The Journal of Agricultural Science

www.cambridge.org/ags

Predicting and comparative soil quality assessment in organic and conventional hazelnut orchards using XGBoost approach

Orhan Dengiz[1], Pelin Alaboz[2], Nejc Suban[3] and Sena Pacci[1]

[1]Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Ondokuz Mayis University, Samsun, Turkiye; [2]Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Isparta University of Applied Sciences, Isparta, Turkiye and [3]Department of Forest Physiology and Genetics, Slovenian Forestry Institute, Ljubljana, Slovenia

Crops and Soils Research Paper

Cite this article: Dengiz O, Alaboz P, Suban N, and Pacci S (2026). Predicting and comparative soil quality assessment in organic and conventional hazelnut orchards using XGBoost approach. *The Journal of Agricultural Science* **164**(e9), 1–17. https://doi.org/10.1017/S0021859625100476

Received: 22 August 2025
Revised: 10 November 2025
Accepted: 4 December 2025

Keywords:
Agriculture; Black Sea; machine learning; soil properties

Corresponding author:
Pelin Alaboz;
Email: pelinalaboz@isparta.edu.tr

Abstract

The aim of this study was to determine soil quality index (SQI) for hazelnut gardens managed under organic and conventional agricultural systems. Additionally, the predictability of soil quality was evaluated using the XGBoost algorithm. To determine soil quality, a multi-criteria decision-making process was applied to the total data set (TDS) using standard scoring functions (linear and non-linear). Additionally, the minimum data set (MDS) was obtained using principal component analysis (PCA). Then, the model verification process was performed using SQI and yield data. According to the results, although SQI values in conventional agriculture were statistically significantly higher, the correlation between yield and soils under organic agriculture was higher than in conventional agriculture. The SQI averaged 0.4576 in conventionally farmed soils and 0.4417 in organically farmed areas. Root mean square error values obtained for SQI estimation with the XGBoost algorithm using basic soil properties ranged from 0.038 to 0.065. The mean error rate was approximately 8%. Lin's concordance correlation coefficients for the SQI estimated by MDS and TDS were 0.60 and 0.61, respectively. The most effective basic soil properties for estimating SQI with the XGBoost algorithm were N, K, organic matter, and P. It was concluded that the XGBoost algorithm can be evaluated for soil quality prediction. In addition, the spatial distribution patterns of the values predicted by this algorithm and of the observed values were similar. The exclusive use of soil analyses in the study can be considered a limiting factor for the model. More comprehensive studies are planned using reflectance measurements from remote sensing technologies.

Introduction

Ever since the Green Revolution, agricultural production around the globe has undergone intensification. The development of chemical fertilizers and hybrid plant varieties led to higher yields, which were supported by the use of pesticides to combat unwanted weeds, pests, and diseases (Dornbush and von Haden, 2017). Hazelnut producers in Türkiye are expected to face challenges because hazelnut production in other countries has steadily increased in recent years. High-input agriculture methods in the USA and France, as well as regions with better growing conditions (e.g. Georgia), can produce yields per hectare up to twice those in Turkey (Bozoglu *et al.*, 2019). The modernization of agriculture brought about an enormous increase in food production capabilities, which enabled unprecedented global population growth (Philip Robertson *et al.*, 2014). The shift towards more convenient and efficient but also short-term and profit-oriented agriculture frequently came at the expense of natural ecosystems and human health (Dornbush and von Haden, 2017). In the face of the growing global population and the effects of intensive agriculture, such as soil fertility loss, nutrient leaching, loss of biodiversity, and environmental pollution, conventional agricultural practices should be reviewed for their appropriateness and modified to achieve more sustainable food production (Philip Robertson *et al.*, 2014). One of the well-defined and commonly adopted alternative agricultural systems that strive to establish more sustainable agricultural production is organic farming (Willer *et al.*, 2022). The main hazelnut-producing countries are Turkey, Italy, the USA, and Azerbaijan. Turkey holds a significant position in the global hazelnut market. In 2019, it accounted for 66% of global hazelnut production and 76% of total global exports (FAO, 2022).

The principles and basic standards of organic agriculture, formulated by IFOAM (Federation of Organic Agriculture Movement), mainly focus on improving soil quality, fostering crop diversity, protecting the health of consumers and farmers, and reducing the environmental impact of production. Some practices commonly used in organic plant production include crop rotation, use of crop residues, animal and green manure, compost, and biological control of diseases and pests (Reganold and Wachter, 2016). Numerous studies have been conducted to

© The Author(s), 2025. Published by Cambridge University Press.

investigate the environmental effects of the aforementioned system of practices compared with conventional agricultural production practices. The results of these comparisons do not always unequivocally support the thesis that organic farming improves soil quality, since not all the environmental conditions and sets of farming practices included in the study led to improvements in soil properties (Lorenz and Lal, 2016). Soil quality assessment employs a set of attributes encompassing the soil's chemical, physical, and biological properties. To assess soil fertility, enzymatic activity serves as a good indicator of soil ecosystem functioning (Maurya *et al.*, 2020). Since it is impossible to collect all individual parameters for individual soils, a minimum data set (MDS) of parameters that represents the links between soil and plant properties and management practices is selected by means of expert judgement or, more commonly, by statistical methods such as principal component analysis (PCA), redundancy analysis (RDA), discriminant analysis, and multiple regression (Bünemann *et al.*, 2018). Currently, with advances in technology, generating spatial distribution maps using geostatistical approaches rather than relying on point sampling is essential for sustainable soil management. In many studies, regional conditions are assessed using spatial distribution maps. It is also recommended to take precautions in anticipation of adverse future scenarios (Alaboz and Dengiz, 2023).

The use of machine learning algorithms in soil science has become widespread, and researchers are testing the predictive accuracy of different algorithms for various soil properties (Alaboz and Işıldar, 2019; Alaboz *et al.*, 2023; Şenol *et al.*, 2020). Yu *et al.* (2024) combined machine learning methods with a habitat patch approach to more accurately predict spatial variation of soil organic carbon (SOC) in a karst region. The researchers divided the study area into subregions based on their environmental similarity and trained separate XGBoost and Random Forest models for each subregion. XGBoost achieved the best performance in some patches. Zhao *et al.* (2024) constructed machine-learning models of soil salinity (soil salt content, SSC) from spectral reflectance and texture features derived from drone-based multi-spectral imagery. Four variable-input combinations and three algorithms (XGBoost, Random Forest, and BPNN) were used. The study found that XGBoost yielded higher prediction accuracy than Random Forest (RF) and BPNN. XGBoost is a machine learning algorithm first described by Tianqi Chen and Carlos Guestrin in 2016 (Chen and Guestrin, 2016). XGBoost is an optimized implementation of the gradient boosting algorithm. XGBoost algorithms use predictions from an ensemble of decision trees to produce accurate classifications. These classifiers can also be used effectively to rank and select variables with the highest discriminatory accuracy across target classes (Zhou *et al.*, 2023). XGBoost uses the maximum depth value when building the tree. If the created tree exhibits excessive downward growth, pruning is performed. Over-fitting of the model is prevented. The parallel processing feature enables results to be obtained in less time than with other algorithms (Kelle and Yüce, 2022). Soil properties, such as bulk density (BD), soil organic carbon, texture, and shear strength, are estimated with high accuracy using the XGBoost algorithm (Hikouei *et al.*, 2023; Momeni *et al.*, 2023; Yang *et al.*, 2023). The XGBoost algorithm is a recent method that has yielded positive results for some soil properties. However, the algorithm has not been applied to soil quality estimation, which underscores the study's originality.

The main objective of this study is to evaluate the soil quality of hazelnut plantations in the selected area of Çarşamba district, Turkey and to investigate the effects of traditional and organic management practices on selected soil-quality indicators. Therefore, the present study aims to develop a comprehensive framework for assessing soil quality under different management systems of hazelnut plantations. Specifically, the study seeks to identify a MDS of soil indicators critical for estimating the Soil Quality Index (SQI) for hazelnut productivity and calculate SQI values for the study area. This research aims to quantify how management systems influence soil health and crop yield by comparing these indicators and SQI values under traditional and organic management practices. Furthermore, the study evaluates the predictive capacity of soil quality using the XGBoost algorithm and assesses the relationship between SQI values and yield data. Finally, the most suitable geostatistical interpolation model is selected to accurately predict the spatial distribution of soil quality indicators and SQI values across the region. It is hypothesized that organic management practices will enhance soil quality and yield stability compared to traditional systems.

Material and methods

General description of the study area

The study area covers approximately 4,386 ha and is located at between 36° 38' 45" -41° 05' 42" (WGS 84, Zone 37, UTM) (Figure 1). The selected study area lies south of the town in the Samsun-Çarşamba district on the edge of the Çarşamba delta plain, including both lowland and hills, and extends around the accumulation lake on the Abdal River that passes through the middle of our study.

According to meteorological data collected by the Turkish State Meteorological Service (DMI) between 1960 and 2017, the climate in the Çarşamba district is semi-humid, with average annual precipitation and evapotranspiration levels of 1,023 mm and 772 mm, respectively. The average annual temperature is 14.4°C; the average temperatures in July and January are 23.3°C and 6.4°C, respectively. Based on calculations by Miháliková and Dengiz (2019), the Newhall simulation model classifies the study area's soil climate as having a mesic temperature regime and an udic (subgroup: dry tempudic) moisture regime. Bölük (2016) reports that the study area is classified as 'very moist' (precipitation activity index = 66.64), based on Erinc's macroclimate regions for Turkey.

As shown in Figure 2, the varied terrain is reflected in the diversity of geological parent materials. The geological materials of the study area consist of alluvial deposits, including pebbles, sandstone, mudstone, and basalt-andesite pyroclastic rocks. The elevation distribution was computed from the digital elevation model (DEM) raster layer using a triangulated irregular network model (TIN). Elevation in the area ranges from 18 to 290 m above sea level (Figure 2).

The Slope of the terrain was also calculated from the DEM raster layer using the 'Slope' Spatial Analyst function. Slope percentage values were classified using the Jenks natural breaks method. The vast majority of the terrain has gentle slopes between 0% and 20%, a mean slope of 13%. Values range from 0.01% to 92.58% (Figure 3).

