Measurement 201 (2022) 111706

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Three-dimensional mapping of soil organic carbon using soil and environmental covariates in an arid and semi-arid region of Iran

Seyed Roohollah Mousavi [a], Fereydoon Sarmadian [a,*], Mahmoud Omid [b], Patrick Bogaert [c]

[a] Soil Science and engineering Department, Faculty of Agricultural Engineering and Technology, College of Agriculture & Natural Resources, University of Tehran, Karaj, Iran
[b] Agricultural Machinery Engineering Department, Faculty of Agricultural Engineering and Technology, University of Tehran, Karaj, Iran
[c] Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium

ARTICLE INFO

Keywords:
Soil organic carbon
Digital soil mapping
Partial dependency
Machine learning
Hybridization

ABSTRACT

This study focus on modeling and mapping soil organic carbon (SOC) at high spatial resolution and at four standard depths in an arid and semi-arid region of Iran. The SOC data includes 850 soil samples collected from 278 observation profiles. In parallel, a wide range of environmental covariates (n = 62) were obtained from multiple sources. Six individual machine learning (ML) algorithms were compared to modeling and predicting SOC. Two scenarios were investigated. The first one accounts for soil and environmental covariates (S1) while the second one only accounts for environmental covariates (S2). Our results show that accounting for soil variables in the prediction (S1) leads to a twofold increase of R^2 for all ML algorithms, while random forest (RF) outperformed the other ML approaches at all depths. Whenever possible, using additionally the soil variables that are at hand in a study area is thus beneficial for improving SOC predictions.

1. Introduction

Soil organic carbon (SOC) is a key indicator of soil health, quality, and fertility. SOC content is influenced by various factors including climate, vegetation, soil characteristics, soil water regimes, land use, topography, and type of management [1]. It is an important indicator for sustainable agricultural management, and it also plays a vital role in mitigating global warming and increasing crop production. The latter can be achieved by enhancing knowledge on SOC status to spatially manage soil fertility by using fertilizers in a better way [2,3]. In parallel, assessing the spatial variability of SOC at a regional scale may help to establish suitable agricultural management, land-use practices, and ecosystem services [4].

Given the global importance of SOC, the application of digital soil mapping (DSM) techniques to predict SOC has received considerable attention over the past decades [5,6]. Current approaches describe the spatial distribution of SOC by considering its relationships with environmental covariates [7,8]. These covariates often include easily accessible information sources such as topographic attributes derived from digital elevation models (DEM) [9], remotely sensed (RS) data [10,11] and climatic parameters [12]. Recently, machine learning (ML) algorithms that are making use of such covariates have been popular for predicting SOC [10]. Numerous ML algorithms have been proposed, i.e., random forest (RF) [13,14], artificial neural network (ANN) [15,16], support vector regression (SVR) [17], cubist (CB) regression [18], boosted regression tree (BRT) [13], and multiple linear regression (MLR) [19], among others. By comparison with classical linear geostatistical methods, ML algorithms are more easily accounting for possible nonlinear relationships between SOC and covariates [20], with the drawback of being black box models due to their limited interpretability [21].

As it is well known that soils properties are usually varying with depth [22], understanding how the performances of ML algorithms and the effect of environmental covariate is influenced by this depth is an important topic, especially if the goal is to predict SOC both spatially and vertically (see, e.g., [23–25]). Among other studies, [16] compared SVR, neural network and RF methods to predict SOC at two depths (0–15 and 15–30 cm) using soil, climate, topography, and Landsat 8 covariates; they showed that SVR outperforms other ML algorithms, while soil nitrogen, aspect, slope gradient, precipitation, and NDVI index are the most important covariates. Gomes et al. [26] compared CB, RF, generalized linear model (GLM), and SVR to predict SOC storage at 0–5, 5–15, 15–30, 30–60 cm standard depths; they concluded that RF was the best method. Taghizadeh-Mehrjardi et al. [17] and Zeraatpisheh et al. [27] separately modeled the vertical (up to 1 m depth) and surface (0–20 cm) SOC using various ML algorithms in a

* Corresponding author.
E-mail address: fsarmad@ut.ac.ir (F. Sarmadian).

https://doi.org/10.1016/j.measurement.2022.111706
Received 24 February 2022; Received in revised form 15 July 2022; Accepted 30 July 2022
Available online 5 August 2022

semi-arid and arid region of Iran; they identified ANN and CB as the best methods.

Clearly, the various ML algorithms may have distinct performances, as measured using classical statistics as mean error (ME), root mean square error (RMSE), determination coefficient (R^2) or Lin's concordance correlation coefficient (CCC). However, when the prediction residuals exhibit a spatial dependence, this dependence can be used for subsequently enhancing the predictions. Using ML learning methods and improving the final results using prediction residuals is commonly referred to as hybridization. Many studies [28,29] have successfully applied residual kriging to improve the prediction performances of the ML models. Silatsa et al. [30] used RF and generalized boosted regression (GBR) models with their hybridized version for predicting SOC storage at five soil depths in Cameron; they concluded that hybridization indeed improves the predictions when some spatial information content is still contained in the prediction residuals, as estimated from a spatial covariance function or variogram.

By relying on these previous results and findings, our study is focusing on the vertical (i.e., at four standard depths) and lateral (i.e., spatial mapping) prediction of SOC content using a set of ML algorithms and is addressing the following questions: (i) beyond the classical use of environmental covariates (as obtained from DEM, RS variables and climatic data), what is the benefit of additionally accounting for soil variables that could be measured at the soil profiles along with SOC content, and (ii) what is the additional benefit of hybridization based on the spatial information still contained in the prediction residuals. Clearly, question (i) is of paramount importance for the sampling campaign, as measuring additionally specific soil properties (beside SOC content) should be accounted for at the planning stage of a campaign. Using the most beneficial subset of these soil variables in parallel with environmental covariates could lead to increased prediction performances at a moderate additional cost.

2. Materials and methods

2.1. Description of the study area

This study was conducted at the regional scale in the northwest of Iran. The study area covers 57,540 hectares and is partly located in the Alborz (north of Nazarabad city) and Qazvin (south of Abyek city) provinces (Fig. 1). It extends from longitude 50°15′35.06″ to 50°29′25.53″ E and from latitude 35°54′38.30″ to 36°54′38.83″ N. Mean annual precipitation is 280 mm and mean annual air temperature is 14 ° C (Qazvin synoptic station, 1970–2019), with weak Aridic, Dry Xeric, Aquic soil moisture, and thermic temperature regimes [31].

Main soil parent materials are basaltic volcanic rocks, shale, volcanic lava, green tuffs and alluvial/colluvial deposits that relate to Tertiary and Quaternary. The lithological information was gathered from digitized 1:100,000 geology maps of Abyek, Karaj, Eshtehard and Shakeran areas from the Geological Survey of Iran. The soils are classified under three taxonomic orders, namely Aridisols, Inceptisols, and Entisols. The area is characterized by five main typical landscapes from north to south, including Mountain, Hilland, Peneplain, Piedmont, and Plain, based on geomorphic landscape classification [32]. The study area is rather flat, with elevation ranging from 1141 m a.s.l. close to the plain landscape up to 1747 m a.s.l. on the mountain side. Slopes are lower than 5% for about 70% of the area, reaching up to 15% in Peneplain, Hilland and Mountain landscapes.

The Qazvin plain is one of the strategic plains for agricultural productions. As it is under continuous cultivation and exploitation for a long time, soils are at risk of salinization and reduced fertility. Dominant land uses are irrigated crops, saline pasture, dry farming, and non-saline pastures. Winter wheat, barley, alfalfa, maize and orchards are the main agricultural productions.

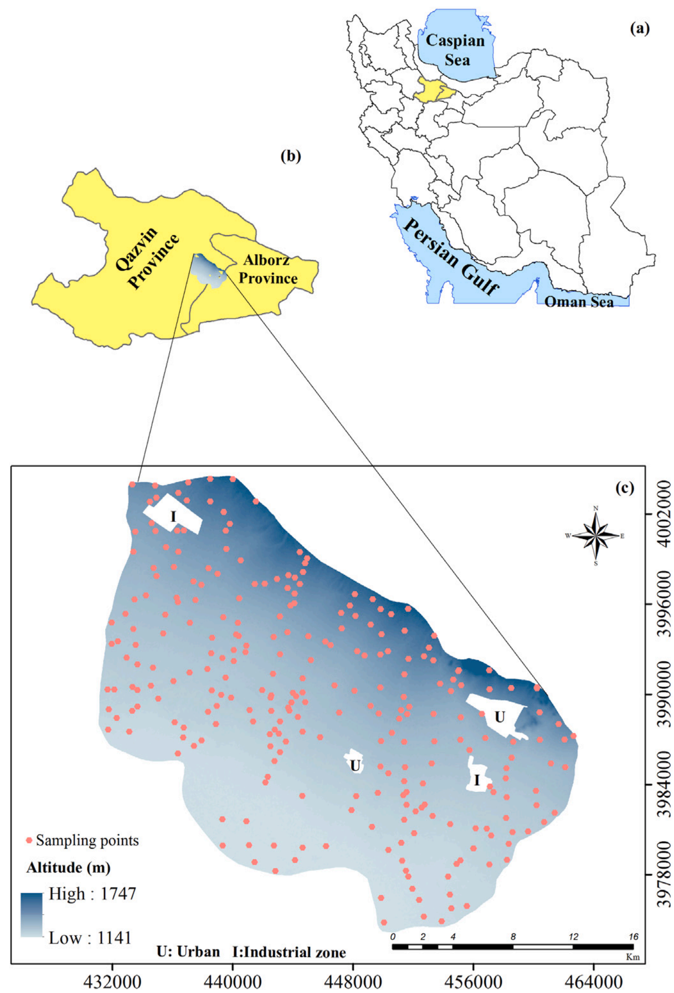

Fig. 1. Location of the Qazvin and Alborz provinces in Iran **(a)**, with a focus on the Qazvin plain **(b)** and the spatial location of the sampled soil profiles in the study area **(c)**.

2.2. Soil sampling and soil variables

There is no prior knowledge about spatial variability of SOC in this area, neither from legacy soil maps nor from worldwide products such as SoilGrids 250 m [33], that has no observation points in the area. Instead, soil data have been gathered from various local sampling campaigns that were conducted from 2016 till 2020 over various parts of the study area. Soil profiles were typically selected using random sampling for each of these campaigns. A total number of 278 soil profiles were gathered, with an average distance of 1200 m between profiles (Fig. 1).

All profiles were dug to a depth of 200 cm or above the lithic or para-lithic contact layer. They were described according to their morphology and identified pedo-genetic horizons. Disturbed soil samples were collected from each genetic horizon. Soil samples were air-dried and passed through a 2 mm sieve. SOC was measured using wet oxidation [34]. Besides SOC measurements, some soil properties that are considered as key ones in arid and semi-arid regions from expert knowledge [35] were also measured, namely pH, bulk density (BD), soil texture fractions (Clay, Silt, Sand), calcium carbonate equivalent (CCE), and electrical conductivity (EC).

