Catena 232 (2023) 107392

Contents lists available at ScienceDirect

Catena

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

Cause-effect relationships using structural equation modeling for soil properties in arid and semi-arid regions

Seyed Roohollah Mousavi [a], Fereydoon Sarmadian [a,*], Marcos Esteban Angelini [b,c], Patrick Bogaert [d], Mahmoud Omid [e]

[a] Soil Science and Engineering Department, Faculty of Agricultural, College of Agriculture & Natural Resources, University of Tehran, Karaj, Iran
[b] Global Soil Partnership (GSP), Food and Agriculture Organization of the United Nations (FAO), Rome, Italy
[c] Universidad Nacional de Luján – Luján, Buenos Aires, Argentina
[d] Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[e] Agricultural Machinery Engineering Department, Faculty of Agricultural, University of Tehran, Karaj, Iran

ARTICLE INFO

Keywords:
Multivariate analysis
Modification method
Soil system
Environmental covariates
Conceptual model

ABSTRACT

Predicting soil properties and evaluating their functions along with their related driving factors is useful for providing useful geographical information for soil management, which is especially important in arid and semi-arid regions. This study investigates the use of a structural equation modeling (SEM) approach for assessing the effects of soil forming factors through environ-mental proxies on three key soil properties, namely soil organic carbon (SOC), calcium carbonate equivalent (CCE) and clay content (clay) in an arid and semi-arid region of Iran. Using a set of 259 soil profiles collected over years 2016–2020 in the Qazvin plain, the cause-effect relationships were estimated between these soil properties and nine environmental factors derived from a digital elevation model and from satellite images. Focusing on two main horizons A and B, it was shown that normalized difference vegetation index, midslope position, elevation, multi-resolution valley bottom flatness, and saga wetness index are impacting these soil properties. Inside each horizon, the effect of CCE and clay on SOC was also evidenced, but to an extent that depends on the horizon. For each soil property, we were able to clearly identify the relationships between the two horizons. Although our SEM approach proved to be useful for identifying and estimating the cause-effect relationships, it failed to provide a good predictive model as required for a relevant digital soil mapping of these soil properties. However, as the SEM approach allows combining soil science knowledge inside a model that accounts for soil forming factors, external factors, and soil system at the same time, it permits an investigation of their potential cause-effect relationships in a rich theoretical framework. The SEM methodology is thus potentially useful for soil scientists that are studying various soil properties in other parts of the world, even if it cannot be advocated as an efficient digital soil mapping method in general.

1. Introduction

Soils are an essential component of the natural environment and they play a critical role in the ecosystem health (Adhikari and Hartemink, 2016). Interactions between soil physical and chemical properties can be used to determine soil quality and fertility properties. Predicting soil qualities and evaluating their functions can offer precise geographical information about soil management, which is especially important in arid and semi-arid regions (Jeong et al., 2017). The consideration of soil information is crucial for conservation and sustainable management practices, as it is essential for the development of sustainable agricultural strategies and monitoring the impacts of improper resource utilization (Carvalho Junior et al., 2016; Costa et al., 2020).

Over the last few decades, digital soil mapping (DSM) has made significant advances and has several key advantages over traditional soil mapping methods (Malone et al., 2009) due to the use of innovative techniques, including predictive algorithms (Zeraatpisheh et al., 2022a; Padarian et al., 2019), models combining machine learning (ML) and geostatistics (Taghizadeh-Mehrjardi et al., 2020; Poggio and Gimona, 2017), along with hybridization techniques making use of soil properties maps (Matinfar et al., 2021; Rostaminia et al., 2021). Traditional soil mapping methods are often reliant on field sampling, which can be time-

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

https://doi.org/10.1016/j.catena.2023.107392
Received 13 April 2023; Received in revised form 4 July 2023; Accepted 12 July 2023
Available online 21 July 2023

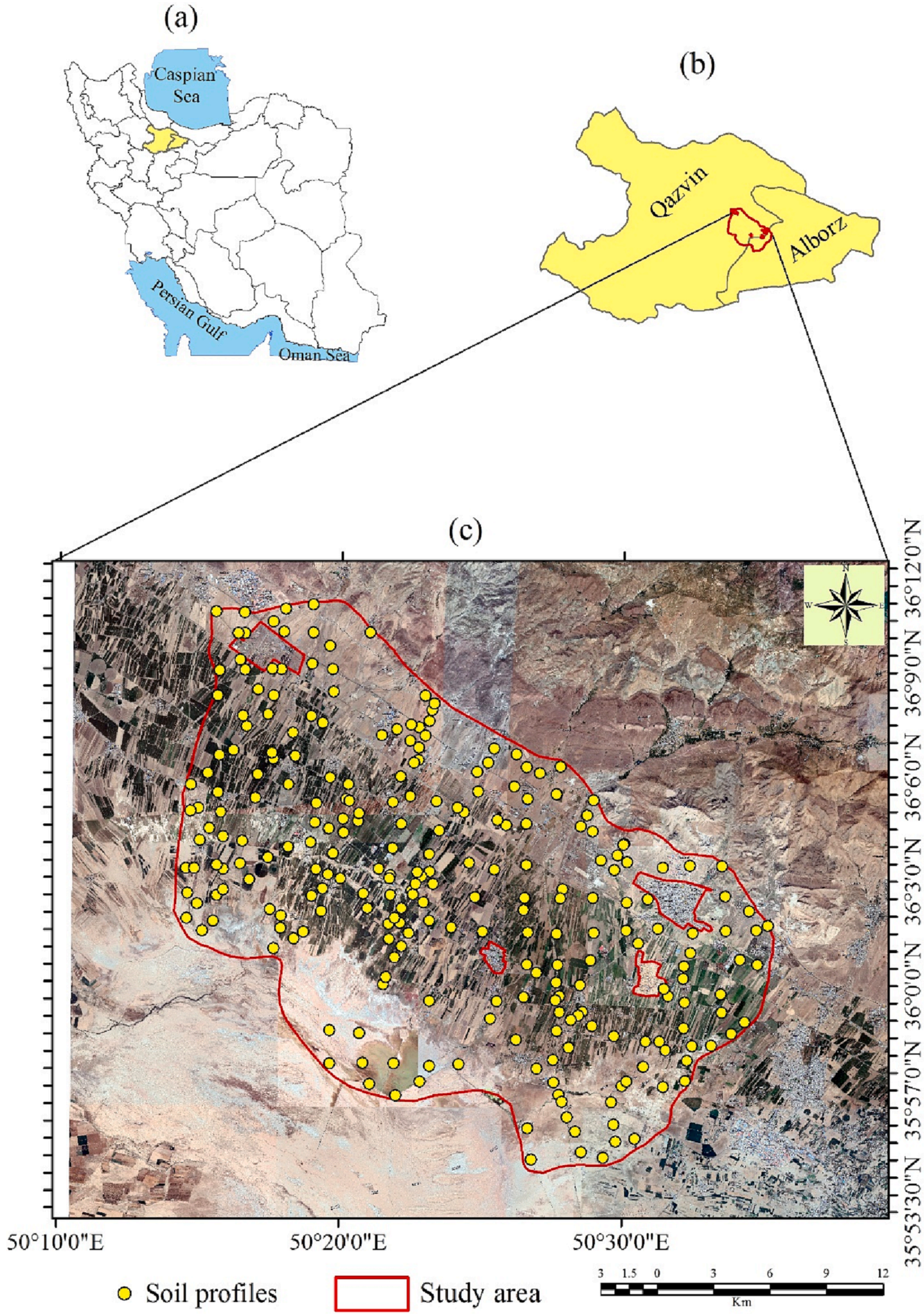

Fig. 1. Study area and locations of the soil profiles; (a): map of Iran; (b) Qazvin and Alborz regions; (c) boundaries of the study area and locations of the soil profiles.

consuming, expensive, and limited in scope, requiring the expertise of individual soil scientists (Minasny and McBratney, 2016). In contrast, DSM can provide more up-to-date information on soil properties using high-resolution topographic attributes and remote or proximal sensing data. This has led to improved performance in soil data modeling and spatial prediction (Padarian et al., 2019).

Although machine learning (ML) algorithms have proved to be useful for predicting soil properties and can lead to higher predictive accuracy by comparison with more conventional approaches, they are also considered as black box models (Wadoux et al., 2020), thus limiting the insight that soil scientists gain from the results of the modeling process. Miller (2019) defined interpretability as the degree to which a human can understand the cause of a decision, and the necessity for ML algorithms to be interpretable is considered an important requirement by Doshi-Velez and Kim (2017). As an example, when applied to the prediction of lateral and vertical variations of soil properties (Zeraatpisheh et al., 2020), the prediction results might be physically unrealistic, and causal relationships between soil properties cannot be accounted for (Angelini et al., 2020). In addition, interpretability of a model refers to the ability to understand how the model is making predictions and the underlying relations between system variables. The interpretability of a model is important for soil scientists to gain insights into how the model is making predictions and which factors are most important in determining soil properties (Molnar, 2020; Wadoux et al., 2021). This can be achieved through various methods. In new machine learning techniques, feature importance analysis provides some insights about variable relationships, but the model randomness makes interpretation of underlying processes cumbersome (Breiman, 2001). Cause-and-effect relationships, instead, are more easily analyzed under linear regression models. Structural equation modeling (SEM) is a statistical

Table 1
Collected environmental covariates over the study area.

Covariate	Abbreviation	Source	Spatial resolution
Digital elevation model	DEM	ALOS PALSAR radar	12.5 m
Midslope position	Midslope	Digital elevation model	12.5 m
SAGA wetness index	SWI	Digital elevation model	12.5 m
Multiresolution index of valley bottom flatness	MrVBF	Digital elevation model	12.5 m
Wind Effect	WE	Digital elevation model	12.5 m
Diffuse insulation	Diffuse	Digital elevation model	12.5 m
Mean annual precipitation	MAP	Mean monthly climate data	250 m
Mean annual temperature	MAT	Mean monthly climate data	250 m
Normalized difference vegetation index	NDVI	Median of 5 years NDVI	30 m

technique based on multivariate linear regression. While this cannot prove cause-effect relations, it can be used to test conceptual models based on cause-effect relationships, which can be useful in understanding the underlying mechanisms that determine soil properties (Bollen, 1989; Pearl, 1995).

In order to address some of these issues, researchers have recently proposed using structural equation modeling (SEM) approaches. Typically, SEM involves the construction of a model that aims at representing how various properties are causally and structurally related to each other. It generally relies on a multivariate regression modeling approach that allows the user to simultaneously account for a network of potential relationships and that permits testing their relevance using a statistical framework (Hoyle, 2012). Scientists can thus design and evaluate the relevance of their hypotheses using a general scheme that accounts for the relationships between various variables in this model (Dion, 2008).