Soil sampling and analysis

Our study area was delineated to include the following villages: Çayvar, Esençay, Esentepe, Gökçakmak, Gökçeli, Kemer, Koldere, Kumköy and Şehgüven (Figure 1). From a list of registered hazelnut producers in the Çarşamba district, we selected 116

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

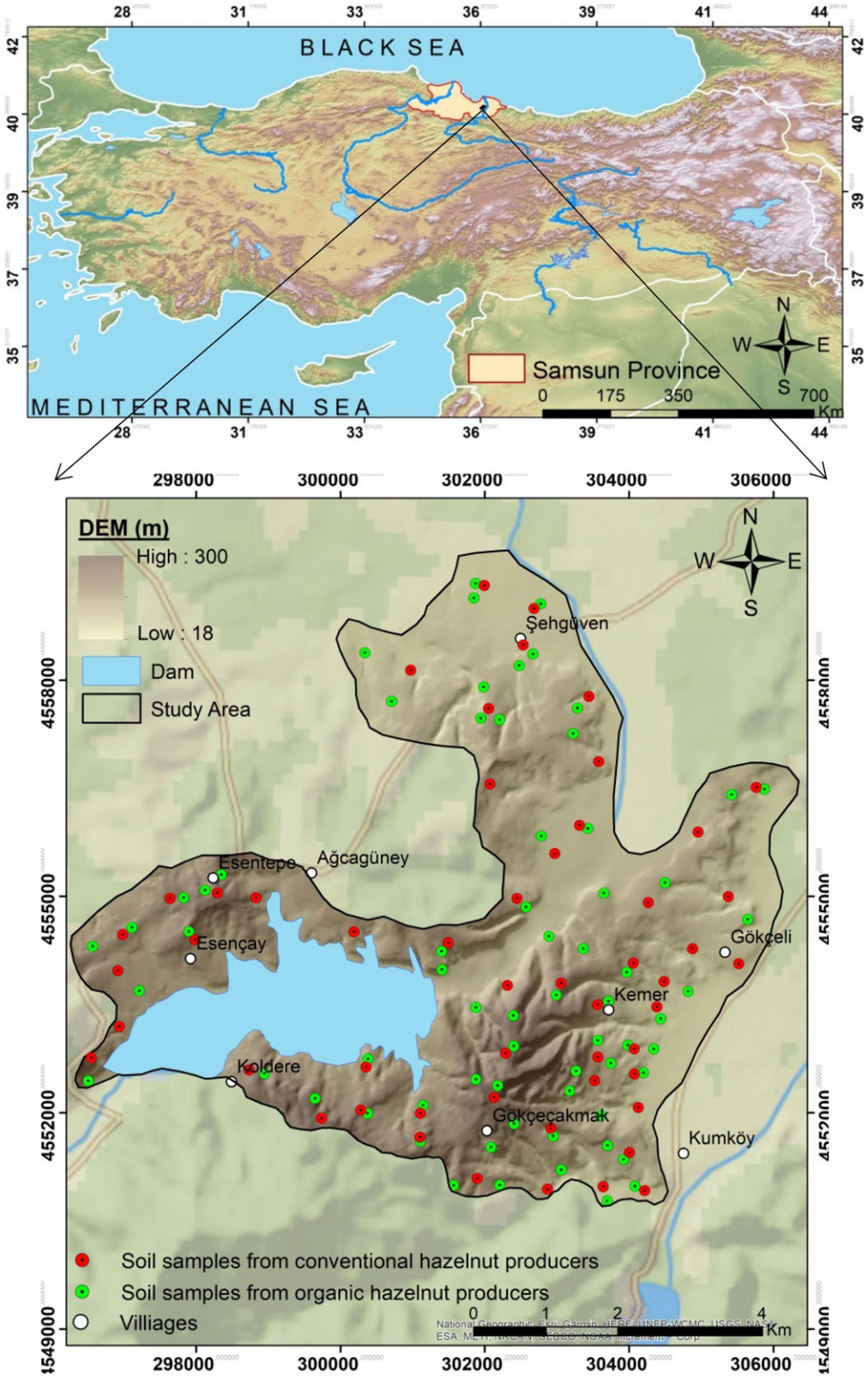


Figure 1. Location and soil sampling pattern maps of the study area.

parcels with surface areas greater than 1 hectare. In addition, we selected 50 plantations managed under traditional agricultural practices for hazelnut production and 66 plantations managed organically for more than 10 years. Soil sampling was carried out in 2022 after the hazelnut harvest to ensure that other applications, such as fertilization, did not affect the soil properties. Additionally, we considered different topographic locations and land use/land cover types (hazelnut parcels) when collecting the soil samples.

After cleaning, drying (air-dry), and grinding, the soil samples were passed through a 2 mm sieve. The soil's physical,

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

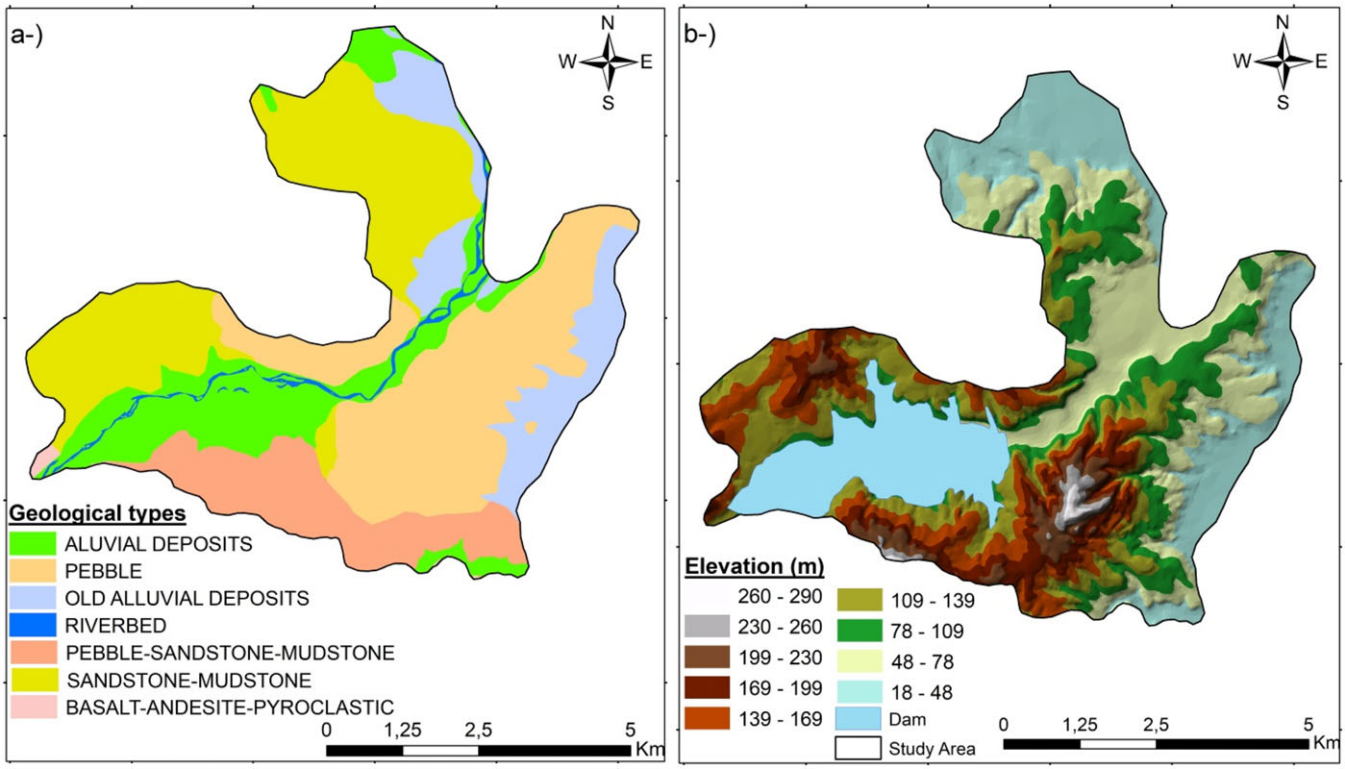


Figure 2. Geological type and elevation maps of the study area.

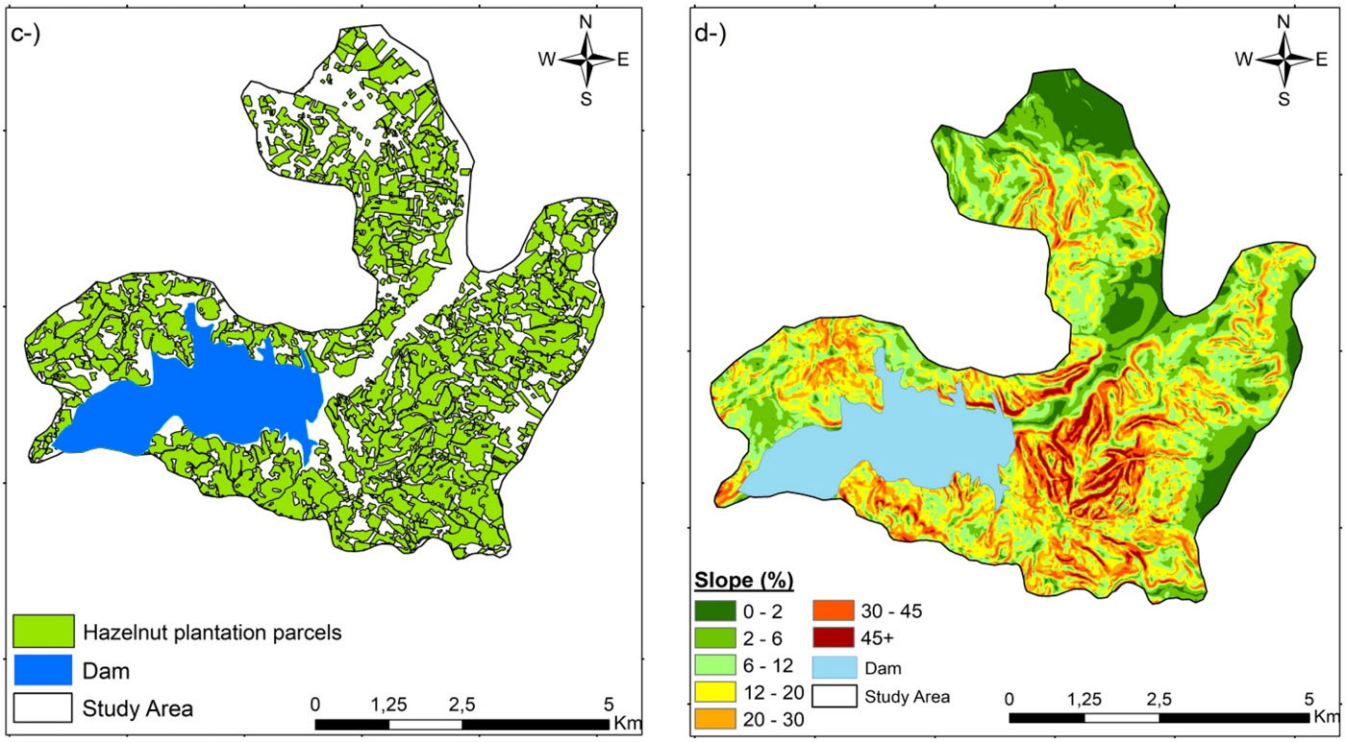


Figure 3. Land use classification and Slope of the study area.

chemical, and biological properties were determined, and soil quality was calculated. A total of 27 soil properties were examined. In the SQI for hazelnut cultivation, the indicators are grouped under four categories and the physicochemical and biological analyses and methods of the soils used in the study are stated in Table 1.

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

Table 1. Protocol measurements for indicators selected

Parameters	Unit	Protocol	Reference
Physical			
Aggregate stability (AS)	%	Wet sieving	Kemper and Rosenau (1986)
Dispersion ratio (DR)	%	DR= (a/b)* 100	Lal and Elliot (1994)
Texture	%	hydrometer method	Bouyoucos (1951)
Available water content (AWC)	%	Computation with mathematical model	Saxton *et al.* (1986)
Bulk density (BD)	g/cm^3	Computation with mathematical model	Saxton *et al.* (1986)
Field capacity (FC)	%	Computation with mathematical model	Saxton *et al.* (1986)
Hydraulic conductivity (HC)	mm/h	Computation with mathematical model	Saxton *et al.* (1986)
Wilting point (WP)	%	Computation with mathematical model	Saxton *et al.* (1986)
Chemical			
Organic matter (OM)	%	Walkley-Black wet digestion	Nelson and Sommers (1982)
pH	1:2.5	(w:v) soil-water suspension	Soil Survey Staff (1993)
EC	d/Sm	(w:v) soil-water suspension	Soil Survey Staff (1993)
Lime ($CaCO_3$)	%	Scheibler calcimeter	Soil Survey Staff (1993)
Fertility			
Available P	mg/kg	Bray and Kurtz, Olsen method	Olsen *et al.*, (1954)
Total N	%	Kjeldahl method	Bremner and Mulvaney (1982)
Extractable K, Ca, Mg, Na	cmol/kg	Ammonium acetate extraction, flame spectrometry detection	Soil Survey Staff (1993)
Extractable Cu, Fe, Mn, Zn	mg/kg	DTPA extraction, AAS detection	Lindsay and Norvell (1978)
Biological			
Microbial biomass carbon (MBC)	µg C/g dry soil	Substrate-induced respiration method at 22° C	Anderson and Domsch (1978)
Basal respiration (BR)	µg CO_2-C/g dry soil	Net CO_2 production at 22° C without addition of glucose	Isermayer, 1952
Metabolic quotient (qCO_2)	unitless	µgCO_2/ µgCmic/d = BR/MBC	Anderson and Domsch (1978)

EC, electrical conductivity; P, phosphorus; N, nitrogen; K, potassium; Ca, calcium; Mg, magnesium; Na, sodium; Cu, copper; Fe, iron; Mn, manganese; Zn, zinc; CO_2, carbon dioxide a is the percentage of silt plus clay in suspension, b is the percentage of silt plus clay dispersed with chemical agent.