For obtaining soil properties along the vertical profiles at all possible depths, equal-area quadratic splines were used to obtain these soil variables as continuous functions of depth using the "GSIF" package in R. The values for SOC, pH, EC, CCE, BD, Sand, Silt, Clay where finally

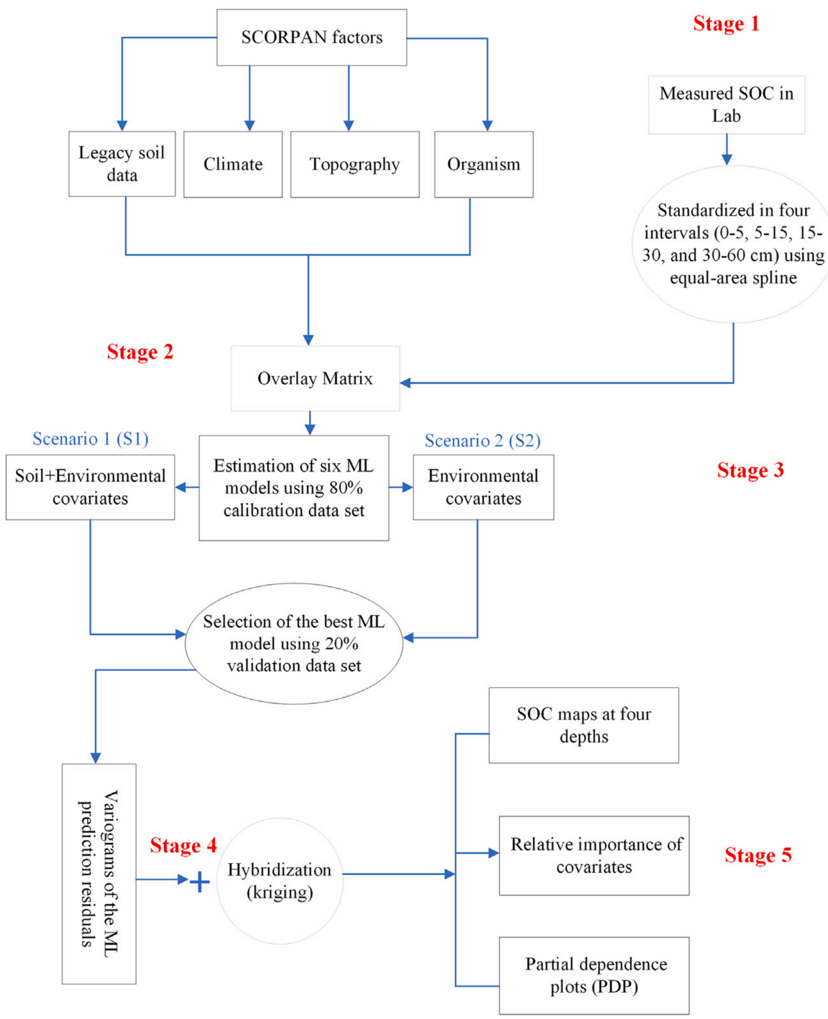

Fig. 2. Workflow for predicting SOC content at the four standard soil depths.

obtained at four standard depths (0–5, 5–15, 15–30, and 30–60 cm). A detailed mathematical description of equal-area quadratic splines can be found in [23,36]. Pearson's correlation coefficient (*r*) was used to measure correlation between SOC and the various soil properties (pH, EC, CCE, BD, Sand, Silt, Clay) at the sampled locations (Fig. 3).

2.3. Environmental covariates

A total of 55 covariates were selected as potentially related to soil forming factors in the SCORPAN model [37]. These covariates were gathered from different sources such as DEM, RS images, climate data and soil properties. Six of the most important covariates are presented in Fig. 4. For modeling purposes, all covariates were resampled at a 15 m spatial resolution using a bilinear resampling method in the SAGA-GIS software [38].

2.3.1. Topography

DEM was obtained from ALOS PALSAR (https://asf.alaska.edu/) with an initial spatial resolution of 12.5 m and was used to derive 30 topographic attributes. Geomorphometric attributes (e.g., elevation, aspect, valley depth, wetness index, modified catchment area, diffuse insolation, total insolation, etc.) were produced using the SAGA-GIS software [39].

2.3.2. Remotely sensed images

A total of 20 RS indices and spectral band values were extracted from Landsat 8 images (OLI/TIRS) with a cloud cover lower than 10%. The optical data were downloaded from USGS (https://www.usgs.gov/) between summer and autumn 2016 to 2020, thus covering the dates of the various soil sampling campaigns that were gathered for this study. ERDAS IMAGINE software was used for radiometric corrections. Several individual bands were used (bands 2, 3, 4, 5, 6, 7, 10 and 11) along with ratios of vegetation and spectral indices, i.e., Difference Vegetation Index (DVI), Normalized Difference Vegetation Index (NDVI), Ratio Vegetation Index (RVI), Transformed Vegetation Index (TVI), Corrected Transformed Ratio Vegetation Index (CTVI), Soil Adjusted Vegetation Index (SAVI), Green Normalized Difference Vegetation Index (GNDVI), Modified soil-adjusted vegetation index (MSAVI). In parallel, Clay index, Gypsum index, Salinity index, and Carbonate index were also obtained.

2.3.3. Land use

As there was no legacy land use map at hand for the study area, we prepared it from a set of 2000 observation points where land use was known. Five main land use classes were defined, namely irrigated land (the dominant land use covering 52% of the area), saline pasture (22%, scattered and mainly occurring in the low relief zones), dry farming (16%), non-saline pasture (8%), and urban and industrial zone (2%). Using four Landsat 8 spectral bands (bands 3, 4, 5 and 6) and the DEM, land use was predicted using a RF algorithm, with a resulting overall

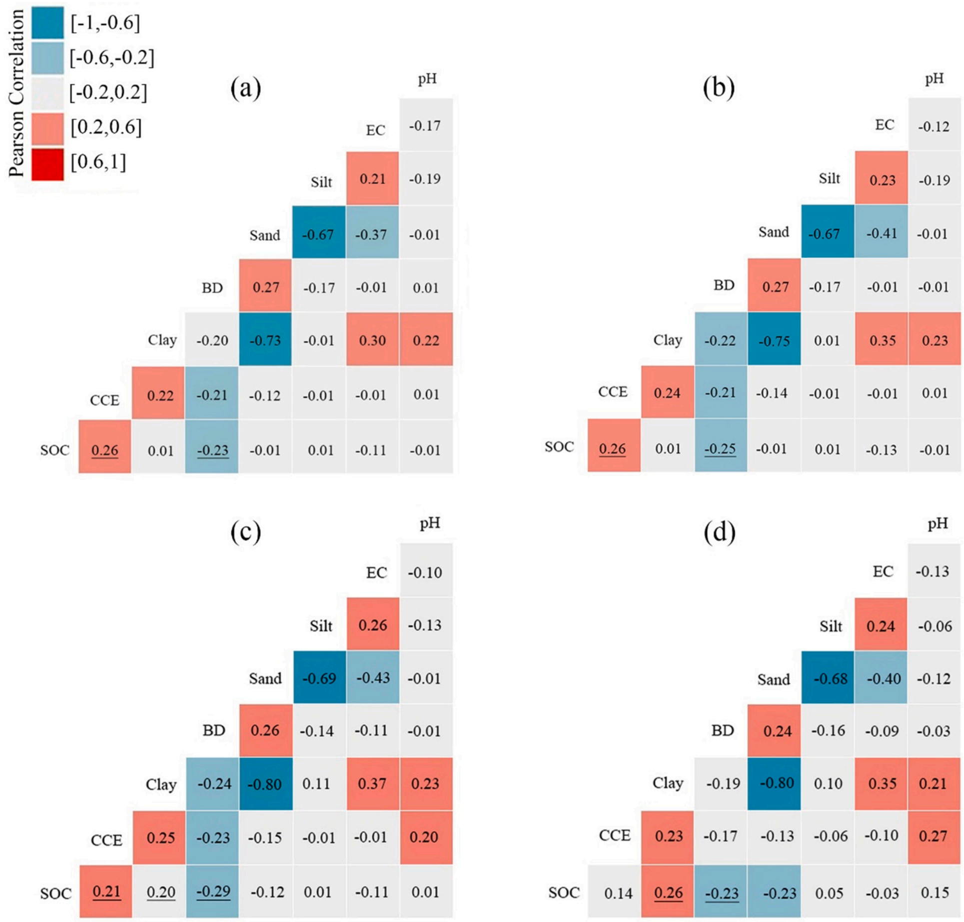

Fig. 3. Correlations between SOC and the various soil measured physico-chemical properties at the four standard soil depths, i.e., 0–5 cm **(a)**, 5–15 cm **(b)**, 15–30 cm **(c)** and 30–60 cm **(d)**. Underlined correlation values between SOC and other soil variables are significantly different from zero.

accuracy of 94% for these classes. Using this fitted RF model, the land use map was finally predicted over a regular grid at the required spatial resolution.

2.3.4. Climate

The monthly climate data obtained from 24 weather stations were gathered for generating maps of maximum, minimum and mean annual temperature (MaxT, MinT and MAT) and precipitation (MAP) based on the Empirical Bayesian kriging (EBK) interpolation method using Arc GIS software. The EBK method was selected because (i) it accounts for the most difficult aspects of building a valid kriging model, (ii) in comparison with other kriging methods, EBK is more accurate for the prediction of relatively non-static data, particularly when limited data are at hand [40]. The original maps were obtained at 250 m spatial resolution and were resampled at 15 m spatial resolution using a bilinear resampling method.

2.4. Scenarios and processing workflow

Two scenarios were investigated for predicting and mapping SOC using various ML algorithms. The first one (S1) is accounting for all DEM, RS, climate covariates and for soil variables that were also measured for the sampled profiles at the various depths, in parallel to SOC measurements. The second scenario (S2) only accounts for DEM, RS, climate covariates by considering that soil variables other than SOC are not available for the modeling and prediction process. Comparing results for S1 and S2 allows assessing the benefit of accounting for soil properties, while traditionally ML algorithms tend to discard them, either because they were not measured in parallel with SOC or because they are not initially available at the nodes of the prediction grid.

The workflow for modeling and predicting SOC at the four selected standard depths is detailed in Fig. 2. The mains stages are : (1) collection of soil samples and measurement of SOC content, along with other soil variables and environmental covariates; (2) creation of an overlay matrix for SOC and the various SCORPAN factors at the four standard depth (0–5 cm, 5–15 cm, 15–30 cm, and 30–60 cm); (3) calibration of the six individual ML models using two scenarios (S1 and S2); (4) prediction of SOC content using the various ML models, by additionally accounting for the spatial correlation of the prediction residuals (hybridization) using ordinary kriging; (5) comparison and interpretation of the SOC maps at the four depths, using relative importance and partial dependence analyses.