In a wide variety of fields that include psychology (Martens, 2005), health sciences (Christ et al., 2014), agricultural sciences (Sarkar et al., 2021) and natural resources (Tuominen et al., 2022), the SEM approach has proved to be a convenient way to test the complex relationships between natural variables by overcoming the limitations of more traditional analysis methods (Bollen, 1989; Grace et al., 2012). SEM can be viewed as a semi-mechanistic modeling approach (Liu et al., 2021). For soil sciences, Angelini et al. (2016) showed that how relationships between soil properties can be included in the SEM modeling process in a prediction framework, while Angelini et al. (2017) applied SEM for multi-layers and multivariate soil mapping and tested the functionality of SEM for the improvement of soil properties modeling. Liu et al. (2021) reported that SEM yields reasonable results and is an efficient method to simulate soil development using ancillary data. Sales et al. (2017) used SEM to investigate the relationships between the phosphorus cycle and soil properties. They reported that their conceptual model was consistent with the hypotheses about the relationships between phosphorus, organic matter storage, and physicochemical properties of the soil. At the watershed scale, Zhu et al. (2016) explored the relative influence of soil chemistry and topography attributes on the availability of micronutrients (Fe, Mn, Cu, Zn, and B) using SEM. They concluded that soil chemistry and topographic attributes affected the availability of these micronutrients both in direct and indirect ways.

By relying on these previous results and findings, this paper aims at quantifying the relationships between soil organic carbon (SOC) and soil properties such as calcium carbonate equivalent (CCE) and clay content (denoted clay hereafter), along with environmental covariates. Improving our knowledge about the spatial distribution of SOC content and its relationships with other soil properties and environmental covariates is essential for efficient soil management and environment protection (Guo et al., 2013). In a previous study by Mousavi et al. (2022), ML algorithms and environmental covariates were used to identify the most influential factors impacting SOC at four separate standardized depths in the Qazvin plain (Iran), without identifying the causal effects and the direct or indirect relationships between SOC and these covariates. Although partial dependent plots and relative importance functions where used to evidence the relationships between SOC con-tent and environmental covariates, there is still a knowledge gap about the way the soil system and these covariates are interacting.

The specific aim of this study is to use structural equation modeling (SEM) to predict three key soil properties (SOC, CCE, and clay) in the A and B horizons of an arid and semi-arid region of Iran using nine environmental covariates. The secondary objectives were to identify the main controlling factors for each soil property and horizon, investigate the cause-effect relationships between soil properties and environmental covariates, and evaluate the efficiency of SEM for predicting and mapping these soil properties over the study area. The ultimate goal was to determine the potential usefulness of SEM methodology for soil scientists studying soil properties in other parts of the world, despite its limitations in terms of efficiency for digital soil mapping.

2. Material and methods

2.1. Study area

The study area covers 57, 540 ha and is located in the Qazvin plain, Iran, between 50°15′–50°29′ E and 35°54′–36°54′ N (Fig. 1). It is one of the most strategic plains for agricultural production and it has been under continuous cultivation and exploitation for a long time. Major land uses are irrigated and dry farming (68%), saline and non-saline pastures (30%), and urban and industrial zones (2%).

The three main soil types in this area are Typic Calcixerepts, Typic Haploxerepts, and Fluventic Haplxerepts (Soil Survey Staff, 2014). The dominant geological structures are basaltic volcanic rocks and alluvial or colluvial deposits in the northern and central parts, with silt and clay flat and mud, and salt flat in the southern part. The study area consists of four main topographic units, including Mountainous areas, Hilland, Piedmont, and Plains (Mousavi et al., 2022). The northern parts of the region consist of Mountains and Hilland, the central parts have the largest area of Piedmont, and the southern parts are limited to the appearance of Plains. Overall, the average slope variation in the region ranges from 1% to 12%. The soil parent material relates to Tertiary and Quaternary. Rainfall is mainly occurring in fall and winter, with a total mean annual precipitation of 280 mm over the year. Mean annual temperature ranges from 13.3 °C to 15.5 °C according to the Iran Meteorological Organization, and the soil moisture and temperature regimes are classified as Dry Xeric, weak Aridic, Aquic, and thermic, respectively.

2.2. Soil samples

The soil data were gathered from various surveys conducted from 2016 to 2020. A total of 278 soil profiles were drilled (Fig. 1c) based on stratified random method and samples were collected from all soil genetic horizons. After sampling, all samples were transferred to the laboratory, air-dried, and then measured based on standard methods. Soil organic carbon (SOC) percentage was measured using the wet oxidation method (Walkley and Black, 1934). Calcium carbonate equivalent (CCE) percentage was measured using titration method (Loeppert and Suarez, 1996). The clay percentage was measured using the hydrometric method (Gee and Bauder, 1986). These three key soil properties were selected as they are considered as having a large impact on the evolution of soils for arid and semi-arid regions (Zeraatpisheh et al., 2019). These soil properties were modeled in the two major soil horizons A and B after grouping the original soil horizons. For the A horizon, A and Ap or any subdivision of these were grouped, while Bw, Bk, By or any subdivision

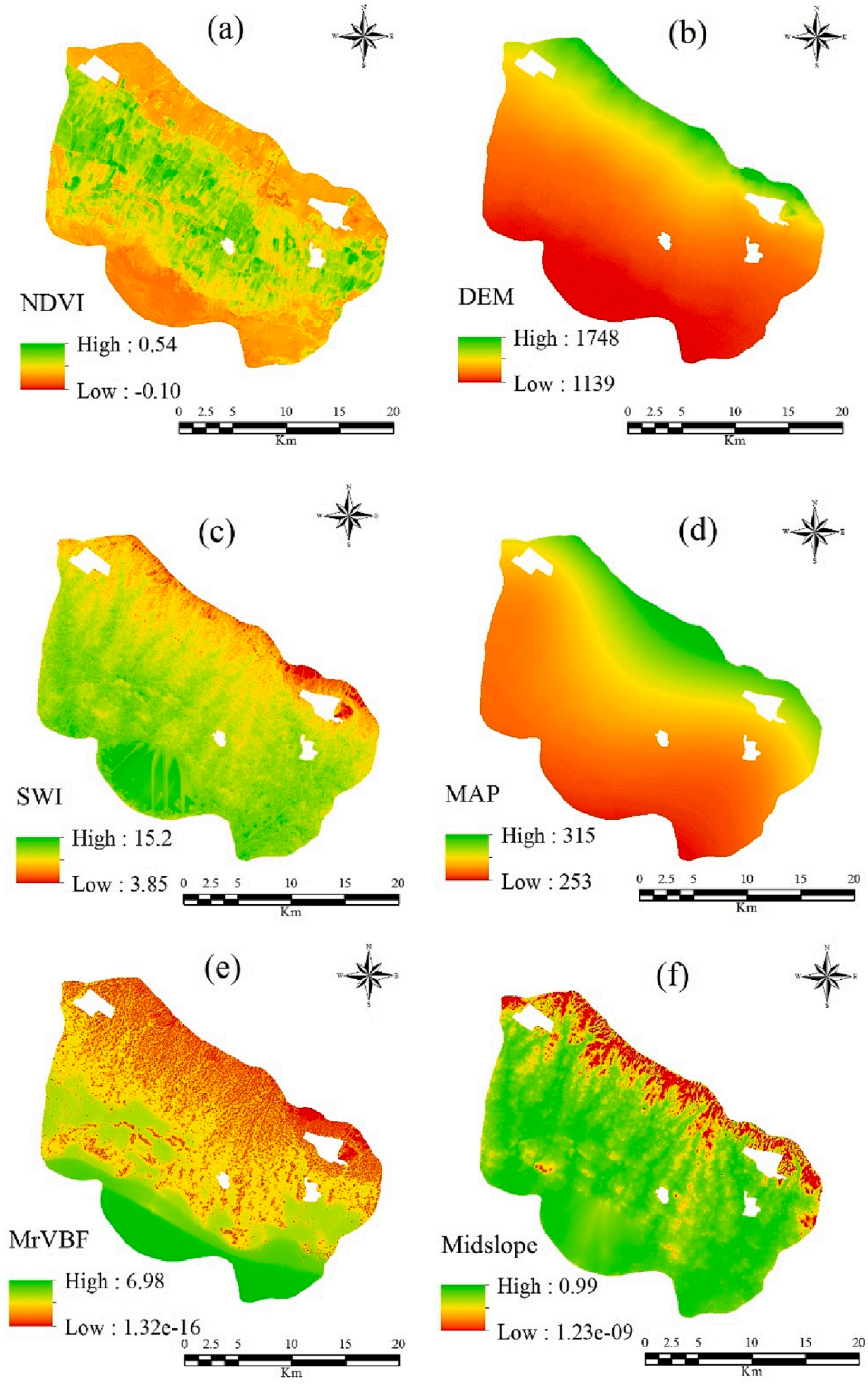

Fig. 2. Maps of six covariates (among the nine at hand) over the study area (NDVI: normalized difference vegetation index, DEM: digital elevation model, SWI: SAGA wetness index, MAP: mean annual precipitation, MrVBF: multiresolution index of valley bottom flatness, Midslope: midslope position).

of these were grouped for the B horizon. Most of the A horizon depths range from 0 to35 cm, while B horizon depths range from 35 to 150 cm. Soil profiles including transitional horizons such as AB or BC were not included. Soil profiles that contained a C horizon were excluded from the analysis, as only a few of them included this horizon. The final dataset used in this study thus consists of 259 soil profiles sampled both for A and B horizons, as defined above.