Determination of soil quality

The examined features were converted into unitless scores (xi) between 0 and 1, which can be compared with each other using linear and nonlinear scoring functions. The Analytic Hierarchy Process (AHP) was used to determine the weights (Wi) of the features. Soil quality was assessed using the linear combination method.

AHP is a multi-criteria decision-making method that enables numerical, hierarchical organization of parameters regarding their respective importance for solving the problem or achieving the objective of our interest. The pairs of individual indicators are compared to one another using arbitrarily assigned scales in a pairwise comparison matrix ranging from 1/9 to 9; where 9 represents the maximum importance of one indicator over another, 1 represents equal importance between indicators, and values less than 1 indicate a lesser importance of one indicator relative to another. The assigned values are based on experts' experience. The upper triangular matrix is filled with assigned values from 1 to 9, whereas the lower triangular matrix is filled with the reciprocal values of the pairwise comparisons. The comparison scale thus includes the values: 9, 8, 7, . . . , 1/7, 1/8, and 1/9.

The matrix is then normalized by dividing each pairwise comparison value by the sum of the values in the corresponding column. The row averages of the normalized pairwise comparison matrix constitute the priority vector (weight scores) of the respective indicators. Lastly, the consistency of the obtained eigenvector should be checked by calculating the Consistency Ratio (CR). We multiply the pairwise comparison matrix by the eigenvector of the normalized matrix to yield a new vector. The second new vector is obtained by dividing each component of the second vector by its corresponding eigenvalue of the normalized matrix. The Random Index (RI) is also required to calculate the Consistency Ratio (CR). The RI values are generated by filling 100 matrices with dimensions corresponding to the total number of weighted indicators in our study and calculating the average of their coherence indicators (Saaty, 1980). A smaller CR indicates a more consistent comparison matrix. As a general rule, CR must be less than 0.1 for the indicator weightings to be considered consistent. The consistency is improved by better expert judgement when assigning importance scores, by fewer indicators, and by greater homogeneity of assigned scores (Saaty, 1993). Weights derived from soil quality indicators are shown in Table 2.

Scoring of indicator values

Due to the different units of soil quality indicators, we normalized their values using standard scoring functions (SSF) before aggregating

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

Table 2. Weights of soil indicators with the Analytic Hierarchy Process (AHP) method

Rank	Criterion	Subcriterion rank	Subcriterion	Criterion weight (%)	Subcriterion weights (%)	Combined weights (criterion × subcriterion)
1	Physical indicators			0.4161		
		1	AS		0.244	0.102
		2	AWC		0.2006	0.083
		3	Clay		0.1401	0.058
		4	FC		0.1145	0.048
		5	WP		0.0733	0.031
		6	BD		0.0682	0.028
		7	HC		0.0642	0.027
		8	DR		0.041	0.017
		9	Sand		0.0296	0.012
		10	Silt		0.0245	0.010
2	Chemical indicators			0.2828		
		1	OM		0.5048	0.143
		2	pH		0.2876	0.081
		3	Lime		0.1431	0.040
		4	EC		0.0645	0.018
3	Biological indicators			0.189		
		1	MBC		0.558	0.105
		2	BR		0.32	0.060
		3	qCO_2		0.122	0.023
4	Fertility indicators			0.112		
		1	N		0.2517	0.028
		2	P		0.2166	0.024
		3	K		0.1466	0.016
		4	Fe		0.0997	0.011
		5	Ca		0.0818	0.009
		6	Zn		0.0596	0.007
		7	Cu		0.0506	0.006
		8	Mg		0.0439	0.005
		9	Mn		0.0312	0.003
		10	Na		0.0183	0.002

AS: Aggregate stability, AWC: Available water content, FC: Field capacity, WP: Wilting point, BD: Bulk density, HC: Hydraulic conductivity, DR: Dispersion ratio, OM: Organic matter, EC: electrical conductivity, MBC: Microbial biomass carbon, BR: Basal respiration, qCO_2: Metabolic quotient, N: nitrogen, P: phosphorus, K: potassium, Fe: iron, Ca: calcium, Zn: zinc, Cu: copper, Mg: magnesium, Mn: manganese, Na: sodium.

them into the SQI. The normalization assigns unitless scores ranging from 0 and 1 to the indicators using different scoring functions depending on each indicator's functional effect on soil quality (Andrews *et al.*, 2004). A 'more-is-better' (MB) function is assigned to indicators for which higher values indicate better soil quality. A 'less is better' (LB) function describes the opposite relationship where low indicator values contribute to better soil quality. For some of the indicators, an optimal range was delimited by assigning them MB and LB functions depending on whether the indicator value was above or below the optimal range (Andrews *et al.*, 2004).

We used linear and nonlinear approaches for standard scoring. For linear SSF, we followed the equations used in Fernández *et al.* (2020) (Equations 1 and 2). The nonlinear approach using a sigmoid function, with a curve slope of −2.5, was adapted from the paper by Bastida *et al.* 2006 (Equations 3 and 4). The SSF equations are presented below:

$$MB = \frac{0,9 \times (x - L)}{(U - L)} + 0,1 \quad (1)$$

$$LB = \frac{1 - 0,9 \times (x - L)}{(U - L)} \quad (2)$$

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

$$MB_{NL} = \frac{1}{1 + (x/\overline{x})^{-2,5}} \quad (3)$$

$$LB_{NL} = \frac{1}{1 + (x/\overline{x})^{2,5}} \quad (4)$$

where MB represents linear 'more is better' function, MB_{NL} represents non-linear 'less is better' function, LB represents linear 'less is better' function, LB_{NL} represents non-linear 'less is better' function, x is indicator value, is mean indicator value, U is upper limit of indicator values and L lowest indicator values.

The assignment of the types of SSF to the respective soil indicators, as well as their mean, upper, and lower values, is presented in Table 3. All the indicators were calculated using both linear and non-linear approaches.

Integration of indicators with soil quality index

The indicator values were combined into a single index ranging from 0 to 1, with higher values indicating greater soil quality. The SQI was calculated using the weighted additive method, as presented in the formula below (Equation 5):

$$SQI = \sum_{i=1}^{n} (S_L \times W_i) \quad (5)$$

where SQI represent Soil Quality Index with weights assigned by AHP and scores calculated with linear and nonlinear SSF, S_L represents sub-criteria's normalization score calculated with linear and non-linear SSF respectively, W_i is a weight value of the parameter previously determined by AHP method and n is a total number of parameters.

Geostatistical interpolation and statistical approaches

Spatial distribution maps of indicator-value data and of the values for four SQI were generated using ArcGIS v10.7. The spatial prediction from the sampling points' data was interpolated across the delimited study area using inverse distance weighting (IDW) geostatistical model (power = 2) and a radial basis function (thin-plate spline). To separate spatially predicted values into several classes of maximal dissimilarity, the Jenks natural breaks classification method (Jenks, 1967) was applied.

The PCA was applied to the data to create a MDS of soil properties. When determining parameters for inclusion in the MDS, component loadings from PCA, correlation-load totals, and inter-data correlation analyses will be considered. Descriptive statistics for soils and PCA will be performed using IBM SPSS 23 software package. In the PCA, soil properties will be standardized, and the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity will be examined to assess the suitability of the data for factor analysis. Principal components (PC) with eigenvalues greater than 1 will be selected and the MDS will be created by examining the correlation matrices. Tukey's multiple-comparison test was used to compare the soil quality approaches. The boxplot and clustering dendrogram were obtained using the IBM SPSS 23 software package. The correlation matrices among the studied soil-quality models were calculated using the METHANE package in R.

The XGBoost package in R was used to implement the XGBoost algorithm to estimate soil quality. The model can be expressed mathematically (Equation 6):

Table 3. The types of standard scoring functions (SSF) assigned to each of the soil quality indicator

Criterion	Subcriterion	SSF	mean	L	U
Physical indicators					
	AS	MB	70.10	38.24	96.44
	AWC	MB	11.79	7.90	19.40
	Clay	MB (<45.1%). LB (>45%)	46.65	15.74	68.44
	FC	MB	40.25	26.70	45.70
	WP	MB	28.46	15.70	34.90
	BD	LB	1.28	1.11	1.49
	HC	MB	4.77	0.06	28.88
	DR	LB	16.95	5.88	38.43
	Sand	LB	27.42	11.54	76.62
	Silt	LB	25.93	3.74	40.05
Chemical indicators					
	OM	MB	5.65	1.99	11.61
	pH	MB (>4.27). LB (>7.49)	6.14		
	Lime	MB (<11.01%). LB (11%)	6.95	1.60	27.91
	EC	LB	185.04	47.22	472.50
Biological indicators					
	MBC	MB	23.30	2.78	125.90
	BR	MB	39.81	1.17	128.43
	qCO_2	LB	1.20	0.12	12.80
Fertility indicators					
	N	MB	0.13	0.04	0.23
	P	MB	1.34	0.02	8.58
	K	MB	0.89	0.15	2.95
	Fe	MB	811.97	21.34	3905
	Ca	MB	37.74	6.75	187.31
	Zn	MB	45.17	7.02	685.30
	Cu	MB	41.54	2.64	89.10
	Mg	MB	22.91	5.26	443.20
	Mn	MB	509.61	38.72	1976
	Na	LB	1.17	0.15	3.58

MB refers to 'more is better' SSF, LB refers to 'less is better' SSF, mean represents mean indicator values, L is minimum and U the maximum indicator value. AS: Aggregate stability, AWC: Available water content, FC: Field capacity, WP: Wilting point, BD: Bulk density, HC: Hydraulic conductivity, DR: Dispersion ratio, OM: Organic matter, EC: electrical conductivity, MBC: Microbial biomass carbon, BR: Basal respiration, qCO_2: Metabolic quotient, N: nitrogen, P: phosphorus, K: potassium, Fe: iron, Ca: calcium, Zn: zinc, Cu: copper, Mg: magnesium, Mn: manganese, Na: sodium.

$$\widehat{y}\, i = \sum_{k=1}^{k} f_k(x_i),\ f_k \in F \quad (6)$$

where F is the set of all potential regression trees, K is the number of trees, and f is a function in the functional space F. Consequently, the optimization's goal can be expressed as follows (Equation 7):