2.5. Machine learning methods

Six machine learning (ML) algorithms were compared in this study, namely multiple linear regression (MLR), random forest (RF), decision tree regression (DTr), k-Nearest Neighbor (k-NN), support vector regression (SVR), and cubist (CB). All ML methods were applied for predicting SOC from the set of environmental covariates for the two scenarios S1 and S2 at the four standard depths. It is worth noting that MLR is a linear method, and it is used here for comparison purpose with the results for the five other nonlinear ML methods. All methods used in this paper were implemented using various packages of the R statistical software 3.5.1 [41].

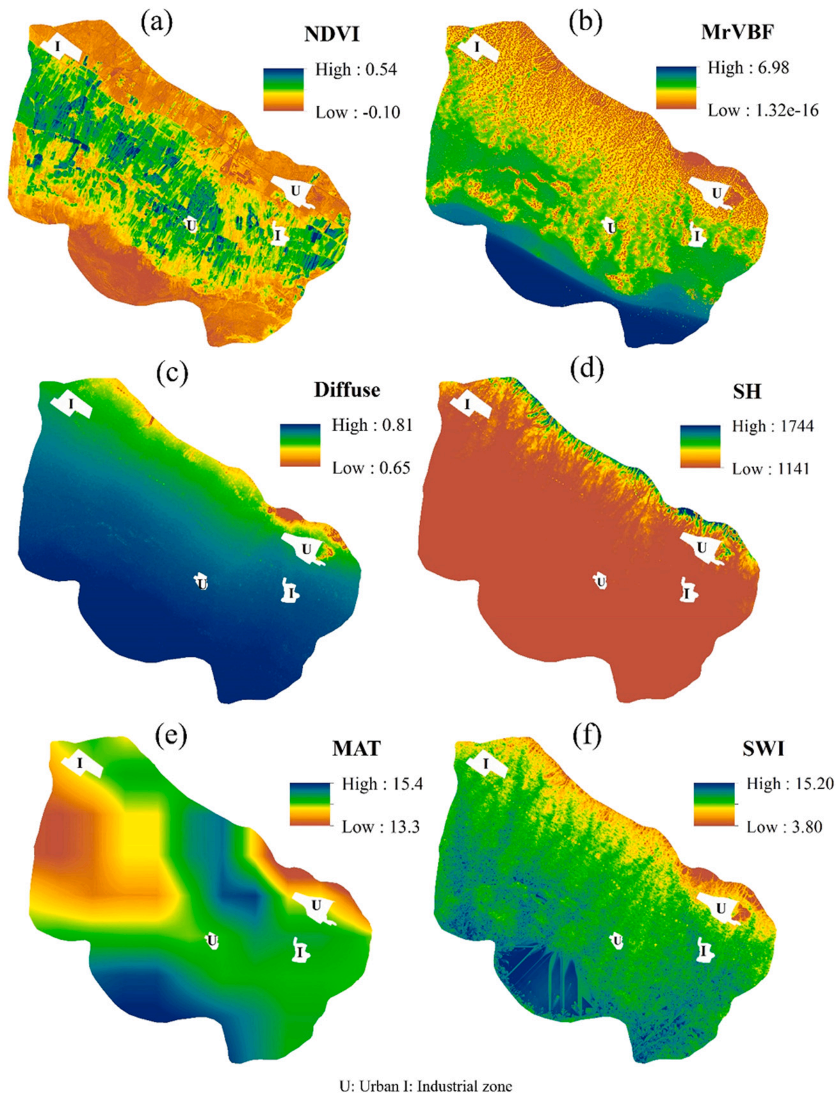

Fig. 4. Most influential environmental covariates derived from DEM, climate data, and RS indices: NDVI **(a)**, Multi-resolution valley bottom flatness index **(b)**, Diffuse Insolation **(c)**, Standardized Height **(d)**, Mean annual temperature **(e)**, and SAGA Wetness Index **(f)**.

2.5.1. Multiple linear regression (MLR)

MLR is a linear statistical model that offers the benefit of a straightforward estimation of its parameters [42]. The MLR regression model for SOC is given by

$$\mathrm{SOC} = b_0 + \sum_{i=1}^{n} b_i x_i + \varepsilon_i \tag{1}$$

where n is the number of predictors, x_i's are the covariates (i.e., environmental covariates and soil properties) and ε_i's are the regression residuals, where the unknown parameters are the intercept b_0 and the slopes b_i's of the hyperplane along the x_i's axes. MLR can be viewed as the simplest ML method among those that are presented here, thanks to the fact that it merely a linear combination of covariates, where the b_i's are easily estimated using a least squares criterion.

2.5.2. Decision tree regression (DTr)

Decision tree builds regression or classification models in the form of a tree structure. A decision tree is incrementally developed by breaking down a dataset into smaller and smaller subsets. The final result is a tree with decision nodes and leaf nodes. Creating a DTr consists of two steps. The first step is the creation and growth of the tree that involves bonding and branching. The second one is the stopping and pruning in order to minimize the prediction error. Partitioning stops when CART (classification and regression tree) detects that no further gain can be made, or when some pre-set stopping rules are met. For a regression tree, the terminal node can be a single value or a regression model. DTr requires tuning two parameters that are the complexity (Cp) and the tree size [43]. The "rpart" package of the R software was used for this goal.

2.5.3. Random forest (RF)

Random forest is an extended version of the regression and classification tree models [44]. It is one of the most used ML models in DSM. The RF output is the mean of all regression trees and the ensemble model is built by averaging multiple models based on different bootstrap sample datasets [45]. RF requires tuning two parameters, that are the number of randomly sampled variables at each split (mtry) and the number of regression trees grown in the forest (ntree). For each depth, the tuning of these parameters was done by iterating mtry values (i.e., the number of covariates) from 1 to 16 and ntree values from 500 to 3,000 with an increment of 100 [46]. The "randomForest" and "caret" packages [47] of the R software were used for this goal.

2.5.4. k-Nearest Neighbor (k-NN)

The k-NN algorithm does not use any predefined mathematical functions to estimate a target soil property. Instead, a reference dataset is searched for covariates that are the most similar to the target property based on the selected input covariates. The similarity between the target soils properties and the known instances is measured by the Euclidean distance, with

$$d_i = \sqrt{\sum_{j=1}^{x} \Delta a_{ij}^2} \tag{2}$$

where d_i is the "distance" of the ith soil from the target soil properties, Δa_{ij} is the difference of the ith soil from the target soil properties in the jth soil properties, and x is the number of soil properties considered for the model. In this study, the k-NN algorithm was adapted from the variant developed by [48] and was implemented using the "caret" package in the R software. The optimal number of k-neighbors was estimated by minimizing the RMSE using cross-validation.

2.5.5. Support vector regression (SVR)

SVR is a kernel-based learning regression method proposed by [49]. It is based on the computation of a linear regression function in a multidimensional feature space [50], where modeling a linear regression hyperplane for nonlinear relationships is possible using this feature space. SVR was implemented using the "caret" package of the R software. A Radial Basis Function (RBF) kernel method was used here.

2.5.6. Cubist (CB)

Cubist is a piecewise linear tree model that uses a recursive partitioning of the predictor variables space [51]. It uses a divide-and-conquer strategy and seeks to minimize the intra subset variation at each node. CB models take the form: if [conditions] then [linear model] approach. If the predictor variables associated with an observation satisfy a set of conditions, the linear model is used to predict the response. The advantage of the condition set in each rule is that they enable interactions to be handled automatically by allowing different linear models to capture the local linearity in various parts of the predictor variables space. This can often lead to smaller trees and better prediction accuracy than regression trees [51]. CB requires tuning two parameters that are the optimal number of committees and neighbors. The "caret" package of the R software was used for this goal.

2.6. Hybridization of SOC predictions

Hybridization is the process of combining at least two methods or algorithms to obtain a new hybrid method or algorithm that improves the prediction accuracy [30]. Ordinary kriging (OK) was used here for the hybridization, as it allows us to account for the spatial correlation of the prediction errors. As OK predicts values at unsampled locations by taking advantage of the spatial dependence, it provides an estimation of the prediction errors at arbitrary locations. See, e.g., [52] for a more detailed presentation of kriging in general. In this study, the empirical semivariogram $\hat{\gamma}_\varepsilon(\|\mathbf{h}\|)$ was estimated first from the prediction residuals $\varepsilon(.)$ at the various profiles locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with

$$\hat{\gamma}_\varepsilon(\|\mathbf{h}\|) = \frac{1}{2N(\|\mathbf{h}\|)} \sum_{N(\|\mathbf{h}\|)} \left[\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_i + \mathbf{h})\right]^2 \tag{3}$$

where $\mathbf{h}$ is the distance vector between two profile locations and $N(\|\mathbf{h}\|)$ is the number of pair of locations separated by the Euclidean distance $\|\mathbf{h}\|$. A combination of a nugget and exponential variogram models was then fitted using a least squares procedure, where the variogram model is

$$\gamma_\varepsilon(\|\mathbf{h}\|; \theta_1, \theta_2, \theta_3) = \theta_1 \delta(\|\mathbf{h}\|) + \theta_2 \left(1 - \exp(-3\|\mathbf{h}\|/\theta_3)\right) \tag{4}$$

where $\delta(\|\mathbf{h}\|)$ is the Kronecker delta, θ_1 is the nugget effect (i.e. the spatially uncorrelated part of the prediction errors), $\theta_1 + \theta_2$ is the total sill (i.e. the variance of the prediction errors), $\theta_1/(\theta_1 + \theta_2)$ is the nugget-to-sill ratio (i.e. a relative measure of the spatial dependence importance), and θ_3 is the range (i.e., the extent of the spatial dependence). See, e.g., [53] for more details about variogram interpretation, estimation and modeling. This fitted variogram model was finally used in OK to predict the errors over the grid used by the various ML algorithms.

2.7. Features selection and partial dependence plots

Due to the large number of potential covariates, it was necessary to reduce their number for limiting collinearity issues [54]. This was done using the variance inflation factors (VIF), with

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} \tag{5}$$

where R_j^2 is the determination coefficient of the multiple linear regression model when predicting the jth covariate using the remaining $(j-1)$ ones. All covariates with VIF values above 10 were considered as highly correlated and were discarded [55]. Collinearity between soil variables and these selected environmental covariates were also assessed using VIF, but all of them where lower than 5, so no serious collinearity issue was identified. After removal of the highly correlated environmental covariates, remaining ones were further subject to Recursive Feature Elimination (RFE) based on a backward selection, thus avoiding the need of fitting again the models at each step of the search [56]. See [35] for more detailed information about the RFE method in general.

When using ML algorithms, due to the high nonlinearity of the methods, the relationship between the input variables and the output prediction is difficult to assess [57]. In order to gain some insight about the respective importance of each input variable in the prediction process, the use of partial dependence plots (PDP) has been proposed [58]. In a SOC mapping context, this approach has been used by [59–61]. The "pdp" package of the R software was used for that goal.