2.3. Environmental covariates

The environmental covariates were collected from different sources and are detailed in Table 1. Terrain attributes were obtained from a digital elevation model (DEM) at a 12.5 m spatial resolution downloaded from the Alaska Satellite Facility (ASF) Distributed Active Archive Center (DAAC) (https://search.asf.alaska.edu/).All-terrain attributes were obtained using the SAGA GIS (Conrad et al., 2015) terrain analysis toolbox after a pre-processing step based on a Fill Sink

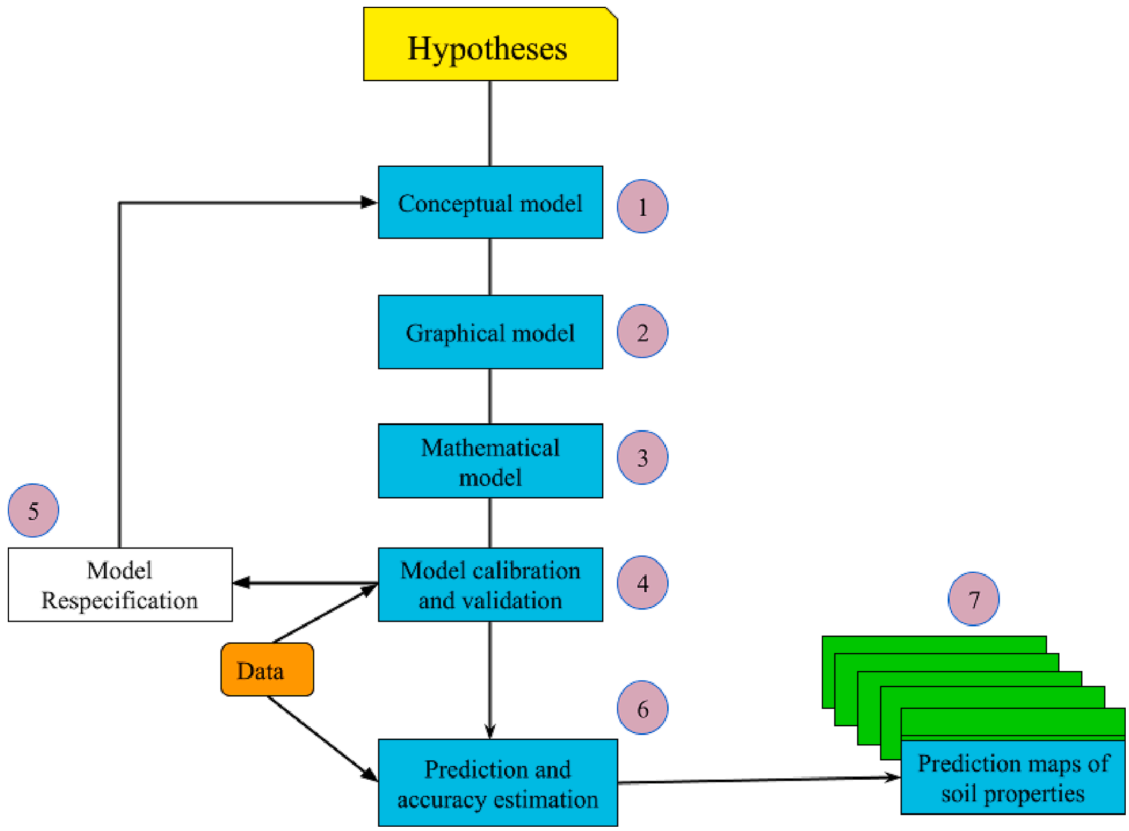

Fig. 3. Workflow of the SEM framework, where the successive steps are numbered from 1 to 7.

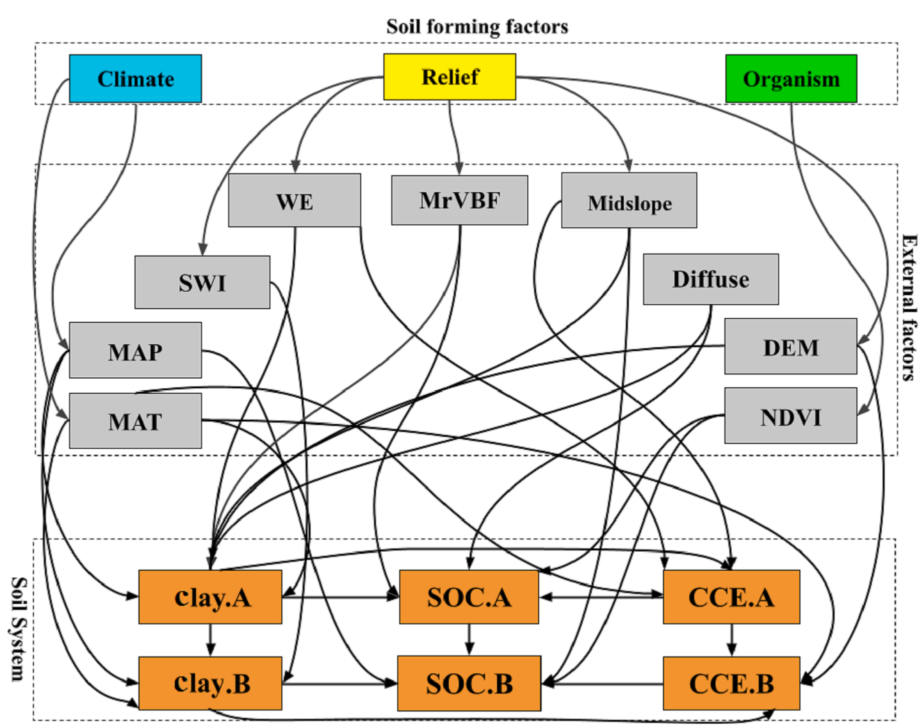

Fig. 4. Graphical model of the SEM using the various soil forming factors (i.e., Climate, Relief, and Organism), the external factors (i.e., the nine selected environmental covariates here), and the soil system (i.e., clay, SOC, and CCE in A and B horizons here).

algorithm (Wang and Liu, 2006) for reducing artifacts and noise. These attributes are midslope position (Midslope), SAGA wetness index (SWI), multi-resolution index of valley bottom flatness (MrVBF), wind effect (WE), and diffuse insulation (Diffuse). This extracted from the digital elevation model, which have high efficiency even in areas with low relief (Zeraatpisheh et al., 2019; Guo et al., 2021). In parallel, the

Table 2
Descriptive statistics for the soil profiles (n = 259) in A and B horizons.

Soil property	Min	Max	Mean	SD	CV (%)	Skewness	Kurtosis
SOC.A (%)	0.10	3.40	0.93	0.55	59.2	1.37	2.52
SOC.B (%)	0.01	1.40	0.46	0.25	55.2	1.60	4.24
CCE.A (%)	1.0	50.0	13.1	7.99	61.1	1.40	2.85
CCE.B (%)	2.0	59.0	16.5	9.52	57.7	1.13	1.75
clay.A (%)	1.0	70.7	27.0	13.1	48.3	0.67	0.14
clay.B (%)	5.0	65.0	29.1	13.6	46.7	0.48	−0.38

Min: minimum, Max: maximum, SD: standard deviation, CV: coefficient of variation, SOC: soil organic carbon; CCE: calcium carbonate equivalent, clay: clay content.

median of the normalized difference vegetation index (NDVI) was computed from monthly Landsat 8-time series (available via U.S. Geological Survey web portals at (https://earthexplorer.usgs.gov/) by combining periods of 16 days each, each period corresponding to the dates of the soil surveys that took place between 2016 and 2020. Mean annual precipitation (MAP) and mean annual temperature (MAP) raster maps were obtained from mean monthly climate data provided by the weather stations located in the study area. Finally, all raster maps were harmonized at a 15 m spatial resolution using a bilinear re-sampling method. Six of them are presented in Fig. 2 for the sake of illustration. For modeling purposes, all soil properties and environmental covariates were subsequently standardized by subtracting their mean and dividing by their standard deviation. Also, the selection of covariates was based on previous research conducted in the study area by Mousavi et al. (2022) and Neyestani et al. (2021), as well as expert opinion.

2.4. Structural equation model (SEM)

The general workflow for modeling and predicting SOC, CCE, and clay using the SEM approach is presented in Fig. 3. This process was carried out based on successive steps that are detailed hereafter. SEM is a statistical methodology for developing and testing hypotheses about relationships in a system. While in most cases it is used for testing hypotheses, it can also be used for prediction. SEM includes both graphical and equational forms, with the graphs representing the cause–effect relations and the equations specifying the relationships between the variables. SEM is based on conventional regression techniques, but it is more flexible in several ways. For example, SEM can explicitly incorporate measurement error, it can jointly analyze direct and indirect effects, and it can model multiple dependent variables simultaneously (Grace and Keeley, 2006). These advantages make SEM a powerful tool for analyzing complex systems. The SEM process typically begins with the development of general hypotheses where one represents the knowledge about a system. Then, it is adapted to a specific conceptual model to make the system relations more explicit. The conceptual model is then converted into a graphical model and a mathematical model,

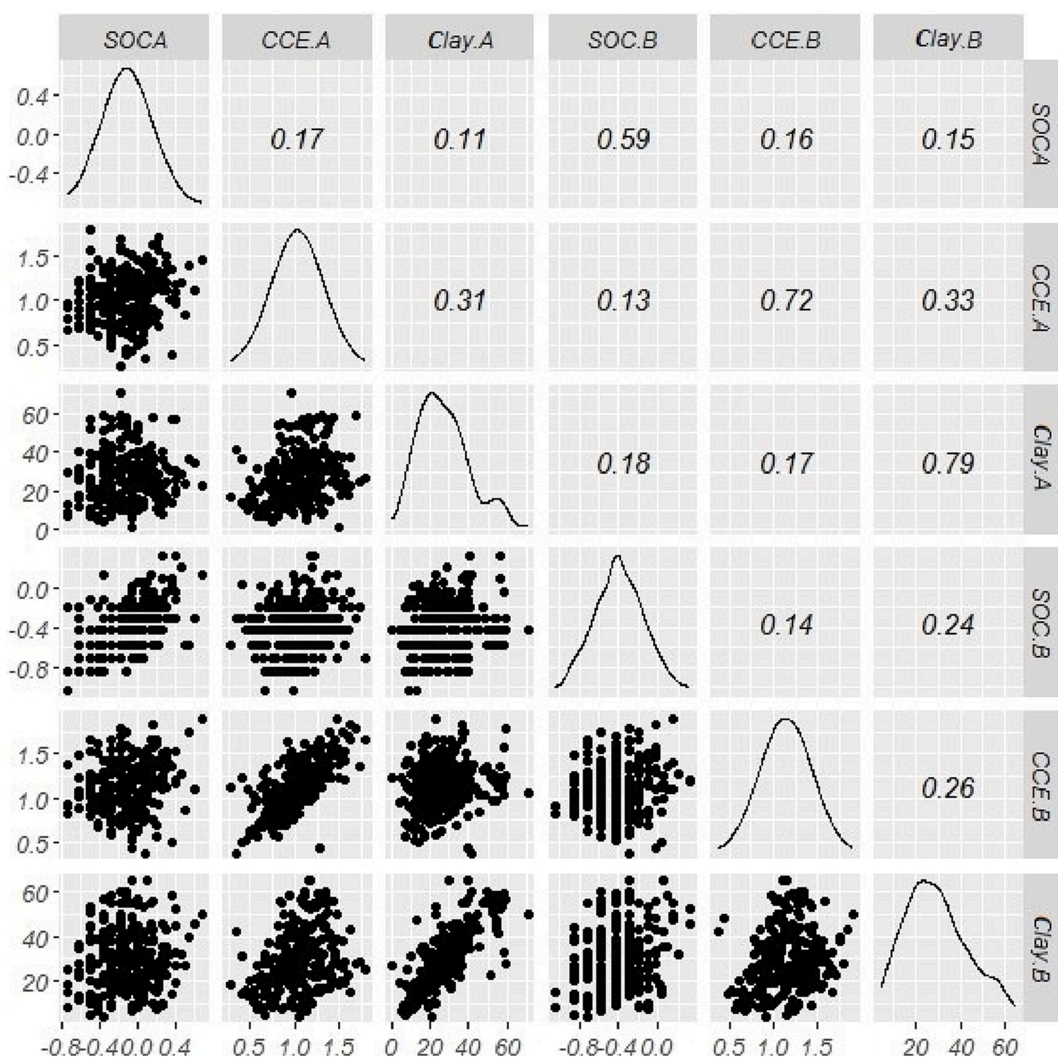

Fig. 5. Histograms, correlations ($p = 0.05$), and scatter plots for SOC, CCE, and clay in A and B horizons (on a logarithmic scale for SOC and CCE in both horizons).