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

Table 4. Statistical description of soil indicators

Indicator	mean	σ	σ^2	CV	Skewness	Kurtosis	Minimum	Maximum
Clay %	46.65	10.68	113.98	0.23	−0.34	−0.36	15.74	68.44
Silt %	25.93	6.43	41.39	0.25	−0.13	1.12	3.74	40.05
Sand %	27.42	11.29	127.55	0.41	1.29	2.57	11.54	76.62
AS %	70.10	10.07	101.34	0.14	0.05	1.10	38.24	96.44
AWC %	11.79	1.95	3.80	0.17	0.50	0.92	7.90	19.40
BD g/cm^3	1.28	0.07	0.005	0.06	0.38	1.08	1.11	1.49
DR %	16.95	5.69	32.35	0.34	0.89	1.67	5.88	38.43
FC %	40.25	3.40	11.55	0.08	−1.45	2.87	26.70	45.70
HC mm/hr	4.77	4.76	22.65	1.00	2.36	7.00	0.06	28.88
WP %	28.46	4.66	21.70	0.16	−0.48	−0.55	15.70	34.90
OM %	5.65	1.73	3.01	0.31	0.61	0.68	1.99	11.61
$CaCO_3$ %	6.95	3.86	14.91	0.56	3.41	13.61	1.60	27.91
pH	6.14	0.87	0.75	0.14	0.14	−0.66	4.27	7.82
EC d/Sm	185.04	98.16	9635	0.53	1.20	0.92	47.22	472.50
N %	0.13	0.04	0.00	0.29	0.25	−0.03	0.04	0.23
P mg/kg	1.34	1.38	1.90	1.03	2.58	8.50	0.02	8.58
K cmol/kg	0.89	0.53	0.28	0.60	1.73	3.42	0.15	2.95
Ca cmol/kg	37.74	26.23	687.82	0.69	3.42	16.57	6.75	187.31
Mg cmol/kg	23.02	40.10	1607	1.74	10.18	107.42	5.26	443.20
Na cmol/kg	1.17	0.60	0.36	0.51	1.01	2.02	0.15	3.58
Fe mg/kg	811.97	664.02	440924	0.82	1.78	4.88	21.34	3905
Cu mg/kg	41.54	15.86	251.51	0.38	0.53	0.63	2.64	89.10
Zn mg/kg	45.17	108.60	11794	2.40	4.98	24.55	7.02	685.30
Mn mg/kg	509.61	377.71	142665	0.74	1.27	1.93	38.72	1976
MBC µg C/g	23.30	20.90	436	89.69	3.85	15.48	2.78	3.85
CO_2 µg CO_2-C/g	39.81	30.44	926	76.47	1.14	0.38	1.17	1.14
qCO_2	1.19	2.00	4.02	167.67	4.01	17.79	0.11	4.01

Indicators' values including mean value (mean), standard deviation (σ), variance (σ^2), coefficient of variation (CV), skewness, kurtosis, range, minimum and maximum values. AS: Aggregate stability, AWC: Available water content, BD: Bulk density, DR: Dispersion ratio, FC: Field capacity, HC: Hydraulic conductivity, WP: Wilting point, OM: Organic matter, $CaCO_3$: lime, EC: electrical conductivity, N: nitrogen, P: phosphorus, K: potassium, Ca: calcium, Mg: magnesium, Na: sodium, Fe: iron, Cu: copper, Zn: zinc, Mn: manganese, MBC: Microbial biomass carbon, CO_2-BR: Basal respiration, qCO_2: Metabolic quotient.

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) \sum_{k=1}^{k} f_k(f_k) \quad (7)$$

where the regularization term is $\Omega(fk)$ and the training loss function is $I(yi, yi)$. The XGBoost model's objective is to minimize obj($\emptyset$). Several tests revealed that shrinkage (eta), maximum depth (max_depth), number of iterations (nround), and alpha, set to 0.8, 10, 200, and 0.2 respectively, stabilized prediction error in the XGBoost algorithm. To improve learning, the k-fold cross-validation method was also used, with validation performed on 70% of the training data. This validation method prevented overfitting of the model to the training data set. Following multiple randomized tests, 10-fold cross-validation was found to be the most successful method. Basic soil properties (sand, silt, clay, pH, EC, organic matter (OM), lime content, N, P, and K) were used as input to the models.

Root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), coefficient of determination (R^2), and Lin's concordance correlation coefficient (LCC) were used to evaluate the models.

Results

Physico-chemical, productivity and biological properties of soils

The soil characteristics of 116 hazelnut plantations sampled in our study area vary due to the diverse parent material, environmental factors, different degrees of soil formation and leaching processes, and plantation management activities (e.g. fertilization, liming, soil cultivation, etc.). Descriptive statistics for soil properties are presented in Table 4. Values of most physical indicators were normally distributed except for hydraulic conductivity (HC), whose high kurtosis indicates a heavy-tailed distribution caused by outliers. Soil particle-size distribution classes in the samples range from heavy to moderate clay, coarse to fine silt, and very coarse to fine sand

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

texture classes. The mean soil texture-class values indicate a heavy texture for clay and a moderate texture for silt and sand. The maximum relative contents of clay and sand in the soil samples were 68.44% and 76.62%, respectively. The mean values for AS, BD and DR are 70.1%, 1.28 g/cm^3 and 16,95% which correspond to high AS, low BD and slight dispersion percentage rate as suggested by Hazelton and Murphy (2016). All analysed samples exhibit AS values rated as highly stable. Water retention properties vary with sample texture. The ranges were 7.90–19.4% for available water content (AWC), 26.7–45.7% for field capacity (FC), 15.7–34.9% for wilting point (WP), and 0.06–28.88 mm/h for HC; their mean values were 11.79% for AWC, 40.25% for FC, 28.46% for WP, and 4.77 mm/h for HC. The mean HC value is rated very low according to Hazelton and Murphy (2016). Chemical indicator values range from 1.99–11.61% for OM, 1.99–27.91% for lime content, pH 4.27–7.82, and EC 47.22–472.5 µS/cm. Their values, with the exception of lime content, are normally distributed. The mean OM content is 5.65%, which, according to the rating proposed by Hazelton and Murphy (2016), is considered extremely high. The OM content in all samples ranged from moderate to extremely high levels. pH values range from mildly alkaline to very strongly acidic, with a mean pH of 6.14 (slightly acidic). The lime content, however, was found to be between 1.6% and 27.91%, with a mean of 6.75%, which is considered moderate. The EC values are rated non-saline to slightly saline (Hazelton and Murphy 2016); the mean value is 185.04 µS/cm.

Indicator values of nutrients in the soil were generally normally distributed, except for Ca, Mg, and Zn, which exhibited heavy-tailed, skewed distributions. Total nitrogen in the soil ranged from very low (0.04%) to medium (0.23%), with the mean value being low (0.13 mg/kg) (Hazelton and Murphy, 2016). Available phosphorus levels were very low (0.02 mg/kg) and low (8.58 mg/kg), and its mean value of 1.34 mg/kg was considered very low (Hazelton and Murphy, 2016). Potassium values between 0.15 and 2.95 cmol/kg correspond to content levels ranging from low to very high. A mean potassium concentration of 0.89 cmol/kg is considered high. The values of the remaining macronutrients (Ca, Mg, and Na) range from 6.75 to 187.31 cmol/kg, 5.26 to 443.2 cmol/kg, and 0.15 to 3.58 cmol/kg, respectively. The levels of Ca range from moderate and very high, with its mean value (37.74 cmol/kg) at the very high level. The Mg content is between high and very high, while the Na content is between low and very high, with its mean value being at a high level (Hazelton and Murphy, 2016). The values of micronutrients are unusually high and reach levels that can be harmful to the plants, as the case of the maximum Zn value of 685.3. The values of biological factors range from 2.78 to 125.9 µg C/g soil for MBC and from 1.17 to 128.43 µg C/g soil for BR. Adapting the rating proposed by Andrews *et al.* (2004) and Vasconcellos *et al.* (2013), all the MBC values are rated as low; basal respiration values range from low to high, whereas the mean BR value is considered low.

Generating a minimum data set

To ensure the sustainability of hazelnut cultivation and to determine the SQI, selecting the most appropriate soil indicators is essential. In this study, we evaluated 27 factors in the total data set (TDS) to best represent soil physical, chemical, biological, and fertility properties. We used PCA to generate this MDS (Nabiollahi *et al.*, 2017; Qi *et al.*, 2009; Saygın *et al.*, 2023). The PCA performed to derive a MDS of soil properties is presented in Table 5. During the creation of the MDS using the PC method, the KMO sphericity test was applied. The results were determined to be statistically significant (KMO > 0.5). In the evaluation using PCA, PCs with an eigenvalue ≥1.0 were retained to include the indicators in the MDS. With an eigenvalue ≥1.0, eight PC were considered. A total of 73.95% of the variance was explained by 8 PC. In each PCA with an eigenvalue ≥1.0, indicators falling within 10% of the highest weight were evaluated for the MDS. When more than one variable from the same PC that met the indicator selection criteria were entered into the data set and the correlation between those variables was ≥0.60, the variable with the highest weight was selected for the MDS. The features with the highest weights in PCA1 were identified as clay and WP. The selected features showed high correlation (≥0.60) with each other. Clay was determined as the feature with the highest weight and was considered as the indicator included in the MDS. In PCA2, pH and CO_2 have high weights. It exhibited a high correlation ($r = 0.68$) with pH-CO_2. The feature with the highest weight is pH, and pH-related features are included in the MDS. The features with the highest weights in PCA3 and PCA4 were identified as BD and Zn and were included in the MDS. In PCA5, MBC and qCO_2 were selected as features with high weights; because they were highly correlated, qCO_2 was eliminated and MBC was added to MDS. Features with high loadings on other PC were not identified. 27 indicators were subjected to PCA, resulting in a MDS comprising nine indicators.

Soil quality and statistical comparisons

The boxplots showing the minimum and maximum values of the soil-quality approaches are shown in Figure 4. In addition, the results of cluster analysis of SQIs obtained using different approaches are presented in Figure 5.

SQI values ranged from 0.331 to 0.577 for TDS-ORG-L and from 0.334 to 0.587 for the nonlinear method. Using the MDS, MDS-ORG-L ranged from 0.211 to 0.591, and the non-linear method ranged from 0.290 to 0.580. In conventional agricultural areas, SQI values obtained using the TDS with linear and non-linear methods were 0.314-0.622 and 0.304–0.587, respectively; using the MDS they were 0.211–0.606 and 0.290–0.589, respectively. The lower and upper outliers with the greatest deviation from the normal distribution were detected by the MDS-ORG-NL method. A 96.38% statistical similarity was detected between the total data set linear SQI for conventional hazelnut cultivation (TDS-CON-L) and the total data set non-linear SQI (TDS-CON-NL), while a 91.60% statistical similarity was detected between the total data set linear SQI for organic hazelnut cultivation (TDS-ORG-L) and the total data set non-linear SQI (TDS-ORG-NL) (Figure 5). In addition, a 65.55% statistical similarity was observed between conventional and organic hazelnut cultivation.

The changes in SQIs obtained by TDS and MDS, using linear and nonlinear methods for conventional and organic cultivation are presented in Table 6. The average soil quality in conventional hazelnut cultivation was 0.4576, compared with 0.4417 in organic hazelnut cultivation; this difference was statistically significant. In addition, SQI values determined by linear and nonlinear methods for conventional hazelnut cultivation and for the TDS were statistically different (0.4824–0.4310), whereas SQI values for organic hazelnut farming determined by linear and nonlinear methods were statistically similar (0.4501–0.4534). When SQI values obtained from MDSs were examined using linear and nonlinear methods, they yielded similar results.