2.8. Model evaluation

For the aim of external validation of the various ML algorithms, the data set was split in two parts, with 80% of them used for calibration and 20% of them used for validation. For tuning the RF, DTr, k-NN, SVR, CB and MLR parameters, a 10-fold cross-validation was repeated 10 times [62] using only the calibration data set. Common performance indices, i.e., mean error (ME), root mean square error (RMSE), coefficient of determination (R^2) and Lin's concordance correlation coefficient (CCC) were used to evaluate each model based on the validation data set, with

$$\mathrm{ME} = \frac{1}{n} \sum_{i=1}^{n} (P_i - O_i) \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} \tag{8}$$

$$CCC = \frac{2r\sigma_{pi}\sigma_{oi}}{\sigma_{pi}^2 + \sigma_{oi}^2 + (\bar{P}_i - \bar{O}_i)^2} \tag{9}$$

with P_i and O_i the predicted and observed values, $\bar{O}$ the average of the observed values over the n measurements, r the correlation coefficient between the predicted and observed values, and σ_P^2, σ_O^2 the variance of the predicted and observed values.

3. Results and discussion

3.1. Descriptive statistics and correlation analysis

Descriptive statistics of the SOC content for the 278 sampled profiles at the four standard depths are presented in Table 1. As seen from these results, SOC decreases with depth, as expected in this study area [63], with coefficients of variation close to 60% at all depths (to be considered as high values based on [64] classification). Correlations between SOC and other soil properties are given in Fig. 3. It can be seen that BD is significantly correlated with SOC at all depths, while CCE is significantly correlated with SOC except at 30–60 cm depth. In addition, Clay is significantly correlated with SOC only for 15–30 cm and 30–60 cm depths. Accordingly, only BD, CCE and Clay will be considered as useful soil properties for predicting SOC at the various depths.

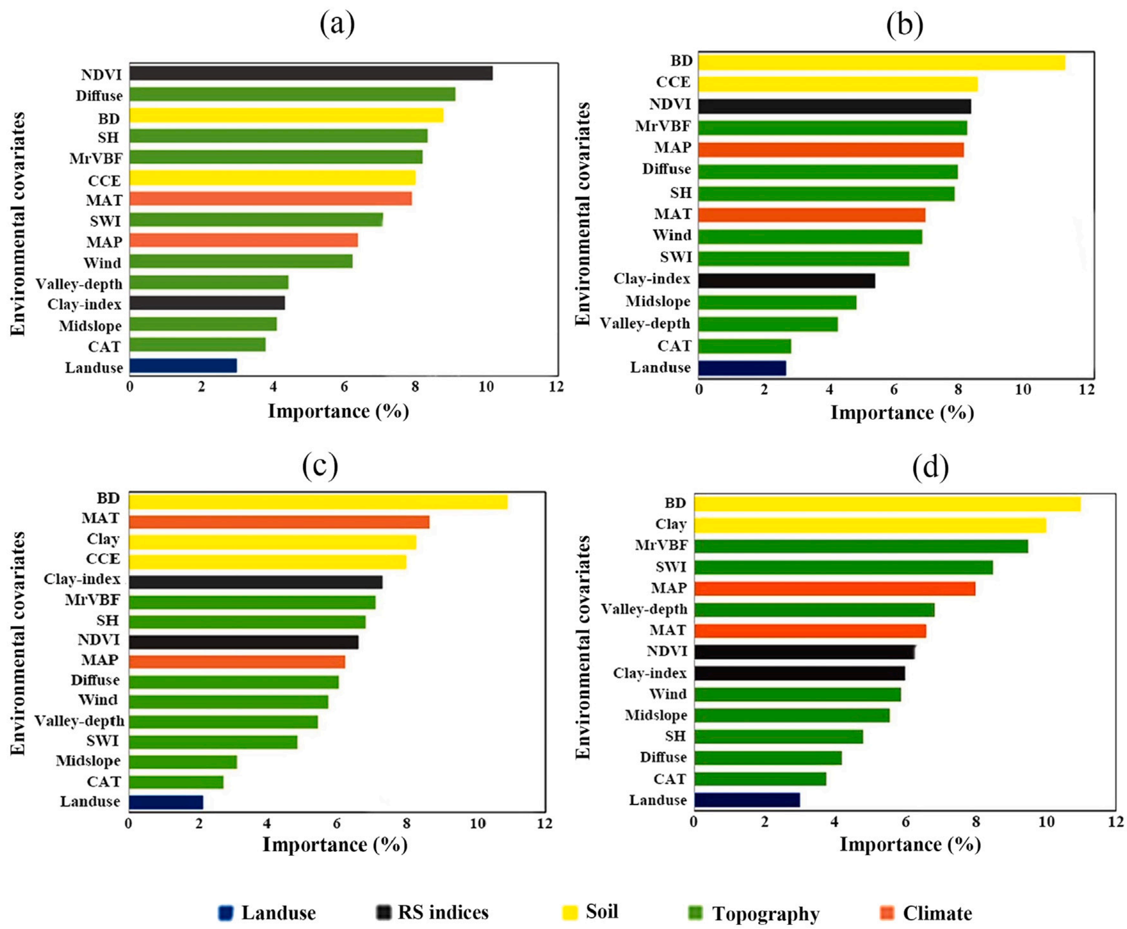

Fig. 5. Relative importance of the covariates and soil variables for SOC prediction using the hybridized RF at the four standard depths, where **(a)**, **(b)**, **(c)** and **(d)** refer to 0–5, 5–15, 15–30, and 30–60 cm, respectively. NDVI: Normalized different vegetation index, Diffuse: Diffuse insolation, BD: Bulk density, SH: Standardized Height, MrVBF: Multiresolution index of valley bottom flatness, Clay: Clay content, CCE: Calcium carbonate equivalent, MAT: Mean annual temperature, SWI: SAGA Wetness Index, MAP: Mean annual precipitation, Wind: Wind effect, Valley depth: Valley depth, Midslope: Midslope position, CAT: Catchment area, Land use.

Table 1
Descriptive statistics of SOC (in %) at the four standard soil depths (in cm) for the $n = 278$ soil profiles.

	Soil depth	Min	Max	Mean	Median	SD	CV (%)
SOC (%)	0–5	0.16	3.83	1.01	0.97	0.61	59.99
	5–15	0.12	3.29	0.96	0.86	0.56	58.59
	15–30	0.08	2.42	0.78	0.46	0.44	56.17
	30–60	0.03	1.80	0.51	0.37	0.32	62.06

SD: standard deviation, CV: coefficient of variation

3.2. Selected features using RFE

Using Recursive feature elimination, a total of 13 environmental covariates were selected among the 55 initial ones. These selected covariates can be grouped in four categories, i.e., topography, climate, RS, and soil properties. As seen from Table 2, the selected covariates include eight topography attributes derived from the DEM, two climate covariates (MAP and MAT), and three RS covariates (NDVI, Land use and Clay index). As expected, numerous DEM-derived covariates were identified as relevant for predicting SOC, as most changes in topography attributes explain much of the variability of SOC at the various depths [65]. Six of the most influential environmental covariates derived from DEM, climate, and RS indices are shown in Fig. 4 for the sake of illustration.

3.3. Performance of individual and hybridized ML models

All ML algorithms were tuned with respect to their set of hyperparameters. These tuned values are given in Table 3. The predictive performances of the various ML algorithms were assessed using ME, RMSE, R^2 and CCC. As seen from Tables 4 and 5, results for the two scenarios S1 and S2 show that S1 (that includes topography, RS, climate covariates and soil variables) is the most efficient one for predicting SOC content. For scenario S2 (without the soil variables), the performance of ML algorithms as measured by R^2 and CCC are divided by about two, for all ML algorithms and at all depths. Based on these results, it can be concluded that accounting for soil properties is clearly improving the prediction performance.

At all depths, RF is ranked as the best performing ML algorithm among those that are compared in this study, while MLR is ranked as the worst one. Globally, sorted in decreasing order of performance, the ML algorithms are RF, SVR, CB, k-NN, DTr and MLR. The best performances for RF are in agreement with the findings by [13,18], who compared various ML approaches for predicting SOC. Similarly, [16,17] concluded that RF yields acceptable prediction accuracy for SOC. As emphasized by [66], higher prediction accuracies are a major advantage of the RF approach for DSM.

For the other ML algorithms, it is worth remarking that, in general (i) SVR yields prediction performances that are close to RF, while DTr prediction performances are close to MLR; (ii) CB and k-NN yield intermediate prediction performances, that are very close to each other whatever the scenario and the soil depth. Overall, this is in agreement

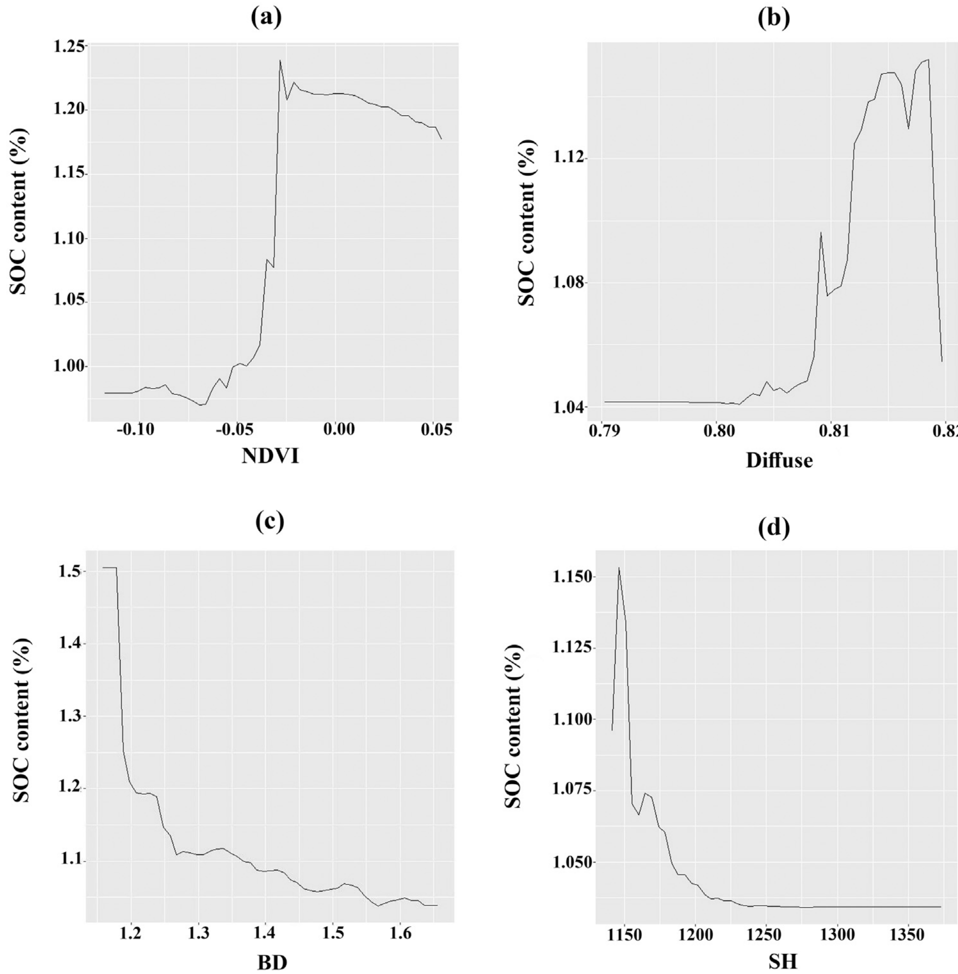

Fig. 6. Partial dependence plots for the four most influential covariates at the 0–5 cm (surface) depth for SOC content with respect to Normalized Difference Vegetation Index **(a)**, Diffuse insolation **(b)**, Bulk Density **(c)** and Standardized Height **(d)**.