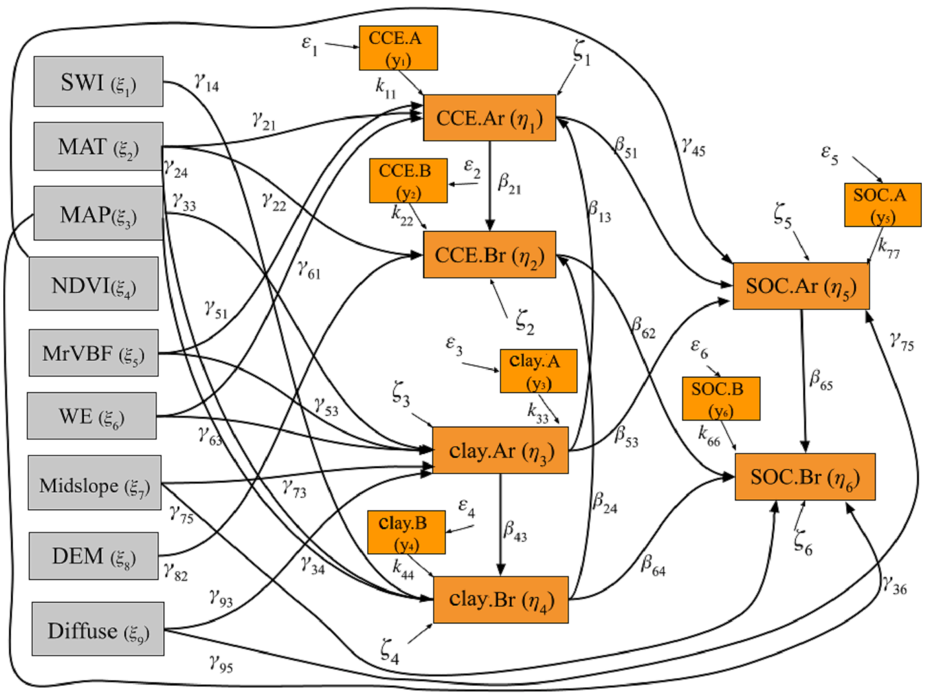

Fig. 6. SEM graphical model with its parameters, showing the relationships between environmental covariates (grey boxes) and soil properties (orange boxes). The A and B suffixes refer to the A and B horizons, respectively, while the absence or presence of the "r" suffix refers to the latent or observed value of the corresponding soil property, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Goodness-of-fit indices of the estimated SEM at the various respecification steps.

Step	χ^2	df	p − value	CFI	GFI	RMSEA	SRMR
First	272.5	38	0.000	0.702	0.791	0.101	0.080
Second	219.3	35	0.000	0.858	0.823	0.090	0.071
Third	150.6	32	0.002	0.902	0.875	0.060	0.051
Fourth	66.75	25	0.004	0.970	0.930	0.048	0.030

Table 4
Successive SEM respecifications based on the suggestions from the lavaan package. The operator refers to the kind of relationship that links the soil property to the environmental covariate (i.e., ~ means "regressed on" here).

Step	Soil property	Operator	Covariate
Second	CCE.Ar	~	MAP
	CCE.Ar	~	DEM
Third	clay.Br	~	Diffuse
	clay.Br	~	DEM
	clay.Ar	~	SWI
Fourth	SOC.Ar	~	Midslope
	CCE.Br	~	Midslope

where the expected relations between the variables are hypothesized providing details how they link each other. The mathematical model is then calibrated using a dataset (Grace et al., 2012). A model is considered appropriate when the model relationships represent accurately all significant relationships present in the data, which mean that the model fits the data well. As a result, SEM provide a graphical model where the arrows that connect variables represent the cause-effect direction and magnitude, and a system of equations that formalize the mathematical model. Although SEM can be used to test causal hypotheses, it is important to remember that it cannot to prove causality (Pearl, 1988,2009). More details about the SEM methodology applied to soil mapping can be found in Angelini et al. (2017). SEM modeling was done using the lavaan package in R software (R Core Team, 2022), where parameters are estimated using a maximum likelihood procedure.

2.4.1. Conceptual model

The first step of the SEM methodology is to propose a set of relationships that convey the major knowledge about the interactions between the selected soil properties (i.e., SOC, CCE and clay here) and environmental covariates (i.e., those presented in Table 1). This knowledge may come from soil science theory (Brady and Weil, 2008), main soil genetic models (Jenny, 1994), expert opinions, empirical evidence or knowledge gained from previous studies (Azizi et al., 2023; Neyestani et al., 2021) in similar areas subject to the same conditions (see Mousavi et al. (2021); Esmaeili et al. (2021) for the knowledge gained about the area used in this study). Some of these relationships may also correspond to specific research hypotheses that need to be assessed using the SEM. The final result is a conceptual model proposing a potential set or oriented links between the various components, that are justified on experimental and theoretical grounds or that are hypothesized from a research question. For our study, the relationships that were accounted for in the conceptual model are only briefly detailed here for SOC, CCE, and clay.

SOC is a key soil property that influences chemical, physical, and biological processes, particularly fertility and plant development (Shahriari et al., 2012). In arid and semi-arid regions, soil texture, CCE, and clay mineralogy that are all linked to the parent material have a major effect on SOC content in soils (Zeraatpisheh et al., 2022). CCE is related to the inorganic part of the soil carbon. It provides essential information about soil mineralogy and chemical properties, especially when pH and cation exchange capacity are considered (Grinand et al., 2012). The soils of the study area have carbonate-rich parent materials, especially in the central part. Their development is highly related to the accumulation of CCE with a secondary form. In arid and semi-arid regions, CCE may accumulate in soils through pedogenic processes or may be inherited from calcareous parent materials. Subsequently, due to a low organic carbon entry in this area, most carbon is in the mineral form

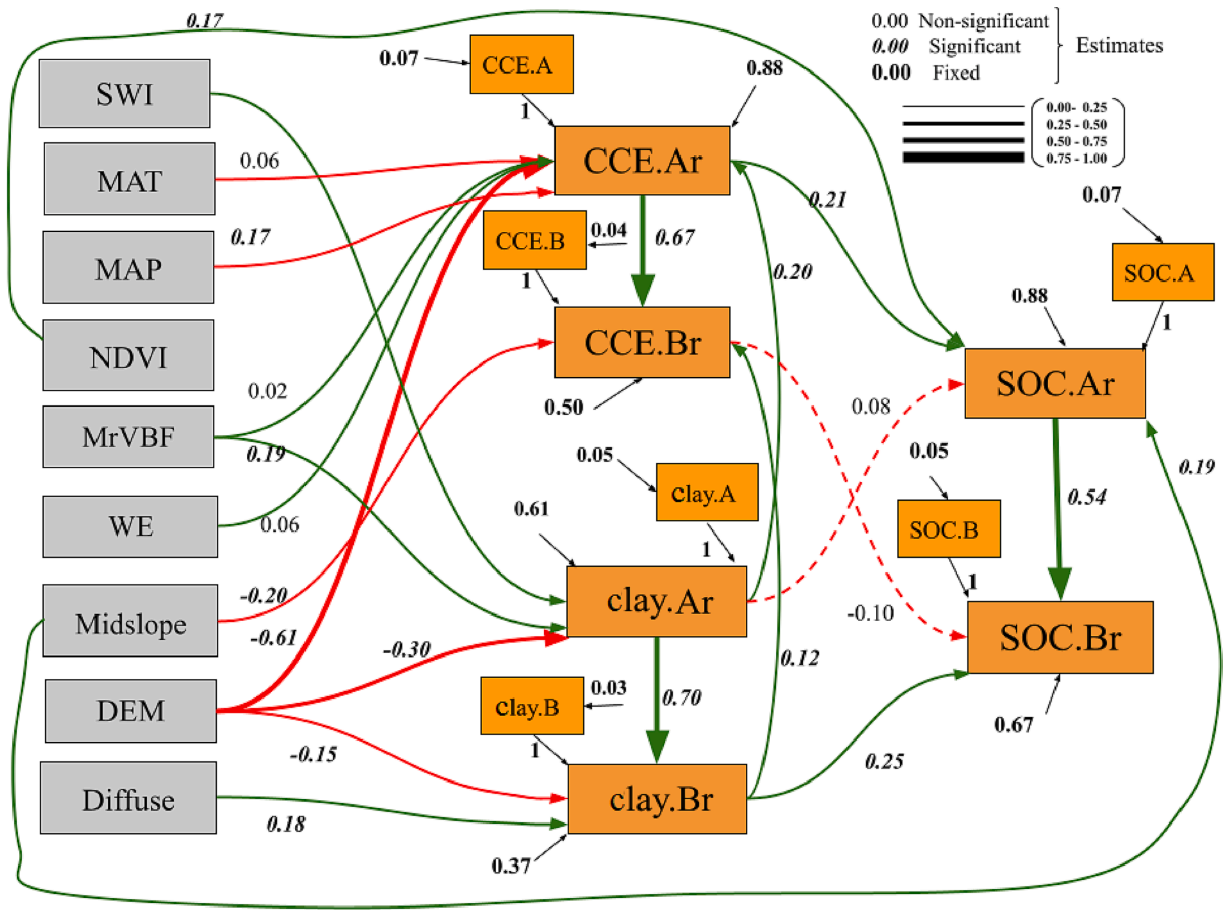

Fig. 7. Final SEM model after respecification, along with its estimated parameters. Arrow thickness is proportional to the magnitude of the coefficient while color corresponds to its sign (green for positive and red for negative). Dashed arrows represent model error correlations. Bold italic numbers refer to statistically significant estimates (*p-value* < *0.05*), bold non-italic numbers are fixed coefficients, while non-bold non-italic numbers are non-significant estimates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Accuracy and precision for the predicted soil properties using the final SEM.