Correlation matrices showing the relationships between SQI values obtained by different methods and yield values are shown in Figures 6 and 7. Correlation matrices indicated a positive relationship between SQI values obtained by different methods

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

Table 5. Principal component analysis matrix of soil properties

	Principal Component (PC)							
	1	2	3	4	5	6	7	8
C	**−0.974**	0.113	0.053	−0.019	0.010	−0.038	0.013	0.012
Si	0.294	0.239	0.808	−0.147	0.002	0.161	0.006	−0.133
Sa	0.744	−0.242	−0.510	0.102	−0.011	−0.056	−0.016	0.065
BD	0.115	−0.051	**−0.932**	−0.054	−0.060	0.027	0.100	0.017
AS	−0.426	−0.264	0.039	−0.058	−0.156	**−0.622**	0.121	0.102
HC	0.774	−0.048	0.437	0.046	0.001	−0.068	−0.007	−0.054
WP	**−0.967**	0.044	0.044	0.012	0.015	−0.062	−0.030	0.056
AWC	0.789	−0.004	0.252	0.036	−0.002	0.114	−0.194	−0.055
DR	0.365	0.420	−0.126	−0.074	0.133	0.585	0.158	−0.036
FC	−0.886	0.058	0.204	0.037	0.019	−0.020	−0.153	0.045
AW	0.812	−0.014	0.393	−0.026	0.048	0.163	−0.189	−0.053
OM	0.055	−0.188	0.569	0.473	0.165	−0.143	0.226	0.124
$CaCO_3$	0.005	0.590	0.217	−0.105	−0.186	−0.296	0.272	−0.059
pH	0.009	**0.883**	0.006	0.039	−0.057	0.182	−0.032	−0.026
EC	−0.258	0.681	0.068	0.231	0.122	0.132	−0.227	0.000
N	0.116	−0.180	0.513	0.448	−0.142	0.119	0.035	0.104
P	0.175	−0.026	0.083	0.660	−0.164	−0.100	−0.198	−0.086
K	−0.087	0.246	−0.150	0.668	−0.001	−0.100	0.101	−0.023
Ca	−0.226	0.498	0.011	0.055	−0.146	0.001	0.289	0.486
Mg	−0.040	0.009	−0.063	−0.044	0.040	−0.054	−0.062	**0.866**
Na	0.011	0.007	−0.095	0.157	0.049	−0.011	**0.804**	−0.082
Fe	0.196	−0.670	0.275	0.210	−0.104	0.071	0.037	0.169
Cu	−0.192	−0.162	0.220	−0.027	−0.036	0.209	0.455	0.302
Zn	−0.069	−0.053	0.082	**0.794**	0.050	0.087	0.190	0.014
Mn	−0.116	−0.268	0.249	−0.125	−0.186	**0.650**	0.096	0.023
MBC	0.010	0.085	0.082	0.009	**0.895**	0.082	−0.010	0.040
CO_2	0.009	**0.839**	0.126	0.045	−0.173	−0.048	−0.064	0.211
qCO_2	−0.013	−0.280	−0.024	−0.112	**0.849**	−0.086	0.043	−0.051
Eigen value	6.083	3.725	3.239	2.146	1.752	1.454	1.244	1.065
Proportion	21.724	13.305	11.569	7.663	6.258	5.192	4.443	3.802
Cumulative	21.724	35.029	46.597	54.260	60.519	65.710	70.153	73.955

C: clay, Si: Silt, Sa: sand, BD: Bulk density, AS: Aggregate stability, HC: Hydraulic conductivity, WP: Wilting point, AWC: Available water content, DR: Dispersion ratio, FC: Field capacity, OM: Organic matter, $CaCO_3$: lime, EC: electrical conductivity, N: nitrogen, P: phosphorus, K: potassium, Ca: calcium, Mg: magnesium, Na: sodium, Fe: iron, Cu: copper, Zn: zinc, Mn: manganese, MBC: Microbial biomass carbon, CO_2-BR: Basal respiration, qCO_2: Metabolic quotient.

(linear and non-linear) and yield. A statistically significant positive correlation was observed between the SQI values from the TDS of conventional hazelnut cultivation soils and yield ($r = 0.51$, $p < 0.05$). In addition, the same results were obtained for soils used for organic hazelnut cultivation. Moreover, the highest positively correlation was found between TDS-SQI-ORG-L and yield values ($r = 0.56$, $p < 0.001$)

Prediction and spatial distribution of SQI

Model evaluation parameters are stated in Table 7. The lowest LCC and R^2 values when estimating soil quality using basic soil properties (sand, silt, clay, lime, OM, pH, EC, N, P, K) with the XGBoost algorithm were observed in the MDS-L data set: 0.458 and 0.32, respectively. Although MAPE differed by approximately 5% when soil quality was estimated using linear versus nonlinear methods with the MDS data set, error levels in the TDS data set were very similar. RMSE values for the predictions obtained with the MDS-NL and TDS-L data sets were 0.043. The most influential features for model prediction were identified as N (0.189), K (0.020), OM (0.010), P (0.008), and pH (0.004) (Figure 8).

Spatial distribution maps of the SQI values for TDS and MDS, generated using linear and non-linear approaches for both conventional and organic hazelnut farming, are presented in

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

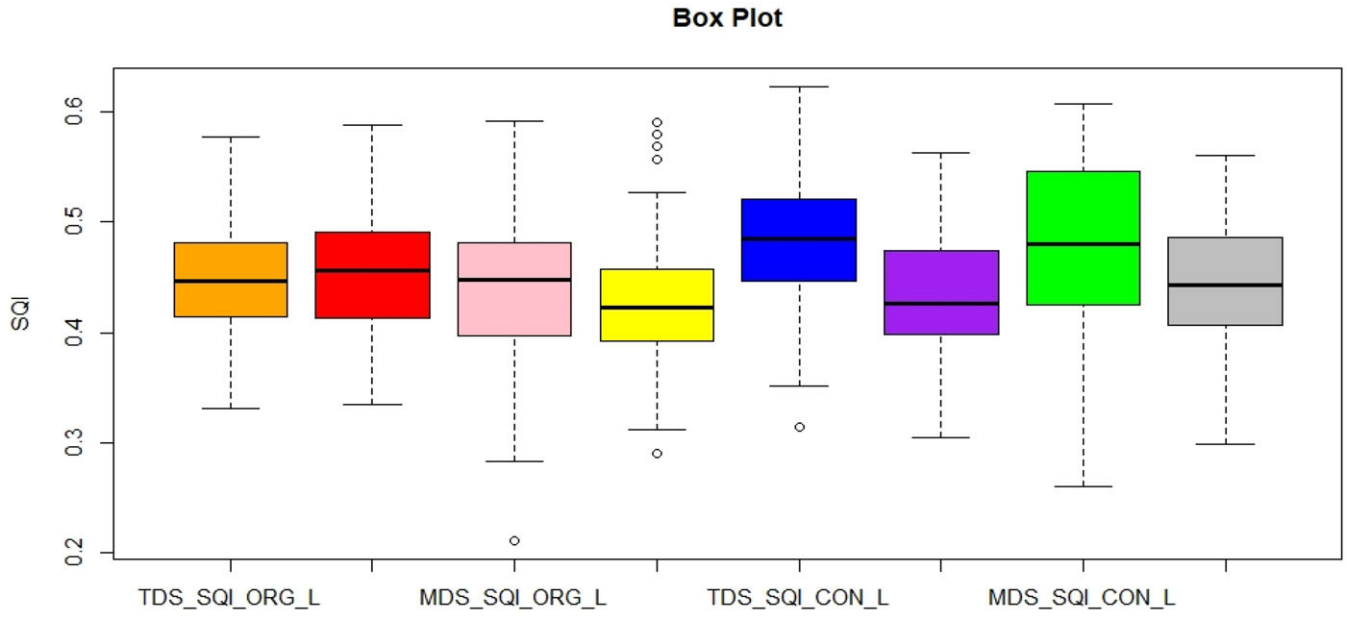


Figure 4. Box-Plot graphics. TDS: total data set, MDS: minimum data set, L: Linear, NL: Non-Linear, CON: Conventional hazelnut farming, ORG: Organic hazelnut farming.

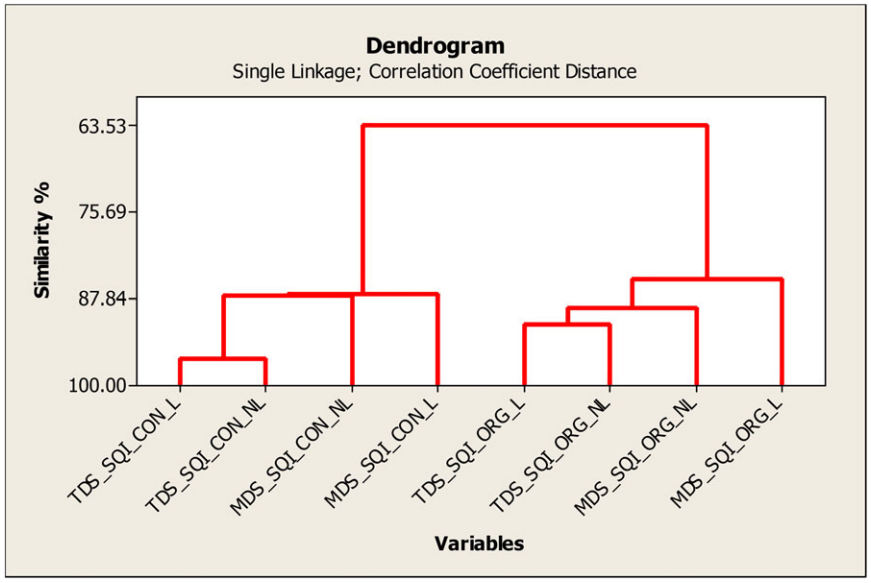


Figure 5. Clustering analysis. SQI: soil quality index, TDS: total data set, MDS: minimum data set, L: Linear, NL: Non-Linear, CON: Conventional hazelnut farming, ORG: Organic hazelnut farming.

Table 6. Soil quality index (SQI) variation due to the difference in cultivation and approaches (Linear and Non-linear)

	Conventional		Organic	
	L	NL	L	NL
TD	0.4824 a*	0.4310 b	0.4501 ab	0.4534ab
MD	0.4781a	0.4390b	0.4361 b	0.4271b
Mean	0.4576 **A****		0.4417 **B**	

* The change due to the difference in the method and data set is statistically significant ($P < 0.01$), ** The change in SQI due to the difference in cultivation is statistically significant ($P < 0.05$), L: Linear, NL: non-linear, TD: total data set, MD: minimum data set.

Figure 9. In the study, the obtained results indicated that TDS-CON-L, MDS-CON-NL, and MDS-CON-L approaches showed a similar pattern to each other (except for the map of TDS-CON-NL). This is also apparent from the clustering dendrogram in Figure 5. In addition, the distribution maps of SQI values (TDS-ORG-L and TDS-ORG-NL) exhibited a similar pattern, whereas those of MDS-ORG-L and MDS-ORG-NL exhibited a different pattern. High-quality soils in hazelnut cultivation are located in the northern and northwestern parts of the study area (Şengüven, Esentepe, and Aşağıgüney villages), whereas low-quality soils in hazelnut-growing areas are found in the mountainous southern part of the study area.

Spatial distribution maps of observed and predicted SQI values for TDS and MDS, obtained using linear and nonlinear approaches, are shown in Figure 9. The distribution maps for TDS-OBS-L, TDS-PRE-L, and TDS-PRE-NL exhibit similar spatial patterns across the study area, except for the TDS-OBS-NL map. The MDS-PER-L, MDS-PER-NL and MDS-OBS-NL maps have almost identical distributions, except for the MDS-OBS-L map.

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

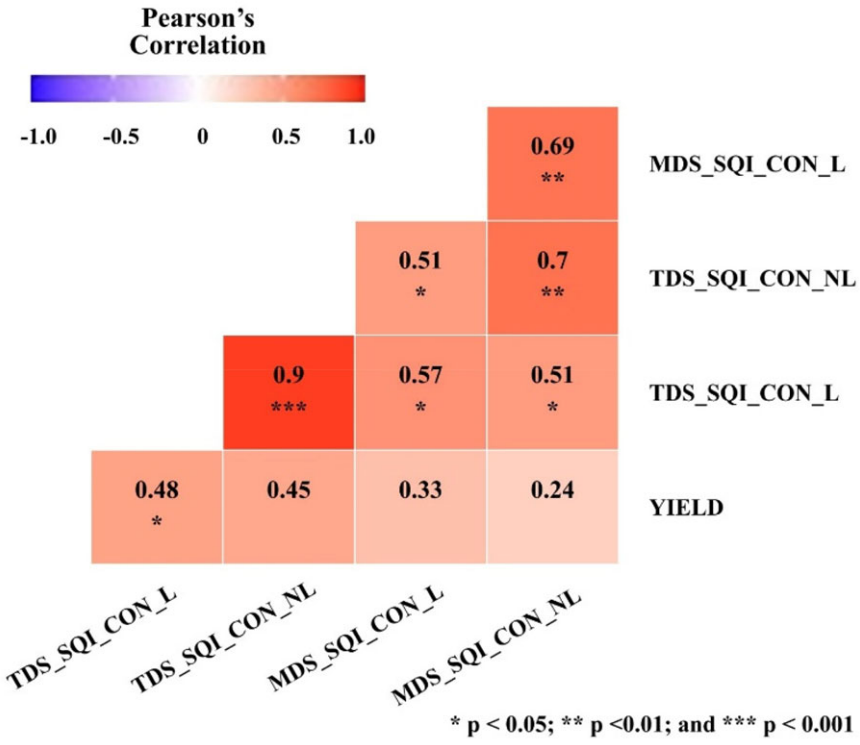


Figure 6. Relationship between SQI values of conventional hazelnut cultivation and yield Values. SQI: soil quality index, TDS: total data set, MDS: minimum data set, L: Linear, NL: Non-Linear, CON: Conventional hazelnut farming.

