Earth Science Informatics (2023) 16:3811–3838
https://doi.org/10.1007/s12145-023-01125-1

RESEARCH

Exploring soil property spatial patterns in a small grazed catchment using machine learning

Jesús Barrena-González[1] · V. Anthony Gabourel-Landaverde[1] · Jorge Mora[2] · J. Francisco Lavado Contador[1] · Manuel Pulido Fernández[1]

Received: 6 September 2023 / Accepted: 11 October 2023 / Published online: 26 October 2023

Abstract

Acquiring comprehensive insights into soil properties at various spatial scales is paramount for effective land management, especially within small catchment areas that often serve as vital pastured landscapes. These regions, characterized by the intricate interplay of agroforestry systems and livestock grazing, face a pressing challenge: mitigating soil degradation while optimizing land productivity. This study aimed to analyze the spatial distribution of eight topsoil (0–5 cm) properties (clay, silt, sand, pH, cation exchange capacity, available potassium, total nitrogen, and soil organic matter) in a small grazed catchment. Four machine learning algorithms—Random Forest (RF), Support Vector Machines (SVM), Cubist, and K-Nearest Neighbors (kNN)—were used. The Boruta algorithm was employed to reduce the dimensionality of environmental covariates. The model's accuracy was assessed using the Concordance Correlation Coefficient (CCC) and Root Mean Square Error (RMSE). Additionally, uncertainty in predicted maps was quantified and assessed. The results revealed variations in predictive model performance for soil properties. Specifically, kNN excelled for clay, silt, and sand content, while RF performed well for soil pH, CEC, and TN. Cubist and SVM achieved accuracy in predicting AK and SOM, respectively. Clay, silt, CEC, and TN yielded favourable predictions, closely aligning with observations. Conversely, sand content, soil pH, AK, and SOM predictions were slightly less accurate, highlighting areas for improvement. Boruta algorithm streamlined covariate selection, reducing 23 covariates to 10 for clay and 4 for soil pH and AK prediction, enhancing model efficiency. Our study revealed spatial uncertainty patterns mirroring property distributions, with higher uncertainty in areas with elevated content. Model accuracy varied by confidence levels, performing best at intermediate levels and showing increased uncertainty at extremes. These findings offer insights into model capabilities and guide future research in soil property prediction. In conclusion, these results urge more research in small watersheds for soil and territorial management.

Keywords Environmental covariate reduction · Predictive modelling · Spatial variability · Uncertainty assessment

Introduction

Catchments play a vital role in assessing various physical and biological ecosystem processes and variables. These hydrological catchments provide a comprehensive context within which numerous distinct ecological and hydrological processes operate at local scales (Feng et al. 2013). Among the many process-influencing as well as context-related environmental factors, the soil properties can exhibit considerable variation within catchments, influenced by factors such as land use, topography, and geology (Terefe et al. 2020). Understanding the distribution of soil properties over these areas is crucial for effective land management and decision-making. Accurately mapping soil properties in small catchments is, hence, vital for addressing localized environmental challenges, designing appropriate land management strategies, and promoting sustainable land use practices (Tang et al. 2015). Soil mapping at small catchment scales can provide valuable information when assessing soil health, erosion-prone areas, nutrient leaching or runoff potential, as some of the many issues of interest at these spatial scales

Communicated by: H. Babaie

✉ Jesús Barrena-González
jesusbarrena@unex.es

[1] Instituto Universitario de Investigación para el Desarrollo Territorial Sostenible (INTERRA), Grupo de Investigación GeoAmbiental, Universidad de Extremadura, 10071 Cáceres, Spain

[2] Center of Applied Ecology and Sustainability (CAPES), Pontificia Universidad Católica de Chile (PUC), Avda. Vicuña Mackenna 4860, 7820436 Santiago de Chile, Chile

(Altaf et al. 2014; Khosravi Aqdam et al. 2022; Pulleman et al. 2000).

Nowadays, harnessing the power of machine learning is pivotal for precise soil property mapping across various scales. These innovative techniques, as exemplified by Adeniyi et al. (2023), have become indispensable tools. By maximizing the use of collected data, these techniques enable the analysis of complex and non-linear patterns, as those of the soil properties and many other environmental-related aspects, that would otherwise be challenging to assess (Forkuor et al. 2017; Hastie et al. 2009). Machine learning allows for the efficient processing of large volumes of geospatial data, such as satellite imagery and topographic data, by integrating them with existing field data (Poggio et al. 2021). This integration facilitates accurate and detailed analysis of the influences of multiple factors on soil properties and identify complex correlations and relationships between variables, thereby enhancing the modeling capabilities and improving the accuracy of the resulting maps (Beguin et al. 2017; Khaledian & Miller 2020). Some of the most used algorithms for mapping soil properties, such as soil organic carbon stocks (Mishra et al. 2020), hydraulic conductivity (Araya & Ghezzehei 2019), pH (Xiao et al. 2023), or soil aggregate stability (Bouslihim et al. 2021), include regression trees, cubist, random forest (RF), gradient boosting machines (GBM), multivariate adaptive regression splines (MARS), and support vector machines (SVM). These machine learning algorithms have proven to be useful in predictive soil mapping and have been applied in various studies (Padarian et al. 2019).

However, it's imperative to quantify the uncertainty inherent in soil maps, as highlighted by Ramcharan et al. (2018) and Wadoux et al. (2020). When it comes to this topic, it's worth noting that not all machine learning algorithms provide built-in mechanisms for uncertainty quantification. Only select algorithms, such as the quantile regression forest (QRF) method (Poggio et al. 2021), offer uncertainty quantification. In such cases, complementary techniques like bootstrapping become indispensable. Bootstrapping, a statistical resampling method, offers a robust approach to estimate uncertainty in the context of machine learning algorithms. Bootstrapping offers accurate results without making assumptions about data distribution and allows flexibility for different data types and models, providing estimation in confidence intervals despite its computational cost for large datasets (Malone et al. 2017; Szatmári & Pásztor 2019).

Despite the abundance of published literature and the growing use of machine learning algorithms for soil property mapping at different scales (Behrens et al. 2018), there is still a lack of knowledge on the spatial distribution of soil properties for certain areas at near-detail scales, such as the small catchments. This is particularly noteworthy for environmental contexts, as those of the Mediterranean catchments, where different pressures, including climate change, are leading changes in land use and management, fostering soil degradation and, therefore, compromising the sustainability and resilience of the whole soil system (Montanarella et al. 2016). Therefore, even when studies exist that focuses on soil property mapping, as soil thickness assessed in agricultural catchments (Li et al. 2017), CO_2 emissions over large catchments (Bailey et al. 2009), or soil erosion quantification (Fitria & Kurniawan 2021; Wang et al. 2022); the cartography of soil properties at the scale of small catchments in the Mediterranean region has received limited attention, despite its significant importance. So far, several studies have been carried out in the catchments located within agroforestry areas of Extremadura. These studies have focused on examining the spatial distribution of soil moisture and, more importantly, investigating the soil degradation issues that these environments suffer from due to inadequate soil management practices (Alfonso-Torreño et al. 2021; Gómez Gutiérrez et al. 2009; Lavado Contador et al. 2006). This situation underscores the urgent need to acquire a more profound insight into the spatial distribution of soil properties in these degraded environments and to better comprehend the environmental determinants influencing their distribution. However, it is worth noting that previous research addressing the spatial distribution of other soil properties, such as total nitrogen (TN), P (phosphorus), available potassium (AK), or SOM (%) (SOM), is currently limited to a regional level.

The scarcity of research in this area has resulted in a significant gap in our understanding of the spatial distribution and variability of soils within this specific context. Considering the critical role of soils in Mediterranean ecosystems and their profound influence on the sustainability of agro-silvopastoral systems, which are particularly abundant in the southern Mediterranean region, there is a pressing need for obtaining accurate and detailed soil property maps in these areas. In regions such as Extremadura, a substantial part of the regional surface is occupied by small agroforestry catchments with livestock farming, which is also a common feature in various other Mediterranean countries. Understanding the spatial behavior of soil properties in these intricate and ecologically significant environments is paramount for addressing soil conservation, enhancing productivity, and developing tailored management strategies that promote sustainability.

Given the limited research conducted in the mentioned agroforestry systems, especially regarding detailed soil property mapping at the scale of small catchments, this study represents a valuable effort to bridge this significant knowledge gap. It aims to provide comprehensive maps of eight essential soil properties within the topsoil layer (0–5 cm) in a specific study area encompassing a small catchment within an agroforestry system situated in Extremadura, Spain. To achieve this, four distinct and state-of-the-art machine learning algorithms, namely k-Nearest Neighbors (KNN), Random Forest (RF), Cubist, and Support Vector Machines

(SVM), were deployed. The specific objectives of the study were: a) to evaluate the accuracy of each of the proposed machine learning algorithms in predicting soil properties, b) to quantify the uncertainty associated with the predictive maps, and c) to identify the key environmental covariates influencing the predicted models.

Materials and methods

Study area, soil survey and environmental covariate analysis

This study was conducted in a small catchment located in the southwestern region of Spain known as Extremadura (Fig. 1a). The catchment belongs to the agroforestry system called "dehesa" (Fig. 1b-1c). Both the catchment and the entire farm follow a livestock approach under a conventional management system, without designed rotation plans and with high livestock stocking rates (1.59 Animal Units ha^{-1}) (Pulido et al. 2018). The catchment covers a total area of 99.5 hectares and exhibits typical topographic characteristics of the Trujillano-Cacereña Peneplain where it is located. The average slope is 8%, although the valley bottoms can be completely flat, while the steeper slopes can reach a 12% slope. The predominant soils are classified as Cambisols and Leptosols, with some Regosols found in the valley bottoms (IUSS Working Group WRB 2015). These soils are generally shallow and have low nutrient and organic matter concentrations. The prevailing climate in this area is Mediterranean continentalized (Csa) (Peel et al. 2007), characterized by mild winters with average temperatures above 0 °C and hot summers with average temperatures above 22 °C. The average annual temperature is 16 °C. Rainfall is scarce and mainly concentrated in autumn and spring, with an average annual rainfall in this catchment of 513 mm.

The soil properties considered for mapping were clay, silt, sand percentage (%) pH, CEC (cmol kg^{-1}), AK (cmol kg^{-1}), TN (%), and soil organic matter (SOM). A comprehensive representation of the catchment heterogeneity was ensured through simple field random sampling by a man. These samples were subsequently transported to the laboratory, where they underwent meticulous analysis following the processes

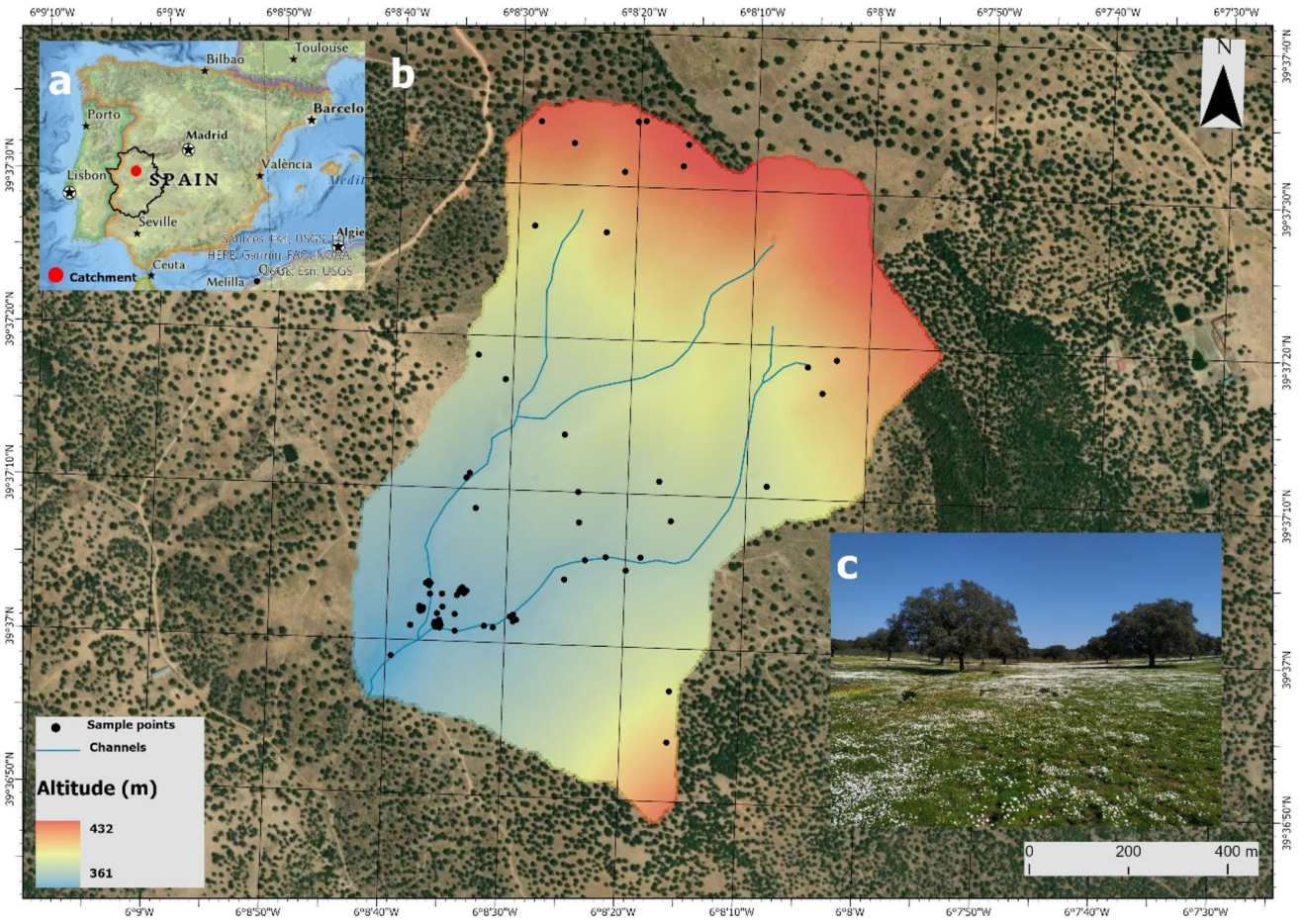

Fig. 1 Location of the Extremadura region within the Iberian Peninsula (**a**). Study area and sampling points (**b**) and photography of the catchment (**c**)

of drying and sieving. In line with prior research, sampling was specifically carried out within the 0–5 cm depth interval. Other studies in the shallow soils of Extremadura's rangelands have consistently revealed that this top layer (0–5 cm) holds the highest concentration of essential nutrients (Pulido et al. 2018). Therefore, focusing on this interval is paramount for capturing changes influenced by factors such as tillage, fertilizer application, or crop rotation. The total number of soil samples for topsoil layer was 80.

