



# Digital mapping and spatial modeling of some soil physical and mechanical properties in a semi-arid region of Iran

Pegah Khosravani · Majid Baghernejad ·  
Ali Akbar Moosavi · Meisam Rezaei

Received: 26 June 2023 / Accepted: 11 October 2023 / Published online: 24 October 2023  
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

**Abstract** The soil's physical and mechanical (SPM) properties have significant impacts on soil processes, such as water flow, nutrient movement, aeration, microbial activity, erosion, and root growth. To digitally map some SPM properties at four global standard depths, three machine learning algorithms (MLA), namely, random forest, Cubist, and  $k$ -nearest neighbor, were employed. A total of 200-point observation was designed with the aim of a field survey across the Marvdasht Plain in Fars Province, Iran. After sampling from topsoil (0 to 30 cm) and subsoil depths (30 to 60 cm), the samples were transferred to

the laboratory to determine the mean weight diameter (MWD) and geometric mean diameter (GMD) of aggregates in the laboratory. In addition, shear strength (SS) and penetration resistance (PR) were measured directly during the field survey. In parallel, 79 environmental factors were prepared from topographic and remote sensing data. Four soil variables were also included in the modeling process, as they were co-located with SPM properties based on expert opinion. For selecting the most influential covariates, the variance inflation factor (VIF) and Boruta methods were employed. Two covariate dataset scenarios were used to assess the impact of soil and environmental factors on the modeling of SPM properties including SPM and environmental covariates (scenario 1) and SPM, environmental covariates, and soil variables (scenario 2). From all covariates, nine soil and environmental factors were selected for modeling the SPM properties, of which four of them were the soil variables, three were related to remote sensing, and two factors had topographic sources. The results indicated that scenario 2 outperformed in all standard depths. The findings suggested that clay and SOM are key factors in predicting SPM, highlighting the importance of considering soil variables in addition to environmental covariates for enhancing the accuracy of machine learning prediction. The  $k$ -nearest neighbor algorithm was found to be highly effective in predicting SPM, while the random forest algorithm yielded the highest  $R^2$  value (0.92) for penetration resistance properties at 15–30 depth. Overall, the

## Highlights

- Machine learning algorithms were used for mapping soil physical and mechanical properties at four soil depths in a semi-arid region of Iran.
- Among the three machine learning algorithms that were compared, a  $k$ -nearest neighbor was identified as the best algorithm.
- Soil variables were recognized as the most critical SCORPAN factor in comparing topographic and remote sensing indices.
- The Boruta algorithm was used to select the best features among the pool of environmental covariates.

P. Khosravani · M. Baghernejad (✉) · A. A. Moosavi  
Department of Soil Science, College of Agriculture, Shiraz  
University, Shiraz, Iran  
e-mail: majidbaghernejad@yahoo.co.uk

M. Rezaei  
Soil and Water Research Institute (SWRI), Agricultural  
Research, Education and Extension Organization  
(AREEO), Karaj, Iran

approach used in this research has the potential to be extended beyond the Marvdasht Plain of Fars Province, Iran, as well as to other regions worldwide with comparable soil-forming factors. Moreover, this study provides a valuable framework for the digital mapping of SPM properties, serving as a guide for future studies seeking to predict SPM properties. Globally, the output of this research has important significance for soil management and conservation efforts and can facilitate the development of sustainable agricultural practices.

**Keywords** Cubist · Environmental covariate ·  $k$ -nearest neighbor · Machine learning algorithms · Random forest · Spline function

## Introduction

Soil physical and mechanical (SPM) properties are crucial parameters to evaluate soil quality and health, as well as to determine soil aggregate stability. They play a significant role in land degradation and land management in the surrounding environment across the world, especially the areas similar to Iran's climate condition (Rezaee et al., 2020a, 2020b; Mozaffari et al., 2021, b, 2022, Mozaffari et al., 2022; Zahedifar, 2023a, b). These properties also provide valuable information on the water infiltration and nutrients cycle and play crucial role in soil ventilation, microbial activity, and tillage performance (Mustafa et al., 2020). Therefore, knowing the variability of SPM in the landscape is necessary for determining the fertilizer requirement of agricultural crops, water, and cultivation management such as tillage toward sustainable production (Moosavi and Sepaskhah, 2012; Brevik et al., 2015).

Traditional mapping of SPM methods is labor intensive, timely, and costly, as they require highly dense soil observation, high field survey, and laboratory activates (Shahabi et al., 2017, Kazemi Garajeh et al., 2022). Moreover, they rely on expert interpretation (expert opinion) of environmental covariates involving the key soil-forming factors (Gorji et al., 2015). Furthermore, conventional soil mapping methods are unable to provide quantitative result of spatial soil maps in the term of accuracy and uncertainty analysis about variation of soil properties in soil survey projects (Zahedi et al., 2017). To prevail the limitations of conventional

approach, novel methods like the digital soil mapping (DSM) approach have been more applied by researchers in recent decades. DSM employs mathematical and statistical methods for establishing the correlation soil properties and environmental factors that are representative of soil formation factors (McBratney et al., 2003). The output of DSM consists of the spatial prediction maps along with their quantitative validation that can help reduce the cost and time for soil science surveys (Esfandiarpour-Boroujeni et al., 2020).

In current DSM studies, pedometricians typically derive topographic attributes from digital elevation model (DEM) (Wang et al., 2018) and remote sensing (RS) data (Xiao et al., 2019) which are easily accessible information sources of environmental covariate. The use of grid proxy of topographic attributes and RS indices has been confirmed in different studies in the field of SPM prediction (Mashalaba et al., 2020; Camera et al.; 2017; Ugbaje and Reuter, 2013). In this regard, Mashalaba et al. (2020) reported that topography feature had the most important effect in soil property prediction in central Chile, while environmental covariates have been widely applied worldwide, where they may not be sufficient for predicting soil properties. In two different studies, Mousavi et al. (2022) and Khosravani et al. (2023) investigate the estimation of SOC and soil properties in two scenarios which considers both soil and environmental covariates and only considers environmental covariates. Their results showed that including soil variables along with environmental covariates improved the accuracy of MLAs compared to the scenario without soil variables. Similarly, Zeraatpisheh et al. (2021) demonstrated that soil and RS variables were recognized as the most important driving of soil aggregate stability. In this regard, Mozaffari et al. (2022) believed that using the primary soil properties for modeling particles size distribution could lead to acceptable accuracy.

Recently, applied ML algorithms for modeling soil properties by the aid of environmental factors in the DSM have attracted more attention by pedometricians. Selection of the appropriate ML algorithms can have a significant impact on the accuracy of the produced maps (Khaledian and Miller, 2020). In a comprehensive review of research conducted over the last 10 years, Khaledian and Miller (2020) assess the capability of six ML algorithms in soil mapping, namely, multivariate linear regression (MLR),  $k$ -nearest neighbor ( $k$ -NN), support vector machine

regression (SVR), Cubist (CB), random forest (RF), and artificial neural network (ANN). According to their review, the RF algorithm was known to outperform in modeling of aggregate stability (Bouslihim et al., 2021). Also, Yamaç et al., (2020) confirm that the  $k$ -NN algorithm had the high  $R^2$  value (0.8) for predicting permanent wilting point (PWP) among other ML algorithms in calcareous soils.

In recent years, most DSM studies have focused on applying MLA to predict soil properties in top soil (Zeraatpisheh et al., 2019; Parsaie et al., 2021), whereas the potential of applied ML algorithms to predict SPM along with depth has not been deeply explored (Hengl et al., 2017). Some case studies have mostly focused on SOC prediction at top soil and subsoil simultaneously (Taghizadeh-Mehrjardi et al., 2014; Mousavi et al., 2022). According to the literature until writing this paper, few research has been conducted on the on the vertical variation of SPM using soil depth functions such as spline, ML algorithms, environmental covariates, and soil variables.

Therefore, limited studies have mapped spatial variation of SPM at the surface and subsurface (vertical and horizontal dimensions) by considering environmental covariates along with soil variables. Thus, the current research was conducted with the main aim of modeling SPM attributes including GMD, mean weight diameter of aggregates (MWD), shear strength (SS), and penetration resistance (PR) certainly using environmental covariates as a first scenario (S1) and accounting basic soil variables and environmental covariates as a second scenario (S2) in the southwest of Iran. Furthermore, we evaluated the capability of three ML algorithms of RF, CB, and  $k$ -NN in preparing spatial estimation maps of GMD, MWD, SS, and PR at increment four depths of 0–5, 5–15, 15–30, and 30–60 cm, to provide more accurate and detailed maps of SPM which can be used in land and water management strategies, soil erosion control, and improving soil stability.

## Materials and methods

### Research workflow

The general framework of this study is designed in six main steps and presented in Fig. 1. The main steps are presented in the following order: (1) designing

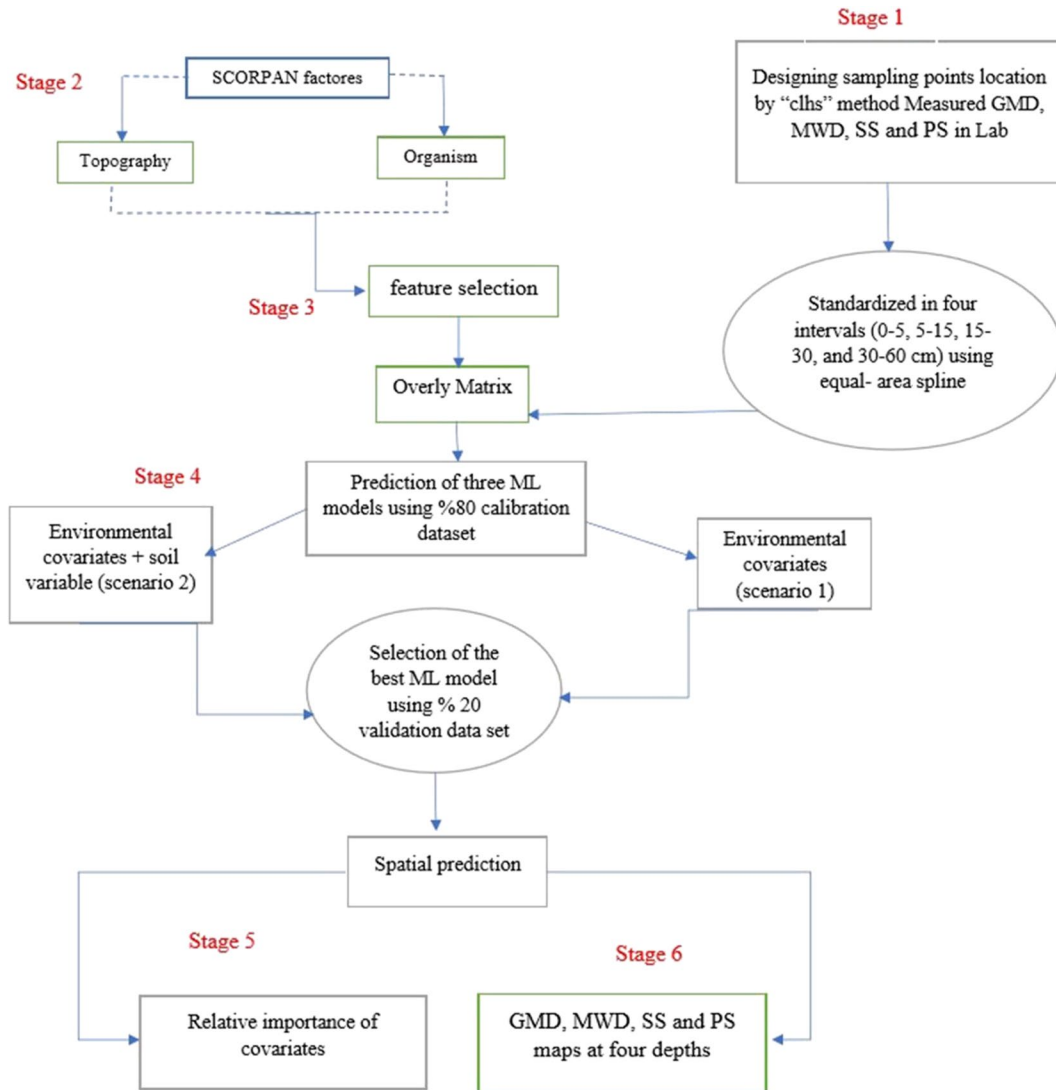
sampling point locations using the “clhs” package in R statistical software, collection of soil samples from 0 to 30 and 30 to 60 cm, and standardizing soil depth by spline depth function, (2) preparing/collecting all possible environmental covariates from RS indices and DEM as representative soil forming factors, (3) selecting the most appropriate environmental covariates for predicting SPM properties, (4) evaluating three ML algorithms (RF,  $k$ -NN, and CB) at the four standard soil depths (spatial modeling of soil properties) based on two scenarios (S1: environmental factors, S2: environmental factors + soil variables), (5) determining the relative importance (RI) of covariates; and (6) preparing prediction maps of SPM.