Table 2
List of the selected covariates for predicting SOC at the four standard depths.

SCORPAN factors	Parameter	Abbreviation	Source
s	Clay content (%)	Clay	LAB
s	Bulk density (g cm^{-3})	BD	LAB
s	Calcium carbonate equivalent (%)	CCE	LAB
c	Mean annual precipitation (mm)	MAP	CLIM
c	Mean annual temperature (°C)	MAT	CLIM
o	Land use	RS	DEM
o	Normalized diff. vegetation index = $\frac{(NIR-RED)}{(NIR+RED)}$	NDVI	RS
s,p	Clay mineral index = $\frac{(SWIR_a-SWIR_b)}{(SWIR_a+SWIR_b)}$	Clay–index	RS
r	Standard height (–)	SH	DEM
r	Valley depth (m)	Valley–depth	DEM
r	Multiresolution index of valley bottom flatness (–)	MrVBF	DEM
r	SAGA wetness index (–)	SWI	DEM
r	Midslope position (–)	midslope	DEM
r	Catchment area (m^2)	CAT	DEM
r,c	Diffuse Insolation (kWh m^{-2} yr^{-1})	Diffuse	DEM
r,c	Wind Exposition Index (–)	Wind	DEM

s = soil properties; o = organism; r = relief; p=parent material; c = climate, LAB = laboratory analysis; CLIM = climate data; DEM = Digital elevation model; RS = Remote sensing; RED = visible red reflectance; NIR = near infrared reflectance; $SWIR_a$, $SWIR_b$ = short-wave infrared reflectance of bands 6 and 7 of Landsat 8 Operational Land Imager, respectively.

with similar studies about SOC using SVR [67,68] or using CB and k-NN [14,69,70].

Given the fact that SOC prediction using scenario (S1) performs better than scenario (S2) at all depths, the hybridization of the individual ML algorithms was only done for S1. Using the respective prediction residuals of each ML algorithms, variograms were estimated and fitted using a combination of nugget and exponential models, with corresponding parameters given in Table 6. As seen from the high nugget-to still ratio and the limited range in general, it can already be understood that hybridization is not expected to lead to dramatic

Table 3
Tuned hyper parameters for the various ML methods for predicting SOC at the four standard soil depths (in cm) based on scenario S1.

ML model	Soil depth	mtry	ntree	RMSE
RF	0–5	5	500	0.20
	5–15	6	650	0.22
	15–30	7	750	0.24
	30–60	10	800	0.26
	Soil depth	Sigma	C	RMSE
SVR	0–5	0.02	0.25	0.09
	5–15	0.07	0.25	0.12
	15–30	0.08	0.50	0.15
	30–60	0.08	0.50	0.19
	Soil depth	Committees	Neighbors	RMSE
CB	0–5	20	0	0.45
	5–15	20	5	0.48
	15–30	10	9	0.52
	30–60	10	0	0.54
	Soil depth	–	k-Neighbors	RMSE
k-NN	0–5		9	0.28
	5–15		9	0.38
	15–30		9	0.39
	30–60		9	0.59
	Soil depth	CP	tree size	RMSE
DTr	0–5	0.061	2	0.98
	5–15	0.074	1	1
	15–30	0.094	1	1
	30–60	0.051	2	0.96

RF = Random forest; SOC = soil organic carbon; mtry = number of environmental covariates in each random subset; ntree = number of trees in the forest; RMSE = root mean square error; SVR = Support vector regression; Sigma = kernel parameter; C = Penalty parameter; CB = Cubist; k-NN = k-Nearest Neighbor; DTr = Decision tree regression; CP = complexity parameter.

improvements in this study. The nugget-to-sill ratios are also in agreement with the global performance of the various ML algorithms, with values slightly above 0.7 for RF and SVR at all depths, intermediate values around 0.6 for CB and k-NN, while lower values are obtained for DTr and MLR. The corresponding results for the hybridized ML models are given in Tables 4 and 5. They show that hybridization indeed increases the accuracy of SOC prediction maps at all depths. However, the benefit is relatively moderate here, with a R^2 increase that ranges from 2% to 11% depending on the ML algorithm and the soil depth. These results agree with those of [71] who conclude using RF and gradient boosting (GB) algorithms that combining ML algorithms with OK increased the prediction performance from 1% to 34%. Similar findings were documented by [29,30].

3.4. Relative importance and partial dependence plot (PDP) analysis

Based on the hybridized RF model, the respective importance of the various covariates and soil variables was assessed. We propose to use here the relative importance of each covariate and its partial dependence plot (PDP) as main tools for this goal. While the relative importance of a covariate quantifies the part of the SOC variability that one can attribute to this covariate, a PDP shows how SOC content tends to change as a function of this covariate.

3.4.1. Topsoil layer (0–5 cm)

For topsoil SOC prediction (0–5 cm depth), the four most important covariates are NDVI, Diffuse, BD and SH, explaining together 36% of SOC variability (Fig. 5a). The results of the PDP for these four covariates are shown in Fig. 6, where it can be seen that SOC content increases with NDVI and Diffuse, while it decreases with BD and SH.

NDVI is ranked as the most important covariate. This agrees with the results by [72], who reported that SOC variations were mostly affected by vegetation indices and soil covariates under the same climate conditions. Similarly, [70] identified NDVI as the most important predictor of SOC variability, while [73] reported NDVI as one of the main factors controlling the topsoil spatial pattern of SOC. From Fig. 6**a**, increasing value of SOC with NDVI is an agreement with the fact that NDVI is a proxy for vegetation cover that reflects the ecosystem productivity and that impacts the input rate of organic carbon in soils [29,74].

From Fig. 6**b**, SOC content is increasing with Diffuse. In agreement with [75], solar radiation and NDVI are key environmental variables for SOC distribution, where solar radiation can indirectly control the SOC content by affecting soil moisture dynamics and plant growth.

From Fig. 6**c**, the negative impact of BD on SOC is confirmed (see Section 3.4.3 for the corresponding interpretation). It was already mentioned that BD it is negatively correlated with SOC at all depths (see Fig. 3). Similar results were found by [76] between BD and SOC both at the surface (0–25 cm) and subsurface (25–100 cm) soil depths in Canada.

Finally, from Fig. 6**d**, it is seen that SH has a negative impact on SOC. Similarly, [77] identified SH as an effective topographic proxy for SOC because of its role in the stability and storage of SOC through soil erosion and sedimentation processes [29,78].

3.4.2. Intermediate soil layers (5–15 and 15–30 cm)

When considering the four most important covariates for intermediate soil layers (Fig. 5**b** and 5**c**), two and three of them are related to soil properties, respectively. BD is ranked by far as most important one. According to Fig. 5**b**, BD, CCE, NDVI and MrVBF are the four most important covariates that explain 36% of SOC content in the 5–15 cm layer. For the 15–30 cm layer, BD, MAT, Clay, and CCE are the four most important covariates that explain 34% of SOC content (Fig. 5**c**). From this, it can be concluded that soil properties (BD, CCE, Clay) play an important role for predicting SOC at intermediate depths. Similarly, [27,37] reported that soil texture, CCE, and clay mineralogy that are linked to the parent material tend to increase the potential for carbon conservation in soil. In parallel, MAT has a profound effect on the accumulation and decomposition of organic matter in the soil by affecting plant growth and net primary production [9,72]. Finally, CCE can be related to carbonate-rich parent material in the study area.

3.4.3. Lower soil layer (30–60 cm)

For the subsoil layer (30–60 cm), the four most important covariates are BD, Clay, MrVBF and SWI, that explain together 40% of SOC variability (Fig. 5**d**). The two first covariates are related to soil properties while the two last ones are related to topography. As for the intermediate layers, BD is reported as the most important covariate. On the opposite, covariates related to RS play here a much minor role by comparison with the other soil layers. Similarly, [79] mentioned that BD has frequently been linked to SOC in soils that are storing large quantities of organic matter, while [80] identified BD, salinity and sand content as the most important ones to explain carbon and nitrogen pools in an area of the north Nile Delta, Egypt.

From the PDP presented in Figs. 6**c** and 7**a**, one can see that the results for BD are similar to those for the top layer, with a decrease of SOC with respect to BD. Clearly, BD is a soil physical indicator which is strongly influenced by land management. Kizilkaya and Dengiz [81] found that successive tillage operations and land use changes led to joint significant BD increase and SOC decrease. Celik [82] believed that the reduction of SOC induces a breakdown of soil aggregates, where fine particles are transported by water erosion and fill the soil pores. Therefore, after the destruction of the soil structure, soil porosity decreases and consequently BD increases.

For Clay, it was already mentioned that it is positively correlated with SOC at all depths (see Fig. 3). This behavior is confirmed from Fig. 7**b**, where highest SOC values are observed for clay content above 20–30%. Soils with high clay content have typically high

Table 4
Validation results for the individual SOC prediction at the four standard depths (in cm) using scenarios S1, S2 and for the hybridized modeling.

Scenarios	Soil depth	0–5		5–15		15–30		30–60	
		R^2	CCC	R^2	CCC	R^2	CCC	R^2	CCC
S1	RF	0.79	0.85	0.76	0.80	0.72	0.76	0.68	0.76
	SVR	0.75	0.85	0.72	0.80	0.68	0.76	0.62	0.72
	CB	0.55	0.66	0.50	0.61	0.47	0.50	0.45	0.50
	k-NN	0.54	0.61	0.48	0.58	0.45	0.47	0.40	0.47
	DTr	0.38	0.55	0.36	0.52	0.34	0.35	0.30	0.35
	MLR	0.37	0.40	0.33	0.45	0.29	0.28	0.27	0.28
S2	RF	0.40	0.45	0.38	0.50	0.34	0.37	0.30	0.34
	SVR	0.37	0.42	0.32	0.42	0.22	0.18	0.15	0.21
	CB	0.30	0.35	0.27	0.41	0.24	0.18	0.17	0.26
	k-NN	0.30	0.31	0.28	0.40	0.23	0.32	0.19	0.27
	DTr	0.32	0.30	0.26	0.35	0.23	0.26	0.20	0.23
	MLR	0.19	0.26	0.18	0.25	0.14	0.19	0.15	0.22
Hybrid	RF	0.81	0.87	0.80	0.82	0.76	0.81	0.74	0.78
	SVR	0.77	0.85	0.75	0.80	0.73	0.77	0.69	0.71
	CB	0.60	0.71	0.56	0.65	0.54	0.63	0.53	0.56
	k-NN	0.57	0.68	0.53	0.65	0.51	0.58	0.48	0.50
	DTr	0.46	0.63	0.44	0.60	0.41	0.45	0.39	0.53
	MLR	0.45	0.50	0.43	0.53	0.40	0.43	0.38	0.39

Table 5
Validation results for the individual SOC prediction at the four standard depths (in cm) using scenarios S1, S2 and for the hybridized SOC (%) modeling.