In this study, 23 environmental covariates were calculated for the digital mapping of different soil properties (Table 1). Using the digital elevation model (DEM) and satellite images, 19 geomorphological indices and 3 vegetation indices were computed. The DEM was downloaded from the National Center of Geographic Information (https://centrodedescargas.cnig.es/) with a spatial resolution of 5 m x 5 m. The SAGA (System for Automated Geoscientific Analyses) software (Gerstoft 2001) was then used to calculate various topographic and geomorphological parameters. Sentinel satellite images (Level-2A product) were utilized for the calculation of vegetation indices, and the Google Earth Engine platform was employed for data processing. The mean values of the Normalized Difference Vegetation Index (NDVI), Soil-adjusted Vegetation Index (SAVI), and Enhanced vegetation index (EVI) for the past 5 years were calculated for the study area. Additionally, to harmonize the spatial resolution with other variables for seamless integration into the digital mapping of soil properties, the satellite image underwent resampling using the ArcGIS Pro's resample tool. It's essential to note that while this resampling operation did not introduce additional information, it ensured that all variables shared the same spatial resolution, a critical requirement for the mapping process.

Covariates selection process

Before executing the predictive models, a preliminary environmental covariates selection process was conducted using the Boruta algorithm implemented with the *Boruta* package in R (Kursa & Rudnicki 2010). This algorithm is particularly well-suited for feature selection in predictive modeling tasks. Boruta operates by comparing the importance of each predictor variable to that of a shadow variable, essentially a randomized version of the original variable. Variables that significantly outperform their shadow counterparts in terms of predictive power are retained, while those that do not meet this criterion are discarded. This approach not only ensures that only the most informative covariates, those that genuinely contribute to predicting the target property, are included in the subsequent modeling process but also indirectly helps mitigate multicollinearity by eliminating redundant or highly correlated variables.

Table 1 List of environmental covariates used in the predictive models

Environmental Covariates	Abbreviation	Data source
Altitude	Altitude	DEM
Aspect	Aspect	DEM
Slope	Slope	DEM
Profile Curvature	Profile Curvature	DEM
Plan Curvature	Plan Curvature	DEM
Maximum Curvature	Maximum Curvature	DEM
Minimum Curvature	Minimum Curvature	DEM
Multiresolution Index of Valley Bottom Flatness	MRVBF	DEM
Multiresolution Index of Ridge Top Flatness	MRRTF	DEM
LS Factor	LS Factor	DEM
Terrain Ruggedness Index	TRI	DEM
Valley Depth	Valley Depth	DEM
Topographic Wetness Index	TWI	DEM
Topographic Position Index	TPI	DEM
Total Catchment Area	TCA	DEM
Relative Slope Position	RSP	DEM
Convexity	Convexity	DEM
Convergence Index	Convergence Index	DEM
Channel Network Distance	CND	DEM
Analytical Hillshading	AH	DEM
Normalized Difference Vegetation Index	NDVI	Sentinel-2A
Soil-adjusted Vegetation Index	SAVI	Sentinel-2A
Enhanced Vegetation Index	EVI	Sentinel-2A

Statistical analysis

Statistical parameters like minimum, maximum, mean, and standard deviation were employed to characterize the available dataset for spatial analysis of the study properties. The computations were conducted using RStudio software. Likewise, ArcGIS Pro software was used to calculate the statistical parameters for the various final maps generated for each property.

Machine learning models

Four widely known machine learning algorithms, namely Random Forest (RF), Support Vector Machines (SVM), Cubist, and k-Nearest Neighbors (kNN), were utilized in this study to spatially predict eight soil properties. The selection of these algorithms was based on their effectiveness with small datasets. RF mitigates overfitting in small datasets by employing an ensemble of decision trees. These trees are constructed based on different subsets of the data, reducing the risk of overfitting (Aqdam et al. 2022). RF further enhances its performance by randomly selecting features at each node, effectively addressing dimensionality issues in small datasets while prioritizing informative features. This adaptability enables RF to capture intricate relationships even when working with limited data. SVM leverage wide margins to identify the hyperplane that optimally separates classes while maximizing the margin (Li et al. 2009). In small datasets, SVM is adept at finding well-separating hyperplanes without overfitting, making it an excellent choice for delineating clear class boundaries. Moreover, SVM excels in low-dimensional spaces, making it effective for datasets with fewer features. Cubist introduces an advanced modeling approach well-suited for small datasets. It provides interpretable coefficient estimates, facilitating a deeper understanding of the relationships between predictors and soil properties (Quinlan 1992). This interpretability is particularly valuable in scenarios with limited data. kNN, a non-parametric algorithm, does not assume a specific data distribution, making it flexible and suitable for small datasets that do not meet the assumptions of other models. Additionally, kNN makes decisions based on the nearest instances, which can be advantageous in small datasets by leveraging all available information and capturing local patterns effectively. All models were implemented using the *caret* package in RStudio.

Model deployment and uncertainty quantification

In all the models developed to predict the soil properties of interest in this study, the dataset underwent a random division into calibration and validation subsets. The training subset, constituting 90% of the data, was utilized to construct multiple models through resampling (bootstrap) with 25 iterations (Malone et al. 2017; Sharififar 2022) following a preliminary covariate selection using the Boruta algorithm. In this case, 70 samples were assigned to the training set, while 10 were designated for validation. In each iteration, a random subset of the training set was sampled with replacement, generating a diverse set of models trained on different data subsets. This approach effectively addresses the challenge of limited data availability in small datasets and mitigates the risk of overfitting.

Furthermore, a randomized search for optimal hyperparameters was performed in each model (Table 2). Each model was fitted to the prediction formula using a selection of predictor variables, including vegetation indices, and a series of topographic and geomorphological parameters. Additionally, variable importance metrics were employed to evaluate the relative contribution of each predictor in predicting each soil property. The importance analysis allowed for the identification of the most relevant features in the prediction process and highlighted which topographic, geomorphological parameters or vegetation indices significantly influenced the studied soil properties.

After training the models, their performance was assessed using key goodness-of-fit metrics, namely the coefficient of determination (R^2), Lin's concordance correlation coefficient (CCC), and root mean squared error (RMSE). These metrics provided a comprehensive evaluation of the models' fit to both the training and validation datasets, enabling a robust measurement of accuracy and generalization capability. In determining the superior model, the focus was solely on achieving the highest CCC value and the lowest RMSE value, thus prioritizing models with the strongest explanatory power and the smallest prediction errors.

To quantify the uncertainty in spatial predictions of soil properties, a multi-step process was employed. Firstly, predictions of soil properties for the testing dataset were retained across 25 replications. Subsequently, the standard deviation of these predictions was calculated, incorporating the average Generalized Mean Squared Error (GMSE) across the bootstrapped models. The GMSE represents the overall variance between the predicted values and the observed values. The standard deviation was then multiplied by quantiles of the normal distribution to determine the upper and lower

Table 2 List of environmental covariates used in the predictive models

Model	Abbreviation	Hyperparameters
Random Forest	RF	mtry, splitrule, min.node.size
Cubist	none	committees, neighbors
Support Vector Machine	SVM	sigma, C
k-nearest neighbor	kNN	k

prediction limits for various confidence levels. Specifically, these limits were calculated as the predicted mean value ± z * standard deviation, where z corresponds to the quantiles. This approach provided prediction intervals for each confidence level (e.g., 90% confidence level).

To assess the quality of the uncertainty estimates, the proportion of measured soil values falling within these prediction intervals (referred to as Prediction Interval Coverage Probability or PICP) was calculated for each confidence level. A higher PICP indicates better-quality uncertainty estimates.

Results and discussions

Descriptive statistics

The data reveal a relative variability in the soil properties as present in the descriptive statistics (Table 3) analyzed in this study. For example, clay content exhibits remarkable variability, ranging from a minimum of 1.7% to a maximum of 22.96%. This variability underscores the significance of considering the spatial distribution of clay in the management of Extremadura's dehesas. Higher clay values can enhance water and nutrient retention in the soil. This finding aligns with previous research emphasizing the spatial variability of clay in various watersheds (Tesfahunegn et al. 2011; Wei et al. 2008). Similarly, the content of silt and sand, though with different variabilities, plays a crucial role in soil structure and water infiltration. Silt varies from a minimum of 19.37% to a maximum of 62.7%, while sand ranges from a minimum of 25.3% to a maximum of 71.35%. Wang et al. (2010) demonstrated that the granular and less cohesive nature of sand could lead to a more even distribution in the soil. In this study, the soil texture (loam) with a higher clay content slightly differs from the texture found in previous studies in similar environments where soils were characterized by higher sand and silt content, resulting in a loamy sand texture (Pulido-Fernández et al. 2013; Reyna-Bowen et al. 2020).

The soil pH exhibits a mean value of 5.48 with a CV of 19.12%, indicating relative stability and falling within the acidity range of soils in this environment due to their parent material primarily composed of sandstones and granites (Gazol et al. 2021; Schnabel et al. 2013). On the other hand, CEC content displays a wide range, ranging from 0.81 to 24.1 cmol kg^{-1}, showing significant variability in the soil's ability to retain and release cations, which may be linked to its relationship with clay content (Saidi et al. 2022; Seybold et al. 2005).

The AK spans a substantial range (0.03 – 3.51 cmol kg-1), as reflected by the CV (119%), indicating the high variability of this property in such environments (Pulido et al. 2017). This behavior may be influenced by historical management practices, as well as the spatial characteristics of vegetation and topography. The TN (%) content shows relatively low mean values (0.20%), which is a characteristic situation of the dehesas soils (Plieninger et al. 2003). Also, the TN (%) content shows a notable range, with values ranging from 0.02% to 0.65%. The CV for N is 55.9%, suggesting relatively high variability in the distribution of this nutrient. This could be due to areas of TN (%) accumulation resulting from animal resting areas or N scarcity on slopes due to leaching processes (Hassan-Vásquez et al. 2022; Lassaletta et al. 2021; Pulido-Fernández et al. 2013). Regarding SOM (%), it exhibits moderate mean values (3.34%) for this type of environment, suggesting that the soils in this study area have a certain amount of SOM compared to others in similar systems (Pulido-Fernández et al. 2013; Reyna-Bowen et al. 2020). However, the high CV (51.56%) indicates significant variability in organic matter concentration, which could influence soil fertility and water retention capacity (Simón et al. 2013).

Table 3 Descriptive statistics of soil properties included in the study

Property	Unit	Min	Max	Mean	CV (%)
Clay	%	1.70	22.96	15.02	39.60
Silt	%	19.37	62.70	38.62	30.54
Sand	%	25.30	71.35	46.37	21.12
Soil pH		4.68	7.00	5.48	19.12
CEC	cmol kg^{-1}	0.81	24.10	10.64	41.61
AK	cmol kg^{-1}	0.03	3.51	0.37	119.60
TN	%	0.02	0.65	0.20	55.90
SOM	%	0.28	9.08	3.34	51.56

Min Minimum value, *Max* Maximum value, *CV* Coefficient of Variation.

Model performances

The evaluation of model performance provides valuable insights into the predictive capabilities of different machine learning algorithms for soil properties. In this study, we assessed four distinct models—kNN, Cubist, RF, and SVM—across eight soil properties (Table 4). Among these properties, clay content emerged as particularly challenging to predict accurately. However, the kNN model demonstrated notable success with a CCC value of 0.61, emphasizing its ability to capture fine-scale variations, especially in clay-rich regions. Silt content also saw commendable performance with a CCC of 0.63 by the kNN model, highlighting its capacity to capture variations in silt distribution. Conversely, sand content presented more difficulty for all models, with the kNN model achieving a CCC of 0.30. These findings underscore the challenge of predicting sand content, which may be influenced by more complex factors beyond spatial patterns. The superior performance of the k-NN model can be attributed to its effective capture of local spatial patterns, which is particularly advantageous in the presence of a strong

spatial autocorrelation, and its non-parametric nature, making it well-suited for soil texture predictions. Nevertheless, like clay, our results differ from those reported by Kasraei et al. (2021) in terms of the effectiveness of the kNN model for soil particle size predictions.