### Description of study area

Here, the interest area is limited to longitude of  $52^{\circ} 41' 35.82''$  to  $52^{\circ} 57' 1.07''$  E and latitude  $30^{\circ} 2' 14.72''$  to  $29^{\circ} 48' 35.02''$  N, covering about 48,963 ha in Marvdasht, which is located in Northern part of Shiraz (Fig. 2). The slope gradient is varying from 0 to 12% with mean altitude of 1605 msl. Most of this area has low physiographic intensity, and over 85% of the land has a slope of less than 5%. The mean annual precipitation and temperature are 287 mm and  $17.5^{\circ}\text{C}$ , and according to the closest climatic station, July and January are the hottest and coldest months, respectively. Also, the xeric and thermic are soil moisture and temperature regimes of the study area, respectively. Marvdasht Plain is a main agricultural region for crops like irrigated winter wheat, barley, alfalfa, and canola. Therefore, preparing digital maps of the key soil properties in this region offers valuable insights into the soil condition, and the maps can serve as a useful tool for evaluating and adjusting land management practices.

### Field survey and laboratory activity

For the field survey, the location of 200 sampling points was determined by Conditional Latin Hypercube Sampling (CLHS), a random stratified method that selects sampling points based on initial information pertaining to a suite of environmental factors in an interest area (Minasny & McBratney, 2006), using the open-source R statistical software (4.0.3 version). The location is shown in Fig. 2c. After fixing the sampling locations, soil shear strength (SS) and penetration resistance (PR)



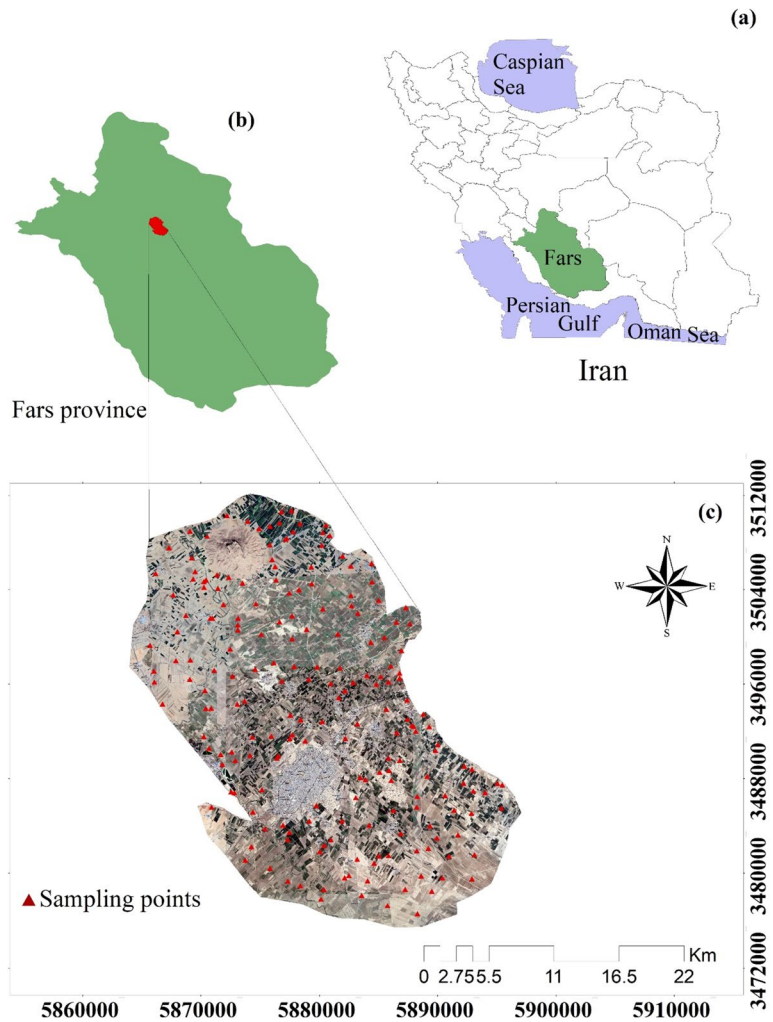
**Fig. 1** Flowchart of the research in the study area at soil standard depth (0–5, 5–15, 15–30, and 30–60 cm). Geometric mean diameter of aggregates (GMD), mean weight diameter of

aggregates (MWD), shear strength (SS), penetration resistance (PR), conditioned Latin hypercube sampling (clhs), machine learning model (ML)

were directly determined by Vane shear resistance meter and pocket cone penetrometer (ELE algorithm), respectively, during field survey (Fig. 3a, b). The SS was measured using a Torvin resistance tester with three replications around each sampling point. Furthermore, PR was measured using a hand-held penetrometer on intact soils with three replications around each point in the study area. The instrument had a narrow cylindrical rod of 6 mm diameter and 5.7 cm length. The penetrometer was pushed into the soil up to the marked part (about 6 mm), and the required pressure

(kPa) was recorded. It should be noted that the average values of three measurements (replications) were used to determine the soil shear strength (SS) and penetration resistance (PR) at each sampling point. The measurements were taken at points at equal distances on the side of a circle with a radius of about 0.5 m (Fig. 3c). After sampling, soil samples were transferred to the laboratory for measuring the aggregate stability using wet sieving method (Kemper and Rosenau, 1986). In other words, stability of the soil aggregates against water were measured using the standard sieving method

**Fig. 2** **a** Location of Fars province in Iran, **b** location of Marvdasht plain in Fars province, and **c** location of the soil samples (red circle) in the study area



with seven group sizes of sieves including > 2, 2, 1, 0.5, 0.25, 0.125, 0.053, and < 0.053 mm opening diameter (Kemper and Rosenau, 1986; Le Bissonnais, 2016). Then, for quantifying the structural stability of the soil aggregates, the geometric mean diameter (GMD) and mean weight diameter (MWD) values were calculated based on the results of the wet sieving method using the following equations:

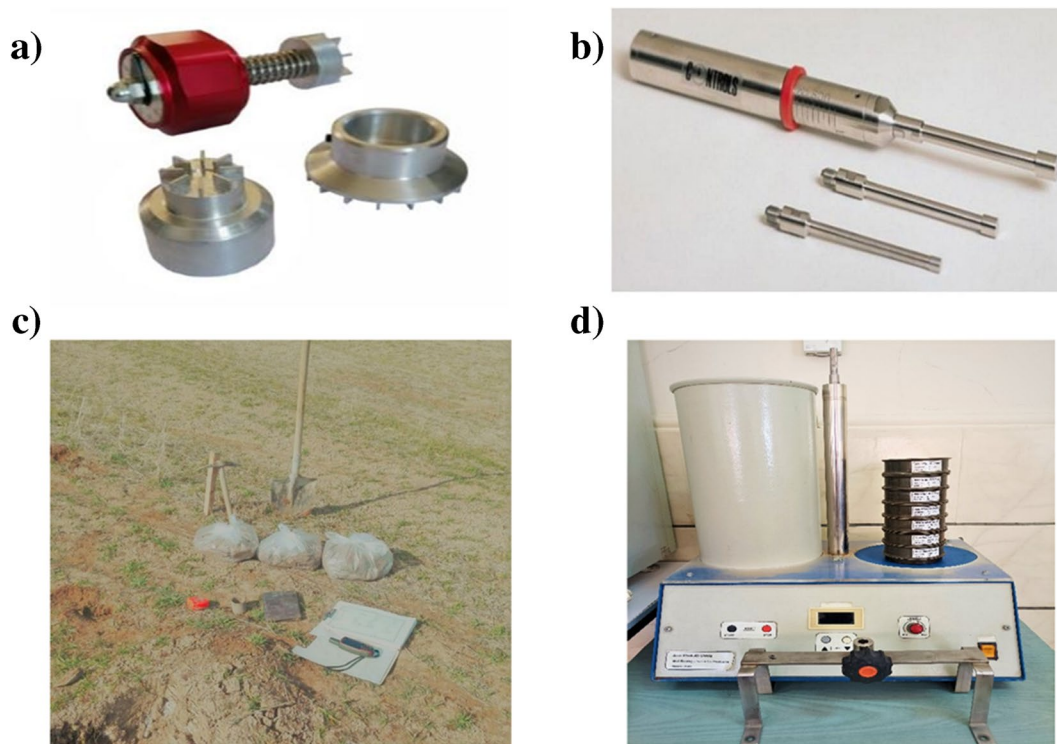
$$GMD = \exp \left[ \sum_{i=1}^n w_i \text{Log}(\bar{d}_i) \right] \tag{1}$$

$$MWD = \sum_{i=1}^n w_i \bar{d}_i \tag{2}$$

where  $\bar{d}_i$  is the mean diameter of two consecutive sieves (mm) and  $W_i$  is the weight of particles in that size range as a percentage of the total sample.

Moreover, auxiliary soil properties including soil organic matter (SOM) and soil textural components (i.e., sand, silt, and clay contents) were measured using the wet oxidation (Nelson and Sommers, 1996) and hydrometer (Gee and Bauder, 1986) methods, respectively. Furthermore, in order to prepare continuous maps of the mentioned soil attributes for use in further steps, the interpolation geostatistical approach of Ordinary Kriging was used to estimate the values of the aforementioned soil attributes at unknown points using their corresponding measured values along with their modeled spatial structure.





**Fig. 3** **a** Shear strength (SS), **b** penetration resistance (PR), **c** field sampling point, and **d** wet sieving tool

The detailed descriptions on the estimation procedure using the mentioned geostatistical approach can be seen in the literature (Moosavi and Sepaskhah, 2012; Moradi et al., 2016; Azizi et al., 2022).