Soil property	ME	RMSE	Mean SSPE	Median SSPE	R^2	CCC
SOC.A	−0.001	0.520	1.017	0.319	0.075	0.150
SOC.B	− 0.001	0.250	1.034	0.310	0.045	0.100
CCE.A	0.015	7.74	1.056	0.443	0.055	0.140
CCE.B	0.056	9.48	1.086	0.482	0.004	0.060
clay.A	0.058	10.34	1.035	0.331	0.370	0.600
clay.B	0.038	11.24	1.023	0.449	0.315	0.500

SOC: soil organic carbon, CCE: calcium carbonate equivalent, clay: clay content, ME: mean error, RMSE: root mean square error, SSPE: mean and median of standardized squared prediction error, R^2: coefficient of determination, CCC: concordance correlation coefficient.

(McGinnies, 1981). In such regions where particulate matter is transported by wind from source areas, the accumulation of secondary carbonates within the soil profile is a common occurrence. Finally, clay is an essential soil physical property too, as it regulates wilting point, field capacity, and saturated hydraulic conductivity in the soil, which in turn regulates water fluxes in the unsaturated zone (Shabou et al., 2015). It has an important effect on SOC storage, as clay particles lead to a higher accumulation of SOC than sand and silt particles (Jagadamma and Lal, 2010). Based on data from several short-term litter decomposition studies, Giardina et al. (2001) also reported a decline in SOC decomposition rates with increasing clay content. Accordingly, clay affects the capacity, magnitude, and rate of SOC storage (Feng et al., 2013).

2.4.2. Graphical model

In order to formally express the relationships between variables that are identified from the conceptual model, the graphical model depicts the many possible ways these variables are influencing each other. As seen from Fig. 4, this graphical model consists of three major boxes, namely the soil forming factors (i.e., climate, relief, and organism), the external factors (i.e., the various covariates that were collected) and the soil system (i.e., SOC, CCE, and clay in the A and B horizons). The graphical model identifies plausible causal relationships between soil properties and environmental covariates as well as the interactions between them. More precisely, the arrows on the graph are indicating the plausible cause-effect relationships between these variables (Angelini et al., 2017). The first set of relationships is linking the three soil-forming factors to the nine external factors. A second set is linking these external factors to the three soil properties. Finally, the soil properties in the A and B horizons are linked to each other, while each soil property is linked between the A and B horizons too.

2.4.3. Mathematical model

As it directly results from the graphical model, the mathematical model can be generated automatically and is summarized by three sets of linear equations (Bollen, 1989; Angelini et al., 2016), with

$$X = \Lambda\xi + \delta \quad (1)$$

$$Y = K\eta + \epsilon \quad (2)$$

$$\eta = B\eta + \Gamma\xi + \zeta \quad (3)$$

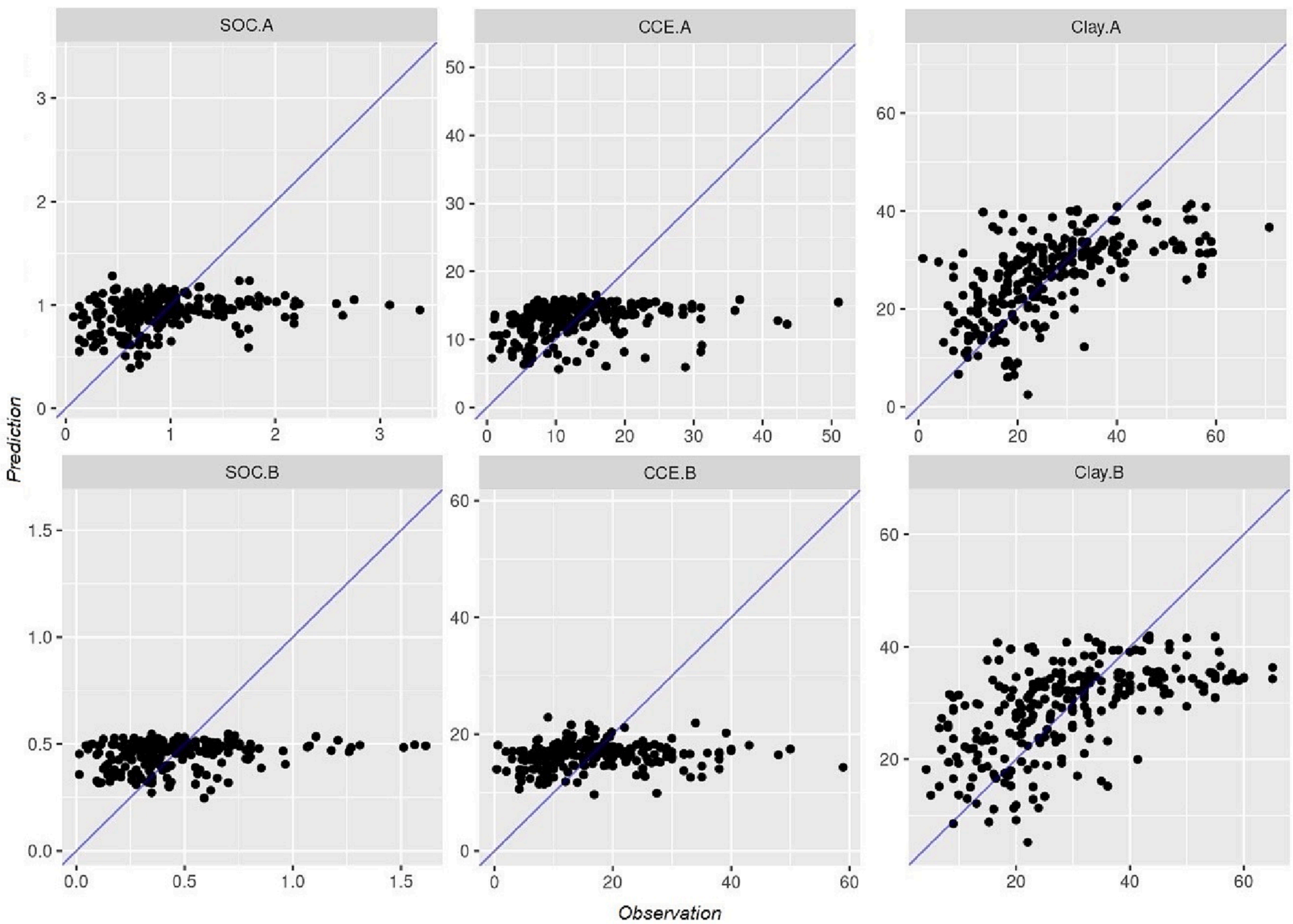

Fig. 8. Scatter plots of predicted (vertical axis) versus observed (horizontal axis) values of SOC, CCE, and clay at the profile locations in A and B horizons, where the blue line is the 1:1 line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Eqs. (1) and (2) define the so-called measurement model, where X is a column vector of q observed exogenous variables (i.e., the nine external factors), Y is a column vector of p observed endogenous variables (i.e., the three soil properties), ξ and η are column vectors referring to n latent exogenous and m latent endogenous variables, respectively, while Λ and K are q × n and p × m matrices of coefficients λij and kij, respectively, that linearly link the observed variables to the latent ones.

The column vectors δ and ∊ are vectors of measurement errors, that are mutually independent and zero-mean Gaussian distributed. Eq. (3) corresponds to the so- called structural model, where B and Γ are m × m and m × n matrices of coefficient βij and γij for the endogenous and exogenous relationships, respectively, while ζ is again a vector of measurement errors (Grace et al., 2012). It is worth noting that the diagonal elements βii in B are forced to be equal to zero so that soil properties cannot depend on themselves.

2.4.4. Parameters estimation and model validation

Parameters appearing in Eqs. (1) to (3) are generally estimated using a maximum likelihood procedure based on a comparison between the estimated covariance matrix of the data contained in X and Y with the covariance matrix derived from the measurement model definition (see Angelini et al. (2017) for the details). A close examination of the estimated parameters is required to check that their sign is consistent with the conceptual model and that their magnitude is in agreement with those that can be rationally accepted (see Bollen (1989)). The fit of the model with respect to the data can be assessed using various statistics, i.e. χ2 test, p-values, goodness-of-fit Index (GFI), comparative fit index (CFI), standardized root mean square of residuals (SRMR) and root mean square error of approximation (RMSEA) (Bayram, 2013). For CFI, values are in the [0, 1] range, and values above 0.9 indicate a good fit of the model (Schermelleh-Engel et al., 2003). Similarly, GFI values are in the [0, 1] range with values above 0.9 indicating a good fit. RMSEA is a useful statistic as its values are largely invariant with respect to the sample size, where values lower than 0.05 indicate a very good fit.

2.4.5. Model respecification

A poor fit of the model might result from a conceptual or a graphical model that does not account for all relevant relationships between variables. In these cases, a respecification of the conceptual and graphical models should be considered. The lavaan package includes a modification index function that is used to identify potential improvements in the model fit by testing different restrictions. It is calculated based on the Lagrangian multiplier test, which compares the fit of restrictive to less restrictive models. It estimates the potential improvement in the χ2 statistics if a specific parameter is included as a free parameter in the model. By examining the univariate modification index, one can identify which restriction should be removed to achieve a better model fit (Bollen, 1989, Chapter 7).