Regarding pH and CEC predictions, it's noteworthy that these properties showed different model performance characteristics. RF model exhibited the highest R^2 and CCC values for pH, with a R^2 of 0.06 and CCC of 0.19. These values indicate a relatively low accuracy in predicting pH, which is consistent with the intrinsic variability of pH in soil. The R^2 and CCC of the SVM model for pH were 0.15 and 0.15, respectively, showcasing a slightly better variability explanation than RF. Additionally, the R^2 values for all models are relatively low, indicating that none of the models can explain more than 20% of the spatial variability in soil pH. Similar results were obtained by Kasraei et al. (2021) for predicting soil pH. On the other hand, for CEC, RF emerged as the top-performing model, achieving a R^2 of 0.73 and a CCC of 0.61. This strong performance can be attributed to the ability of RF to capture complex relationships in the data, which may be more prominent in CEC variations. Similar results were found by Zeraatpisheh et al. (2019), where they identified RF as the top-performing model for predicting CEC in Iranian soils. Cubist also performed well, with a R^2 of 0.56 and CCC of 0.60 for CEC, showcasing its capacity to model the non-linear nature of this property.

In the prediction of AK (cmol kg^{-1}) levels, the Cubist model demonstrated the highest performance among the models, achieving a CCC of 0.21. This implies a moderate level of accuracy in forecasting K content. RF and SVM closely trails behind with a CCC of 0.21 and 0.20, indicating a comparable degree of precision. Sharififar (2022) also

Table 4 Validation criteria for predicting soil properties in calibration and validation datasets from best to worst performance. The most accurate method highlighted in bold

			Calibration			Validation		
Property	Unit	Model	R2	CCC	RMSE	R2	CCC	RMSE
Clay	%	kNN	0.44	0.56	4.50	**0.74**	**0.61**	**2.82**
		Cubist	0.87	0.88	2.09	0.61	0.55	3.08
		RF	0.85	0.86	2.55	0.63	0.48	3.85
		SVM	0.77	0.81	2.63	0.35	0.39	3.81
Silt		kNN	0.53	0.65	8.17	**0.75**	**0.63**	**5.65**
		RF	0.88	0.87	4.81	0.78	0.57	5.61
		Cubist	0.82	0.88	5.15	0.45	0.41	7.23
		SVM	0.58	0.64	7.51	0.47	0.40	6.92
Sand		kNN	0.41	0.54	7.85	**0.23**	**0.30**	**5.33**
		SVM	0.53	0.56	6.81	0.31	0.27	4.39
		Cubist	0.64	0.75	6.11	0.20	0.26	5.24
		RF	0.79	0.79	5.25	0.22	0.20	4.77
Soil pH		RF	0.68	0.51	0.79	**0.06**	**0.19**	**0.55**
		SVM	0.21	0.13	1.03	0.15	0.15	0.47
		kNN	0.10	0.16	1.04	0.06	0.13	0.51
		Cubist	0.60	0.64	0.64	0.05	0.08	0.52
CEC	cmol kg^{-1}	RF	0.78	0.79	2.33	**0.73**	**0.61**	**2.32**
		Cubist	0.61	0.72	2.83	0.56	0.60	2.79
		SVM	0.46	0.52	3.26	0.58	0.53	2.35
		kNN	0.24	0.39	3.93	0.03	0.11	4.65
AK		Cubist	0.43	0.51	0.32	**0.16**	**0.21**	**0.21**
		RF	0.55	0.54	0.33	0.10	0.21	0.22
		kNN	0.12	0.22	0.43	0.07	0.20	0.22
		SVM	0.32	0.37	0.40	0.06	0.16	0.23
TN	%	RF	0.66	0.70	0.06	**0.64**	**0.49**	**0.12**
		Cubist	0.51	0.63	0.07	0.56	0.44	0.14
		kNN	0.29	0.45	0.08	0.44	0.32	0.15
		SVM	0.54	0.61	0.07	0.33	0.26	0.16
SOM		SVM	0.44	0.45	1.38	**0.16**	**0.26**	**0.87**
		kNN	0.20	0.32	1.60	0.10	0.23	0.99
		RF	0.76	0.75	0.97	0.08	0.19	1.23
		Cubist	0.71	0.81	0.95	0.01	0.06	1.94

found that Cubist, RF, and SVM exhibited similar accuracy in predicting K content in Australian soils. On the other hand, RF exhibited the best performance in predicting TN (%)with a CCC value of 0.49, indicating a relatively strong ability to predict N levels. Cubist also delivered commendable performance with a CCC of 0.44, suggesting solid accuracy. Furthermore, both models explain more than 50% of the spatial variability with R2 values of 0.64 and 0.56. These results align with those obtained by Parsaie et al. (2021), where RF and Cubist outperformed other models in predicting nitrogen content in the topsoil.

The prediction of SOM (%) content posed a particular challenge due to its intricate nature. Among the models, SVM model exhibited the most robust performance, with an R^2 of 0.16 and a CCC of 0.26. These results suggest that the SVM model could capture a portion of the spatial variability in SOM content, which often reflects complex organic matter distribution patterns influenced by vegetation cover, land management practices, and soil formation processes (Forkuor et al. 2017; Khlosi et al. 2016). Similarly, Morellos et al. (2016) also found that SVM provided the best performance in predicting soil organic carbon. However, it's important to note that the relatively low performance metrics indicate the inherent difficulty in predicting SOM content accurately. The substantial variability in SOM levels, as indicated by a CV of 51.56%, underscores the heterogeneity of organic matter distribution within the study area and in other studies carried out in similar environments (Andivia et al. 2015). This heterogeneity arises from factors like land use history and localized inputs of organic material, making SOM a complex property to predict spatially.

Prediction maps of soil properties and environmental covariates importance

The maps generated by the models with the highest performance for each of the study soil properties are presented in Fig. 2. However, predictive maps for each soil property using the models employed are provided in the Appendix section, specifically in Figs. 6-13.

The kNN model, excelling in predicting clay, silt, and sand content, uncovers distinct spatial patterns. Clay accumulates in lower valleys, silt in mid-slope areas, and sand dominates

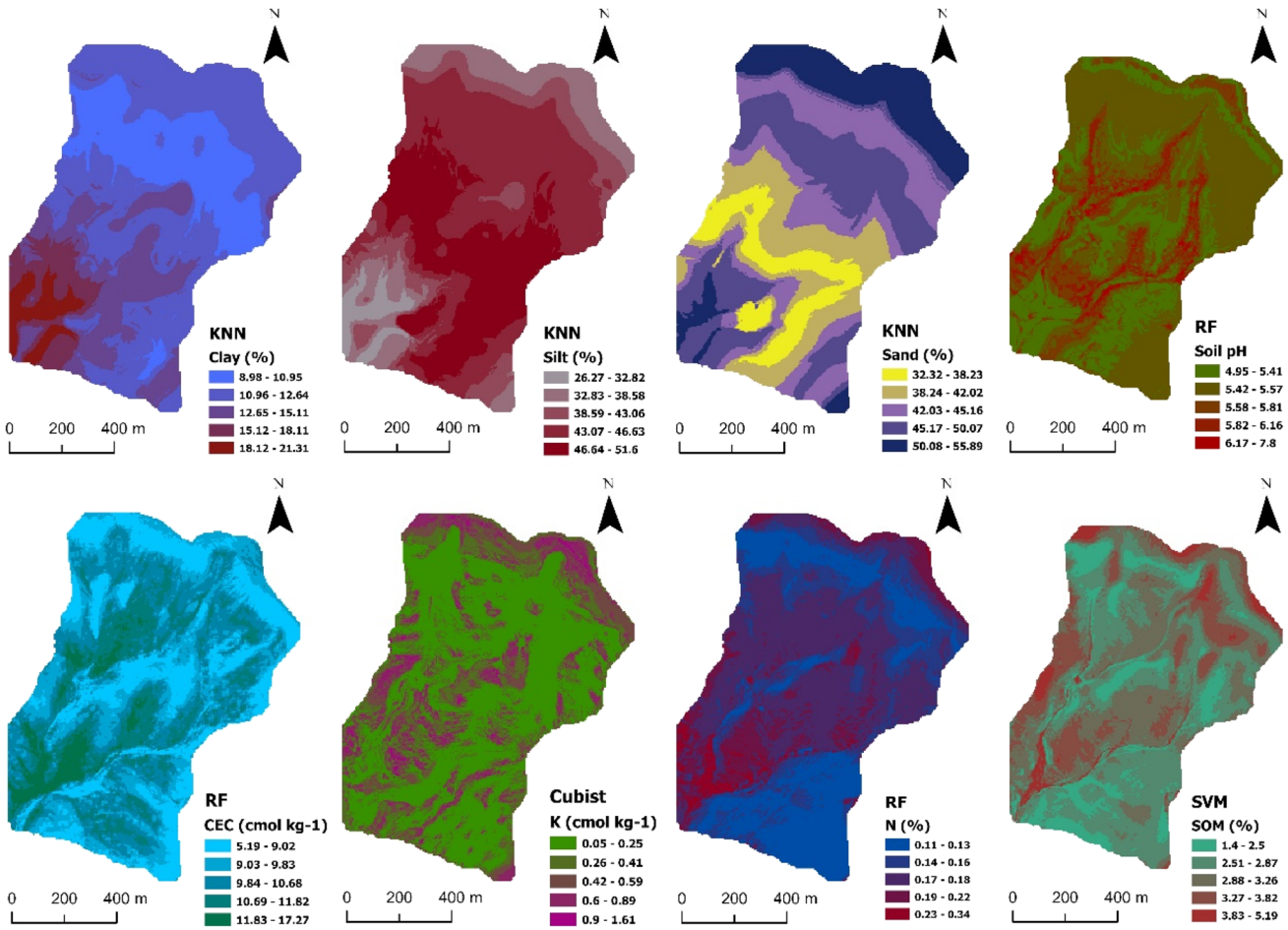

Fig. 2 Maps produced by the models with the best performance for each study property

higher regions, influenced by topography. This spatial pattern could be attributed to the dynamics of soil particle size transport and sedimentation influenced by topographic factors, as demonstrated in previous studies (Gallant & Dowling 2003). Topographic indices like the topographic roughness index (TRI) and valley depth play crucial roles, as shown in Fig. 3. Previous studies also stress their significance (Mishra et al. 2009; Zeraatpisheh et al. 2019). On the other hand, covariate importance for each soil property and predictive model is presented in the Appendix section, in Figs. 14 and 15.

In terms of predicted values, all models, except Cubist, tend to overestimate minimum values and underestimate maximum values when compared to descriptive statistics (Table 3). This suggests that models perform well in capturing mid-range spatial variability but struggle with extreme values. This trend should be considered when interpreting predictive maps, especially in areas with wide soil property value ranges.

The soil pH predictive map, generated using the RF model, reveals higher pH values around channels and in the upper catchment areas. The mean predicted pH value (5.51) shows minimal variation compared to other models (Table 5), confirming the prevalent acidic nature of Extremadura's rangeland soils, as reported previously (Ceballos & Schnabel 1998). Topographic attributes, such as altitude and maximum curvature, are significant drivers of soil pH variation, aligning with earlier research (Mosleh et al. 2016). This underscores the role of land surface features in shaping soil pH levels. Similarly, the CEC predictive map generated by the RF model exhibits patterns resembling soil pH, with higher values in depressed catchment areas. However, it tends to overestimate minimum values (5.17 cmol kg-1) while underestimating maximum and mean values (17.26 and 9.92 cmol kg-1, respectively) compared to dataset values (Table 3). The importance of covariates in predicting CEC highlights topographic attributes like altitude, TPI (Topographic Position Index), and valley depth as influential factors. Previous research consistently shows the strong correlation between these topographic attributes and soil CEC, explaining a substantial portion of its spatial variability (Khaledian et al. 2017). The models' tendency to overestimate minimum and underestimate maximum values for both pH and CEC suggests their effectiveness in capturing mid-range variability while encountering challenges with extreme values.

For soil available potassium (AK) prediction, the Cubist model outperformed others with a range of 0.04 to 1.61 (Table 5). While the mean values closely match the dataset, Cubist exhibits a wider variability range in soil AK, suggesting its ability to capture diverse patterns in the study area. Terrain convexity emerges as the primary factor influencing K distribution (Figs. 2 and 3), impacting water flow and AK, as shown by Bui et al. (2019) and Arabameri et al. (2019).