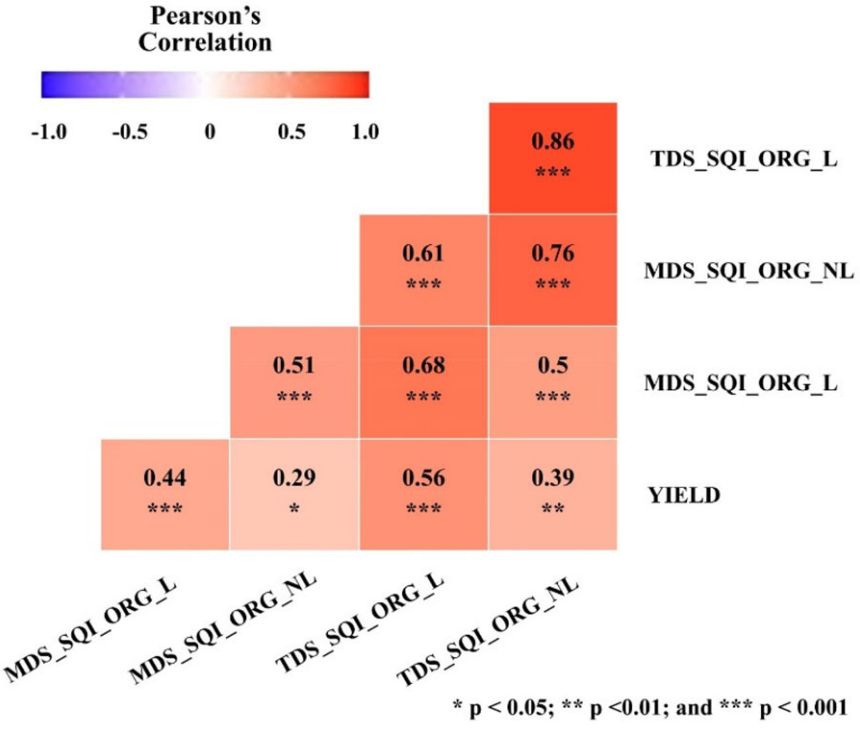


Figure 7. Relationship between SQI values of organic hazelnut cultivation and yield values. SQI: soil quality index, TDS: total data set, MDS: minimum data set, L: Linear, NL: Non-Linear ORG: Organic hazelnut farming.

Discussion

The foundation of all agricultural productivity is soil. Therefore, improving soil fertility and quality is the first step towards achieving the required quantity and quality of a product. Plant nutrients are among the most crucial elements for boosting soil fertility, but it is critical to preserve the soil's chemical, physical, and biological properties (Alaboz *et al.*, 2023; Şenol *et al.*, 2020). The results indicate that OM content and topsoil OM stocks are, on average higher under organic management (Gattinger *et al.*, 2012; Lorenz and Lal, 2016). OM is particularly important for soil quality, as it reduces nutrient leaching and promotes soil aggregation, which improves soil structure, thereby reducing soil erodibility (Lynch, 2014) and increasing water-holding capacity

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

Table 7. Model evaluation parameters

	LCC	R^2	MSE	MAPE %	MAE	RMSE
MDS-L	0.458	0.32	0.004	12.74	0.052	0.065
MDS-NL	0.600	0.50	0.001	7.92	0.033	0.043
TDS-L	0.610	0.51	0.001	7.71	0.034	0.043
TDS-NL	0.640	0.54	0.001	7.42	0.031	0.038

LCC: Lin's concordance correlation coefficient, MSE: mean square error, MAPE: mean absolute percentage error, MAE: Mean absolute error, RMSE: Root mean square error, MDS-L: the minimum data set linear SQI, MDS-NL: the minimum data set non-linear SQI, TDS-L: the total data set linear SQI, TDS-NL: the total data set non-linear SQI.

(Gomiero *et al*, 2011). Although evidence supports the claim that organic agriculture leads to improvements in some soil properties lowers greenhouse gas emissions, enhances air and water quality, and consumes less energy, organic farming areas cover only a small fraction of cultivated land (9.4% in the EU and 1.5% globally in 2020) (Lorenz and Lal, 2016). The major drawback of organic farming is its yields, which are on average 19% lower compared with more intensive agricultural practices. Nonetheless, organic farming holds significant potential to reduce the environmental impact of food production, and its methods could contribute to a sustainable intensification of agricultural systems in the future (Lorenz and Lal, 2016).

The soil is a complex, heterogeneous system characterized by closely interrelated physical, chemical, mineralogical, and pedological properties, as well as individual characteristics. Consequently, its management should encompass all of these aspects. Conversely, the absence of such a mechanism can result in irreversible errors, as previously highlighted by Delgado and Gómez (2024). The efficient and sustainable use of soils is possible only by determining, using appropriate methods, the limiting factors that affect productivity and the properties affecting sustainability. The conventional approach has been predicated on analysis of specific properties and their compatibility with various criteria, employing designated tools for each property and measuring positive or negative correlations among these properties. Despite the development of fertilization programs based on analysis, this practice, in its current form, is inadequate to maintain or improve plant productivity and production. Soil is an intricate system characterized by numerous sources of variability. It is imperative to recognize that evaluating soil solely in terms of nutrient content, while beneficial, is an incomplete assessment that fails to guarantee productivity or sustainability.

The study's assessment of physical and chemical properties revealed variability resulting from dynamic interactions between human activities, such as fertilizer management, and environmental factors, such as the degree of soil development and leaching processes. Sandy clay loam, loam, sandy loam, and clay loam were the texture classes identified in the medium-fine and coarse texture groups of soils used in this study. The FC and AWC values of catchment soils are generally low due to their coarse texture. These results are also consistent with those reported by Mihalikova *et al.* (2016). Furthermore, a range of microbial-response parameters, including microbial biomass carbon, basal respiration, and metabolic quotient, are frequently used to assess the stabilization of soil OM, soil aggregation, and soil quality, and to monitor soil formation (Ashraf *et al.*, 2022). One of the most significant factors in this regard is microbial activity involved in acquiring energy and nutrients, including OM (organic carbon) and elements such as N, P, K, Ca, and Mg. In the present study, MBC demonstrated the greatest variability among the selected biological indicators. Electrical conductivity values were low, indicating 'non-saline' soils, whereas pH values in soil samples ranged from strongly acidic to mildly acidic. Furthermore, the average $CaCO_3$ content in each soil sample is less than 2%. As a result, the $CaCO_3$ level of the soils in the research area was categorized as 'low', while their OM content ranged from 'low' to 'high'. Additionally, several soil samples had adequate micronutrient concentrations, as indicated by their mean concentrations. Elevated mobility of micronutrients was observed in the soil solution at acidic pH. Additionally, it might be related to the very low concentrations of leachable calcium carbonate. Similar results were found in several studies conducted in the Black Sea region, including Sayğın *et al.* (2017).

The XGBoost model demonstrated higher simulation accuracy and greater generalization capability (Liu *et al.*, 2023). Various nonlinear models have been widely used for estimation, including extreme gradient boosting machine (XGBoost), back-propagation neural network (BPNN), and random forest (RF) (Huang *et al.*, 2023, Yang *et al.*, 2021). In terms of their estimation ability, the three estimation models could be ranked as follows: XGBoost > BPNN > RF. In particular, CWT–XGBoost, combined with the six feature-variable algorithms, provided excellent results, with its RPD ranging from 2.14 to 2.78 (Yang *et al.*, 2023). In terms of model selection, Yu *et al.* (2024) provided support for the advantage of XGBoost in capturing complex relationships and working with heterogeneous data. Zhao *et al.* (2024) reported that XGBoost provided high accuracy and stability in soil salinity estimation, demonstrating particularly strong performance when combining spectral and texture data. However, they noted that overfitting, interpretability, and regional generalizability should be considered cautiously, and that testing in different geographies would be a key area for future research.

In this study, nitrogen (N), potassium (K), OM, phosphorus (P), and pH were identified as the most influential variables in model predictions (Figure 10). These variables are known to be key determinants of soil fertility and ecosystem functioning. Nitrogen and phosphorus are essential macronutrients vital for plant growth and microbial activity, while potassium plays a critical role in osmotic balance and enzymatic processes. OM improves soil physical structure, water holding capacity, and nutrient cycling, while pH directly affects plant uptake of nutrients and microbial activity. These findings are consistent with previous studies that conducted similar machine-learning-based soil-quality assessments (Lee and Lee, 2024; Trontelj ml and Chanbersi, 2021), demonstrating that these variables are reliable indicators of the chemical and biological dimensions of soil quality. All nutrients, except nitrogen, largely originate from the lithosphere and enter soil nutrient cycles via various soil-forming processes, while nitrogen enters the soil through precipitation, OM mineralization, biological fixation, and addition as fertilizers. The latter is required, above all, in agricultural systems where the intensive plant production far exceeds the amount of vital nutrients stored in the soil. Generally, the limiting factor for plant production is nitrogen, followed by phosphorus and potassium. These vital elements are normally added to the soil in synthetic form. Since their concentrations and availability can change over small-time scales, they are considered dynamic indicators of soil quality.

The gain metric, which quantifies the relative contribution of each feature to the model, was used to assess the importance of XGBoost features (Liang *et al.*, 2021). It appears that the XGBoost algorithm can forecast local soil variability and account for the

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

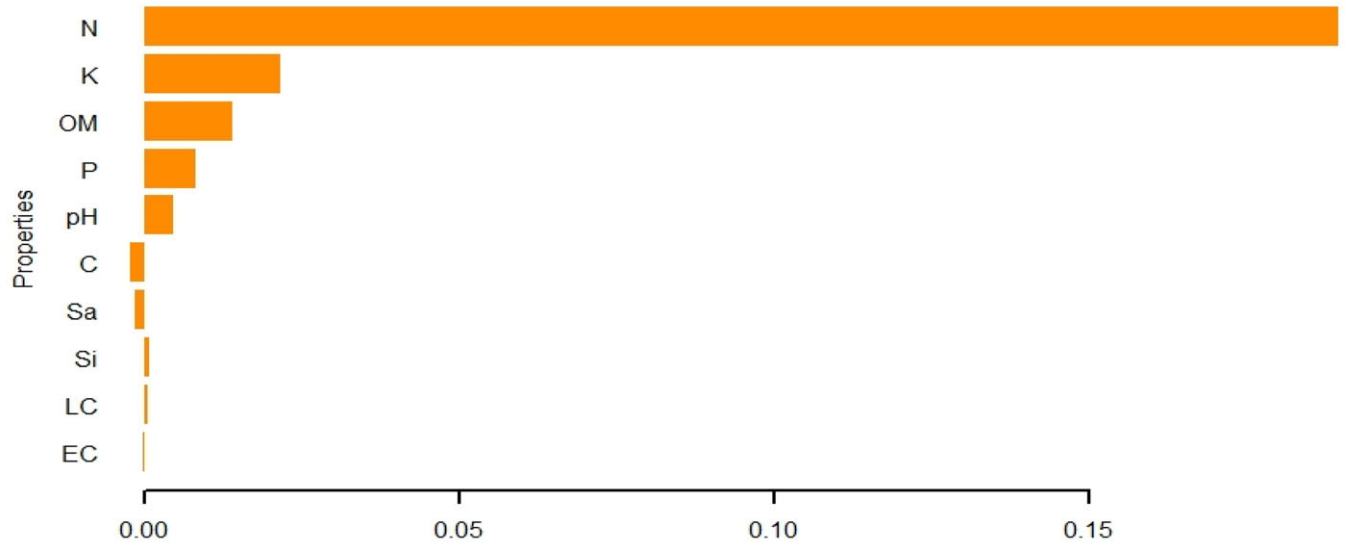


Figure 8. Importance of soil properties OM: organic matter, LC: Lime content, Si: Silt, C: Clay, Sa: Sand, EC: electrical conductivity.