#### Standardization of soil depth

The spline method was used to extract soil properties at standard depths along vertical profiles continuously using the R package “GSIF.” The values for GMD, MWD, SS, and PR were standardized at four depths 0–5, 5–15, 15–30, and 30–60 cm. For more details about splines, see Bishop et al. (1999) and Malone et al. (2009).

#### Feature selection

A total of 79 driving factors were prepared from soil variable RS data and topographic attributes. Here, four soil covariates such as clay, silt, sand, and soil organic matter (SOM) were selected based on the expert opinion and related literature in this field (Celik, 2005; Ayoubi et al., 2012; Zeraatpisheh et al., 2021); 36 remote sensing covariates and individual

band were prepared from Landsat 8 with 30-m spatial resolution after the necessary corrections (radiometry) using ENVI software version 5.3. In addition, 39 topographic attributes were extracted from DEM (ALOS PALSAR satellite) using the topographic analysis method (Wilson, 2018) in SAGA GIS version 7.9.1 software. As mentioned, we prepared the maps of soil variables by the results of Ordinary Kriging interpolation method (Azizi et al., 2022). Finally, the spatial resolution of all covariates was fixed to 30-m in Arc GIS software. The details of the covariates are described in Table 1.

After preparing soil and environmental factors for avoiding the increase of time and model fitting process, the environmental factors were chosen using the variance inflation factor (VIF) method (Akinwande et al., 2015), using the “VIF” package in R software. The VIF is a step way method and eliminates the covariates that have the highest correlation with each other. After applying VIF method, 36 environmental covariates remained. To further select the most appropriate covariates, the Boruta method was applied.

**Table 1** Soil and environmental covariates used in this study to predict soil physical and mechanical (SPM) attributes

SCORPAN factors	Parameter	Abbreviation	Source
S	Clay content (%)	Clay	LAB
S	Soil Organic Matter	SOM	LAB
S	Silt content (%)	Silt	LAB
S	Sand content (%)	Sand	LAB
O	Normalized different vegetation index	NDVI	RS
O	Renormalized difference vegetation index	RDVI	RS
O	Band 2 of Landsat 8	Blue	RS
O	Band 3 of Landsat 8	Green	RS
O	Band 4 of Landsat 8	Red	RS
O	Band 5 of Landsat 8	Near infrared	RS
O	Band 6 of Landsat 8	Short-wave infrared-1	RS
O	Band 7 of Landsat 8	Short-wave infrared-2	RS
O	Wetness brightness difference index	WBDI	RS
O	Normalized difference salinity index	NDSI	RS
O	Normalized difference moisture index	NDMI	RS
O	Brightness Index	BI	RS
O	Gypsum Index	GI	RS
O	Clay Index	CI	RS
O	Carbonate Index	CI	RS
O	Ratio Vegetation Index	RVI	RS
O	Enhanced Vegetation Index	EVI	RS
O	Green-Red Vegetation Index	GVI	RS
O	Differenced Vegetation Index	DVI	RS
O	Infrared Percentage Vegetation Index	IPVI	RS
O	Iron Oxide Ratio	IOR	RS
O	Soil Adjusted Vegetation Index	SVI	RS
O	Modified Soil Adjusted Vegetation Index	MSAVI	RS
O	Perpendicular Vegetation Index	PVI	RS
O	Structure Insensitive Pigment Index	SIPI	RS
O	Modified Normalized Difference Water Index	MNDVI	RS
O	Enhanced Vegetation Index	EVI	RS
O	Soil Adjusted Vegetation Index	SAVI	RS
O	Bare Soil Index	BSI	RS
O	Index-Based built-up Index	IBI	RS
O	Enhanced Vegetation Index 2	EVI 2	RS
O	Visible Atmospherically Resistance Index Green	VARIgreen	RS
O	Green Normalized Difference Vegetation Index	GNDVI	RS
O	Tasseled cap wetness	TCW	RS
O	Tasseled cap greenness	TCG	RS
O	Normalized Ratio Vegetation Index	NRVI	RS
R	Wind Effect	WE	DEM
R	Analytical Hillshading	AH	DEM
R	Aspect	Aspect	DEM
R	Catchment Area	CA	DEM
R	Catchment Slope	CS	DEM
R	Channel Direction	CD	DEM

**Table 1** (continued)

SCORPAN factors	Parameter	Abbreviation	Source
R	Channel Network Base Level	CNBL	DEM
R	Channel Network Distance	CND	DEM
R	Channel Network	CN	DEM
R	Closed Depressions	CD	DEM
R	Convergence Index	CI	DEM
R	Diffuse Insolation	Diffuse	DEM
R	Direct Insolation	DI	DEM
R	Flow Directions	FD	DEM
R	Geomorphons	Geomorphons	DEM
R	Landforms	Landforms	DEM
R	LS Factor	LS Factor	DEM
R	Mass Balance Index	MBI	DEM
R	Modified Catchment Area	MCA	DEM
R	MRRTF	MRRTF	DEM
R	Plan Curvature	PC	DEM
R	Relative Slope Position	RSP	DEM
R	Slope Height	SH	DEM
R	Slope	Slope	DEM
R	Standardized Height	SH	DEM
R	Stream Power Index	SPI	DEM
R	Surface Area	SA	DEM
R	Watershed Basins	WB	DEM
R	Wind Exposition	WE	DEM
R	Valley depth (m)	VD	DEM
R	Multiresolution index of valley bottom flatness	MRVBF	DEM
R	Normalized Height	NH	DEM
R	Midslope position	MS	DEM
R	Texture	Texture	DEM
R	Slope gradient	SG	DEM
R	Vertical Distance to Channel Network	VDCN	DEM
R	Solar radiation or insolation	SROI	DEM
R	Convexity	Convexity	DEM
R	Topographic Wetness Index	TWI	DEM

*S* soil properties, *O* organism, *R* relief, *LAB* laboratory analysis, *DEM* digital elevation model, *RS* remote sensing

The Boruta algorithm for selecting environmental covariates was proposed by Kursa and Rudnicki (2010). This approach is one of the semi-automated supervised methods for feature selection, based on the random forest (RF) algorithm, which selects the most important environmental covariates using the repeatable backward and forward system. Finally, the output of the covariate's selection was done based on

the value of the *Z* factor, which is determined in four general categories. The covariate is unrelated, slightly related, moderately related, or completely related when the *Z* factor is lower than 5, 5 to 10, 10 to 15, and more than 15, respectively (Keskin et al., 2019). Additionally, soil variables were included in the modeling process by the expert opinion.



### Machine learning (ML) algorithms

In this study, we evaluated three ML algorithms, RF, *k*-NN, and CB, to predict SPM by using soil and environmental factors and employing two scenarios including S1 (using just environmental factors, i.e., topographic and RS covariates) and S2 (using both the mentioned environmental covariates and soil variables), at four standard depths.

#### Random forest (RF)

Random forest (RF) is one of the non-linear ML algorithms that is widely used in DSM of soil properties. The RF algorithm is easy to implement and requires few parameters to tune (Rahmani et al., 2022). Here, the RF algorithm was applied to predict the surface and depth of the soil’s physical and mechanical properties. The RF algorithm was tuned according to two hyper-parameters: the number of trees (ntree), which was between 100 and 1000 trees with the distance of 100 trees interval, and mtry, which represents the number of environmental covariates that can be used to grow at each tree according to the minimum error (Breiman, 2001).

#### *k*-nearest neighbor (*k*-NN)

The *k*-nearest neighborhood (*k*-NN) algorithm is one of the non-linear methods. This operates based on calculating the Euclidean distance between the desired soil sample and other observation points. The *k*-NN method then weighs *k* numbers of adjacent observation samples based on their distance to the desired sample. In addition, based on the weight of each sample in a set of *k* number of samples, an estimate of the desired data is made according to the minimum error in that set (Nemes et al., 2006).

#### Cubist (CB)

The Cubist algorithm is a regression tree algorithm that generates various algorithms using training data. Each algorithm comprises multiple rules, which are summarized by one or more conditions (Holmes et al., 1999). When all the conditions of a rule are met, the corresponding linear relationship is utilized to forecast the SPM. The algorithm’s rules are ranked through the Cubist algorithm’s decreasing importance process. This implies that the first rule has the highest

contribution to the algorithm’s accuracy, while the last rule has the least. The algorithm predicts the target variable’s value based on influential variables, and the number of rules is adjusted using the best-fitting regression algorithm. To optimize the algorithm, it was fine-tuned by adjusting two hyper-parameters: the number of committees and the number of neighbors (Ma et al., 2017).

### Assessment of prediction performance

For assessment of the ML algorithms (RF, *k*-NN, and CB), all data was split to the training and testing subset which consisted of 80 and 20%, respectively. Four statistical indices included the coefficient of determination ( $R^2$ ), normalized root means square error (nRMSE), and Nash-Sutcliffe coefficient (NS) that is a statistical measure commonly applied to assessment of the performance of ML algorithm predictions. Also, the mean standardized squared prediction error (MSSPE) was applied for assessment of the uncertainty of ML algorithms. It is defined as the mean squared prediction error (MSSPE) of an algorithm divided by the average MSSPE of a set of benchmark algorithms (Rossel and McBratney, 2008). The mentioned statistical measures were calculated using the following equations:

$$R^2 = \frac{\left(\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})\right)^2}{\sum_{i=1}^n (O_i - \bar{O})^2 \sum_{i=1}^n (P_i - \bar{P})^2} \tag{3}$$

$$\text{nRMSE (\%)} = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2}}{\bar{O}} \times 100 \tag{4}$$

$$\text{NS} = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \tag{5}$$

$$\text{MSSPE} = (1/n)^* \sum (O_i - P_i)^2 / (S^2) \tag{6}$$

where  $O_i$  and  $P_i$  are the observed and predicted values, respectively;  $n$  is the amount of data; and  $S^2$  is the variance of the observed values. As SPM varies on different scales, the nRMSE is a suitable statistical index for quantifying the algorithm accuracies in this

study. The nRMSE values range from 0 to 100, where values close to zero showed excellent performance, and values above 0.3 show poor algorithm validation (Bannayan and Hoogenboom, 2009).

## Results and discussion

### Summary statistic

The results of statistical indices of SPM for 200 soil samples at the four standards of Lapuee plain are presented in Table 2. The results showed that both GMD and MWD decreased by increasing the depth, while SS and PR have irregular trends with depth. According to the coefficient of variation, CVs, all four soil properties at four standard depths showed high variability according to the classification by Wilding (1985). One of the reasons for the high CV values of SPM may be attributed to the agricultural activities and land management (Heydari et al., 2020). According to the findings, the average SPM at a standard depth decreased from top- to sub-soils (Table 2). Higher amount of SOM was observed at upper layers which is same with Mousavi et al. (2023) findings. Based on the pedological theories,

the SOM increases the porosity and ventilation and reduces soil compaction (Soane, 1990; Elbasiouny et al., 2014). Therefore, it seems that SOM has a significant effect on SPM.