Regarding total nitrogen (TN) prediction, the RF model produced values ranging from 0.10 to 0.33, deviating from dataset values of 0.02 to 0.65 (Table 5). Despite slight differences in minimum and maximum values across models, mean values remain consistent. Higher TN values

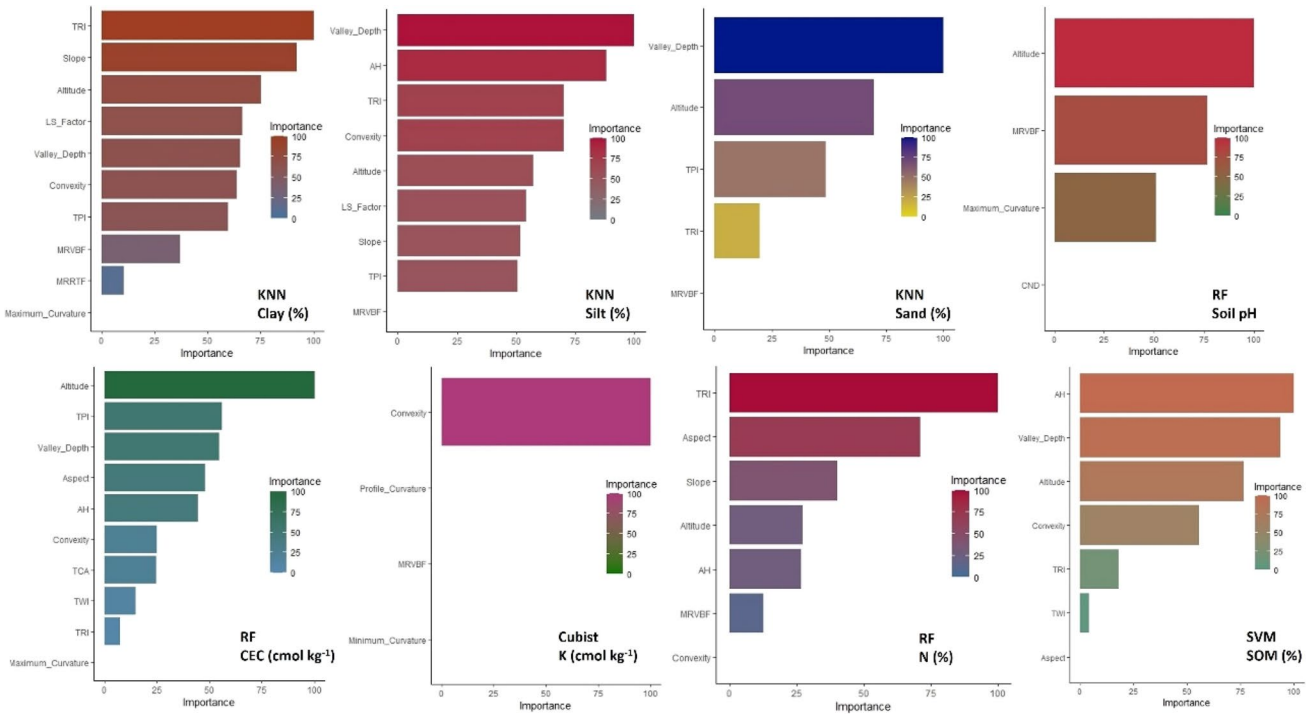

Fig. 3 Importance of environmental covariates in predicting the different soil properties using the model with the best performance

Table 5 Descriptive statistics for soil properties prediction and 90% prediction interval by predictive model

Soil property	Unit	Model	Soil Prediction Maps				90% Prediction Interval Maps			
			Min	Max	Mean	SD	Min	Max	Mean	SD
Clay	%	**kNN**	**8.96**	**21.31**	**12.47**	**2.34**	**9.58**	**9.85**	**9.65**	**0.05**
		Cubist	2.15	23.54	12.47	3.82	10.23	12.14	10.61	0.24
		RF	6.90	21.27	12.22	2.03	12.81	13.06	12.91	0.04
		SVM	5.04	21.93	12.32	2.43	12.80	17.06	12.91	0.11
Silt	%	**kNN**	**26.25**	**51.50**	**43.50**	**4.98**	**19.49**	**20.45**	**19.73**	**0.08**
		RF	25.95	56.04	42.95	3.69	18.80	19.11	18.99	0.04
		Cubist	20.97	75.44	44.02	6.36	24.16	25.76	24.41	0.13
		SVM	25.92	57.43	41.72	3.72	23.49	27.77	23.77	0.22
Sand	%	**kNN**	**32.32**	**55.89**	**44.76**	**5.22**	**18.08**	**18.88**	**18.22**	**0.13**
		SVM	34.76	61.56	46.23	3.15	15.31	19.07	15.58	0.23
		Cubist	17.45	74.22	43.63	7.04	17.49	18.22	17.81	0.09
		RF	30.82	59.72	44.42	3.97	16.14	16.72	16.33	0.07
pH		**RF**	**4.95**	**7.80**	**5.51**	**0.18**	**2.36**	**2.68**	**2.41**	**0.02**
		SVM	5.07	5.89	5.37	0.06	2.19	2.58	2.19	0.01
		kNN	5.02	6.16	5.58	0.21	2.25	2.37	2.30	0.02
		Cubist	5.11	6.56	5.41	0.26	2.31	2.46	2.33	0.03
CEC	cmol kg^{-1}	**RF**	**5.17**	**17.26**	**9.92**	**1.11**	**7.51**	**7.94**	**7.68**	**0.03**
		Cubist	2.07	18.71	9.81	1.81	9.31	10.17	9.52	0.06
		SVM	6.32	14.22	9.31	1.04	8.02	8.56	8.13	0.04
		kNN	7.49	14.33	10.34	2.01	15.42	15.54	15.47	0.03
AK	cmol kg^{-1}	**Cubist**	**0.04**	**1.61**	**0.32**	**0.19**	**0.71**	**1.70**	**0.80**	**0.07**
		RF	0.16	1.19	0.33	0.14	0.76	1.01	0.82	0.04
		kNN	0.17	0.65	0.40	0.17	0.79	0.92	0.85	0.05
		SVM	-0.35	1.34	0.28	0.14	0.77	0.96	0.82	0.03
TN	%	**RF**	**0.10**	**0.34**	**0.16**	**0.03**	**0.47**	**0.56**	**0.50**	**0.01**
		Cubist	0.01	0.44	0.16	0.05	0.45	0.64	0.53	0.02
		kNN	0.13	0.29	0.16	0.03	0.55	0.60	0.56	0.01
		SVM	0.04	0.31	0.16	0.03	0.58	0.62	0.59	0.01
SOM	%	**SVM**	**1.39**	**5.19**	**2.92**	**0.44**	**2.86**	**5.76**	**3.29**	**0.11**
		kNN	2.19	4.95	3.05	0.30	3.39	3.70	3.49	0.03
		RF	1.53	6.54	3.12	0.53	4.26	4.69	4.36	0.04
		Cubist	0.66	9.17	3.24	0.93	6.43	6.77	6.53	0.04

In bold, the best-performing model

Min Minimum value, *Max* Maximum value, *SD* Standard Deviation.

concentrate in valleys' bottoms, while lower values occur in flow areas and steep slopes. Topographic indices like TRI, aspect, slope, and altitude (Fig. 3) play crucial roles in predicting soil TN, reflecting landscape features that influence nutrient dynamics. Luizão et al. (2004) and Hawthorne and Miniat (2018) have previously illustrated how these attributes affect water retention, drainage patterns, and microbial activity, ultimately influencing TN availability.

For soil organic matter (SOM) content, the SVM model predicted values ranging from 1.39% (minimum) to 5.19% (maximum), with an average of approximately 2.92% (Table 5). These predictions differ from the initial dataset values, which ranged from 0.28% to 9.08%, averaging 3.34% (Table 3). SVM tends to provide higher predictions for SOM. The predicted maps highlight higher SOM values in valley bottoms and flatter areas, with key variables being analytical hillshading, valley depth, altitude, and convexity. These terrain attributes capture topographic and geomorphological characteristics influencing SOM distribution and accumulation (Adhikari et al. 2018; Guo et al. 2019; Taghizadeh-Mehrjardi et al. 2016). For example, analytical hillshading reflects local slope and orientation, impacting microclimates, soil moisture, and organic matter decomposition rates (Guo et al. 2019). Valley depth indicates landscape morphology, affecting water and organic matter retention in lower-altitude areas, influencing SOM content. High convexity or concavity areas alter water flow dynamics, nutrient distribution, and subsequently, SOM content (Mahmoudzadeh et al. 2020).

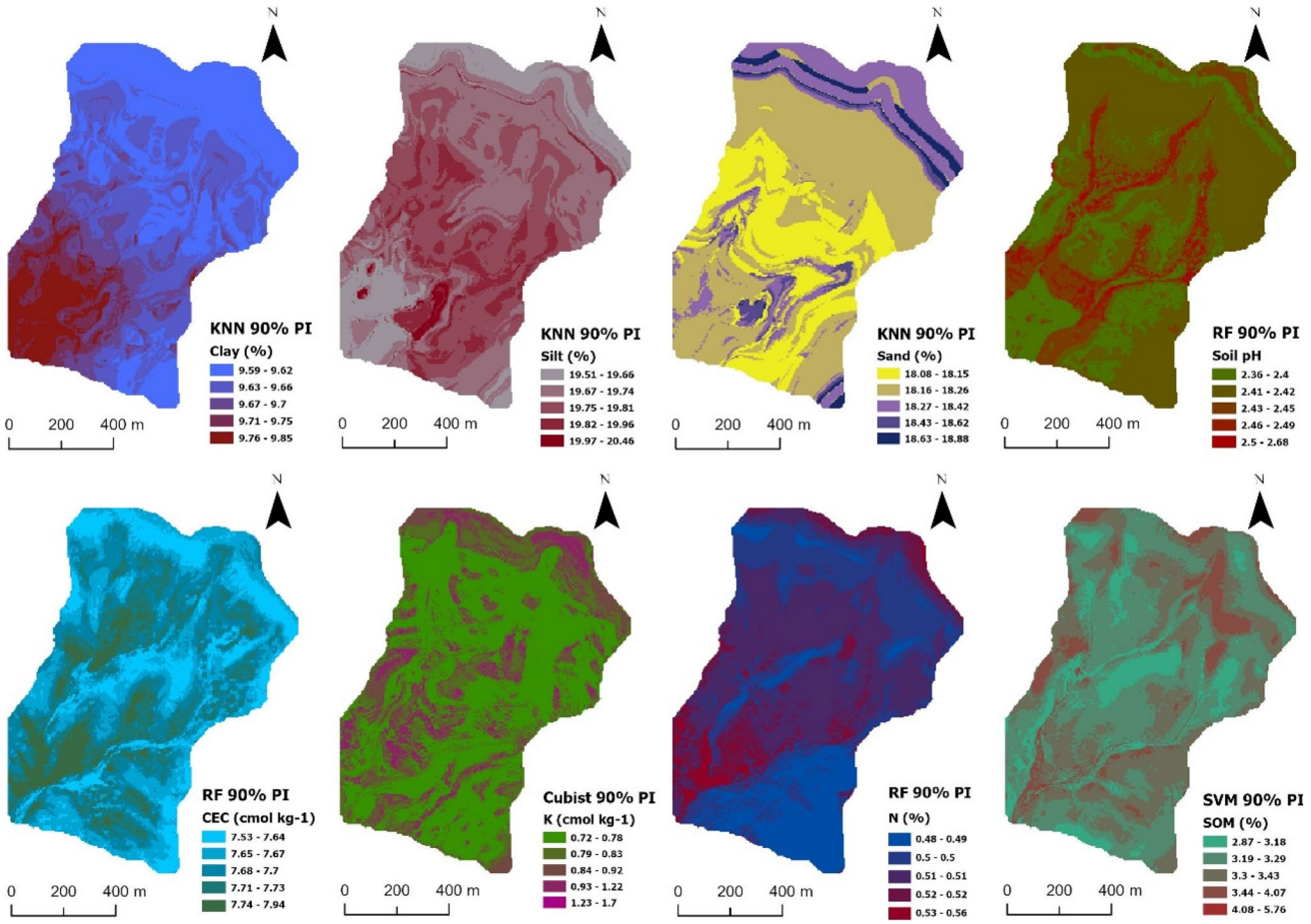

Fig. 4 Uncertainty maps displaying the 90% prediction interval using the best-performing model for each studied soil property

The analysis of variable importance (Fig. 3) underscores the dominance of topographic and geomorphological attributes over remote sensing data in model generation, following variable selection with the Boruta algorithm. These attributes exhibit greater influence, particularly at smaller scales, in capturing subtle landscape variations. While remote sensing data, such as satellite imagery, offer valuable insights at larger scales, they may struggle to detect microtopographic features, especially in regions with intricate surface heterogeneities (Cresto Aleina et al. 2015). To address the spatial variability of small-scale soil processes effectively, consideration of higher-resolution spatial information, including vegetation indices, becomes crucial. This finer-grained data allows for a more comprehensive characterization of the landscape's influence on soil properties, with previous studies highlighting the intricate connection between small-scale soil parameter variability and geological, climatic, and biological processes, with vegetation playing a significant role (Agam et al. 2007; Isermann 2005).

Additionally, the implementation of the Boruta algorithm significantly reduces the number of variables in the models, streamlining model complexity while preserving predictive accuracy. For example, the variable count for clay content decreased from 23 to 10, and for pH and AK, it reduced to 4. This highlights the importance of the selected attributes in achieving accurate predictions.

