Zhang Y, Sui B, Shen H, Wang Z (2018) Estimating temporal changes in soil ph in the black soil region of northeast China using remote sensing. Comput Electron Agric 154:204–212
Zhu AX, Moore A, Burt JE (2006) Prediction of soil properties using fuzzy membership. In: 2nd Global Workshop on Digital Soil Mapping, Rio de Janeiro, Brazil, vol 4
Zhu AX, Liu F, Li BL, Pei T, Qin CZ, Liu GH, Wang YJ, Chen YN, Ma XW, Qi F, Zhou CH (2010) Differentiation of soil conditions over low relief areas using feedback dynamic patterns. Soil Sci Soc Am J 74:861–869
Zhu AX, Liu J, Du F, Zhang SJ, Qin CZ, Burt J et al (2015) Predictive soil mapping with limited sample data. Eur J Soil Sci 66:535–547. https://doi.org/10.1111/ejss.12752