2.4.6. Prediction and accuracy estimation

Using the SEM equations as given by Eqs. (1) to (3), it is possible to predict the values of dependent variables based on the measured independent variables. These estimated values $\hat{\eta}$ are given by (see Angelini et al. (2016) for the details) Eq. 4:

$$\hat{\eta} = (I - B)^{-1}\Gamma\Lambda^{-1}x \tag{4}$$

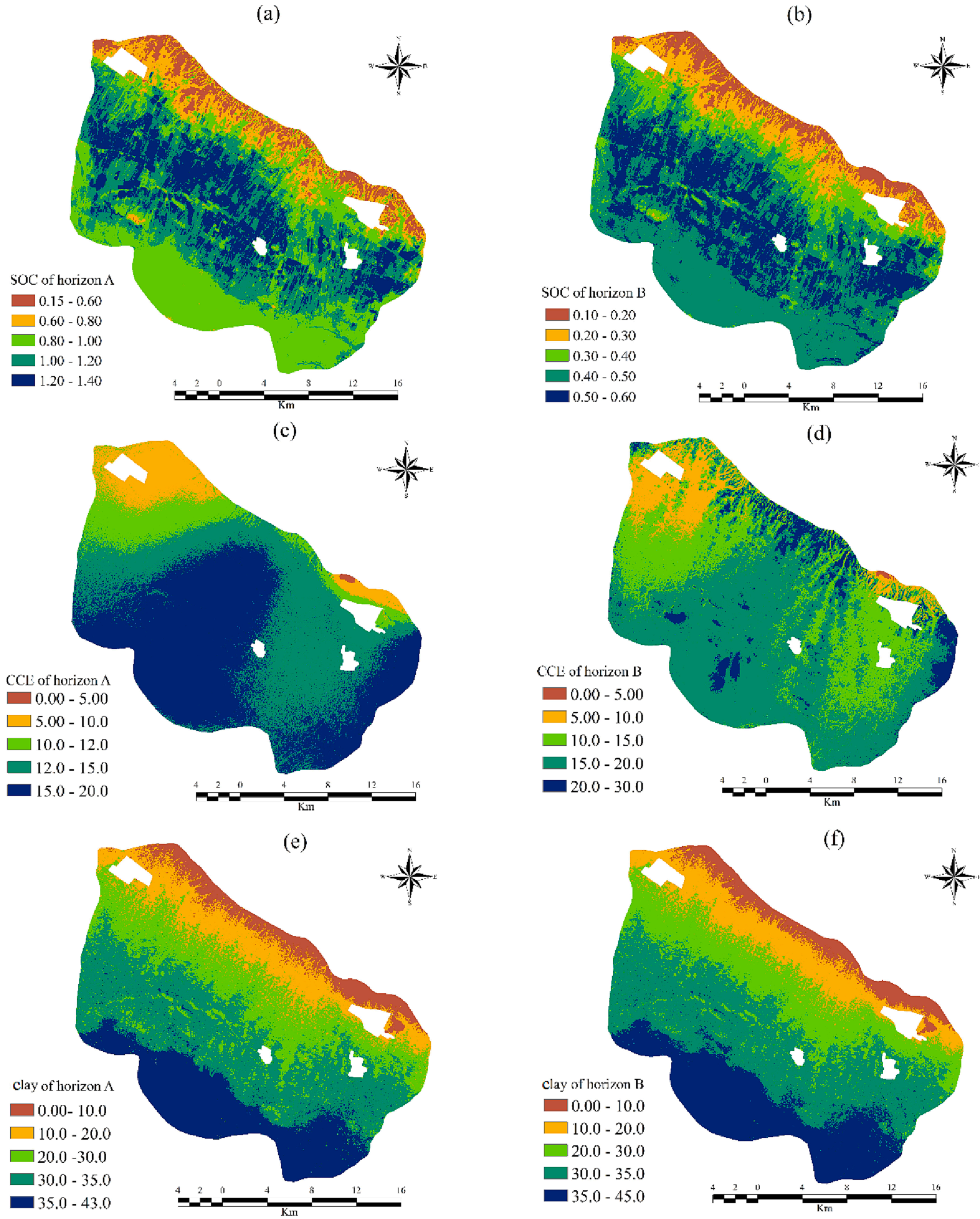

Fig. 9. Maps of the predicted soil properties using SEM, where (a) and (b) refer to SOC (in %), (c) and (d) refer to CCE (in %), (e) and (f) refer to clay (in %). Left and right columns refer to A and B horizons, respectively.

The accuracy of the predictions obtained using the SEM model was assessed using a leave-one-out-cross-validation procedure, where model parameters are re- estimated each time (Kline, 2016).

The quality of the predictions can be measured using standard statistical indices such as the coefficient of determination (R^2), concordance correlation coefficient (CCC), mean error (ME), root mean square error (RMSE), and, with Eqs. (5), (6), respectively.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i\text{-}y_j) \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i\text{-}y_j)^2} \tag{6}$$

where n is the number of observations, yi is the observed value, y_j is the predicted value based on the SEM model. The mean and median standardized squared prediction error (SSPE) were also calculated as they indicate the correct assessment of the model prediction uncertainty (Lark, 2000).

3. Results and discussion

3.1. Descriptive statistics

Descriptive statistics of the sampled soil properties in A and B horizons as estimated from the soil profiles are presented in Table 2. These results show that soils of the study area have a low SOC content (from 0.10% to 3.40%) in A horizon (SOC. A) and even less in B horizon (SOC. B). On average, CCE content is slightly higher in B horizon (CCE.B) than in A horizon (CCE.A), but it exhibits large variations over the study area. Similarly, clay largely varies, ranging from 1 to 70% in A horizon (clay. A), but with mean values that are quite similar in both horizons (clay. A, and clay. B). Differences in particle size distributions are one of the main reasons for soil variability in this area. As the northern part is subject to a higher amount of precipitation and has a diverse physiography, water flow is actively eroding the finest particles, so the dominant soil textures are Sandy loam and Loam. These particles are then transported to the central and southern parts of the area where there are deposited as the result of lower slopes, thus increasing the clay content of the soils and leading to Silty clay, Clay loam, and clay dominant soil textures.

All soil properties have coefficients of variation (CV%) that exceed 35% and are thus considered as high variables in space according to the classification by Wilding and Drees (1978). Except for clay in both horizons, SOC and CCE distributions are positively skewed and do not correspond to normal distributions, as previously observed (Mohammadi and Motaghian, 2011). Applying a logarithmic transformation to SOC and CCE values solved this issue, and Fig. 5 shows that both CCE and clay are clearly correlated between A and B horizons in $p = 0.05$ (with correlation coefficients r = 0.79 and 0.72, respectively). However, this correlation between A and B horizons is much less marked for SOC (r = 0.59, $p = 0.05$). Inside the same A or B horizon, the correlation between SOC, CCE, and clay is quite limited (with r ranging from 0.11 to 0.31, $p = 0.05$).

3.2. SEM estimation and respecification

In relation to Eqs. (1)–(3), the initial SEM graphical model and its parameters that need to be estimated are presented in Fig. 6. In agreement with the conceptual model discussed in section 2.4.1, it accounts for the main drivers and processes that are thought to control the values of SOC, CCE, and clay in the study area, as represented by the relationships between soil properties and environmental covariates. From a first examination of the non-significant estimated parameters, the relationships between CCE.Br and SOC.Br and between WE and CCE.Ar was removed from the initial SEM model and the SEM estimation was performed again. The corresponding performance metrics are shown in the first line of Table 3. The low values for CFI and GFI are likely to indicate that some of the relevant relationships were not taken into account in the model specification. Accordingly, three successive model respecifications were considered and the corresponding results are presented in the third last lines of Table 3. Each respecification was based on the recommendations provided by the modification index. The details of the various respecifications are presented in Table 4 and are briefly discussed hereafter. Based on the lavaan package recommendations about the parameters currently set equal to 0 but that would be beneficially accounted for with non-null values in the SEM model (see Angelini et al. (2017) for more details), relationships between CCE.Ar with MAP and DEM were separately added to the SEM model (second step in Table 4). After estimating again, the parameters of this modified SEM model, the performance metrics (second step in Table 3) show that both CFI and GFI are increased. Subsequent respecifications based on the recommendations of the modification index led to the inclusion of relationships between clay in both horizons with Diffuse, DEM, and SWI covariates (the third step in Table 4), along with relationships relating Midslope to SOC.Ar and CCE.Br (fourth step in Table 4). From an examination of the corresponding performance metrics (fourth step in Table 3), it can be seen that both CFI and GFI have considerably increased and are now above 0.9 after these three respecifications, as expected for a relevant SEM model.

3.3. SEM interpretation

Focusing first on the relationships between soil properties and environmental covariates as measured by the estimated parameters of the final SEM formulation (Fig. 7), it can be seen that WE and MAT covariates have no significant influence on the soil properties in the A and B horizons, while DEM and MAP are by far the most influential ones in the A horizon. For CCE, it can be seen that the major influential covariates are DEM and MAP in the A horizon, while Midslope is the only identified influential covariate in the B horizon. For clay, the major influential covariate is DEM in both horizons, with an influence of SWI and MrVBF too in the A horizon, and Diffuse in the B horizon. A strong negative correlation between elevation and clay content was also reported by Riza et al. (2021) for well-drained sandy soils located in the crest and shoulder positions of the landscape, while poorly drained clayey soils were found in the toe-slope and valley positions. For SOC, the identified influential covariates are NDVI and Midslope (but no covariates were identified in the B horizon). This is in agreement with Zhang et al. (2019) who showed that SOC is significantly and positively correlated with NDVI. Clearly, NDVI is an important covariate for SOC as it reflects crop growth and biomass and is thus a good proxy for predicting SOC in plains. About the influence of Midslope, Li et al. (2018) and Tajik et al. (2020) believed that topographic attributes are key variables for surface SOC content as they control erosion and deposition processes. Focusing now on the relationships between soil properties, it can be seen that there is a strong influence of SOC, CCE, and clay from the A horizon to the B horizon, in agreement with pedological knowledge, as soil properties like SOC, CCE, and clay progressively change from surface to subsurface. The SOC.A is identified as positively influenced by CCE, while SOC.B is positively influenced by clay.B. These results are in agreement with Rowley et al. (2018) observed that Calcium carbonate ($CaCO_3$) can influence SOC stability through its role in the stabilization of aggregates, while organic matter intercalated in clays exhibits resistance to oxidation (Briedis et al., 2012; Carmeis Filho et al., 2017; Paradelo et al., 2015). Similarly, Meersmans et al. (2008) found that soils with high clay content have the highest amount of SOC, as fine-textured soils positively influence SOC storage through accumulation and mineralization processes (Pandey and Begum, 2010). Identical results were reported by Côté et al. (2000); Shahandeh et al. (2005); Kokulan et al. (2018) with positive correlations between SOC and clay. As higher clay values imply a larger water-holding capacity for plant growth and biomass production, this leads in turn to a higher SOC content. It is worth noting too that relationships between SOC.A and clay. A and between SOC.B and CCE.B were not statistically significant. A likely explanation is the low SOC values in these horizons, with associated SOC measurements errors that are no negligible and that can impair the proper identification of these relationships.

3.4. Prediction validation

Plugging the estimated parameters of the SEM model in Eq. (4), it is possible to predict the values of SOC, CCE, and clay in both horizons and to compare them with the corresponding observed values. As seen from Table 5 and Fig. 8, the prediction performances are very poor for CCE and clay in both horizons and are weak for SOC in both horizons too. For the sake of illustration only, the resulting maps of predicted SOC, CCE and clay are presented in Fig. 9. These poor predictions could be linked to the fact that the environmental covariates we used in this study are not good proxies for the effective soil-forming factors, or to the influence of land management practices that were not accounted for in this study.