Uncertainty estimation and assessment

Uncertainty maps for each soil property were generated with a 90% prediction interval (Fig. 4), and the quality of uncertainty was assessed using PCIP plots (Fig. 5). Detailed uncertainty maps and plots for each property and model are available in the Appendix (Figs. 16-25). Generally, uncertainty values correspond to the spatial patterns in the predicted maps, with higher uncertainty in areas with elevated property values and lower uncertainty in regions with lower values. Areas with higher property content often involve more complex interactions and factors, posing challenges for prediction models. These high-uncertainty zones require additional attention, data collection, and model refinement to enhance the reliability of predicted maps. Variations in uncertainty maps exist among different models, with relatively consistent prediction intervals across soil properties. However, differences in interval widths are observed (Table 5). Notably, Cubist exhibits a broader range between the

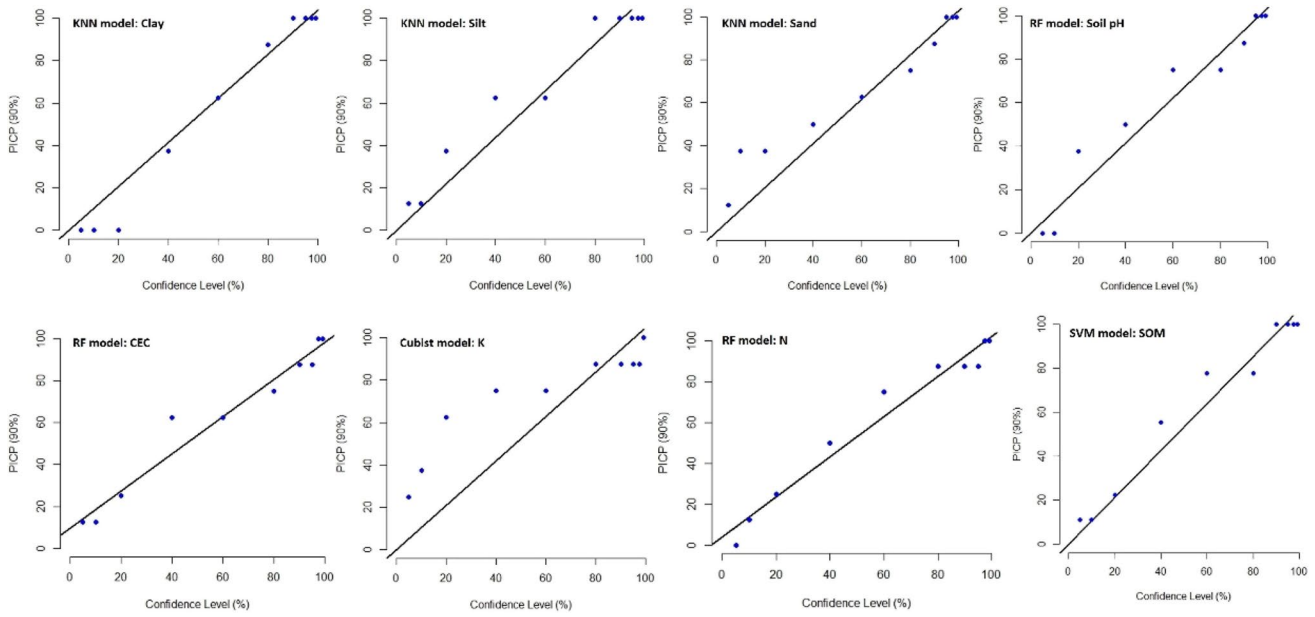

Fig. 5 Prediction interval coverage probability (PICP) plots for uncertainty estimates of soil properties analyzed using the best-performing model

minimum and maximum values for potassium uncertainty (0.71 to 1.70 cmol kg^{-1}). SVM stands out with a substantial range of 2.90% for SOM uncertainty, reflected in a high standard deviation (SD) value of 0.11, the highest among the models. These differences in uncertainty estimation highlight the variability in model performance and emphasize the importance of selecting an appropriate model for specific soil properties.

The assessment of uncertainty quality, evaluated through Percent Interval Prediction (PCIP) plots (Fig. 5), demonstrates a generally consistent trend across all models. As model accuracy improves, there is a noticeable enhancement in fitting. However, it's noteworthy that for percentiles above 80%, the models tend to adopt a moderately conservative approach. Around 87.5% of actual observations fall within the prediction intervals at these confidence levels, reflecting a reliable ability to quantify uncertainty while erring on the side of caution. An intriguing shift occurs at the 60% percentile, where models exhibit a slightly lower rate of real observations within the intervals, approximately 75%. This indicates that models are more accurate and less conservative in this confidence range, suggesting increased reliability in their predictions. However, as percentiles decrease towards 60% and below, the models indicate growing uncertainty and imprecision in their estimates.

Conclusions

In conclusion, this study represents a valuable effort in unraveling the intricate spatial dynamics of eight pivotal soil properties within a grazed small catchment nestled within the agroforestry system known as 'dehesas' in the region of Extremadura, Spain. The analysis has unveiled a rich tapestry of spatial variability in these properties, shedding light on the complex interplay of factors shaping soil characteristics in this environment. This innovative approach underscores the paramount importance of comprehending these spatial intricacies for the purpose of reasonable and sustainable land management practices.

The machine learning algorithms employed exhibited varying levels of performance. k-Nearest Neighbors (kNN) performed exceptionally well in predicting soil particle size (clay, silt, and sand) due to its capacity to capture local spatial patterns effectively, which is advantageous in the presence of strong spatial autocorrelation in these properties. On the other hand, Random Forest (RF) models outperformed others in predicting soil pH, Cation Exchange Capacity (CEC), and nitrogen content, thanks to their ability to capture complex relationships in the data. For these properties, RF's capacity to handle intricate patterns proved beneficial.

Cubist emerged as the top performer for predicting available potassium (AK), and Support Vector Machine (SVM) demonstrated superior accuracy in predicting Soil Organic Matter (SOM) content. Challenges were observed in predicting certain properties, such as sand content, soil pH, potassium, and SOM, particularly in capturing extreme values. These challenges highlight the complexity of these properties and the need for further research and data collection in these areas.

The low variability observed in soil properties such as pH and texture composition in this study can be attributed to the specific characteristics of the study area, which is

characterized by 'dehesas' in Extremadura, Spain. These agroforestry systems often have relatively homogeneous environmental conditions, including similar land use practices, common parent material (sandstones and granites), and similar regional climate patterns. This may explain the limited range of values for certain soil properties.

Regarding the suitability of a lower or higher CV% for predictive models, it depends on the specific objectives and context of the study. A lower CV% indicates less variability in the data, which can be advantageous in predictive modeling when the goal is to achieve more precise and accurate predictions. In such cases, models can perform well because there is less variation to account for, leading to more stable predictions. This is especially true for properties like pH, where relatively consistent values can be expected in specific environments. On the other hand, a higher CV% suggests greater variability in the data, which can be challenging for predictive models. However, higher variability could also be a characteristic of certain properties in diverse landscapes, reflecting the complexity of soil processes. In such cases, models need to be robust and capable of accurately capturing a wide range of values.

Environmental covariates, particularly topographic and geomorphological attributes, played a significant role in model generation. The Boruta algorithm's variable selection helped streamline model complexity by reducing the number of variables, emphasizing the importance of these selected attributes. Uncertainty assessment revealed that models effectively quantify uncertainty, with their degree of conservatism varying with confidence levels. Models exhibited greater accuracy at intermediate confidence levels and a more cautious approach at higher percentiles. This comprehensive analysis advances our understanding of soil property prediction, highlighting the importance of selecting appropriate models based on the specific properties of interest and the underlying spatial patterns.

Appendix

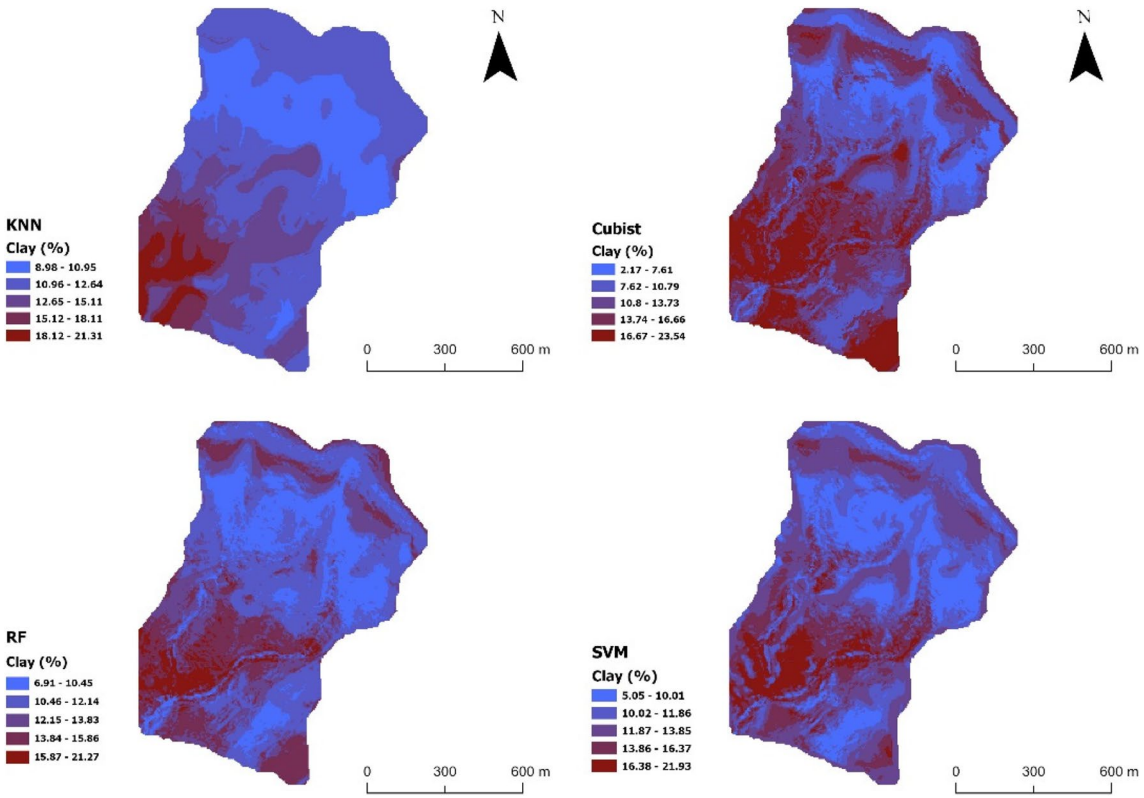

Fig. 6 Predictive maps of clay content (%) using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

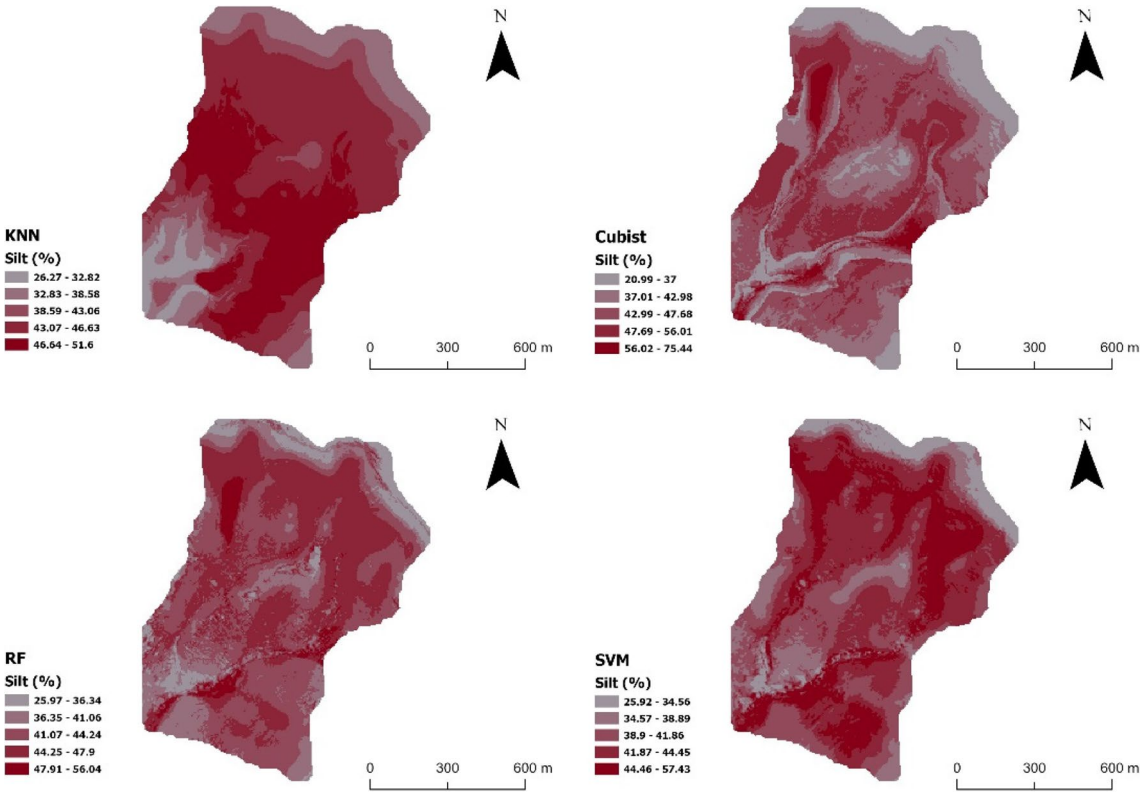

Fig. 7 Predictive maps of silt content (%) using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