### Selected features

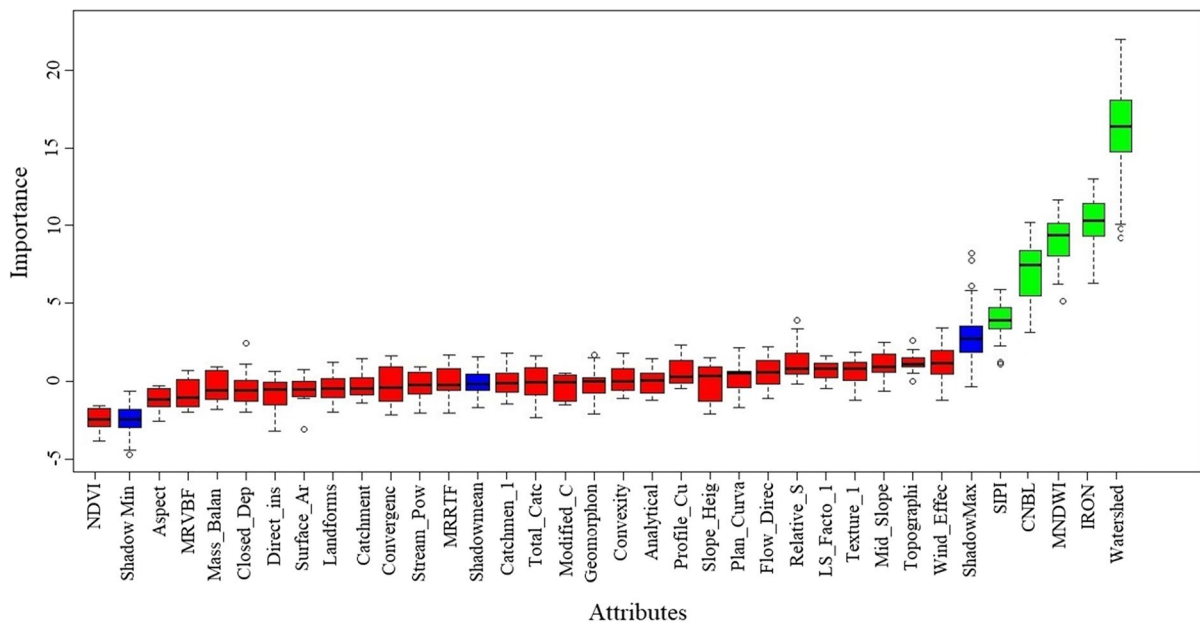
When dealing with a large pool of data, using all of it can be time-consuming and can increase algorithm complexity. Feature selection is a useful method for choosing the appropriate types of covariates which are using for the modeling process (Neyestani et al., 2021).

Based on our aims for selecting the most relevant environmental factors, through the VIF method, the number of environmental factors was reduced from 75 to 36. The Boruta algorithm was also fitted, and five covariates were ultimately selected from the 36 environmental covariates (Fig. 4), in addition to four soil variables, resulting in a total of nine variables used for predicting the SPM properties (Table 3). Among the selected environmental covariates, three of them (SIPI, MNDWI, and IRON) were related to RS indices, while two of them (Watershed Basins (WB) and Channel Network Base Level (CNBL)) were extracted from DEM (Fig. 5). Also, as mentioned in the section of 2.5.1, soil variables of SOM, clay, silt, and sand contents were selected based on

**Table 2** Descriptive statistics of soil physical and mechanical properties at the four standard soil depths for the soil samples ( $n = 200$ )

Soil properties	Depth (cm)	Minimum	Maximum	Mean	Median	CV (%)
GMD (mm)	0–5	0.24	2.76	1.51	1.50	38.4
	5–15	0.21	2.14	1.48	1.49	37.8
	15–30	0.23	2.09	1.37	1.39	36.4
	30–60	0.20	2.01	1.13	1.17	38.0
MWD (mm)	0–5	0.32	2.43	2.06	2.09	38.8
	5–15	0.32	2.35	2.02	2.03	38.6
	15–30	0.33	2.03	1.87	1.93	37.9
	30–60	0.28	2.31	1.53	1.49	40.5
SS (kPa)	0–5	0.14	2.63	2.43	2.60	49.7
	5–15	0.15	2.57	2.43	2.57	48.1
	15–30	0.10	2.39	2.43	2.59	42.3
	30–60	0.11	1.99	2.44	2.61	42.6
PR (kPa)	0–5	0.15	4.45	0.95	0.76	98.9
	5–15	0.17	4.28	0.94	0.75	100
	15–30	0.18	3.79	0.94	0.79	91.4
	30–60	0.10	4.19	0.98	0.75	94.8

CV coefficient of variation, GMD geometric mean diameter of aggregates, MWD mean weight diameter of aggregates, SS shear strength, PR penetration resistance



**Fig. 4** Important variables selection with Boruta algorithm

**Table 3** Select soil and environmental covariates for four properties obtained with Boruta at the four standard depths

SCORPAN factors	Parameter	Abbreviation	Source
O	Structure Insensitive Pigment Index	SIPI	RS
O	Modified Normalized Difference Water Index	MNDWI	RS
O	Iron Oxide Ratio	IRON	RS
R	Watershed Basins	WB	DEM
R	Channel Network Base Level	CNBL	DEM
S	Soil Organic matter	SOM	Soil
S	Clay	Clay	Soil
S	Silt	Silt	Soil
S	Sand	Sand	Soil

O organism, R relief, S soil data, RS remote sensing, DEM digital elevation model, SIPI Structure Insensitive Pigment Index, MNDWI Modified Normalized Difference Water Index, IRON Iron Oxide Ratio, WB Watershed Basins, CNBL Channel Network Base Level

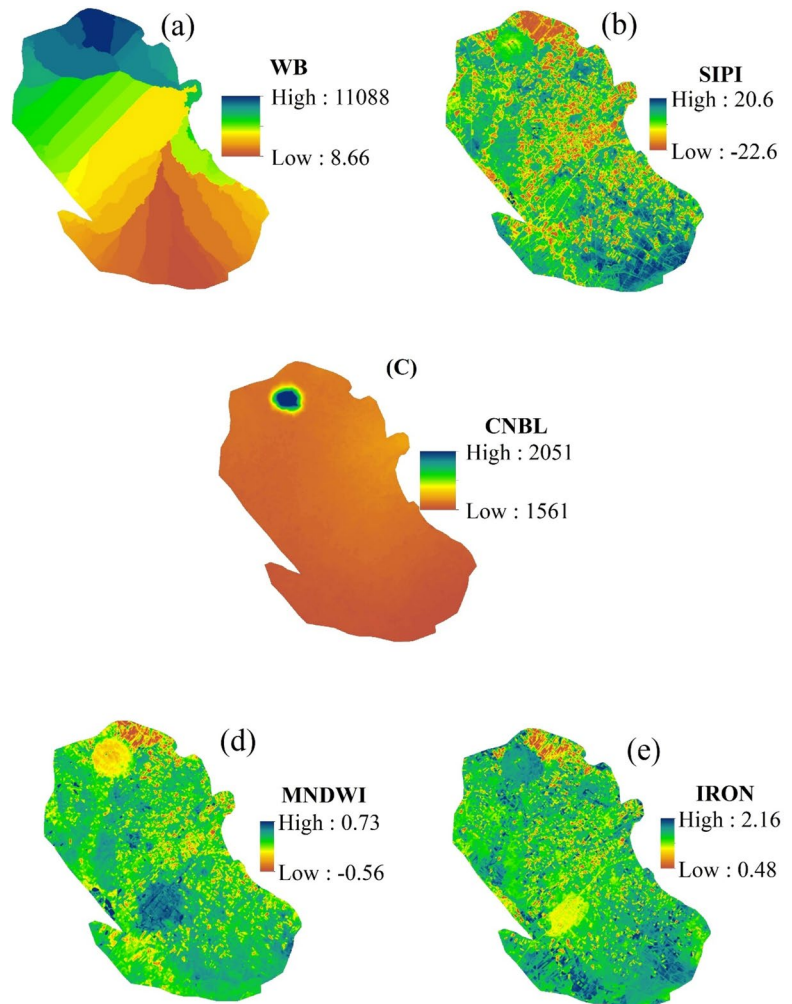
expert opinion (Fig. 6). The most important soil and environmental covariates based on the best scenarios and ML algorithms (Table 4) were applied to predict each SPM.

#### Algorithm performance

The accuracy of the algorithm prediction was evaluated using statistical indices such as  $R^2$ , nRMSE, and NS for GMD, MWD, SS, and PR at four standard depths. The comparison between scenarios S1 and S2, as measured

by R2 and NS, showed that S2 had the highest prediction accuracy for the SPM at the four standard depths. According to the finding, including the soil variable and environmental factors improves the performance of ML algorithms (Mousavi et al., 2022). Tables 5 and 6 list the quantitative results of the scenarios comparison for S1 and S2, respectively, using the ML algorithms. Overall, the validation results for the two scenarios indicate that S2 had higher accuracy, and the subsequent sections will focus on its results.

**Fig. 5** Five environmental covariates were obtained from RS: Structure Insensitive Pigment Index (SIPI), Modified Normalized Difference Water Index (MNDWI), Iron Oxide Ratio (IRON), Watershed Basins (WB), and Channel Network Base Level (CNBL)



#### Validation results of GMD

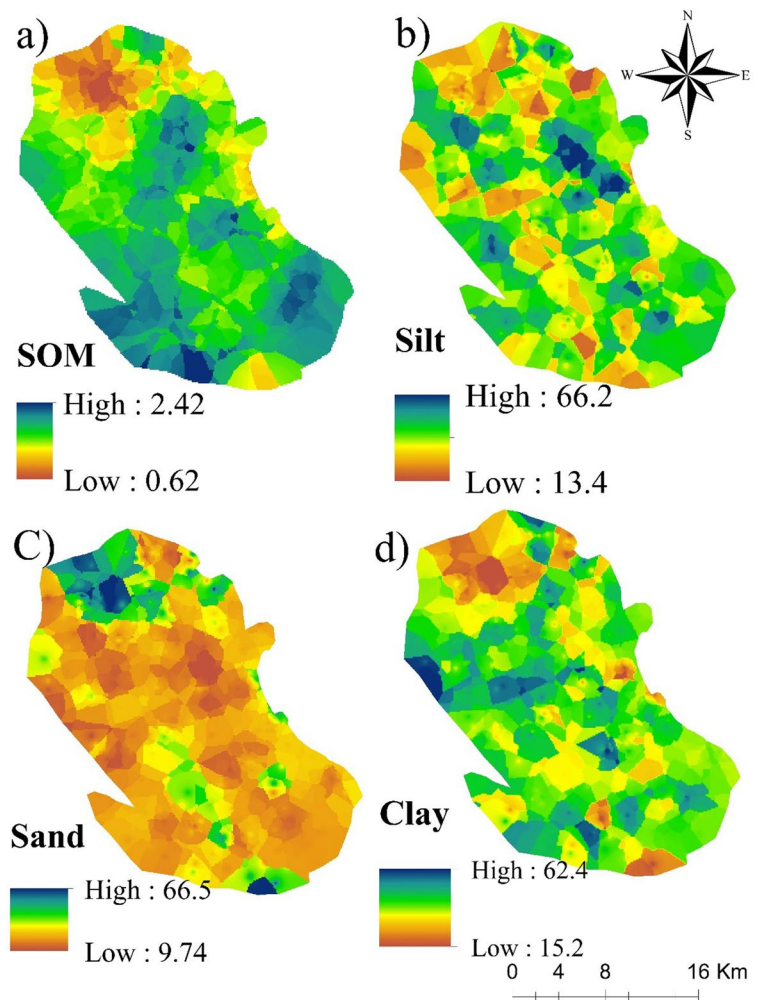
According to Table 6, the RF algorithm showed the best prediction performance in mapping GMD with ( $R^2 = 0.70$  and  $0.68$ ,  $nRMSE = 18.21$  and  $10.21$ , and  $NS = 0.67$ , and  $0.60$ ) in 0–5 and 15–30 cm depths, respectively. The  $k$ -NN algorithm also performed the best for predicting GMD with  $R^2$  of  $0.45$ ,  $nRMSE$  of  $10.17$ , and  $NS$  of  $0.48$  at the depth of 5–15 cm. Finally, CB showed the best predictions of GMD with  $R^2$  of  $0.59$ ,  $nRMSE$  of  $6.19$ , and  $NS$  of  $0.58$  at a depth of 30–60 cm. Furthermore, based on Rossel and McBratney (2008) report, all applied ML algorithms in GMD prediction at four standard depths showed intermediate prediction performance; however, at 0 to 5 cm and 15 to 30 cm depth, the RF outperformed best. Chen et al. (2015) examined surface

and subsurface soil salinity variation and reported that the RF had high capability in prediction verses other ML algorithms. Similarly, Mousavi et al. (2022) and Rahmani et al. (2022) confirm that the RF algorithm has high accuracy and low error for predicting topsoil thickness.