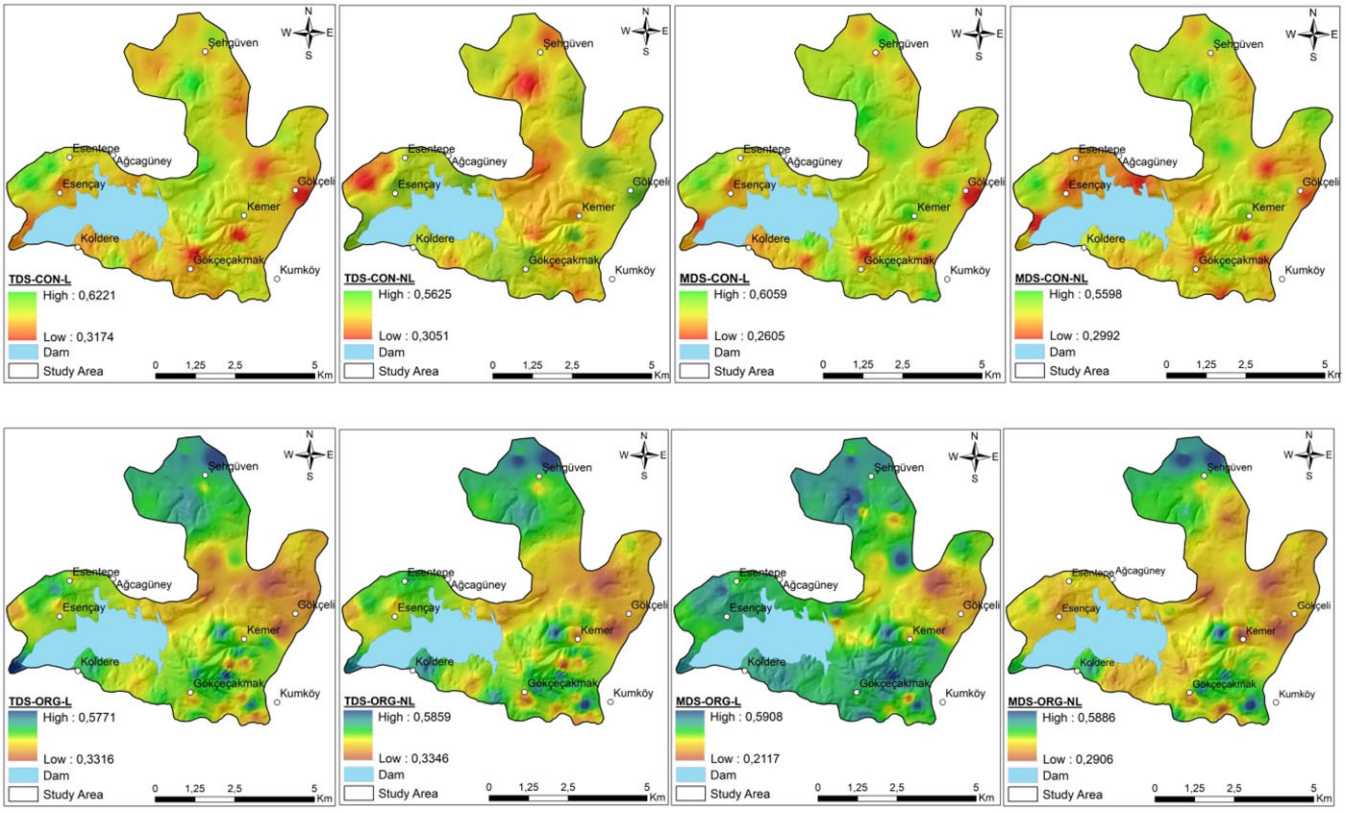


Figure 9. Spatial distribution maps of the Observed (OBS) and Predicted (PRE) values of SQI for total data set (TDS) and minimum data set (MDS) with Linear (L) and Non-Linear (NL) approaches.

impact of aggregation. Like other models, XGBoost is heavily reliant on the resolution and quality of the input data, yet it was only partially successful in modelling variability (Zeitfogel *et al.*, 2023). The high significance of these variables in the XGBoost model may be due to the nonlinear and interactive relationships between these parameters and soil fertility indicators. However, as the relative importance of features may vary depending on soil type, local weather conditions, and land use history, it will be important for future studies to validate these findings across different regions and with larger data sets. In this study, various approaches were tested to improve the model's accuracy. The aim of this study was to determine the feasibility of prediction using

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

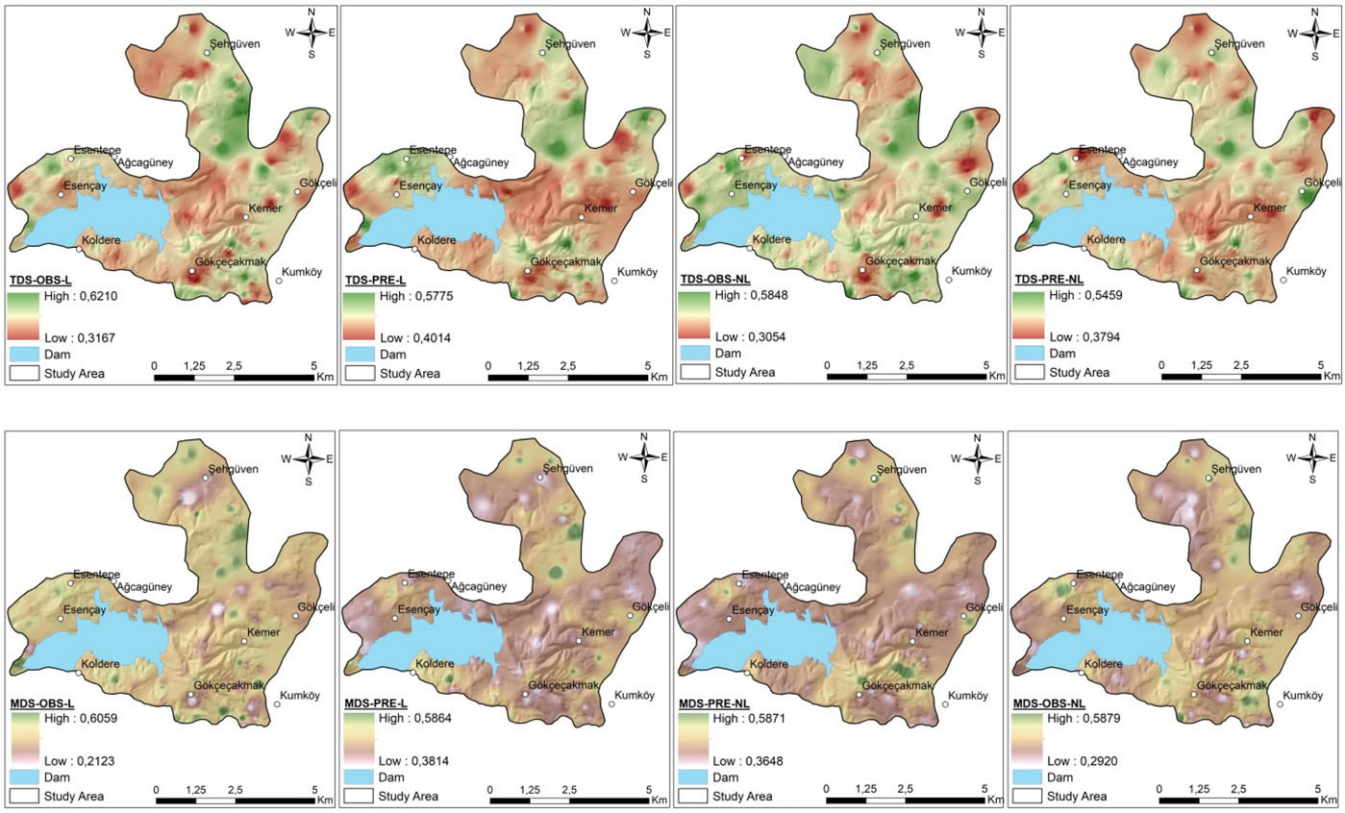


Figure 10. Importance of soil properties. OM: organic matter, LC: Lime content, Si: Silt, C: Clay, Sa: Sand, EC: electrical conductivity.

basic soil properties readily measured in the laboratory. However, the predictive accuracy is expected to increase by evaluating different parameters, especially those from remote sensing. The outcome shows that XGBoost successfully predicted soil attributes using remote sensing indices, indicating that machine learning techniques can be improved by using remote sensing data to enable field-to-regional-scale spatial prediction of soil attributes (Sahabiev *et al.*, 2021). Furthermore, Song *et al.* (2018) noted that regional-scale forecasts of soil property variability are significantly enhanced when multispectral and hyperspectral remote sensing data are combined with machine-learning models.

Conclusions

In this study, soil quality in areas where hazelnuts are produced under conventional and organic agriculture was evaluated. Additionally, the predictability of soil quality was assessed using the XGBoost algorithm. Soil quality in conventional agriculture was found to be higher than in organic farming areas. With the XGBoost algorithm, soil quality can be estimated with approximately 60% accuracy. Additionally, macro elements and OM were considered important features in estimating soil quality. Although soil quality was determined to be higher under conventional agricultural practices in this study, the effects of organic fertilizer applications will persist for many years, allowing the study to continue. In this study, basic soil properties were evaluated to facilitate application in model evaluation. In the study, it can be considered a parameter that limits the accuracy of the model. In a subsequent study, the model's predictive performance will be evaluated across different soil properties. Additionally, the use of only soil analyses in the study can be considered a limiting factor for the model. It is planned to carry out more comprehensive studies using reflectance's using remote sensing technologies.

Data availability statement. Data available on request.

Acknowledgements. An AI tool was used to improve the language.

Funding statement. This research has not found any sources.

Author contributions. O. DENGİZ: Conceptualization, Methodology, Sample collection, Data processing, Writing, Supervision, S. PACCi: Sample collection and analysis, Data processing, Writing P. ALABOZ: Conceptualization, Methodology, Data processing, Writing, N. SUBAN: Sample collection and analysis, Data processing, Writing. All authors read and approved the final manuscript.

Competing interests. The authors declare no competing interests.

References

Alaboz P and Dengiz O (2023) The least limiting water range to estimate soil water content using random forest integrated with GIS and geostatistical approaches. *Journal of Agricultural Sciences* **29**, 933–946

Alaboz P and Işıldar AA (2019) Evaluation of pedotransfer functions (PTFs) for some soil physical properties. *Turkish Journal of Science and Engineering* **1**, 28–34. (In Turkish).

Alaboz P, Odabas MS and Dengiz O (2023) Soil quality assessment based on machine learning approach for cultivated lands in semi-humid environmental condition part of Black Sea region. *Archives of Agronomy and Soil Science* **69**, 3514–3532.

Anderson TH and Domsch KHA (1978) Physiological method for the quantitative measurement of microbial biomass in soils. *Soil Biology Biochemistry* **10**, 215–221.