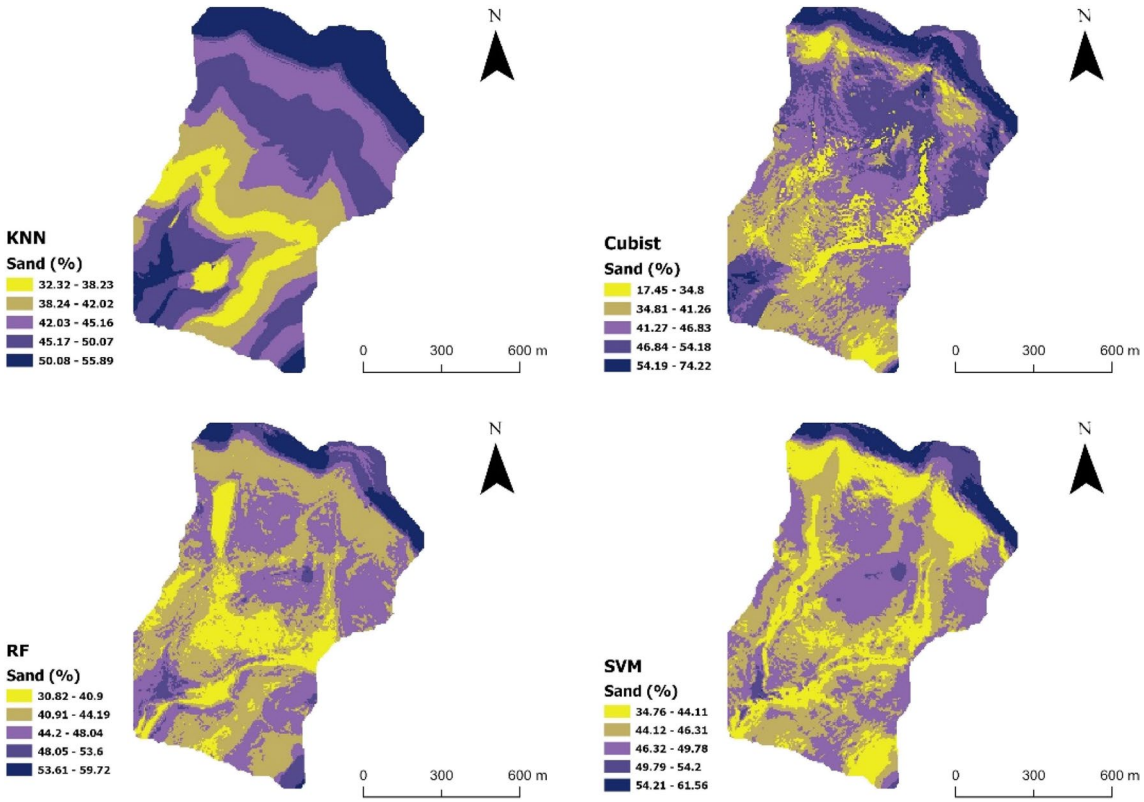

Fig. 8 Predictive maps of sand content (%) using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

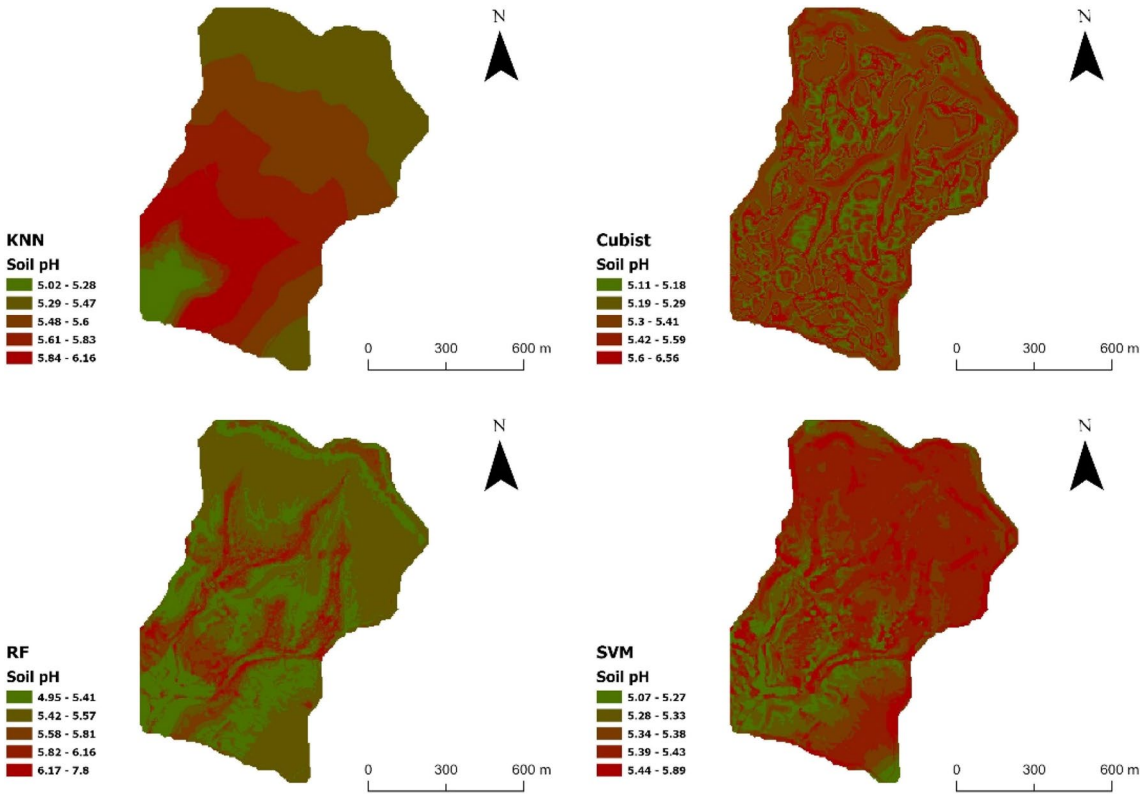

Fig. 9 Predictive maps of soil pH using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

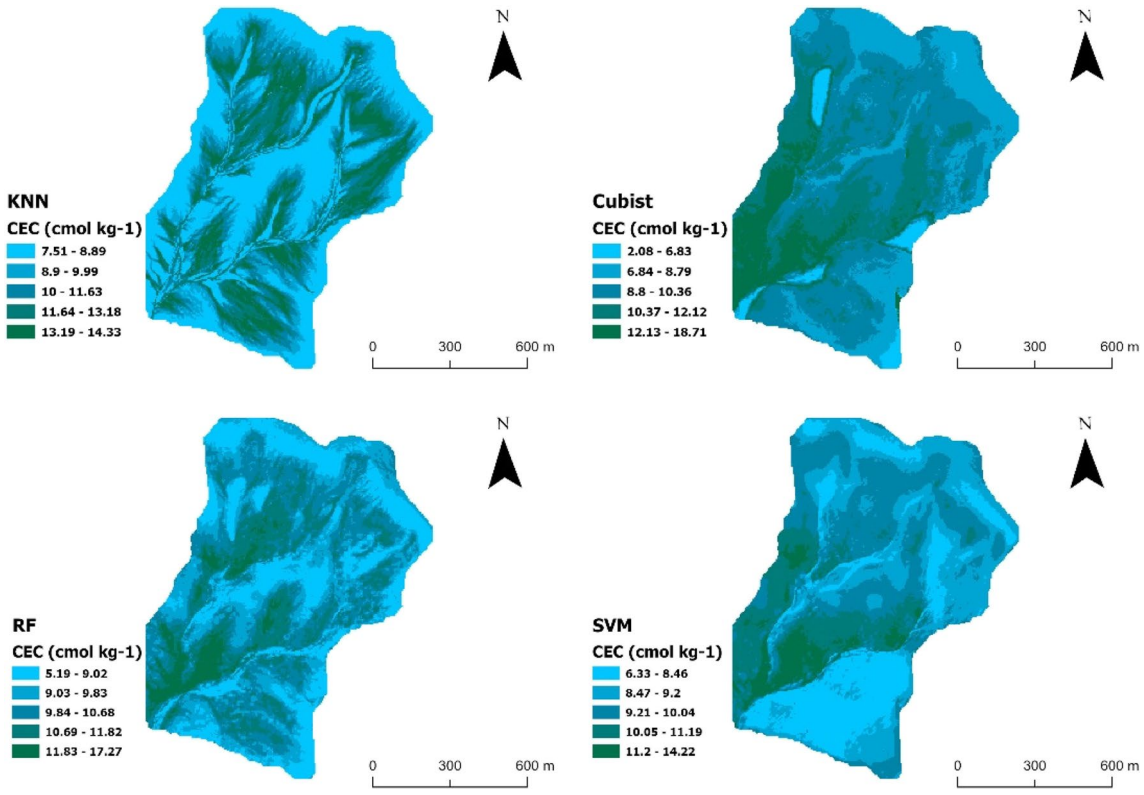

Fig. 10 Predictive maps of Cation Exchange Capacity (CEC) using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

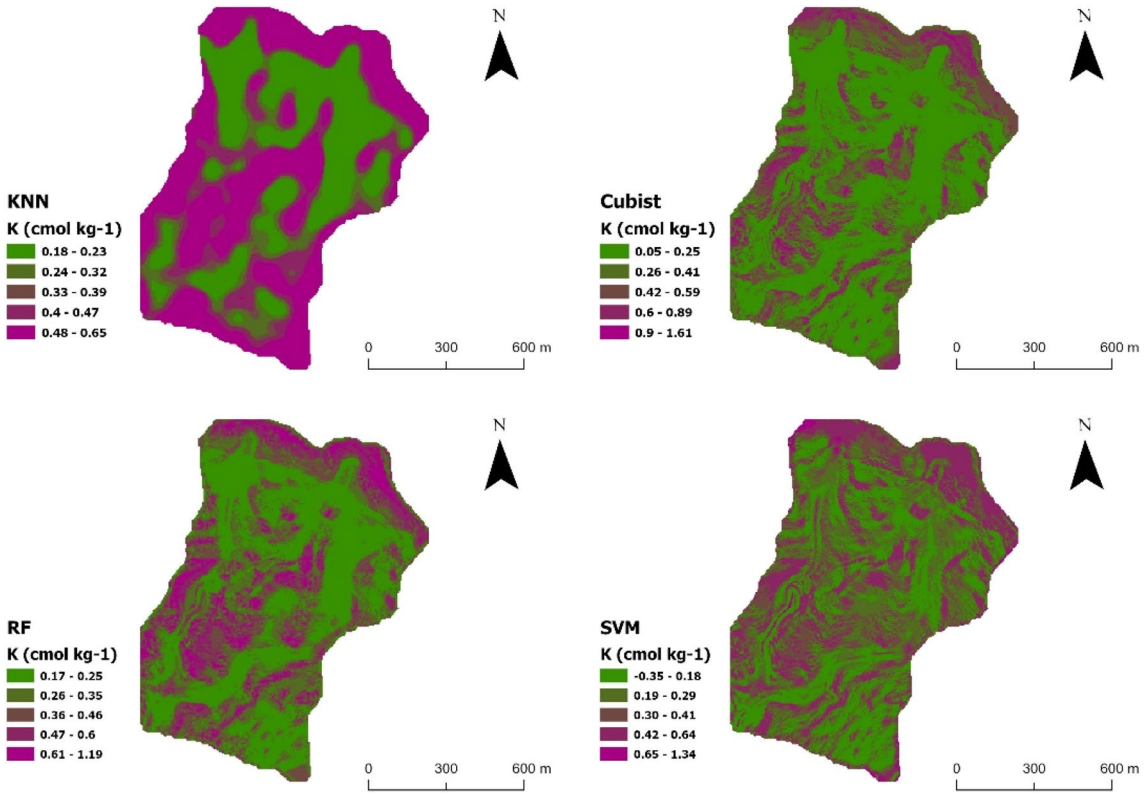

Fig. 11 Predictive maps of AK (cmol kg^{-1}) content using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

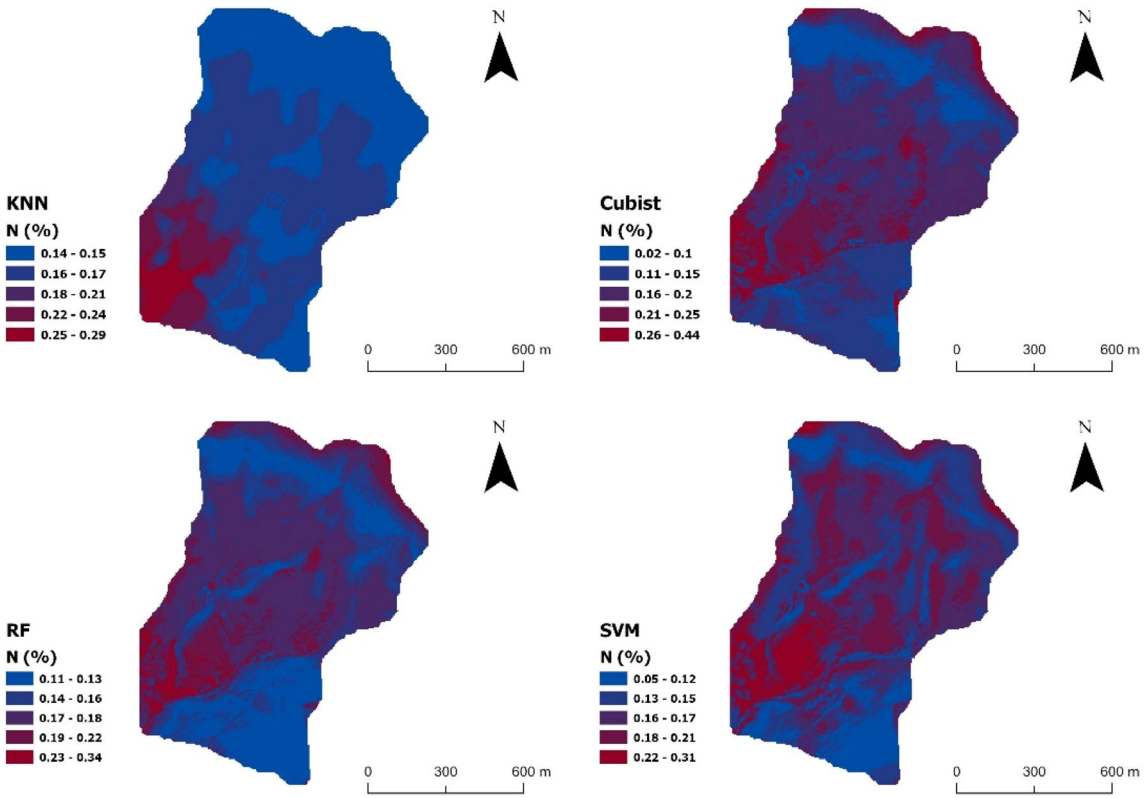

Fig. 12 Predictive maps of TN (%) content using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