#### Validation results of MWD

The  $k$ -NN displayed the well performance in predicting MWD (Table 6). As shown in Table 6, for three depths of 0 to 5, 15 to 30, and 30 to 60 cm, the  $k$ -NN algorithm was the best one ( $R^2$  of  $0.57$ ,  $0.56$ , and  $0.45$  and  $nRMSE$  of  $14.12$ ,  $8.21$ , and  $10.23$ , respectively), while CB showed the best prediction with  $R^2$  of  $0.61$  and  $nRMSE$  of  $8.13$  in 5 to 15 cm depth. In general, the modeling results showed a similar performance of

**Fig. 6** Four variables were obtained from soil analysis: clay, sand, silt, and SOM



prediction for GMD compared to MWD at all depths. Overall, the validation results of the algorithms were intermediate at standard depths; however, the *k*-NN showed better performance compared to other algorithms for MWD; in all standard depths except 5–15 cm depth, CB algorithm showed the best prediction than the *k*-NN algorithm.

*Validation results of SS*

Validation results of the predictive algorithms for SS showed that the best algorithm at all depths was the *k*-NN algorithm with  $R^2$  of 0.65, 0.54, 0.57, and 0.59; nRMSE of 11.14, 12.15, 13.23, and 11.16; and NS of 0.55, 0.53, 0.54, and 0.51 at depths of 0–5, 5–15, 15–30, and 30–60 cm, respectively (Table 6). Furthermore, results showed that all ML algorithms

used in this research had similar performances in predicting SS at surface soils. The highest  $R^2$  value ( $R^2 = 0.65$ ) was obtained for SS prediction in depth of 5–15 cm compared to the other studied depths. Overall, the *k*-NN was the best predictive algorithm for SS in all of the studied depths. In this regard, research conducted by Hengl et al. (2021) on modeling soil fertility properties showed that the RF and CB algorithms had the best accuracy compared to that of the *k*-NN and SVR algorithms. Furthermore, Khosravani et al. (2023) reported that the CB algorithm which followed by RF had the best prediction capability for soil fertility attributes. As regards SPM, similar results were reported by Bouslihimi et al. (2021), while Yamaç et al. (2020) stated that the *k*-NN algorithm was the best predictive algorithm for permanent wilting point (PWP) in calcareous soils.



**Table 4** The most important soil and environmental covariates for prediction of soil physical and mechanical properties based on the best model at each four-standard depth

Soil properties	Depth (cm)	The most important covariates		
GMD (mm)	0–5	Clay (43.4%)	OM (23%)	Silt (18.6%)
	5–15	Clay (44.2%)	OM (20.1%)	Silt (17.3%)
	15–30	Clay (41.5%)	OM (25.3%)	Silt (18.9%)
	30–60	Clay (46.7%)	OM (22.8%)	Silt (16.2%)
MWD (mm)	0–5	Clay (41.3%)	OM (24.8%)	Silt (20.1%)
	5–15	Clay (40.6%)	OM (19.8%)	Silt (18.2%)
	15–30	Clay (49.2%)	OM (21.3%)	CNBL (15.1%)
	30–60	Clay (48.5%)	OM (20.5%)	CNBL (13.8%)
SS (kPa)	0–5	Clay (38.9%)	OM (30.2%)	Silt (23.2%)
	5–15	Clay (39.1%)	OM (28.7%)	Silt (20.9%)
	15–30	Clay (40.2%)	OM (25.6%)	Silt (22.4%)
	30–60	Clay (40.5%)	OM (28.7%)	CNBL (13.4%)
PR (kPa)	0–5	Clay (39.7%)	OM (29.6%)	Silt (21.8%)
	5–15	Clay (41.2%)	OM (27.6%)	Silt (20.45%)
	15–30	Clay (45.5%)	OM (23.7%)	CNBL (12.2%)
	30–60	Clay (44.8%)	OM (26.3%)	CNBL (13.4%)

*WB* Watershed Basins, *CNBL* Channel Network Base Level, *MNDWI* Modified Normalized Difference Water Index, *IRON* Iron Oxide Ratio, *GMD* geometric mean diameter of aggregates, *MWD* mean weight diameter of aggregates, *SS* shear strength, *PR* penetration resistance

### Validation results of PR

The validation results indicated that CB was the best algorithm ( $R^2 = 0.67$ ) for predicting PR in the depth of 0–5 cm, while *k*-NN was the best algorithm ( $R^2 = 0.68$ ) in the depth of 5–15 cm. Also, RF algorithm ( $R^2 = 0.92$  and  $0.86$ ) was the best for predicting PR at the depth 15–30 cm and 30–60 cm (Table 6).

Based on the results obtained from the algorithms' validation for predicting PR, the RF algorithm showed the best performance ( $R^2 = 0.92$ ) compared to other ML algorithms at the depth of 15–30 cm. But the *k*-NN algorithm performed well compared to the other ML algorithms to predict GMD, MWD, SS, and PR properties at different depths. Totally, the best predictive algorithms were RF for GMD and PR properties, and *k*-NN for MWD and SS properties. Zeraatpisheh et al. (2021) reported that the *k*-NN and support vector machine algorithms were performed well in prediction of SOC in different aggregate size. Furthermore, it is shown that the RF algorithm in comparison to the ANN algorithm was better in prediction of soil surface erosion rate (Khosravi Aqdam et al., 2022).

### The importance ranking of factors

The comparison between algorithms' scenarios showed that scenario S2 had the higher accuracy in predicting SPM compared to scenario S1. Therefore, the relative importance was described based on scenario S2. The results indicated that clay and SOM were two important variables in the prediction of SPM at four standard depths. Increasing SOM can improve soil aggregate stability, which may explain the high GMD and MWD values in cultivation land (Lacoste et al., 2014). Also, Mozaffari et al. (2021, b) observed strong relationship among SOM and MWD and GMD in all of their datasets. They believe that the SOM had important role by protecting SAS and decreasing the effect of wind and water erosion. Additionally, Celik (2005) and Ayoubi et al. (2012) reported that SOM directly contributed to soil aggregate formations and stabilities, and also, the level of SOM can define and explain the type of soil aggregates (macro, meso, and micro aggregates). Correlation analysis between MWD, soil properties, and covariates revealed that organic carbon had the highest influence (27.9%) on MWD. Similar result was reported by Tang et al. (2016) and Wang et al. (2019).



**Table 5** Validation results for prediction of the soil physical and mechanical properties at the four standard depths (scenario 1: using environmental covariates)

Soil properties	ML models	Soil depths (cm)															
		0–5				5–15				15–30				30–60			
		$R^2$	nRMSE	MSSPE	NS	$R^2$	nRMSE	MSSPE	NS	$R^2$	nRMSE	MSSPE	NS	$R^2$	nRMSE	MSSPE	NS
GMD (mm)	RF	0.16	28.1	2.03	0.15	0.00	13.6	2.15	0.00	0.01	18.5	2.45	0.05	0.03	26.4	2.18	0.02
	<i>k</i> -NN	0.14	32.3	2.16	0.10	0.05	31.6	2.41	0.02	0.06	26.5	2.87	0.01	0.08	30.4	2.44	0.05
	CB	0.09	28.5	3.01	0.08	0.00	27.6	2.36	0.06	0.09	34.5	2.36	0.03	0.16	18.3	2.10	0.03
MWD (mm)	RF	0.10	10.7	2.65	0.08	0.08	19.7	2.89	0.03	0.08	19.6	2.58	0.01	0.03	25.6	2.78	0.06
	<i>k</i> -NN	0.17	13.7	2.85	0.08	0.13	21.6	2.80	0.08	0.27	26.6	2.67	0.05	0.07	20.6	2.69	0.05
	CB	0.10	26.7	2.47	0.02	0.17	30.6	2.74	0.02	0.02	29.6	2.35	0.09	0.06	19.6	2.13	0.02
SS (kPa)	RF	0.31	17.3	3.12	0.21	0.27	21.9	2.69	0.15	0.37	15.8	3.14	0.27	0.20	22.2	2.33	0.12
	<i>k</i> -NN	0.30	29.9	3.02	0.22	0.34	27.8	2.85	0.24	0.45	34.8	3.19	0.16	0.29	21.0	2.89	0.08
	CB	0.29	19.0	2.98	0.20	0.33	37.8	3.02	0.10	0.36	29.8	2.90	0.20	0.01	26.1	2.75	0.04
PR (kPa)	RF	0.11	29.9	1.02	0.10	0.09	36.8	1.58	0.01	0.90	28.4	1.58	0.70	0.19	20.9	1.41	0.06
	<i>k</i> -NN	0.22	25.8	1.16	0.10	0.21	34.7	1.47	0.03	0.17	20.7	1.26	0.10	0.01	23.9	1.25	0.06
	CB	0.20	30.8	1.25	0.17	0.19	21.7	1.30	0.09	0.15	24.7	1.38	0.01	0.08	21.0	1.50	0.17

*GMD* geometric mean diameter of aggregates, *MWD* mean weight diameter of aggregates, *SS* shear strength, *PR* penetration resistance,  $R^2$  coefficient of determination, *RMSE* root mean square error, *Bias* bias, *RF* random forest, *k*-*NN* *k*-nearest neighbor, *CB* Cubist

**Table 6** Validation results for prediction of the soil physical and mechanical properties at the four standard depths (scenario 2: using both environmental covariates and soil variables)