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

Andrews SS, Karlen DL and Cambardella CA (2004) The soil management assessment framework: A quantitative soil quality evaluation method. *Soil Science Society of America Journal* **68**, 1945–1962.

Ashraf MN, Waqas MA and Rahman S (2022) Microbial metabolic quotient is a dynamic indicator of soil health: Trends, implications and perspectives. *Eurasian Soil Science* **55**, 1794–1803.

Bastida F, Moreno J L, Hernández T and García C (2006) Microbiological degradation index of soils in a semiarid climate. *Soil Biology and Biochemistry* **38**, 3463–3473.

Bölük E (2016) *According to Erinç Climate Classification Turkish Climate*, Ministry of Forestry and Water Management General Directorate of Meteorology, Ankara.

Borůvka L, Vacek O and Jehlička J (2005) Principal component analysis as a tool to indicate the origin of potentially toxic elements in soils. *Geoderma* **128**, 289–300.

Bouyoucous GJ (1951) A recalibration of the hydrometer for making mechanical analysis of soil. *Agronomy Journal* **43**, 434–438.

Bozoğlu M, Başer U, Kilic Topuz B and Alhas Eroglu N (2019) An overview of hazelnut markets and policy in Turkey. *KSU Journal of Agriculture and Nature* **22**.

Bremner JM and Mulvaney CS (1982) Nitrogen total. In Page AL, Miller RH and Keeney DR (eds), *Methods of Soil Analysis, Part II: Chemical and Microbiological Properties*. Madison (WI): ASA, 595–625.

Bünemann EK, Bongiorno G, Bai Z, Creamer RE, De Deyn G, De Goede R and Brussaard L (2018) Soil quality–A critical review. *Soil biology and biochemistry* **120**, 105–125.

Chen T and Guestrin C (2016) Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H and Zhou T (2015) Xgboost: Extreme gradient boosting. *R Package Version 0.4-2* **1**, 1–4.

Delgado A and Gómez JA (2024). The soil: Physical, chemical, and biological properties. In Principles of agronomy for sustainable agriculture (pp. 15-30). Cham: Springer International Publishing.

Dornbush ME and von Haden AC (2017) Intensified agroecosystems and their effects on soil biodiversity and soil functions. In *Soil Health and Intensification of Agroecosytems*. Academic Press, 173–193.

FAO (2019) Food and agricultural commodities production database, food and agriculture organization of the United Nations. http://faostat.fao.org/faostat/en/#data/QC (Access 22 November 2021)

Fernández MP, Keshavarzi A, Rodrigo-Comino J, Schnabel S, Contador JFL, Gutiérrez ÁG ... and Cerdà A (2020). Developing scoring functions to assess soil quality at a regional scale in rangelands of SW Spain. Revista Brasileira de Ciência do Solo, 44, e0200090.

Gattinger A, Muller A, Haeni M, Skinner C, Fliessbach A, Buchmann N and Niggli U (2012) Enhanced topsoil carbon stocks under organic farming. *Proceedings of the National Academy of Sciences* **109**, 18226–18231.

Gomiero T, Pimentel D and Paoletti MG (2011) Environmental impact of different agricultural management practices: Conventional vs. organic agriculture. *Critical Reviews in Plant Sciences* **30**, 95–124.

Hazelton P and Murphy B (2016) *Interpreting Soil Test Results: What do all the Numbers Mean?* CSIRO publishing.

Hikouei IS, Eshleman KN, Saharjo BH, Graham LL, Applegate G and Cochrane MA (2023) Using machine learning algorithms to predict groundwater levels in Indonesian tropical peatlands. *Science of the Total Environment* **857**, 159701.

Huang X, Wang X, Baishan K and An B (2023) Hyperspectral estimation of soil organic carbon content based on continuous wavelet transform and successive projection algorithm in arid area of Xinjiang, China. *Sustainability* **15**, 2587.

Isermayer H (1952) Eine Einfache Methode zur Bestimmung der Bodenatmung und der Karbonate im Böden Z. *Pflanzenaehr Bodenkd* **5**, 56–60.

Jenks GF (1967) The data model concept in statistical mapping. *International Yearbook of Cartography* **7**, 186–190.

Kelle AC and Yüce H (2022) Classification of DoS attacks in MQTT network with machine learning and interpretation of the model with SHAP. *JournalMM* **3**, 50–62

Kemper WD and Rosenau RC (1986) Aggregate stability and size distribution. *Methods of Soil Analysis: Part 1 Physical and Mineralogical Methods* **5**, 425–442.

Lal R and Elliot W (1994) Erodibility and erosivity. *Soil Erosion Research Methods* **1**, 181–208.

Lee W and Lee J (2024) Tree-based modeling for large-scale management in agriculture: Explaining organic matter content in soil. *Applied Sciences* **14**, 1811.

Liang H, Jiang K, Yan TA and Chen GH (2021). XGBoost: an optimal machine learning model with just structural features to discover MOF adsorbents of Xe/Kr. *ACS omega*, **6**(13), 9066-9076.

Lindsay WL and Norvell W (1978) Development of a DTPA soil test for zinc, iron, manganese, and copper. *Soil Science Society of America Journal* **42**, 421–428.

Liu W, Li R, Wu T, Shi X, Zhao L, Wu X and Wang S (2023) Simulation of soil thermal conductivity based on different schemes: An empirical comparison of 13 models. *International Journal of Thermal Sciences* **190**, 108301.

Lorenz K and Lal R (2016) Environmental impact of organic agriculture. *Advances in agronomy* **139**, 99–152.

Lynch DH (2014) Sustaining soil organic carbon, soil quality and soil health in organic field crop management systems. In Martin RC and MacRae R (eds), *Managing Energy, Nutrients and Pests in Organic Field Crops*. CRC Press, 107–132.

Maurya S, Abraham JS, Somasundaram S, Toteja R, Gupta R and Makhija S (2020) Indicators for assessment of soil quality: A mini-review. *Environmental Monitoring and Assessment* **192**, 1–22.

Miháliková M and Dengiz O (2019) Towards more effective irrigation water usage by employing land suitability assessment for various irrigation techniques. *Irrigation and Drainage* **68**, 617–628.

Miháliková M, Özyazıcı MA and Dengiz O (2016) Mapping soil water retention on agricultural lands in central and eastern parts of the Black Sea Region in Turkey. *Journal of Irrigation and Drainage Engineering* **142**, 05016008.

Momeni E, He B, Abdi Y and Armaghani DJ (2023) Novel hybrid XGBoost model to forecast soil shear strength based on some soil index tests. *CMES-Computer Modeling in Engineering & Sciences* **136**(3), 2527–2550.

Nabiollahi K, Taghizadeh-Mehrjardi R, Kerry R and Moradian S (2017) Assessment of soil quality indices for salt-affected agricultural land in Kurdistan Province, Iran. *Ecological Indicators* **83**, 482–494.

Nelson DW and Sommers LE (1982) Total carbon, organic carbon and organic matter. In Page AL, Miller RH and Keeney DR (eds), *Methods of Soil Analysis, Part 2*, vol 9. 539–579.

Olsen SR, Cole CV, Watanabe FS and Dean LA (1954) *Estimation of Available Phosphorus in Soils by Extraction with Sodium Bicarbonate*. US Department of Agriculture, 939.

Philip Robertson G, Gross KL, Hamilton SK, Landis DA, Schmidt TM, Snapp SS and Swinton SM (2014) Farming for ecosystem services: An ecological approach to production agriculture. *BioScience* **64**, 404–415.

Qi Y, Darilek JL, Huang B, Zhao Y, Sun W and Gu Z (2009) Evalu-ating soil quality indices in an agricultural region of Jiangsu Pro-vince, China. *Geoderma*, **149**, 325–334.

Reganold JP and Wachter JM (2016) Organic agriculture in the twenty-first century. *Nature plants* **2**, 1–8.

Saaty TL (1980) *The Analytic Hierarchy Process*. New York: McGrawHill Inc.

Saaty TL (1993) What is relative measurement? The ratio scale phantom. *Mathematical and Computer Modelling* **17**, 1–12.

Sahabiev I, Smirnova E and Giniyatullin K (2021) Spatial prediction of agrochemical properties on the scale of a single feld using machine learning methods based on remote sensing data. *Agronomy* **11**. https://doi.org/10.3390/agronomy11112266

Saxton KE, Rawls W, Romberger JS and Papendick RI (1986) Estimating generalized soil-water characteristics from texture. *Soil Science Society of America Journal* **50**, 1031–1036.

Saygın F, Gürsoy FE, Demirağ Turan İ and Dengiz O (2017) Determination of physical, chemical and productivity characteristics of tea-growing soils on the eastern side of the Çataklı stream basin. *Journal of Agricultural Research of Türkiye* **4**, 143–154.

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press

Saygın F, Şavşatlı Y, Dengiz O, Yazıcı K, Namlı A, Karataş A, Şenol ND, Akça MO, Pacci S, Karapıçak B, Ay A and Demirkaya S (2023) Soil quality assessment based on hybrid computational approach with spatial multi-criteria analysis and geographical information system for sustainable tea cultivation. *The Journal of Agricultural Science* **161**, 187–204. https://doi.org/10.1017/S0021859623000138

Şenol H, Alaboz P, Demir S and Dengiz O (2020) Computational intelligence applied to soil quality index using GIS and geostatistical approaches in semiarid ecosystem. *Arabian Journal of Geosciences* **13**, 1–20.

Soil Survey Staff (1993) *Soil Survey Manuel.* USDA Handbook. No: 18. Washington DC: US Government Printing Office, NRCS-Soil Conservation Service.

Song YQ, Zhao X, Su HY, Li B, Hu YM and Cui XS (2018). Predicting spatial variations in soil nutrients with hyperspectral remote sensing at regional scale. *Sensors (Switzerland)* **18**. https://doi.org/10.3390/s18093086

Trontelj Ml J and Chambers O (2021) Machine learning strategy for soil nutrients prediction using spectroscopic method. *Sensors* **21**, 4208.

Vasconcellos RDF, Bonfim JA, Andreote FD, Mendes LW, Baretta D and Cardoso EJBN (2013) Microbiological indicators of soil quality in a riparian forest recovery gradient. *Ecological Engineering* **53**, 313–320.

Willer H, Meier C, Schlatter B, Travnicek J (2022) The state and evolution of organic fruit and vegetables: production and market at a world scale. In *XXXI International Horticultural Congress (IHC2022): XII International Symposium on Banana: Celebrating Banana Organic 1367*, 1–12.

Yang C, Feng M, Son, L, Wang C, Yang W, Xie Y, Ji B, Xiao L, Zhang M, Song X and Saleem M (2021) Study on hyperspectral estimation model of soil organic carbon content in the wheat field under different water treatments. *Scientific Reports* **11**, 18582.

Yang J, Li X and Ma X (2023). Improving the Accuracy of Soil Organic Carbon Estimation: CWT-Random Frog-XGBoost as a prerequisite technique for in situ hyperspectral analysis. *Remote Sensing*, **15**(22), 5294.

Yu W, Zhou W, Wang T, Xiao J, Peng Y, Li H and Li Y (2024) Significant improvement in soil organic carbon estimation using data-driven machine learning based on habitat patches. *Remote Sensing* **16**, 688.

Zeitfogel H, Feigl M and Schulz K (2023) Soil information on a regional scale: Two machine learning based approaches for predicting saturated hydraulic conductivity. *Available at SSRN 4143185.*

Zhao W, Li Z, Li H, Li X and Yang P (2024) Soil salinity prediction in an arid area based on long time-series multispectral imaging. *Agriculture* **14**, 1539.

Zhou Y, Zhang R, Wang S and Wang F (2018) Feature selection method based on high-resolution remote sensing images and the effect of sensitive features on classification accuracy. *Sensors* **18**, 2013. https://doi.org/10.3390/s18072013.

https://doi.org/10.1017/S0021859625100476 Published online by Cambridge University Press