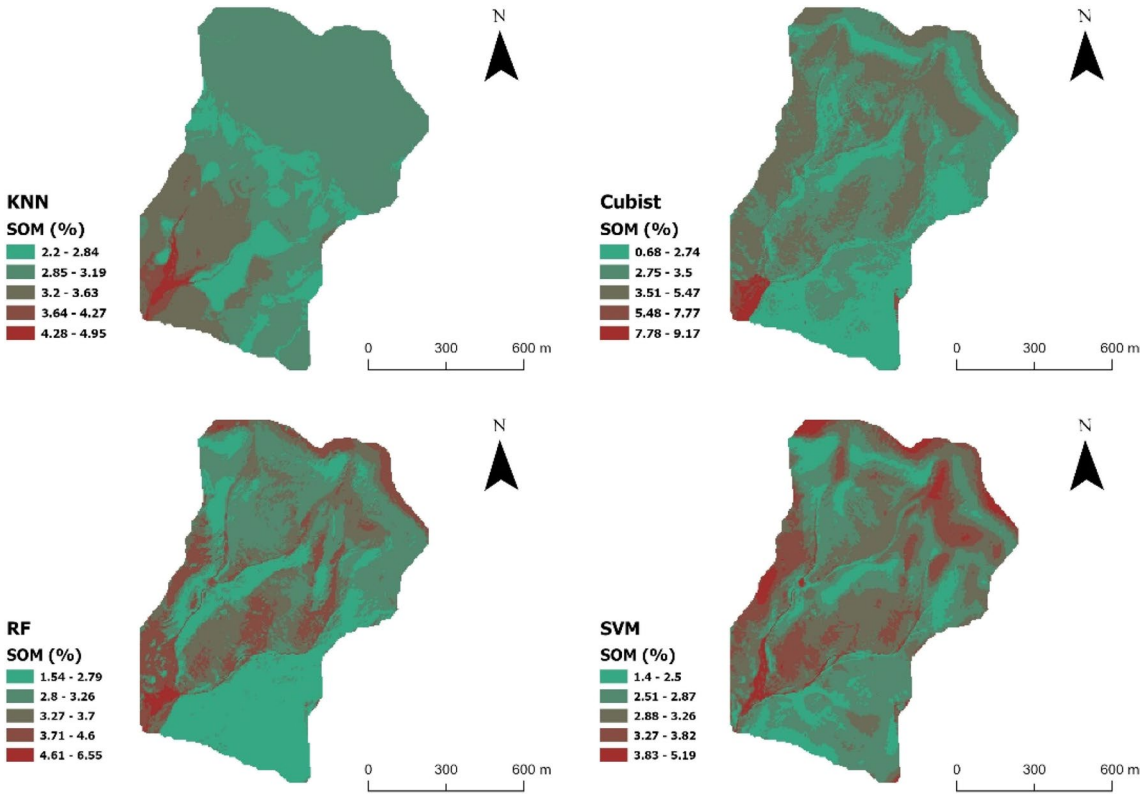

Fig. 13 Predictive maps of SOM (%) content using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

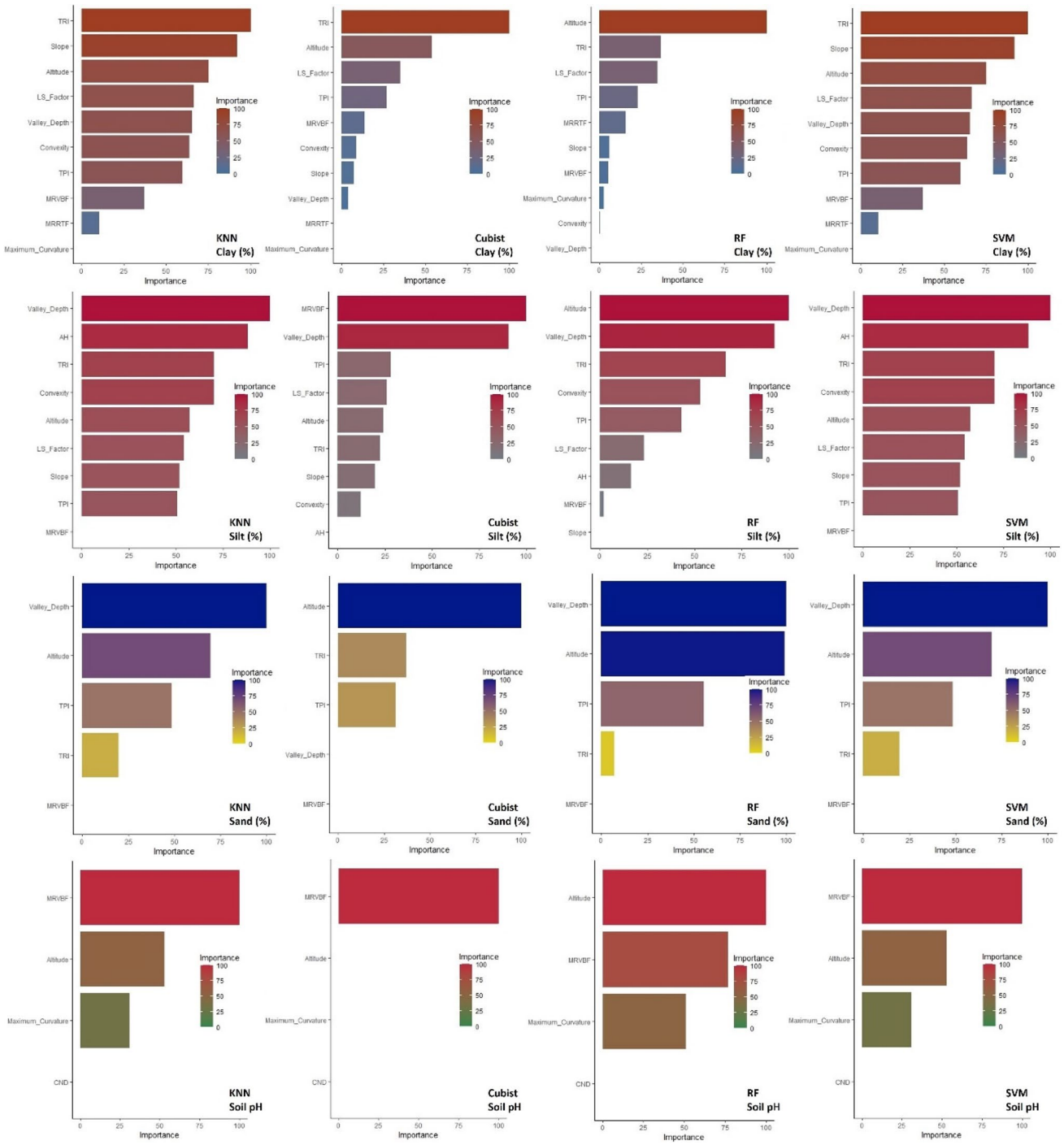

Fig. 14 Importance of environmental covariates used to predict clay, silt, sand, and soil pH (pH) content with k-Nearest Neighbor (kNN), Cubist, Random Forest (RF), and Support Vector Machine (SVM)

Fig. 15 Importance of environmental covariates used to predict Cation Exchange Capacity (CEC), AK (cmol kg^{-1}), TN (%), and SOM (%) content with k-Nearest Neighbor (kNN), Cubist, Random Forest (RF), and Support Vector Machine (SVM)

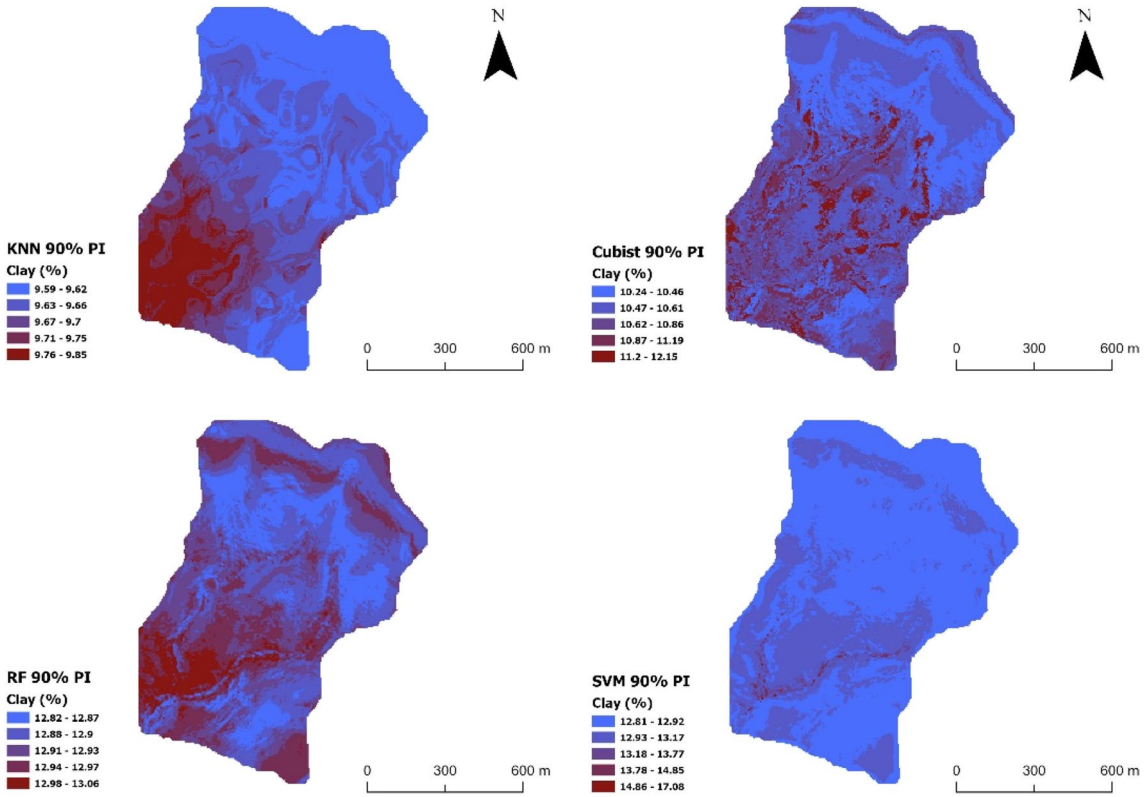

Fig. 16 Uncertainty maps of clay content showing the 90% prediction interval generated using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

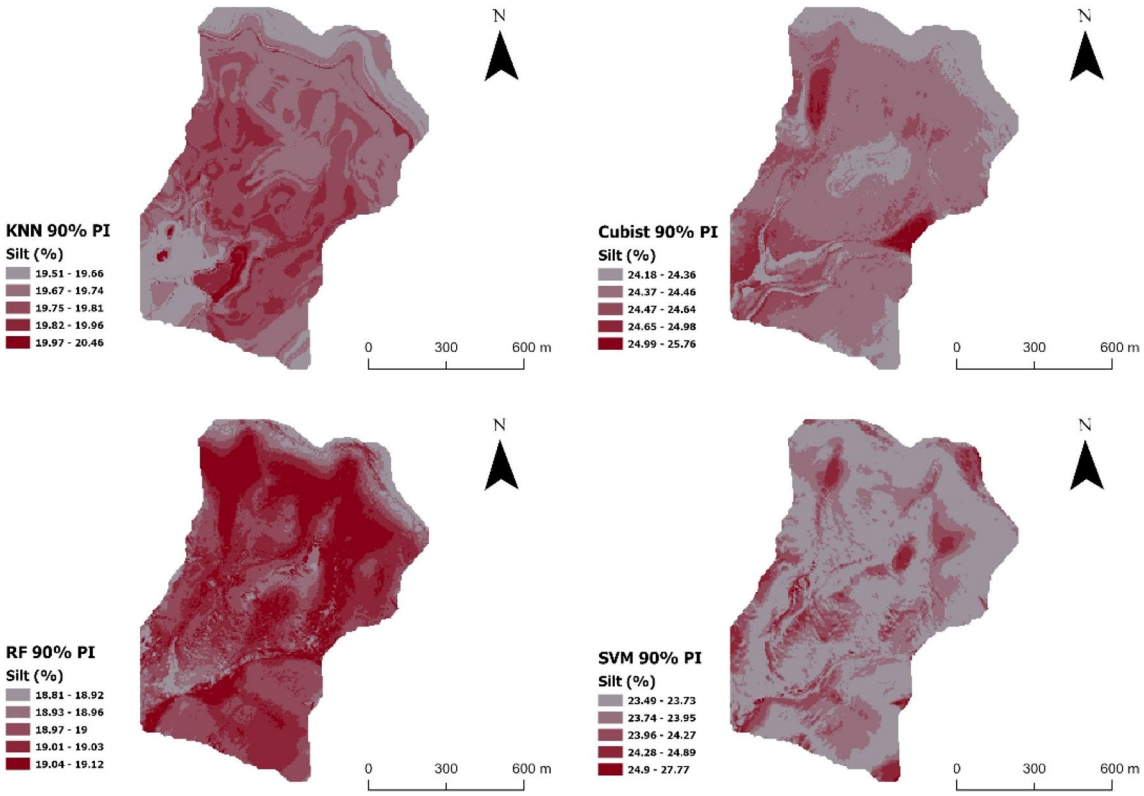

Fig. 17 Uncertainty maps of silt content showing the 90% prediction interval generated using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