Scenarios	Soil depth	0–5		5–15		15–30		30–60	
		RMSE	ME	RMSE	ME	RMSE	ME	RMSE	ME
S1	RF	0.25	−0.02	0.27	0.02	0.30	0.18	0.32	0.01
	SVR	0.28	−0.02	0.30	−0.09	0.32	−0.12	0.34	−0.1
	CB	0.31	0.17	0.35	−0.18	0.37	−0.12	0.39	−0.17
	k-NN	0.32	−0.00	0.34	−0.23	0.42	−0.28	0.45	−0.22
	DTr	0.35	−0.20	0.38	−0.30	0.40	−0.29	0.46	−0.27
	MLR	0.47	−0.00	0.49	−0.32	0.51	−0.30	0.52	−0.31
S2	RF	0.32	0.06	0.34	0.07	0.37	0.22	0.41	0.06
	SVR	0.34	−0.12	0.37	−0.15	0.42	−0.28	0.65	−0.1
	CB	0.39	−0.13	0.42	−0.11	0.48	−0.26	0.59	−0.03
	k-NN	0.42	−0.02	0.63	0.10	0.55	−0.18	0.63	−0.05
	DTr	0.38	−0.10	0.42	0.04	0.54	−0.21	0.57	−0.02
	MLR	0.50	−0.02	0.52	−0.01	0.62	−0.24	0.67	−0.06
Hybrid	RF	0.21	−0.01	0.23	0.01	0.26	0.16	0.28	0.20
	SVR	0.24	0.01	0.26	−0.09	0.27	−0.11	0.29	−0.10
	CB	0.31	−0.11	0.33	−0.13	0.35	−0.10	0.37	−0.16
	k-NN	0.30	0.07	0.31	−0.12	0.33	−0.24	0.35	−0.20
	DTr	0.31	−0.16	0.36	−0.18	0.38	−0.26	0.41	−0.25
	MLR	0.43	0.09	0.45	0.22	0.47	0.34	0.49	0.25

SOC stock [83,84]. In fine-textured soils, the soil organic matter is surrounded by fine mineral particles that reduce its decomposition rate from microbial activity, thus leading to SOC increase in the long run [85].

From Fig. 7**c**, one can see that SOC content increase for highest MrVBF values. As reported by [86], MrVBF is a topographic attribute that reflects the depositional and erosional processes impacting the SOC spatial pattern. Similarly, [87] concluded that higher MrVBF values are associated with higher SOC storage. More generally, topographic attributes such as MrVBF, SWI and Valley depth indirectly influence SOC content by affecting erosion, runoff and drainage [88].

Finally, a PDP of SOC with respect to SWI (Fig. 7**d**) shows that, overall, SWI has a positive impact on SOC content. According to the SWI map (Fig. 4**f**), areas with SWI values higher than 12.5 correspond to saline pastures. As SWI indicates the spatial distribution of moisture and sediment on the soil surface, high soil moisture promotes plant growth and so tends to positively impact SOC concentration [45]. Adhikari et al. [89] evidenced that SWI has a high influence in the prediction of subsoil SOC at lower soil depths. Hamzehpour et al. [61] showed too that high soil moisture with limited soil aeration leads to decreased microbial activity and delayed soil organic matter decomposition.

3.5. Spatial maps of SOC at the four standard depths

As RF with hybridization was the best performing ML algorithm, the SOC maps at the four standard depths were drawn using this method (Fig. 8). It is seen that predicted SOC content is the highest in the central, northern, and western parts of the study area, with maximum SOC values decreasing from 3.05% at the surface (0–15 cm) down to 1.16% at the lowest depth (30–60 cm), as SOC is mainly stored in the topsoil and decreases with depth. Overall, the spatial pattern of SOC is very similar for the 0–5 cm and 5–15 cm depths. Similarities at lower depths are mainly observed for the central part of the study area. Clearly, high SOC content areas are mainly occupied by irrigated crops, dry farming, and non-saline pastures. Lacoste et al. [90] and Taghizadeh-Mehrjardi et al. [17] reported typically higher SOC content in such grassland and cultivated areas. Obviously, application of organic manure could explain the high SOC contents in areas that have been cultivated for a long time. On the opposite and as expected, minimum values of SOC are predicted in the south, southwest and southeast areas, where the southern part is a Plain landscape mainly covered with saline pastures. The main cause of low SOC content in the southern part of the area is the corresponding low vegetation covers, that leads to low input nutrients to the soil without proper management.

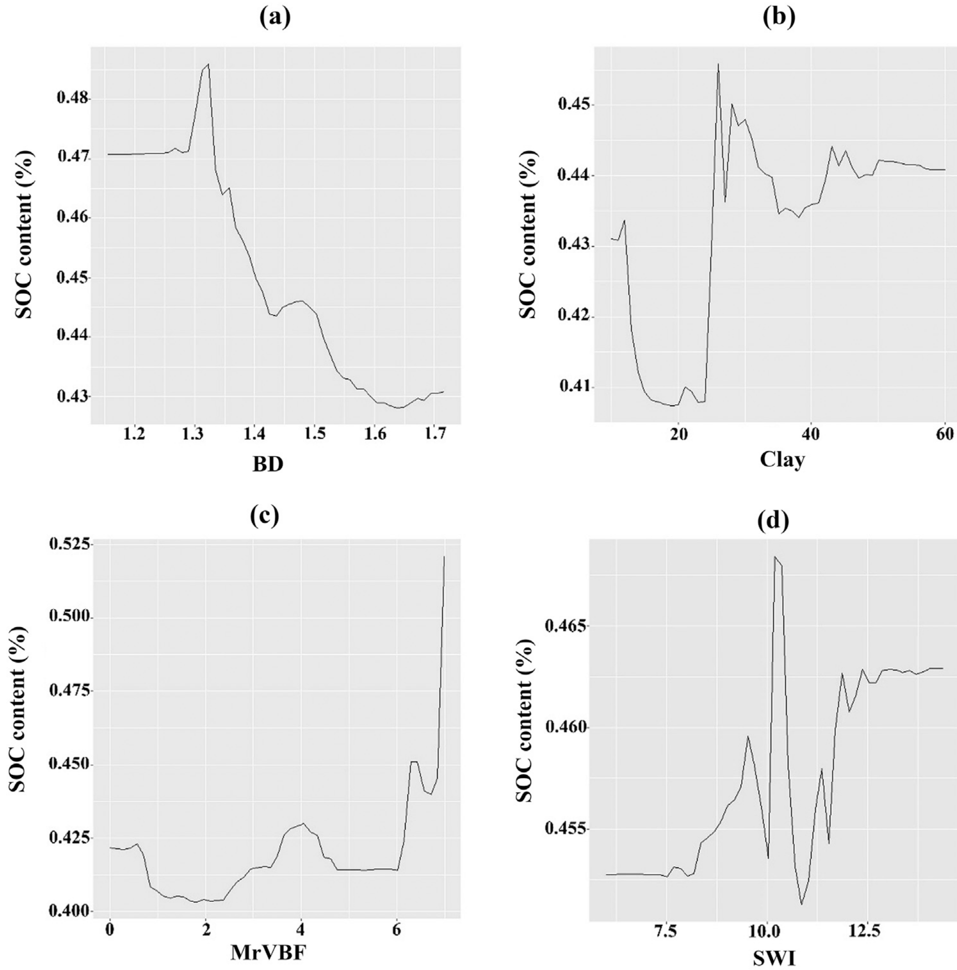

Fig. 7. Partial dependence plots for the four most influential covariates at the 30–60 cm (subsurface) depth for SOC content with respect to Bulk Density **(a)**, Clay content **(b)**, Multiresolution index of valley bottom flatness **(c)** and SAGA Wetness Index (SWI) **(d)**.

Comparing now SOC maps in Fig. 8 to environmental covariates in Fig. 4, it is also clear that none of these covariates are exhibiting the same spatial patterns as for SOC, even at the 0–5 cm depth for which NDVI is the most important covariate. This agrees with the fact that the maps resulting from the ML algorithms are nonlinear combinations of these covariates, so that no obvious similarities can be evidenced from simple visual inspection in general. Moreover, soil variables (not pictured here) also play an important role, especially below the soil surface layer.

4. Conclusions

In this study, SOC predicted maps at four standard soil depths in an arid and semi-arid region of Iran were prepared using a set of environmental covariates (from DEM, RS images and climate data) and soil variables measured at 287 soil profiles. Two scenarios S1 and S2 were considered. While S1 is accounting together for environmental covariates and soil variables that were also measured for the profiles at the various depths (in parallel to SOC measurements for these profiles), S2 discards these extra soil variables in the modeling and prediction processes. Modeling and prediction were conducted using six ML algorithms, thus allowing us to assess their respective performances. Hybridization using kriging was also used to account for the spatial information still contained in the prediction residuals.

For all ML algorithms, prediction performance (as measured by R^2) decreases with soil depth. For all depths and for both scenarios, RF was identified as the best performing ML algorithm. A twofold increase of performance is obtained when considering S1 instead of S2, thus emphasizing the importance of soil variables for properly mapping SOC content. By contrast, hybridization led to a marginal performance increase, which is consistent with the limited spatial information content that remains in the prediction residuals, as measured by the high nugget-to-sill ratio and the limited range of the corresponding variograms.

From a comparison of the relative importance of the environmental covariates and soil variables at the various soil depths, it is seen that soil variables (BD, CCE, Clay) play a major role below the soil surface, with a lower importance for the environmental covariates. This contrasts with the soil surface (0–5 cm), where NDVI and Diffuse were identified as the most important covariates. In parallel, PDP allowed us to investigate the effect of each covariate on the SOC content, with findings that are in good agreement with those from the literature.

In the context of SOC mapping, it can be concluded that using jointly environmental covariates and soil variables can lead to a clear-cut benefit, which is by far higher than the benefit of using refinements like hybridization. This has also important consequences for the proper planning of sampling campaigns: as SOC measurements need to be collected for ML algorithms training and validation, the additional efforts and money needed for measuring in parallel other important soil properties like BD, CCE and Clay are worthy. Even if one cannot offer guarantees that similar findings hold entirely true for other areas, they however deserve a close attention for subsequent studies and should also be beneficial for the digital soil mapping community in general.