Another possible explanation is that SEM is a linear model, while the relationships between environmental covariates and soil properties (and possibly omitted soil forming factors as well) are nonlinear ones. The bended shape of the scatter plots for predicted versus observed values for clay.A and clay.B in Fig. 8 is consistent with this hypothesis. Similarly, Liu et al. (2020) observed low R^2 values and have shown that the linear relationship assumption between soil organic matter and the environmental covariates they used was invalid. In agreement with this, Angelini et al. (2017, 2020) observed low SEM prediction performance for predicting SOC in Argentina, and they attributed this to the typically small amount of SOC in their soils, so that measured SOC values are impacted by large relative measurement errors. While this could be a part of the explanation for SOC in our dataset (with values that do not exceed a few percent in both horizons), this is unlikely for CCE which covers a wide range of values. Whatever the exact reasons, these results preclude the use of our SEM model for sound predictions and for mapping SOC, CCE, and clay over the study area.

In spite of these disappointing prediction results, it is worth emphasizing that the main objective of a SEM model is to analyze whether a model is adequate enough to support the hypotheses that have led to the conceptual model. SEM is commonly and successfully used in various fields like economics, psychology, etc. However, its use for pure prediction or for spatial mapping has received much less attention. As mentioned in the Introduction section, the goal of the SEM approach is to evidence the direct and indirect relationships between soil properties and environmental covariates. Based on the results that were obtained for our case study, our SEM model was indeed beneficial with respect to that goal. The conclusions that are drawn from the estimated SEM model are consistent with the findings of other authors. Besides the fact that SOC, CCE, and clay values are clearly related between A and B horizons, the SEM model is evidencing the direct relationship between SOC.A and CCE.A, and between SOC.B and clay.B, along with a set of direct and indirect relationships between soil properties and environmental covariates (Fig. 7). This was possible thanks to the fact that all soil properties and environmental covariates were used together in the same model, while other studies are restricted to evidence these relationships on a pairwise basis, thus failing to make the distinction between direct and indirect relationships.

4. Conclusions

In this study that was conducted in an arid and semi-arid region of Iran, we investigated the use of a SEM methodology for modeling and predicting the values of three key soil properties (namely SOC, CCE, and clay) in the A and B horizons with the help of nine environmental covariates. Among them, NDVI, Midslope, DEM, MAP, MrVBF, SWI, and Diffuse were identified as the main controlling factors, while DEM and MAP were clearly the most influential ones for CCE.A and clay.A. In contrast, WE and MAT were not significantly related to these soil properties whatever the horizon. In agreement with soil science knowledge, a strong relationship was also evidenced for the values of each soil property between the A and B horizons. Inside each horizon, the effect of CCE and clay on SOC was also evidenced, but to various extents, as CCE was the most influential for SOC.A while it was clay for SOC.B. However, this study highlights too that our SEM was inefficient for predicting SOC and CCE and yielded weak prediction performances for clay, thus preventing its use for mapping these key soil properties over the study area. In spite of this limitation, our SEM was able to account for the cause-effect relationship between soil properties and environmental covariates in the context of the area we studied, which involved complex and intricate soil-forming factors.

As the SEM approach allows combining soil science knowledge inside a model that accounts for all variables (i.e., those related to soil forming factors, external factors, and soil system) at the same time, it permits to investigate of their potential cause-effect relationships in a much richer theoretical framework than usual approaches. We thus believe that the SEM methodology is potentially useful for soil scientists that are studying various soil properties in other parts of the world, even if the corresponding SEM cannot be advocated as an efficient digital soil mapping method in general.

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.

Acknowledgements

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

References

Adhikari, K., Hartemink, A.E., 2016. Linking soils to ecosystem services – a global review. Geoderma 262, 101–111. https://doi.org/10.1016/j.geoderma.2015.08.009.

Angelini, M.E., Heuvelink, G.B., Kempen, B., Morrás, H.J., 2016. Mapping the soils of an Argentine Pampas region using structural equation modelling. Geoderma 281, 102–118. https://doi.org/10.1016/j.geoderma.2016.06.031.

Angelini, M.E., Heuvelink, G., Kempen, B., 2017. Multivariate mapping of soil with structural equation modelling. Eur. J. Soil Sci. 68, 575–591. https://doi.org/10.1111/ejss.12446.

Angelini, M.E., Kempen, B., Heuvelink, G., Temme, A.J., Ransom, M.D., 2020. Extrapolation of a structural equation model for digital soil mapping. Geoderma 367, 114226. https://doi.org/10.1016/j.geoderma.2020.114226.

Azizi, K., Garosi, Y., Ayoubi, S., Tajik, S., 2023. Integration of Sentinel-1/2 and topographic attributes to predict the spatial distribution of soil texture fractions in some agricultural soils of western Iran. Soil Tillage Res. 229, 105681 https://doi.org/10.1016/j.still.2023.105681.

Bayram, N., 2013. Yapısal eşitlik modellemesine giriş [Introduction to structural equation modeling]. Ezgi Yayıncılık, Bursa, Turkey.

Bollen, K.A., 1989. Structural Equations With Latent Variables, vol. 210. John Wiley & Sons.

Brady, N.C., Weil, R.R., 2008. The Soils Around Us. The Nature and Properties of Soils, 14th ed. Pearson Prentice Hall, New Jersey and Ohio, pp. 1–31.

Breiman, L., 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16 (3), 199–231.

Briedis, C., de Moraes Sá, J.C., Caires, E.F., de Fátima Navarro, J., Inagaki, T.M., Boer, A., Neto, C.Q., de Oliveira Ferreira, A., Canalli, L.B., Dos Santos, J.B., 2012. Soil organic matter pools and carbon-protection mechanisms in aggregate classes influenced by surface liming in a no-till system. Geoderma 170, 80–88. https://doi.org/10.1016/j.geoderma.2011.10.011.

Carmeis Filho, A.C., Penn, C.J., Crusciol, C.A., Calonego, J.C., 2017. Lime and phosphogypsum impacts on soil organic matter pools in a tropical Oxisol under long-term no-till conditions. Agr. Ecosyst. Environ. 241, 11–23. https://doi.org/10.1016/j.agee.2017.02.027.

Christ, S.L., Lee, D.J., Lam, B.L., Zheng, D.D., 2014. Structural equation model- ing: a framework for ocular and other medical sciences research. Ophthal. Epidemiol. 21, 1–13. https://doi.org/10.3109/09286586.2013.867508.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for automated geoscientific analyses (saga) v. 2.1. 4. Geosci. Model Dev. 8, 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015.

Costa, E.M., Pinheiro, H.S.K., Anjos, L.H.C.d., Marcondes, R.A.T., Gelsleichter, 475 Y.A., 2020. Mapping soil properties in a poorly-accessible area. Revista 476 Brasileira de Ciência do Solo 44. doi:10.36783/18069657rbcs20190107.

Côté, L., Brown, S., Paré, D., Fyles, J., Bauhus, J., 2000. Dynamics of carbon and nitrogen mineralization in relation to stand type, stand age and soil texture in the boreal mixed wood. Soil Biol. Biochem. 32, 1079–1090. https://doi.org/10.1016/S0038-0717(00)00017-1.

Dion, P.A., 2008. Interpreting structural equation modeling results: a reply to Martin and Cullen. J. Bus. Ethics 83, 365–368. https://doi.org/10.1007/s10551-007-9634-7.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. (2017). arXiv preprint arXiv:1702.08608 27. doi:10.48550/arXiv.1702.08608.

Esmaeili, E., Shahbazi, F., Sarmadian, F., Jafarzadeh, A.A., Hayati, B., 2021. Land capability evaluation using NRCS agricultural land evaluation and site assessment (LESA) system in a semi-arid region of Iran. Environ. Earth Sci. 80, 1–14. https://doi.org/10.1007/s12665-021-09468-y.

Feng, W., Plante, A.F., Six, J., 2013. Improving estimates of maximal organic carbon stabilization by fine soil particles. Biogeochemistry 112, 81–93. doi:10. 1007/s10533-011-9679-7. part 1—physical and mineralogical methods, (methodsofsoilan1). doi:10.2136/sssabookser5.1. 2ed.c15.

Giardina, C.P., Ryan, M.G., Hubbard, R.M., Binkley, D., 2001. Tree species and soil textural controls on carbon and nitrogen mineralization rates. Soil Sci. Soc. Am. J. 65, 1272–1279. https://doi.org/10.2136/sssaj2001. 6541272x.

Grace, J.B., Keeley, J.E., 2006. A structural equation model analysis of postfire plant diversity in California shrublands. Ecol. Appl. 16 (2), 503–514.

Grace, J.B., Schoolmaster Jr, D.R., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M., Schweiger, E.W., 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3, 1–44. https://doi.org/10.1890/ES12-00048.1.

Grinand, C., Barthès, B., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G., Bernoux, M., 2012. Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). Eur. J. Soil Sci. 63, 141–151. https://doi.org/10.1111/j. 1365-2389, 2012.01429. x.

Guo, L., Sun, X., Fu, P., Shi, T., Dang, L., Chen, Y., Linderman, M., Zhang, G., Zhang, Y., Jiang, Q., Zhang, H., 2021. Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas. Geoderma 398, 115118. https://doi.org/10.1016/j.geoderma.2021.115118.

Guo, P.T., Wu, W., Sheng, Q.K., Li, M.F., Liu, H.B., Wang, Z.Y., 2013. Prediction of soil organic matter using artificial neural network and topographic indicators in hilly areas. Nutr. Cycl. Agroecosyst. 95, 333–344. https://doi.org/10.1007/s10705-013-9566-9.

Hoyle, R.H., 2012. Handbook of Structural Equation Modeling. Guilford Press.

Jagadamma, S., Lal, R., 2010. Distribution of organic carbon in physical fractions of soils as affected by agricultural management. Biol. Fertil. Soils 46, 543–554. https://doi.org/10.1007/s00374-010-0459-7.

Jeong, G., Oeverdieck, H., Park, S.J., Huwe, B., Ließ, M., 2017. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. Catena 154, 73–84. https://doi.org/10.1016/j.catena.2017.02.006.

Carvalho Junior, W.d., Calderano Filho, B., Chagas, C.d.S., Bhering, S.B., Pereira, N.R., Pinheiro, H.S.K., 2016. Regressão linear múltipla e modelo random forest para estimar a densidade do solo em áreas montanhosas. Pesquisa Agropecuária Brasileira 51, 1428–1437. doi:10.1590/S0100-204X2016000900041.

Kline, R.B., 2016. Principles and Practice of Structural Equation Modeling. Guilford Publications, 10.4324/9780203807644.