Soil properties	ML models	Soil depths (cm)															
		0–5				5–15				15–30				30–60			
		$R^2$	nRMSE	MSSPE	NS	$R^2$	nRMSE	MSSPE	NS	$R^2$	nRMSE	MSSPE	NS	$R^2$	nRMSE	MSSPE	NS
GMD (mm)	RF	0.70	18.2	1.46	0.67	0.41	11.1	1.59	0.41	0.68	10.2	1.29	0.60	0.43	10.1	1.10	0.42
	<i>k</i> -NN	0.40	13.2	1.89	0.29	0.45	10.1	1.47	0.48	0.55	9.11	1.93	0.40	0.57	9.12	1.56	0.55
	CB	0.46	18.1	2.01	0.40	0.40	8.15	1.68	0.40	0.50	10.0	1.68	0.45	0.59	6.19	1.48	0.58
MWD (mm)	RF	0.41	12.1	2.03	0.40	0.48	11.1	2.67	0.41	0.41	10.01	1.96	0.39	0.44	11.1	1.89	0.36
	<i>k</i> -NN	0.57	14.1	1.83	0.50	0.60	9.20	1.98	0.59	0.56	8.21	1.75	0.51	0.45	10.2	1.76	0.42
	CB	0.42	10.1	2.85	0.41	0.61	8.13	2.05	0.60	0.50	9.24	1.99	0.52	0.42	9.21	1.50	0.41
SS (kPa)	RF	0.62	10.1	2.40	0.54	0.48	10.0	2.30	0.48	0.43	10.0	2.03	0.33	0.52	10.2	2.34	0.50
	<i>k</i> -NN	0.65	11.1	2.37	0.55	0.54	12.1	2.28	0.53	0.57	13.2	2.18	0.54	0.59	11.1	2.06	0.51
	CB	0.60	10.1	2.12	0.57	0.40	9.16	2.63	0.39	0.41	10.1	2.36	0.40	0.57	10.1	2.16	0.54
PR (kPa)	RF	0.58	10.0	0.71	0.53	0.42	12.1	1.87	0.41	0.92	10.2	1.02	0.86	0.54	10.1	0.87	0.53
	<i>k</i> -NN	0.46	11.0	0.94	0.44	0.68	10.1	1.32	0.63	0.39	11.2	0.90	0.30	0.31	9.23	0.99	0.25
	CB	0.67	9.01	0.82	0.60	0.56	9.08	1.34	0.49	0.45	9.30	1.13	0.40	0.41	8.05	0.94	0.37

*GMD* geometric mean diameter of aggregates, *MWD* mean weight diameter of aggregates, *SS* shear strength, *PR* penetration resistance,  $R^2$  coefficient of determination, *RMSE* root mean square error, *Bias* bias, *RF* random forest, *k*-*NN* *k*-nearest neighbor, *CB* Cubist

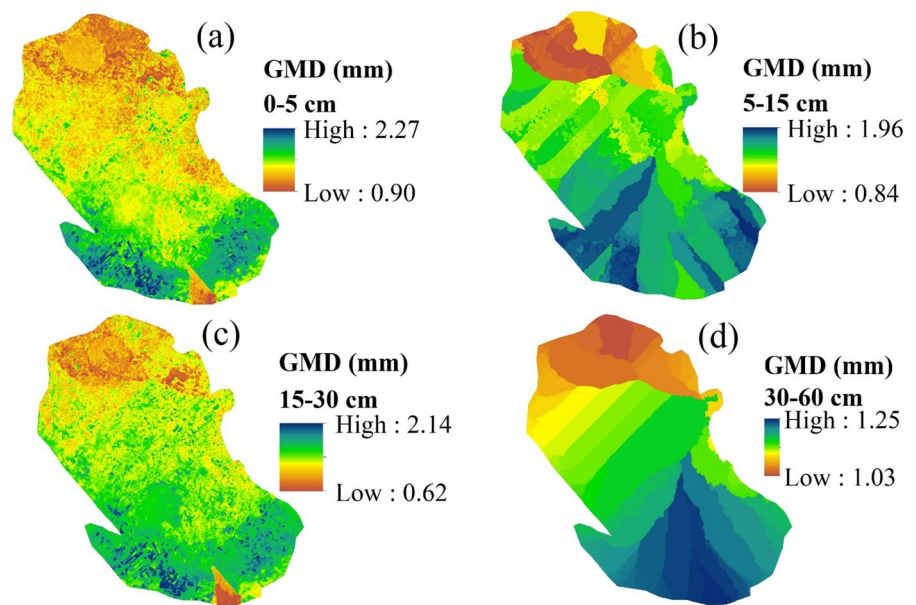
Out of soil variables among environmental covariates, only CNBL as a proxy of topographic attributes was important in the predicting MWD and PR at the depths of 15–30 cm and 30–60 cm and SS only at a depth of 30–60 cm, the soil properties by influencing on the soil climate and hydrology (Wang et al., 2018; Tu et al., 2018; Nsabimana et al., 2020). Forghani et al. (2020) confirmed that topographic features such as CNBL and valley depth are the most influential factors on physical parameters. Also, topographic

attributes, organic matter, and geology data were the most important parameters in the spatial prediction of SAS (Bouslihim et al., 2021). In contrast to soil variables and topographic attributes, the RS indices had a weak effect on SPM prediction.

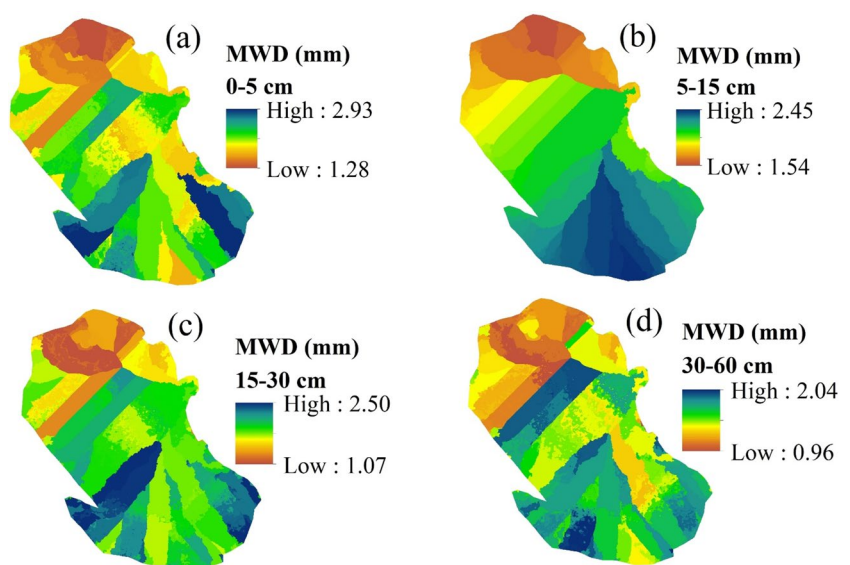
Spatial prediction

In this study, GMD, MWD, SS, and PR maps were prepared based on the best ML algorithm in all four

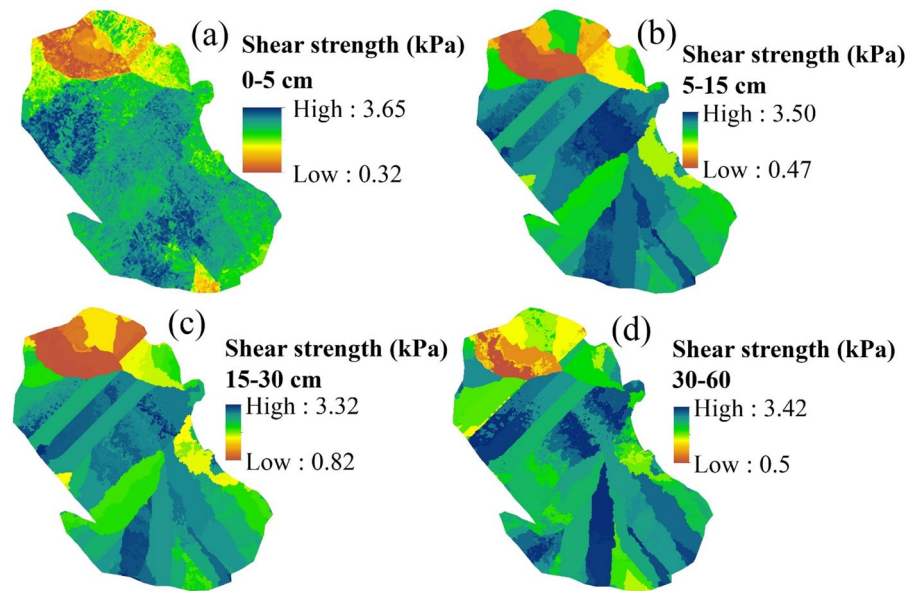
**Fig. 7** Spatial prediction maps of GMD for four standard depths in the Marvdasht area. Map prepared based on the best predictive model. **a** RF for depth of 0–5 cm, **b** k-NN for depth of 5–15 cm, **c** RF for depth of 15–30 cm, and **d** CB for depth of 30–60 cm



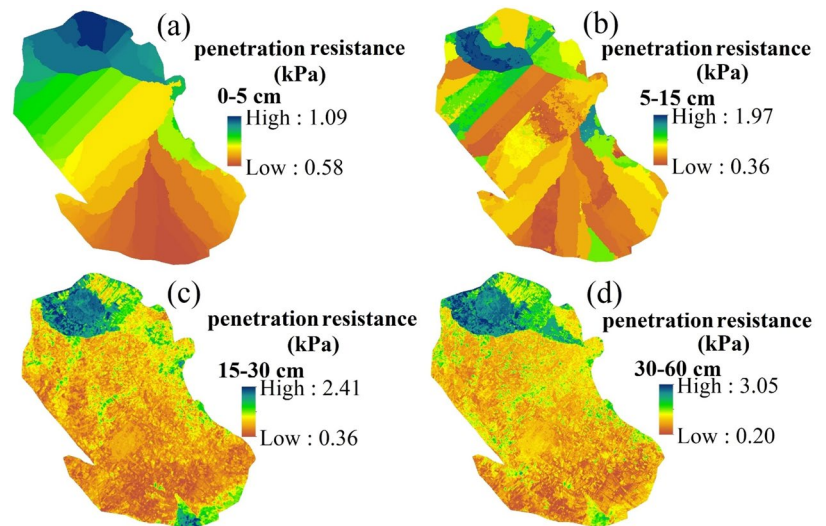
**Fig. 8** Spatial prediction maps of MWD for four standard depths in the Marvdasht area. Map prepared based on the best predictive model. **a** k-NN for depth of 0–5 cm, **b** CB for depth of 5–15 cm, **c** k-NN for depth of 15–30 cm, and **d** k-NN for depth of 30–60 cm



**Fig. 9** Spatial prediction maps of SS for four standard depths in the Marvdasht area. Map prepared based on the best predictive model ( $k$ -NN). **a** For depth of 0–5 cm, **b** for depth of 5–15 cm, **c** for depth of 15–30 cm, and **d** for depth of 30–60 cm