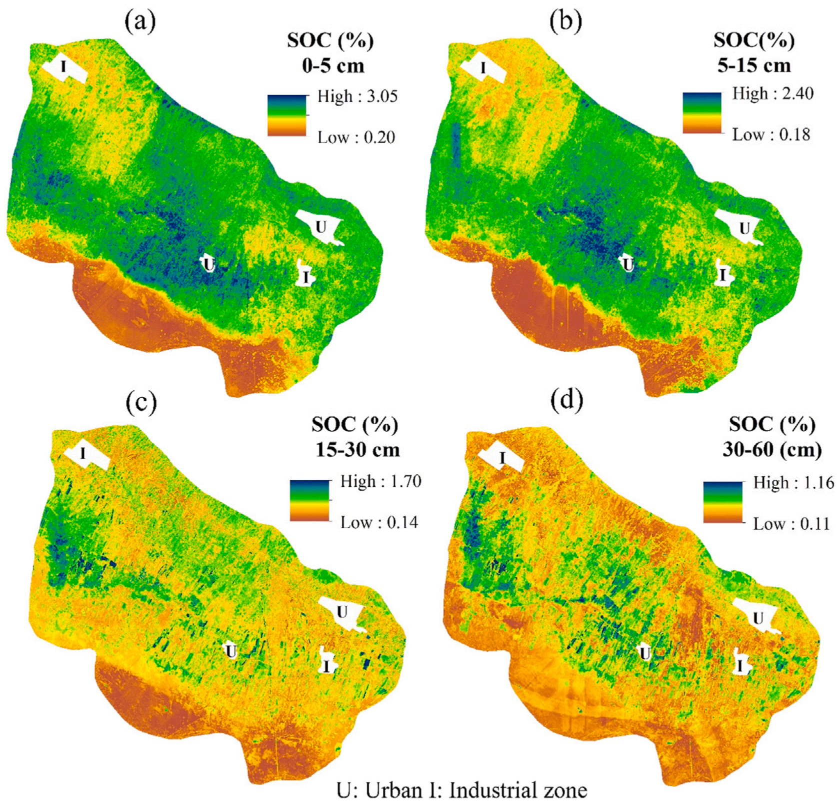

Fig. 8. Predicted spatial distribution of SOC (in %) at the four soil standard depths using the hybridized RF model, where (a), (b), (c) and (d) refer to the SOC at 0–5, 5–15, 15–30 and 30–60 cm, respectively.

Table 6
Fitted parameters for the combined nugget and exponential variogram models, as estimated from the prediction errors associated with the various ML models at the four standard depths (in cm).

ML model	Soil depth	Fitted parameters		
		Total sill	Nugget-to-sill	Range (m)
RF	0–5	0.24	0.75	710
	5–15	0.25	0.73	730
	15–30	0.29	0.72	1000
	30–60	0.42	0.71	2003
SVR	0–5	0.26	0.74	804
	5–15	0.29	0.73	911
	15–30	0.30	0.73	950
	30–60	0.39	0.71	1662
CB	0–5	0.33	0.68	773
	5–15	0.37	0.64	798
	15–30	0.35	0.60	826
	30–60	0.35	0.57	1047
k-NN	0–5	0.33	0.67	790
	5–15	0.35	0.63	915
	15–30	0.39	0.61	945
	30–60	0.41	0.57	1351
DTr	0–5	0.32	0.66	790
	5–15	0.35	0.65	810
	15–30	0.38	0.41	3200
	30–60	0.44	0.57	5238
MLR	0–5	0.44	0.58	817
	5–15	0.48	0.54	878
	15–30	0.49	0.24	6251
	30–60	0.45	0.46	2846

RF = Random forest; SVR = Support vector regression; CB = Cubist; k-NN = k-nearest neighbor; DTr = Decision tree regression; MLR = Multiple linear regression.

CRediT authorship contribution statement

Seyed Roohollah Mousavi: Conceptualization, Investigation, Methodology, Modeling, Writing – original draft, Writing – review & editing. **Fereydoon Sarmadian:** Investigation, Methodology, Editing – original draft, Writing – review & editing, Methodology, Funding acquisition, Resources. **Mahmoud Omid:** Conceptualization, Investigation, Methodology. **Patrick Bogaert:** Methodology, Rewriting – original draft, Writing – review & editing, The revised original draft of the manuscript.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

The three first authors thank the University of Tehran, Iran for its financial and moral support that allowed them to carry out this research project.

References

[1] Z.-P. Wang, X.-G. Han, S.X. Chang, B. Wang, Q. Yu, L.-Y. Hou, L.-H. Li, Soil organic and inorganic carbon contents under various land uses across a transect of continental steppes in Inner Mongolia, Catena 109 (2013) 110–117.

[2] L. Guo, P. Fu, T. Shi, Y. Chen, H. Zhang, R. Meng, S. Wang, Mapping field-scale soil organic carbon with unmanned aircraft system-acquired time series multispectral images, Soil Tillage Res. 196 (2020) 104477.

[3] C. Schillaci, M. Acutis, F. Vesely, S. Saia, A simple pipeline for the assessment of legacy soil datasets: An example and test with soil organic carbon from a highly variable area, Catena 175 (2019) 110–122.
[4] L. Qi, S. Wang, Q. Zhuang, Z. Yang, S. Bai, X. Jin, G. Lei, Spatial-temporal changes in soil organic carbon and pH in the Liaoning Province of China: A modeling analysis based on observational data, Sustainability 11 (13) (2019) 3569.
[5] J. Padarian, B. Minasny, A.B. McBratney, Using deep learning for digital soil mapping, Soil 5 (1) (2019) 79–89.
[6] H. Mahmoudzadeh, H.R. Matinfar, R. Taghizadeh-Mehrjardi, R. Kerry, Spatial prediction of soil organic carbon using machine learning techniques in western Iran, Geoderma Regional 21 (2020) e00260.
[7] A.B. McBratney, U. Stockmann, D.A. Angers, B. Minasny, D.J. Field, Challenges for soil organic carbon research, in: Soil Carbon, Springer, 2014, pp. 3–16.
[8] S. Lamichhane, L. Kumar, B. Wilson, Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review, Geoderma 352 (2019) 395–413.
[9] B. Wang, C. Waters, S. Orgill, A. Cowie, A. Clark, D. Li Liu, M. Simpson, I. McGowen, T. Sides, Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia, Ecol. Indic. 88 (2018) 425–438.
[10] J. Xiao, F. Chevallier, C. Gomez, L. Guanter, J.A. Hicke, A.R. Huete, K. Ichii, W. Ni, Y. Pang, A.F. Rahman, et al., Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years, Remote Sens. Environ. 233 (2019) 111383.
[11] A. Keshavarzi, F. Sarmadian, J. Shiri, M. Iqbal, R. Tirado-Corbalá, E.-S.E. Omran, Application of ANFIS-based subtractive clustering algorithm in soil cation exchange capacity estimation using soil and remotely sensed data, Measurement 95 (2017) 173–180.
[12] F. Veronesi, C. Schillaci, Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation, Ecol. Indic. 101 (2019) 1032–1044.
[13] Y. Wang, L. Deng, G. Wu, K. Wang, Z. Shangguan, Large-scale soil organic carbon mapping based on multivariate modelling: The case of grasslands on the loess plateau, Land Degrad. Dev. 29 (1) (2018) 26–37.
[14] M. Zeraatpisheh, S. Ayoubi, A. Jafari, S. Tajik, P. Finke, Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran, Geoderma 338 (2019) 445–452.
[15] F. Castaldi, S. Chabrillat, C. Chartin, V. Genot, A. Jones, B. van Wesemael, Estimation of soil organic carbon in arable soil in Belgium and Luxembourg with the LUCAS topsoil database, Eur. J. Soil Sci. 69 (4) (2018) 592–603.
[16] K. Were, D.T. Bui, Ø.B. Dick, B.R. Singh, 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 (2015) 394–403.
[17] R. Taghizadeh-Mehrjardi, K. Nabiollahi, R. Kerry, Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran, Geoderma 266 (2016) 98–110.
[18] K. John, I. Abraham Isong, N. Michael Kebonye, E. Okon Ayito, P. Chapman Agyeman, S. Marcus Afu, Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil, Land 9 (12) (2020) 487.
[19] R.V. Rossel, T. Behrens, Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma 158 (1–2) (2010) 46–54.
[20] M. Emadi, R. Taghizadeh-Mehrjardi, A. Cherati, M. Danesh, A. Mosavi, T. Scholten, Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran, Remote Sens. 12 (14) (2020) 2234.
[21] A.M.-C. Wadoux, B. Minasny, A.B. McBratney, Machine learning for digital soil mapping: Applications, challenges and suggested solutions, Earth-Sci. Rev. 210 (2020) 103359.
[22] R. Ponce Hernandez, The use of soil information systems in land planning, (Ph.D. thesis), University of Oxford, 1987.
[23] T. Bishop, A. McBratney, G. Laslett, Modelling soil attribute depth functions with equal-area quadratic smoothing splines, Geoderma 91 (1–2) (1999) 27–45.
[24] B. Minasny, B.I. Setiawan, C. Arif, S.K. Saptomo, Y. Chadirin, et al., Digital mapping for cost-effective and accurate prediction of the depth and carbon stocks in Indonesian peatlands, Geoderma 272 (2016) 20–31.
[25] K. Benke, S. Norng, N. Robinson, K. Chia, D. Rees, J. Hopley, Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content, Geoderma 366 (2020) 114210.
[26] L.C. Gomes, R.M. Faria, E. de Souza, G.V. Veloso, C.E.G. Schaefer, E.I. Fernandes Filho, Modelling and mapping soil organic carbon stocks in Brazil, Geoderma 340 (2019) 337–350.
[27] M. Zeraatpisheh, Y. Garosi, H.R. Owliaie, S. Ayoubi, R. Taghizadeh-Mehrjardi, T. Scholten, M. Xu, Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates, Catena 208 (2022) 105723.
[28] M. Rostaminia, A. Rahmani, S.R. Mousavi, R. Taghizadeh-Mehrjardi, Z. Maghsodi, Spatial prediction of soil organic carbon stocks in an arid rangeland using machine learning algorithms, Environ. Monit. Assess. 193 (12) (2021) 1–17.
[29] H.R. Matinfar, Z. Maghsodi, S.R. Mousavi, A. Rahmani, Evaluation and prediction of topsoil organic carbon using machine learning and hybrid models at a field-scale, Catena 202 (2021) 105258.
[30] F.B. Silatsa, M. Yemefack, F.O. Tabi, G.B. Heuvelink, J.G. Leenaars, Assessing countrywide soil organic carbon stock using hybrid machine learning modelling and legacy soil data in Cameroon, Geoderma 367 (2020) 114260.
[31] A.R. Van Wambeke, The Newhall Simulation Model for Estimating Soil Moisture and Temperature Regimes, Department of Crop and Soil Sciences. Cornell University, Ithaca, NY, 2000.
[32] J. Zinck, The geomorphic landscape: The attributes of geoforms, in: Geopedology, Springer, 2016, pp. 127–163.
[33] L.M. de Sousa, L. Poggio, N.H. Batjes, G.B. Heuvelink, B. Kempen, E. Riberio, D. Rossiter, Soilgrids 2.0: producing quality-assessed soil information for the globe, Soil Discuss. 2020 (2020) 1–37.
[34] A. Walkley, I.A. Black, An examination of the degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method, Soil Sci. 37 (1) (1934) 29–38.
[35] M. Kuhn, K. Johnson, et al., Applied Predictive Modeling, Vol. 26, Springer, 2013.
[36] B.P. Malone, A. McBratney, B. Minasny, G. Laslett, Mapping continuous depth functions of soil carbon storage and available water capacity, Geoderma 154 (1–2) (2009) 138–152.
[37] A.B. McBratney, M.M. Santos, B. Minasny, On digital soil mapping, Geoderma 117 (1–2) (2003) 3–52.
[38] O. Conrad, B. Bechtel, M. Bock, H. Dietrich, E. Fischer, L. Gerlitz, J. Wehberg, V. Wichmann, J. Böhner, System for automated geoscientific analyses (SAGA) v. 2.1. 4, Geosci. Model Dev. 8 (7) (2015) 1991–2007.
[39] V. Olaya, O. Conrad, Geomorphometry in SAGA, Dev. Soil Sci. 33 (2009) 293–308.
[40] K. Krivoruchko, A. Gribov, Evaluation of empirical Bayesian kriging, Spatial Stat. 32 (2019) 100368.
[41] R.C. Team, et al., R: A language and environment for statistical computing. Vienna, Austria; 2014, 2015.
[42] S.H.G. Silva, A.F.d.S. Teixeira, M.D.d. Menezes, L.R.G. Guilherme, F.M.d.S. Moreira, N. Curi, Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF), Ciência E Agrotecnol. 41 (2017) 648–664.
[43] B.P. Malone, B. Minasny, A.B. McBratney, Continuous soil attribute modeling and mapping, in: Using R for Digital Soil Mapping, Springer, 2017, pp. 117–149.
[44] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[45] T. Behrens, K. Schmidt, A.-X. Zhu, T. Scholten, The ConMap approach for terrain-based digital soil mapping, Eur. J. Soil Sci. 61 (1) (2010) 133–143.
[46] T. Hengl, G.B. Heuvelink, B. Kempen, J.G. Leenaars, M.G. Walsh, K.D. Shepherd, A. Sila, R.A. MacMillan, J. Mendes de Jesus, L. Tamene, et al., Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions, PLoS One 10 (6) (2015) e0125814.
[47] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R News 2 (3) (2002) 18–22.
[48] A. Nemes, W.J. Rawls, Y.A. Pachepsky, Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties, Soil Sci. Am. J. 70 (2) (2006) 327–336.
[49] V. Cherkassky, F.M. Mulier, Learning from Data: Concepts, Theory, and Methods, John Wiley & Sons, 2007.
[50] L. Xuemei, L. Jianshe, Measurement of soil properties using visible and short wave-near infrared spectroscopy and multivariate calibration, Measurement 46 (10) (2013) 3808–3814.
[51] Z. Ma, Z. Shi, Y. Zhou, J. Xu, W. Yu, Y. Yang, A spatial data mining algorithm for downscaling TMPA 3B43 V7 data over the Qinghai–Tibet Plateau with the effects of systematic anomalies removed, Remote Sens. Environ. 200 (2017) 378–395.
[52] R. Webster, M.A. Oliver, Geostatistics for Environmental Scientists, John Wiley & Sons, 2007.
[53] N. Cressie, Statistics for Spatial Data, in: Wiley Series in Probability and Statistics, Wiley, 1993, URL https://books.google.be/books?id=4SdRAAAAMAAJ.
[54] M. Meier, E.d. Souza, M.R. Francelino, E.I. Fernandes Filho, C.E.G.R. Schaefer, Digital soil mapping using machine learning algorithms in a tropical mountainous area, Revista Brasileira de Ciência Do Solo 42 (2018).
[55] J.D. Curto, J.C. Pinto, The corrected vif (cvif), J. Appl. Stat. 38 (7) (2011) 1499–1507.
[56] R. Taghizadeh-Mehrjardi, K. Schmidt, A. Amirian-Chakan, T. Rentschler, M. Zeraatpisheh, F. Sarmadian, R. Valavi, N. Davatgar, T. Behrens, T. Scholten, Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space, Remote Sens. 12 (7) (2020) 1095.
[57] S. Araya, A. Fryjoff-Hung, A. Anderson, J. Viers, T. Ghezzehei, Advances in soil moisture retrieval from multispectral remote sensing using unmanned aircraft systems and machine learning techniques, Hydrol. Earth Syst. Sci. Discuss. (2020) 1–33.
[58] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Statist. (2001) 1189–1232.