Kokulan, V., Akinremi, O., Moulin, A.P., Kumaragamage, D., 2018. Importance of terrain attributes in relation to the spatial distribution of soil properties at the micro scale: a case study. Can. J. Soil Sci. 98, 292–305. https://doi.org/10.1139/cjss-2017-0128.

Lark, R., 2000. A comparison of some robust estimators of the variogram for use in soil survey. Eur. J. Soil Sci. 51, 137–157. https://doi.org/10.1046/j.1365-23892000.00280.x.

Li, X., McCarty, G.W., Karlen, D.L., Cambardella, C.A., 2018. Topographic metric predictions of soil redistribution and organic carbon in iowa cropland fields. Catena 160, 222–232. https://doi.org/10.1016/j.catena.2017.09.026.

Liu, E., Liu, J., Yu, K., Wang, Y., He, P., 2020. A hybrid model for predicting spatial distribution of soil organic matter in a bamboo forest based on general regression neural network and interative algorithm. J. For. Res. 31, 1673–1680. https://doi.org/10.1007/s11676-019-00980-3.

Liu, L.A., Yang, R.M., Zhang, X., Zhu, C.M., Zhang, Z.Q., 2021. A mechanistic approach for modeling soil development using remotely sensed data collected from invaded coasts. Remote Sens. (Basel) 13, 564. https://doi.org/10.3390/rs13040564.

Loeppert, R.H., Suarez, D.L., 1996. Carbonate and gypsum. Methods of soil analysis: Part 3 chemical methods 5, 437–474. doi:10.2136/sssabookser5. 3.c15.543.

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. https://doi.org/10.1016/j.geoderma.2009.10.007.

Martens, M.P., 2005. The use of structural equation modeling in counseling psychology research. Couns. Psychol. 33, 269–298. https://doi.org/10.1177/0011000004272260.

Matinfar, H.R., Maghsodi, Z., Mousavi, S.R., Rahmani, A., 2021. Evaluation and prediction of topsoil organic carbon using machine learning and hybrid models at a field-scale. Catena 202, 105258. https://doi.org/10.1016/j.catena.2021.105258.

McGinnies, W.G., 1981. Discovering the Desert: Legacy of the Carnegie Desert Botanical Laboratory. University of Arizona Press.

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., Van Molle, M., 2008. A multiple regression approach to assess the spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 143, 1–13. https://doi.org/10.1016/j.geoderma.2007.08.025.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38. https://doi.org/10.1016/j.artint.2018. 07.007.

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

Mohammadi, J., Motaghian, M., 2011. Spatial prediction of soil aggregate stability and aggregate-associated organic carbon content at the catchment scale using geostatistical techniques. Pedosphere 21, 389–399. https://doi.org/10.1016/S1002-0160(11)60140-9.

Molnar, C., 2020. Interpretable machine learning: A guide for making black box models explainable. <https://christophm.github.io/interpretable-ml-book/>.

Mousavi, S., Sarmadian, F., Omid, M., Bogaert, P., 2021. Modeling the vertical soil calcium carbonate equivalent variation by machine learning algorithms in Qazvin Plain. J. Water Soil 35, 719–734. https://doi.org/10.22067/jsw.2021.71748.1076.

Mousavi, S.R., Sarmadian, F., Omid, M., Bogaert, P., 2022. Three-dimensional mapping of soil organic carbon using soil and environmental covariates in an arid and semi-arid region of Iran. Measurement 201, 111706. https://doi.org/10.1016/j.measurement.2022.111706.570.

Neyestani, M., Sarmadian, F., Jafari, A., Keshavarzi, A., Sharififar, A., 2021. Digital mapping of soil classes using spatial extrapolation with imbalanced data. Geoderm. Reg. 26, e00422.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning for digitalsoil mapping. Soil 5, 79–89. https://doi.org/10.5194/soil-5-79-2019.

Pandey, C., Begum, M., 2010. The effect of a perennial cover crop on net soil N mineralization and microbial biomass carbon in coconut plantations in the humid tropics. Soil Use Manag. 26, 158–166. https://doi.org/10.1111/j.1475-2743.2010.00272.x.576.

Paradelo, R., Virto, I., Chenu, C., 2015. Net effect of liming on soil organic carbon stocks: a review. Agr. Ecosyst. Environ. 202, 98–107. https://doi.org/10.1016/j.agee.2015.01.005.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kauffmann, San Mateo, 10.2307/2026705.

Pearl, J., 1995. Causal diagrams for empirical research. Biometrika 82 (4), 669–688.

Pearl, J., 2009. Causal inference in statistics: an overview. Stat. Surv. <https://doi.org/10.1214/09-SS057>.

Poggio, L., Gimona, A., 2017. Assimilation of optical and radar remote sensing data in 3D mapping of soil properties over large areas. Sci. Total Environ. 579, 1094–1110. https://doi.org/10.1016/j.scitotenv.2016.11.078.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. <https://www. R-project. org/>.

Riza, S., Sekine, M., Kanno, A., Yamamoto, K., Imai, T., Higuchi, T., 2021. Modeling soil landscapes and soil textures using hyperscale terrain attributes. Geoderma 402, 115177.

Rostaminia, M., Rahmani, A., Mousavi, S.R., Taghizadeh-Mehrjardi, R., Maghsodi, Z., 2021. Spatial prediction of soil organic carbon stocks in an arid rangeland using machine learning algorithms. Environ. Monit. Assess. 193, 1–17. https://doi.org/10.1007/s10661-021-09543-8.

Rowley, M.C., Grand, S., Verrecchia, É.P., 2018. Calcium-mediated stabili- sation of soil organic carbon. Biogeochemistry 137, 27–49. https://doi.org/10.1007/s10533-017-0410-1.595.

Sales, M.V., Aleixo, S., Gama-Rodrigues, A.C., Gama-Rodrigues, E.F., 2017. Structural equation modeling for the estimation of interconnections between the P cycle and soil properties. Nutr. Cycl. Agroecosyst. 109, 193–207. https://doi.org/10.1007/s10705-017-9879-1.599.

Sarkar, A., Azim, J.A., Al Asif, A., Qian, L., Peau, A.K., 2021. Structural equation modeling for indicators of sustainable agriculture: Prospective of a developing country's agriculture. Land Use Policy 109, 105638. https://doi.org/10.1016/j.landusepol.2021.105638.

Schermelleh-Engel, K., Moosbrugger, H., Müller, H., et al., 2003. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. Methods Psychol. Res. Online 8, 23–74.

Shabou, M., Mougenot, B., Chabaane, Z.L., Walter, C., Boulet, G., Aissa, N.B., Zribi, M., 2015. Soil clay content mapping using a time series of Landsat TM data in semi-arid lands. Remote Sens. (Basel) 7, 6059–6078. https://doi.org/10.3390/rs70506059.610.

Shahandeh, H., Wright, A., Hons, F., Lascano, R., 2005. Spatial and temporalvariation of soil nitrogen parameters related to soil texture and corn yield. Agron. J. 97, 772–782. https://doi.org/10.2134/agronj2004.0287.

Shahriari, A., Khormali, F., Kehl, M., Ayoubi, S., Welp, G., 2012. Effect of a long-term cultivation and crop rotations on organic carbon in loess derived soils of Golestan Province, Northern Iran. Int. J. Plant Prod. 5 (2), 147–152. https://doi.org/10.22069/ijpp.2012.728.

Soil Survey Staff, 2014. Keys to soil taxonomy. USDA-Natural Resources Conservation Service, Washington, DC.

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020. 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. (Basel) 12, 1095. https://doi.org/10.3390/rs12071095.

Tajik, S., Ayoubi, S., Zeraatpisheh, M., 2020. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Reg. 20, e00256.

Tuominen, L.S., Helle, S., Helanterä, H., Karell, P., Rapeli, L., Richmond, D., Vuorisalo, T., Brommer, J.E., 2022. Structural equation modeling re-veals decoupling of ecological and self-perceived outcomes in a garden box social-ecological system. Sci. Rep. 12, 1–16. https://doi.org/10.1038/s41598-022-10178-z.625.

Wadoux, A.M.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359 https://doi.org/10.1016/j.earscirev.2020.103359.

Wadoux, A.M.C., Heuvelink, G.B., Lark, R.M., Lagacherie, P., Bouma, J., Mulder, V.L., Libohova, Z., Yang, L., McBratney, A.B., 2021. Ten challenges for the future of pedometrics. Geoderma 401, 115155. https://doi.org/10.1016/j.geoderma.2021.115155.

Walkley, A., Black, I.A., 1934. An examination of the degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Sci. 37, 29–38. https://doi.org/10.1097/00010694-193401000-00003.632.

Wang, L., Liu, H., 2006. An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. Int. J. Geogr. Inf. Sci. 20, 193–213. https://doi.org/10.1080/13658810500433453.636.

Wilding, L., Drees, L., 1978. Spatial variability: a pedologist's viewpoint. Div. Soils Trop. 34, 1–12. https://doi.org/10.2134/asaspecpub34.c1.

Zeraatpisheh, M., Ayoubi, S., Sulieman, M., Rodrigo-Comino, J., 2019. Determining the spatial distribution of soil properties using the environmental covariates and multivariate statistical analysis: a case study in semi-arid regions of Iran. J. Arid. Land 11, 551–566. https://doi.org/10.1007/s40333-019-0059-9.

Zeraatpisheh, M., Jafari, A., Bodaghabadi, M.B., Ayoubi, S., Taghizadeh-Mehrjardi, R., Toomanian, N., Kerry, R., Xu, M., 2020. Conventional and digital soil mapping in Iran: past, present, and future. Catena 188, 104424.

Zeraatpisheh, M., Bottega, E.L., Bakhshandeh, E., Owliaie, H.R., Taghizadeh-Mehrjardi, R., Kerry, R., Scholten, T., Xu, M., 2022. Spatial variability of soil quality within management zones: homogeneity and purity of delineated zones. Catena 209, 105835. https://doi.org/10.1016/j.catena.2021.105835.

Zhang, Y., Guo, L., Chen, Y., Shi, T., Luo, M., Ju, Q., Zhang, H., Wang, S., 2019. Prediction of soil organic carbon based on landsat 8 monthly ndvi data for the jianghan plain in hubei province, china. Remote Sens. (Basel) 11, 1683. https://doi.org/10.3390/rs11141683.650.

Zhu, H., Zhao, Y., Nan, F., Duan, Y., Bi, R., 2016. Relative influence of soil chemistry and topography on soil available micronutrients by structural equation modeling. J. Soil Sci. Plant Nutr. 16, 1038–1051. https://doi.org/10.4067/S0718-95162016005000076.654.