**Fig. 10** Spatial prediction maps of PR for four standard depths in the Marvdasht area. Map prepared based on the best predictive model. **a** CB for depth of 0–5 cm, **b**  $k$ -NN for depth of 5–15 cm, **c** map RF for depth of 15–30 cm, and **d**  $k$ -NN for depth of 30–60 cm



standard depths (Figs. 7, 8, 9, and 10). It was shown that high GMD values were observed in the Southwest, South, and Southeast section of the region. In these regions, the value of GMD decreased from 2.27 to 1.25 mm at depths of 0–5 to 30–60 cm, possibly due to the high SOM in the topsoil compared to the subsurface layers and improved soil structure resulting from agricultural activities (Fig. 7). For GMD, MWD, SS, and PR, a decreasing trend was observed from the surface to the subsoil, particularly in the northern zone of the region. Spatial prediction maps

showed that the higher GMD content were concentrated in the Southwest, South, and Southeast parts of regions. The trend of MWD was similar to GMD, and its value decreased from the surface to the deeper layers (Figs. 7 and 8). According to the result of Le Bissonnais (2016), the soils with MWD > 2.0 mm have very stable aggregate, so there is no surface crusting available. The minimum values of GMD and MWD were observed at the Northern boundaries (Fig. 7), where the pastures with low vegetation cover mainly increase erosion rates and thus caused

a weak structure in the top soil. These results can be related to the mountainous conditions (stone fragments). Furthermore, the maximum values of GMD and MWD were obtained in the southwest, south, and southeast, which could be due to the high SOM and clay and good soil structure (Fig. 6). Increasing SOM can help soil aggregate stability to improve, so adding SOM could justify the high GMD and MWD values in the cultivated land (Lacoste et al., 2014). Therefore, applying regular SOM is recommended. In addition, the comparison of Figs. 6, 7, and 8 showed that the SOM and clay contents are the most important covariates influencing the prediction of GMD and MWD. Among the rest of the important environmental covariates, CNBL and WB had the same trend as GMD and MWD in the study area. The CNBL as a proxy of topography has an important role in GMD and MWD, which is correlated with the SOC for soil conservation (Schillaci et al., 2017; Sabetizade et al., 2021). Spatial variability of the SPM including SS and PR shows the soil quality condition and provides useful information for making the appropriate decision for improving the soil fertility conditions. The highest values of SS were observed in the southern, central, and northwestern part of the area (Fig. 9). Unlike, the lowest values of SS and PR, properties were observed in the southern, central, and northwestern zone, whereas the highest values were observed in the northern and northeastern zone (Fig. 10). Also, the amount of SS from surface to depth showed a decreasing trend (3.65 to 3.42 kPa), while the PR showed an increasing trend (1.09 to 3.05 kPa). The most influential soil and environmental covariates in predicting SS and PR are derived from expert opinion and DEM (Table 3). The clay, silt, and SOM showed direct relationship with SS and indirect relationship with P unlike sand (Fig. 6). The WB and CNBL showed a negative relationship with SS, GMD, and MWD, while they, especially WB, showed a positive relationship with PR (Fig. 5). SIPI and MNDWI showed no significant trend with changes in GMD and MWD properties (Fig. 5). Khalil et al. (2011) used topographic attributes such as slope, slope direction, and elevation to predict SS and reported that the use of topographic attributes increases the accuracy of SS prediction maps. The changes in SS and PR as the result of land use and agricultural activities affect vegetation type, SOM, soil structure, and porosity. From the start to

the point of maximum shear, soil shear is related to the soil physical condition, especially soil compaction (Komandi, 1992), and as soil density increases, more force is required to break soil particles (Brevik et al., 2015). Lower SS in the northern parts is attributed to mountains and piedmont physiographic units, whereas in the low relief areas (alluvial plains), higher SS was achieved. Higher level areas have a weak soil structure, low SOM, high erosion rate, and low resistance to cutting. Based on field observations, severe erosion, the presence of stones and gravels, and surface soil runoff can lead to a weak soil structure at depth and a decrease of SS from surface to depth (Castro Filho et al., 2002). Based on Fig. 10, the PR values decreased from the north and northeast parts to the central, southwest, south, and southeast parts of the area. It is also observed that the central and the southern parts have the lowest PR values. The suitable vegetation, minimal tillage operations, appropriate land use, and high organic matter reduce the penetration resistance depending on the condition of the soil structure and the porosity of the topsoil. It has been reported that increasing SOM by using organic matter, vermicompost, and biological sludge creates a strong and stable soil structure; therefore, the crust formation on the soil surface and PR values reduces (Asghari et al., 2010). From the surface to the depth, an increasing trend for PR was observed indicating the lower soil quality and weak soil structure due to a low amount of SOM and a reduction of the soil formation process at the deeper layers. Finally, the prepared maps using the ML algorithms indicated that the variation of GMD, MWD, and SS decreased from surface to depth, while PR had an increasing trend from surface to deeper layers. In contrast to PR variation, GMD, MWD, and SS were increased in the southern and central parts of the study area compared to that of the northern parts. The variation trend of these properties indicates that the southern and central parts of the study area have a favorable soil structure compared to the northern parts, and the quality of soil structure decreased from surface to depth.

## Conclusions

In this study, the DSM maps of SPM were produced at four standard depths by RF, *k*-NN, and CB algorithms in a semi-arid region. Two covariates' scenarios consist of



environmental factors, i.e., RS and topography attributes (S1) and environmental factors plus soil properties (S2) were assessed. In S2 scenario, which accounted for both soil variables and environmental covariates, it was recognized as the better scenario for modeling SPM compared to S1. Based on the relative ranking of soil and environmental features, it was found that SOM and clay played a more important role in predicting SPM at all studied depths than that of the topographic and remote sensing attributes. The validation results revealed that the RF algorithm was the best comparison to other ML algorithms in predicting PR at the depth of 15–30 cm, while the *k*-NN algorithm had the highest prediction frequency. So, *k*-NN model has the high potential for mapping SPM in agricultural area and can help soil scientists for filling the gap of SPM mapping in these areas. Our findings display that the spatial and vertical variation of three soil properties (GMD, MWD, and SS) decreased from the surface to subsurface layer, except for PR.

For SPM spatial distribution, we conclude that including soil and environmental factors can lead to an increase in the accuracy of predicting soil properties. Globally, this research highlights the role of soil properties in DSM research. When soil variables are not measured, it is recommended to use freely available global soil databases, such as Soil Grid products, to account for the role of soil properties along with environmental covariates in the modeling process. The applied method is a promising approach for land use planner and farmers for better management of agricultural zones, especially in areas with highly intensive cultivation activity. Finally, for moving forward, future research could further refine the digital mapping of SPM by incorporating more detailed soil and environmental covariate data and expanding the study to other regions with diverse soil-forming factors. By continuing to advance our knowledge about the SPM spatial variability, we can better inform agricultural management practices and contribute to the sustainability of our planet's natural resources.

**Acknowledgements** We would like to thank Shiraz University for providing the needed facilities.

**Author contribution** PK: investigation, methodology, modeling, and writing—original draft. MB: investigation, methodology, writing—review and editing, and funding acquisition. AAM: methodology, original draft, and editing. MR: methodology and modeling.

**Funding** The research has been funded by Shiraz University.

**Data availability** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Declarations**

**Ethics approval** All authors have read, understood, and complied as applicable with the statement on the Ethical responsibilities of Authors.

**Competing interests** The authors declare no competing interests.

## References

- Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. *Open Journal of Statistics*, 5(07), 754.
- Asghari, S., Neyshabouri, M. R., Abbasi, F., Aliasgharzad, N., & Oustan, S. (2010). Effects of polyacrylamide, manure, vermicompost and biological sludge on aggregate stability, penetration resistance and available water capacity in a sandy loam soil. *Water and Soil Science*, 20(3), 15–29.
- Ayoubi, S., Karchegani, P. M., Mosaddeghi, M. R., & Honarjoo, N. (2012). Soil aggregation and organic carbon as affected by topography and land use change in western Iran. *Soil and Tillage Research*, 121, 18–26. <https://doi.org/10.1016/j.still.2012.01.011>
- Azizi, K., Ayoubi, S., Nabiollahi, K., Garosi, Y., & Gislum, R. (2022). Predicting heavy metal contents by applying machine learning approaches and environmental covariates in west of Iran. *Journal of Geochemical Exploration*, 233, 106921. <https://doi.org/10.1016/j.gexplo.2021.106921>
- Bannayan, M., & Hoogenboom, G. (2009). Using pattern recognition for estimating cultivar coefficients of a crop simulation algorithm. *Field Crops Research*, 111(3), 290–302.
- Bishop, T. F. A., McBratney, A. B., & Laslett, G. M. (1999). Modeling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, 91(1–2), 27–45.
- Bouslihim, Y., Rochdi, A., & Paaza, N. E. A. (2021). Machine learning approaches for the prediction of soil aggregate stability. *Heliyon*, 7(3), e06480.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Brevik, E. C., Cerdà, A., Mataix-Solera, J., Pereg, L., Quinton, J. N., Six, J., & Van Oost, K. (2015). The interdisciplinary nature of SOIL. *Soil*, 1(1), 117–129.
- Camera, C., Zomeni, Z., Noller, J. S., Zissimos, A. M., Christoforou, I. C., & Bruggeman, A. (2017). A high resolution map of soil types and physical properties for Cyprus: a digital soil mapping optimization. *Geoderma*, 285, 35–49.