[59] C. Zeng, L. Yang, A.-X. Zhu, Construction of membership functions for soil mapping using the partial dependence of soil on environmental covariates calculated by random forest, Soil Sci. Am. J. 81 (2) (2017) 341–353.

[60] S. Ottoy, B. De Vos, A. Sindayihebura, M. Hermy, J. Van Orshoven, Assessing soil organic carbon stocks under current and potential forest cover using digital soil mapping and spatial generalisation, Ecol. Indic. 77 (2017) 139–150.

[61] N. Hamzehpour, H. Shafizadeh-Moghadam, R. Valavi, Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture, Catena 182 (2019) 104141.

[62] B. Heung, H.C. Ho, J. Zhang, A. Knudby, C.E. Bulmer, M.G. Schmidt, An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping, Geoderma 265 (2016) 62–77.

[63] S. Wang, Q. Zhuang, Q. Wang, X. Jin, C. Han, Mapping stocks of soil organic carbon and soil total nitrogen in Liaoning Province of China, Geoderma 305 (2017) 250–263.

[64] L. Wilding, Spatial variability: its documentation, accomodation and implication to soil surveys, in: Soil Spatial Variability, Las Vegas NV, 30 November-1 December 1984, 1985, pp. 166–194.

[65] C.-C. Tsui, C.-C. Tsai, Z.-S. Chen, Soil organic carbon stocks in relation to elevation gradients in volcanic ash soils of Taiwan, Geoderma 209 (2013) 119–127.

[66] A. Arabameri, B. Pradhan, H.R. Pourghasemi, K. Rezaei, N. Kerle, Spatial modelling of gully erosion using GIS and r programing: A comparison among three data mining algorithms, Appl. Sci. 8 (8) (2018) 1369.

[67] R. Rossel, A. McBratney, Diffuse reflectance spectroscopy as a tool for digital soil mapping, in: Digital Soil Mapping with Limited Data, Springer, 2008, pp. 165–172.

[68] R. Taghizadeh-Mehrjardi, R. Neupane, K. Sood, S. Kumar, Artificial bee colony feature selection algorithm combined with machine learning algorithms to predict vertical and lateral distribution of soil organic matter in South Dakota, USA, Carbon Management 8 (3) (2017) 277–291.

[69] P. Somarathna, B. Malone, B. Minasny, Mapping soil organic carbon content over New South Wales, Australia using local regression kriging, Geoderma Regional 7 (1) (2016) 38–48.

[70] S. Tajik, S. Ayoubi, M. Zeraatpisheh, Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran, Geoderma Regional 20 (2020) e00256.

[71] P. Tziachris, V. Aschonitis, T. Chatzistathis, M. Papadopoulou, Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters, Catena 174 (2019) 206–216.

[72] T. Zhou, Y. Geng, J. Chen, M. Liu, D. Haase, A. Lausch, Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China, Ecol. Indic. 114 (2020) 106288.

[73] R.-M. Yang, G.-L. Zhang, F. Yang, J.-J. Zhi, F. Yang, F. Liu, Y.-G. Zhao, D.-C. Li, Precise estimation of soil organic carbon stocks in the northeast Tibetan Plateau, Sci. Rep. 6 (1) (2016) 1–10.

[74] Y. Peng, X. Xiong, K. Adhikari, M. Knadel, S. Grunwald, M.H. Greve, Modeling soil organic carbon at regional scale by combining multi-spectral images with laboratory spectra, PLoS One 10 (11) (2015) e0142295.

[75] A. Román-Sánchez, T. Vanwalleghem, A. Peña, A. Laguna, J. Giráldez, Controls on soil carbon storage from topography and vegetation in a rocky, semi-arid landscapes, Geoderma 311 (2018) 159–166.

[76] M. Hossain, W. Chen, Y. Zhang, Bulk density of mineral and organic soils in the Canada's arctic and sub-arctic, Inf. Process. Agricul. 2 (3–4) (2015) 183–190.

[77] A. Amirian Chekan, R. Taghizadeh Mehrjerdi, F. Sarmadian, A. Heidary, Three-dimensional mapping of soil texture using spline depth functions and artificial neural networks, Iranian J. Soil Water Res. 48 (1) (2017) 113–123.

[78] C. Fissore, B.J. Dalzell, A. Berhe, M. Voegtle, M. Evans, A. Wu, Influence of topography on soil organic carbon dynamics in a Southern California grassland, Catena 149 (2017) 140–149.

[79] E. Sakin, Organic carbon organic matter and bulk density relationships in arid-semi arid soils in Southeast Anatolia region, African J. Biotechnol. 11 (6) (2012) 1373–1377.

[80] H. Elbasiouny, M. Abowaly, A. Abu_Alkheir, A. Gad, Spatial variation of soil carbon and nitrogen pools by using ordinary Kriging method in an area of north Nile Delta, Egypt, Catena 113 (2014) 70–78.

[81] R. Kizilkaya, O. Dengiz, Variation of land use and land cover effects on some soil physico-chemical characteristics and soil enzyme activity, Zemdirbyste-Agricul. 97 (2) (2010) 15–24.

[82] I. Celik, Land-use effects on organic matter and physical properties of soil in a southern Mediterranean highland of Turkey, Soil Tillage Res. 83 (2) (2005) 270–277.

[83] A. Mondal, D. Khare, S. Kundu, S. Mondal, S. Mukherjee, A. Mukhopadhyay, Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data, Egypt. J. Remote Sens. Space Sci. 20 (1) (2017) 61–70.

[84] J. Meersmans, F. De Ridder, F. Canters, S. De Baets, M. Van Molle, A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium), Geoderma 143 (1–2) (2008) 1–13.

[85] M.J. Kennedy, D.R. Pevear, R.J. Hill, Mineral surface control of organic carbon in black shale, Science 295 (5555) (2002) 657–660.

[86] J.C. Gallant, T.I. Dowling, A multiresolution index of valley bottom flatness for mapping depositional areas, Water Resour. Res. 39 (12) (2003).

[87] S. Lamichhane, L. Kumar, K. Adhikari, Digital mapping of topsoil organic carbon content in an alluvial plain area of the Terai region of Nepal, Catena 202 (2021) 105299.

[88] R. Grimm, T. Behrens, M. Märker, H. Elsenbeer, Soil organic carbon concentrations and stocks on barro colorado island-digital soil mapping using random forests analysis, Geoderma 146 (1–2) (2008) 102–113.

[89] K. Adhikari, A.E. Hartemink, B. Minasny, R. Bou Kheir, M.B. Greve, M.H. Greve, Digital mapping of soil organic carbon contents and stocks in Denmark, PLoS One 9 (8) (2014) e105519.

[90] M. Lacoste, B. Minasny, A. McBratney, D. Michot, V. Viaud, C. Walter, High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape, Geoderma 213 (2014) 296–311.