- Castro Filho, C. D., Lourenço, A., Guimarães, M. D. F., & Fonseca, I. C. B. (2002). Aggregate stability under different soil management systems in a red latosol in the state of Parana, Brazil. *Soil and Tillage Research*, *65*(1), 45–51.
- Celik, I. (2005). Land-use effects on organic matter and physical properties of soil in a southern Mediterranean highland of Turkey. *Soil and Tillage Research*, *83*(2), 270–277. <https://doi.org/10.1016/j.still.2004.08.001>
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1–4.
- Elbasiouny, H., Abowaly, M., AbuAlkheir, A., & Gad, A. (2014). Spatial variation of soil carbon and nitrogen pools by using ordinary Kriging method in an area of north Nile Delta, Egypt. *Catena*, *113*, 70–78.
- Esfandiarpour-Boroujeni, I., Shahini-Shamsabadi, M., Shirani, H., Moseleh, Z., Bagheri-Bodaghabadi, M., & Salehi, M. H. (2020). Assessment of different digital soil mapping methods for prediction of soil classes in the Shahrekord plain, Central Iran. *Catena*, *193*, 104648.
- Forghani, S. J., Pahlavan-Rad, M. R., Esfandiari, M., & Torkashvand, A. M. (2020). Spatial prediction of WRB soil classes in an arid floodplain using multinomial logistic regression and random forest algorithms, south-east of Iran. *Arabian Journal of Geosciences*, *13*(13), 1–11.
- Gee, G. W., & Bauder, J. W. (1986). Particle size analysis, hydrometer methods. In A. Klute (Ed.), *Methods of soil analysis, Part 1, Physical and mineralogical methods* (pp. 383–411). American Society of Agronomy and Soil Science Society of America.
- Gorji, T., Tanik, A., & Sertel, E. (2015). Soil salinity prediction, monitoring and mapping using modern technologies. *Procedia Earth and Planetary Science*, *15*, 507–512.
- Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al. (2017). SoilGrids250m: global gridded soil information based on machine learning. *PLoS one*, *12*(2), e0169748.
- Hengl, T., Miller, M. A., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., et al. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, *11*(1), 1–18.
- Heydari, M., Zeynali, N., Bazgir, M., Omidipour, R., Kohzadian, M., Sagar, R., & Prevosto, B. (2020). Rapid recovery of the vegetation diversity and soil fertility after cropland abandonment in a semiarid oak ecosystem: an approach based on plant functional groups. *Ecological Engineering*, *155*, 105963.
- Holmes, G., Hall, M., & Prank, E. (1999). Generating rule sets from algorithm trees. In *Australasian joint conference on artificial intelligence* (pp. 1–12). Springer.
- Kazemi Garajeh, M., Blaschke, T., Hossein Haghi, V., Weng, Q., Valizadeh Kamran, K., & Li, Z. (2022). A comparison between Sentinel-2 and Landsat 8 OLI satellite images for soil salinity distribution mapping using a deep learning convolutional neural network. *Canadian Journal of Remote Sensing*, *48*(3), 452–468.
- Kemper, W. D., & Rosenau, R. C. (1986). Aggregate stability and size distribution. *Methods of soil analysis: Part 1 Physical and Mineralogical Methods*, *5*, 425–442.
- Keskin, D. B., Anandappa, A. J., Sun, J., Tirosh, I., Mathewson, N. D., Li, S., et al. (2019). Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature*, *565*(7738), 234–239.
- Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, *81*, 401–418.
- Khalil, M. B., Afyuni, M., Jalalian, A., Abbaspour, K. C., & Dehghani, A. A. (2011). Estimation surface soil shear strength by pedo-transfer functions and soil spatial prediction functions. *Water and Soil (Agricultural Sciences and Technology)*, 187–195. <https://doi.org/10.22067/JSW.V0I0.8520>
- Khosravani, P., Baghernejad, M., Moosavi, A. A., & Fallah-Shamsi, S. R. (2023). Digital mapping to extrapolate the selected soil fertility attributes in calcareous soils of a semiarid region in Iran. *Journal of Soils and Sediments*. <https://doi.org/10.1007/s11368-023-03548-1>
- Khosravi Aqdam, K., Asadzadeh, F., Momtaz, H. R., Miran, N., & Zare, E. (2022). Digital mapping of soil erodibility factor in northwestern Iran using machine learning algorithms. *Environmental Monitoring and Assessment*, *194*(5), 1–13.
- Komandi, G. (1992). On the mechanical properties of soil as they affect traction. *Journal of Terramechanics*, *29*(4-5), 373–380.
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*, 1–13.
- Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., & Walter, C. (2014). High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. *Geoderma*, *213*, 296–311.
- Le Bissonnais, Y. (2016). Aggregate stability and assessment of soil crustability and erodibility: I. Theory and methodology. *European Journal of Soil Science*, *67*(1), 11–21.
- Ma, Z., Shi, Z., Zhou, Y., Xu, J., Yu, W., & Yang, Y. (2017). A spatial data mining algorithm for downsampling TMPA 3B43 V7 data over the Qinghai-Tibet Plateau with the effects of systematic anomalies removed. *Remote Sensing of Environment*, *200*, 378–395.
- Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, *154*(1-2), 138–152.
- Mashalaba, L., Galleguillos, M., Seguel, O., & Poblete-Olivares, J. (2020). Predicting spatial variability of selected soil properties using digital soil mapping in a rainfed vineyard of central Chile. *Geoderma Regional*, *22*, e00289.
- McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1-2), 3–52.
- Minasny, B., & McBratney, A. B. (2006). Latin hypercube sampling as a tool for digital soil mapping. *Developments in Soil Science*, *31*, 153–606.
- Moosavi, A. A., & Sepaskhah, A. R. (2012). Spatial variability of physico-chemical properties and hydraulic characteristics of a gravelly calcareous soil. *Archives of Agronomy and Soil Science*, *58*(6), 631–656. <https://doi.org/10.1080/03650340.2010.533659>
- Moradi, F., Moosavi, A. A., & Khalili Moghaddam, B. (2016). Spatial variability of water retention parameters and saturated hydraulic conductivity in a calcareous Inceptisols

- (Khuzestan province of Iran) under sugarcane cropping. *Archives of Agronomy and Soil Science*, 62, 1686–1699.
- Mousavi, S., R., Sarmadian, F., Omid, M., & Bogart, P. (2022). The application of machine learning algorithms in the spatial estimation of soil phosphorus and potassium in a part of the lands of Dasht Abyek. *Soil Research*, 35(4), 397–411.
- Mousavi, S. R., Sarmadian, F., Angelini, M. E., Bogaert, P., & Omid, M. (2023). Cause-effect relationships using structural equation modeling for soil properties in arid and semi-arid regions. *Catena*, 232, 107392.
- Mozaffari, H., Moosavi, A. A., & Dematte, J. A. (2022). Estimating particle-size distribution from limited soil texture data: introducing two new methods. *Biosystems Engineering*, 216, 198–217.
- Mozaffari, H., Moosavi, A. A., & Sepaskhah, A. R. (2021). Land use-dependent variation of near-saturated and saturated hydraulic properties in calcareous soils. *Environmental Earth Sciences*, 80(23), 769.
- Mozaffari, H., Moosavi, A. A., Sepaskhah, A. R., & Cornelis, W. (2022). Long-term effects of land use type and management on sorptivity, macroscopic capillary length and water-conducting porosity of calcareous soils. *Arid Land Research and Management*, 36, 371–397.
- Mozaffari, H., Rezaei, M., & Ostovari, Y. (2021). Soil sensitivity to wind and water erosion as affected by land use in southern Iran. *Earth*, 2(2), 287–302.
- Mustafa, A., Minggang, X., Shah, S. A. A., Abrar, M. M., Nan, S., Baoren, W., et al. (2020). Soil aggregation and soil aggregate stability regulate organic carbon and nitrogen storage in a red soil of southern China. *Journal of Environmental Management*, 270, 110894.
- Nelson, D. W., & Sommers, L. E. (1996). *Method of soil analysis. Part 3. In: Total carbon, organic carbon, and organic matter* (3rd ed., pp. 961–1010). Am. Soc. Agron. Soil Sci. Soc. Am.
- Nemes, A., Rawls, W. J., & Pachepsky, Y. A. (2006). Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. *Soil Science Society of America Journal*, 70(2), 327–336.
- Neyestani, M., Sarmadian, F., Jafari, A., Keshavarzi, A., & Sharififar, A. (2021). Digital mapping of soil classes using spatial extrapolation with imbalanced data. *Geoderma Regional*, 26, e00422.
- Nsabimana, G., Bao, Y., He, X., Nambajimana, J. D. D., Wang, M., Yang, L., et al. (2020). Impacts of water level fluctuations on soil aggregate stability in the Three Gorges Reservoir, China. *Sustainability*, 12(21), 9107.
- Parsaie, F., Farrokhan Firouzi, A., Mousavi, S. R., Rahmani, A., Sedri, M. H., & Homae, M. (2021). Large-scale digital mapping of topsoil total nitrogen using machine learning algorithms and associated uncertainty map. *Environmental Monitoring and Assessment*, 193(4), 1–15.
- Rahmani, A., Sarmadian, F., & Arefi, H. (2022). Digital mapping of top-soil thickness and associated uncertainty using machine learning approach in some part of arid and semi-arid lands of Qazvin Plain. *Iranian Journal of Soil and Water Research*, 53(3), 585–602.
- Rezaee, L., Moosavi, A. A., Davatgar, N., & Sepaskhah, A. R. (2020a). Soil quality indices of paddy soils in Guilan province of northern Iran: spatial variability and their influential parameters. *Ecological Indicators*, 117, 106566.
- Rezaee, L., Moosavi, A. A., Davatgar, N., & Sepaskhah, A. R. (2020b). Shrinkage-swelling characteristics and plasticity indices of paddy soils: spatial variability and their influential parameters. *Archives of Agronomy and Soil Science*, 66, 2005–2025.
- Rossel, R. A., & McBratney, A. B. (2008). Diffuse reflectance spectroscopy as a tool for digital soil mapping. In *Digital soil mapping with limited data* (pp. 165–172). Springer.
- Sabetzade, M., Gorji, M., Roudier, P., Zolfaghari, A. A., & Keshavarzi, A. (2021). Combination of MIR spectroscopy and environmental covariates to predict soil organic carbon in a semi-arid region. *Catena*, 196, 104844.
- Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappie, M., Märker, M., & Saia, S. (2017). Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: the role of land use, soil texture, topographic indices and the influence of remote sensing data to algorithming. *Science of the Total Environment*, 601, 821–832.
- Shahabi, M., Jafarzadeh, A. A., Neyshabouri, M. R., Ghorbani, M. A., & Valizadeh Kamran, K. (2017). Spatial modeling of soil salinity using multiple linear regression, ordinary kriging and artificial neural network methods. *Archives of Agronomy and Soil Science*, 63(2), 151–160.
- Soane, B. D. (1990). The role of organic matter in soil compactibility: a review of some practical aspects. *Soil and Tillage research*, 16(1-2), 179–201.
- Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., & Malone, B. P. (2014). Digital mapping of soil salinity in Ardakan region, central Iran. *Geoderma*, 213, 15–28.
- Tang, F. K., Cui, M., Lu, Q., Liu, Y. G., Guo, H. Y., & Zhou, J. X. (2016). Effects of vegetation restoration on the aggregate stability and distribution of aggregate-associated organic carbon in a typical karst gorge region. *Solid Earth*, 7(1), 141–151.
- Tu, C., He, T., Lu, X., Luo, Y., & Smith, P. (2018). Extent to which pH and topographic factors control soil organic carbon level in dry farming cropland soils of the mountainous region of Southwest China. *Catena*, 163, 204–209.
- Ugbaje, S. U., & Reuter, H. I. (2013). Functional digital soil mapping for the prediction of available water capacity in Nigeria using legacy data. *Vadose Zone Journal*, 12(4), 1–13. <https://doi.org/10.2136/vzj2013.07.0140>
- Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li Liu, D., et al. (2018). Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. *Ecological Indicators*, 88, 425–438.
- Wang, H., Zhang, G. H., Li, N. N., Zhang, B. J., & Yang, H. Y. (2019). Variation in soil erodibility under five typical land uses in a small watershed on the Loess Plateau, China. *Catena*, 174, 24–35.
- Wang, S., Jin, X., Adhikari, K., Li, W., Yu, M., Bian, Z., & Wang, Q. (2018). Mapping total soil nitrogen from a site in north-eastern China. *Catena*, 166, 134–146.
- Wilding, L. P. (1985). Spatial variability: its documentation, accommodation, and implication to soil surveys. In *Soil*

- spatial variability, Las Vegas NV, 30 November-1 December 1984* (pp. 166-194).
- Wilson, J. (2018). *Environmental applications of digital terrain modeling* (p. 359). John Wiley & Sons.
- Xiao, J., Chevallier, F., Gomez, C., Guanter, L., Hicke, J. A., Huete, A. R., et al. (2019). Remote sensing of the terrestrial carbon cycle: a review of advances over 50 years. *Remote Sensing of Environment*, 233, 111383.
- Yamaç, S. S., Şeker, C., & Negiş, H. (2020). Evaluation of machine learning methods to predict soil moisture constants with different combinations of soil input data for calcareous soils in a semi-arid area. *Agricultural Water Management*, 234, 106121.
- Zahedi, S., Shahedi, K., Rawshan, M. H., Solimani, K., & Dadkhah, K. (2017). Soil depth modeling using terrain analysis and satellite imagery: the case study of Qeshlaq mountainous watershed (Kurdistan, Iran). *Journal of Agricultural Engineering*, 48(3), 167–174.
- Zahedifar, M. (2023a). Assessing alteration of soil quality, degradation, and resistance indices under different land uses through network and factor analysis. *Catena*, 222, 106807.
- Zahedifar, M. (2023b). Feasibility of fuzzy analytical hierarchy process (FAHP) and fuzzy TOPSIS methods to assess the most sensitive soil attributes against land use change. *Environmental Earth Sciences*, 82, 1–17.
- Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., & Finke, P. (2019). Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. *Geoderma*, 338, 445–452.
- Zeraatpisheh, M., Ayoubi, S., Mirbagheri, Z., Mosaddeghi, M. R., & Xu, M. (2021). Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. *Geoderma Regional*, 27, e00440.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.