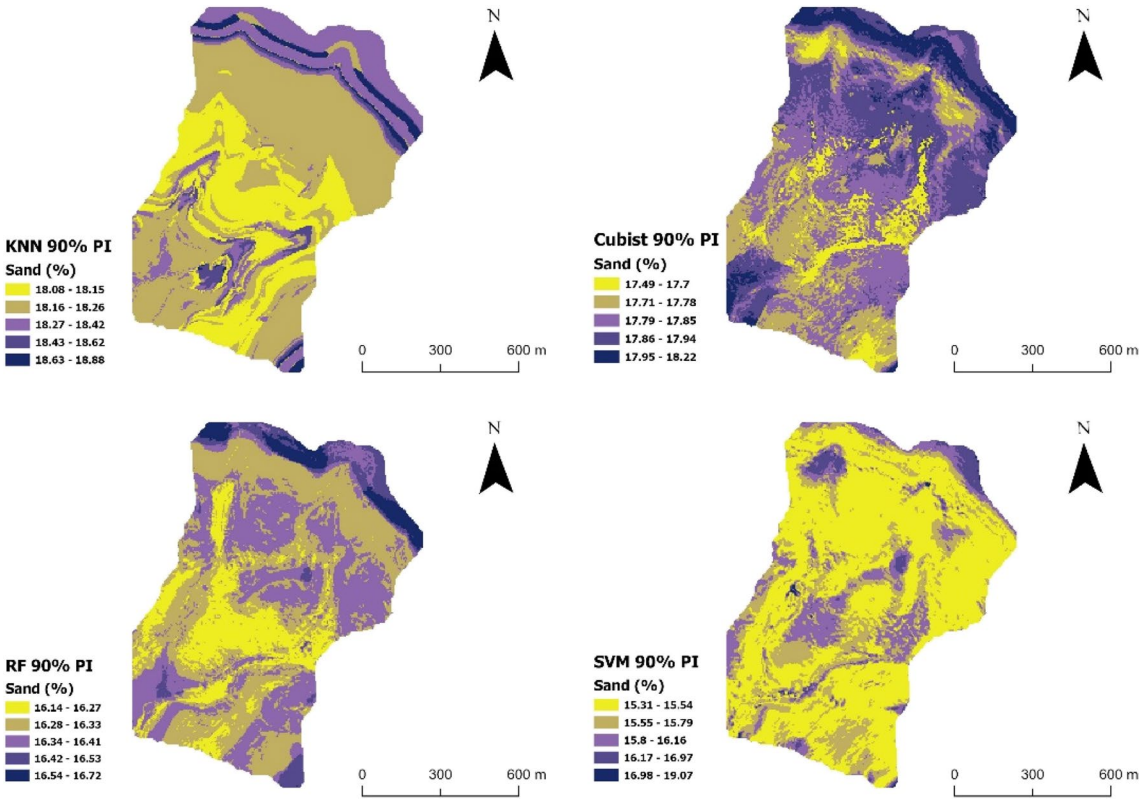

Fig. 18 Uncertainty maps of sand content showing the 90% prediction interval generated using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

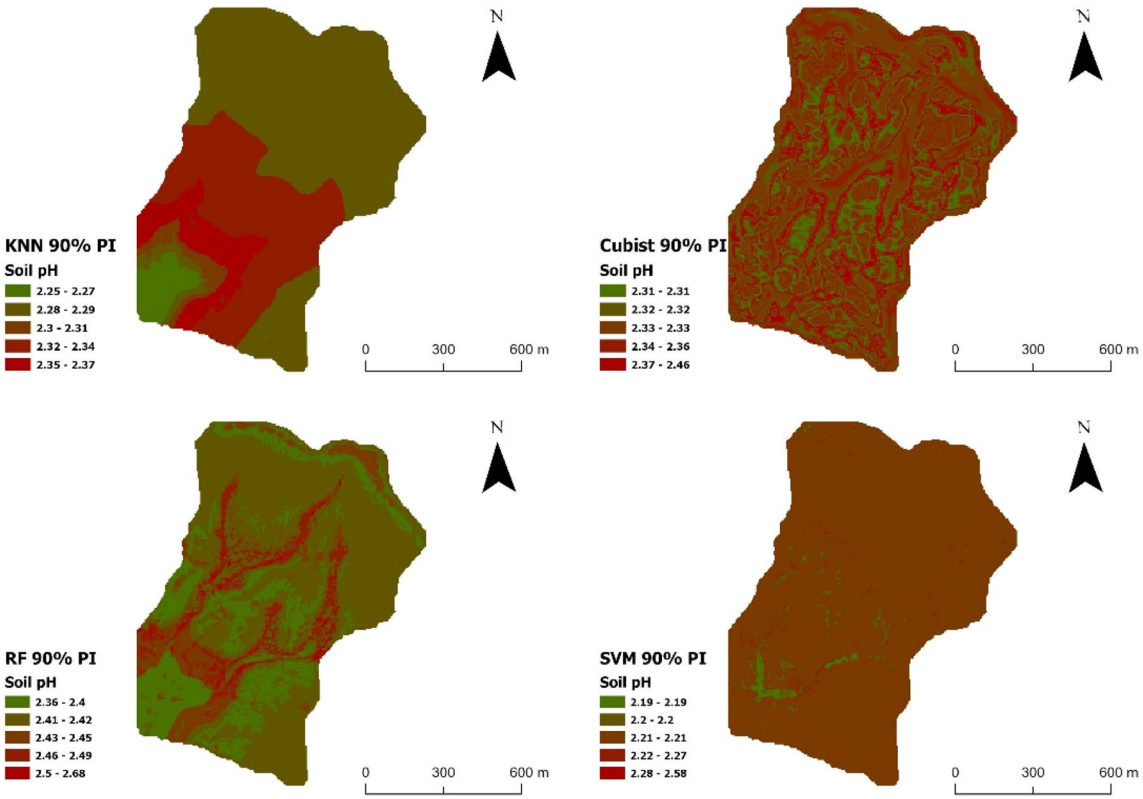

Fig. 19 Uncertainty maps of soil pH showing the 90% prediction interval generated using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

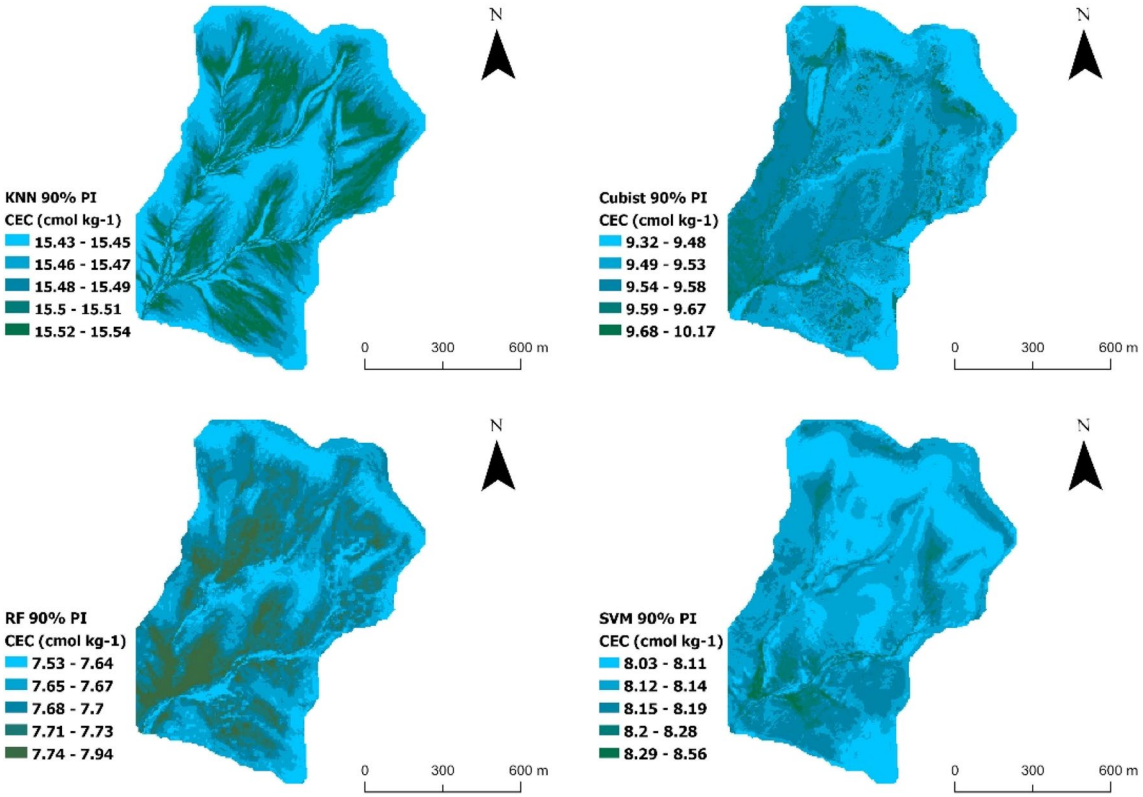

Fig. 20 Uncertainty maps of Cation Exchange Capacity (CEC) showing the 90% prediction interval generated using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

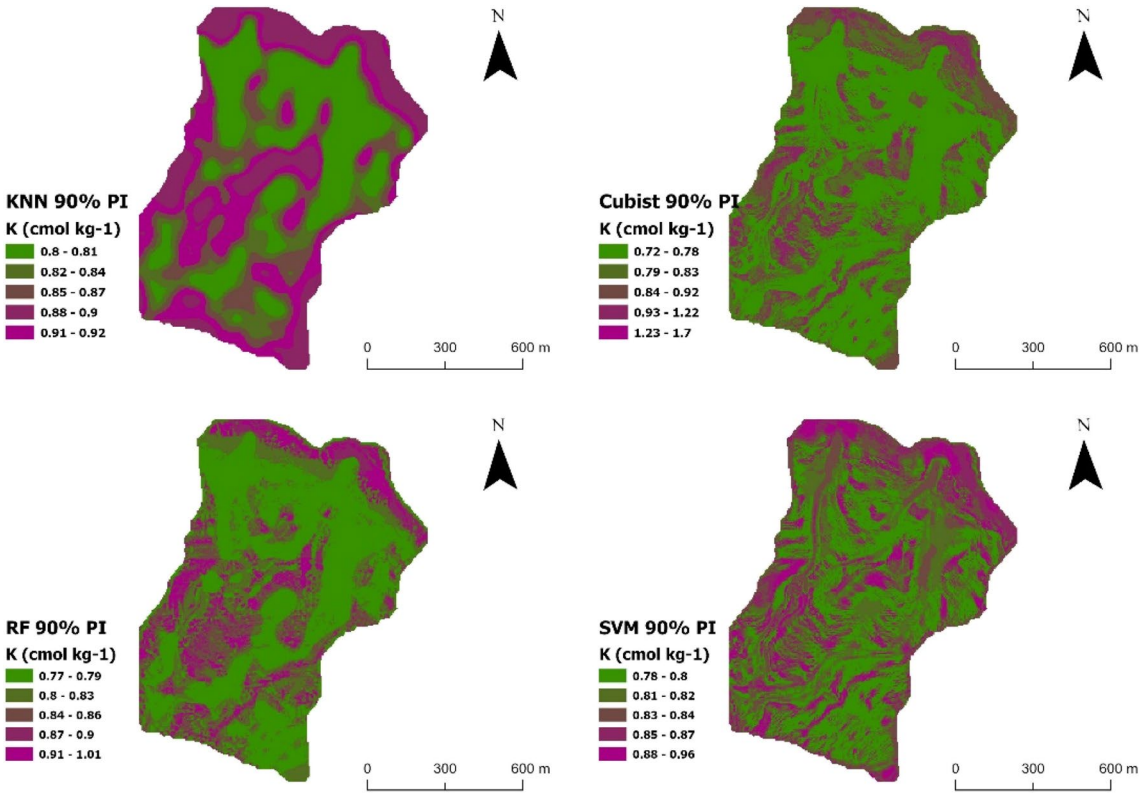

Fig. 21 Uncertainty maps of soil AK (cmol kg^{-1}) content showing the 90% prediction interval generated using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

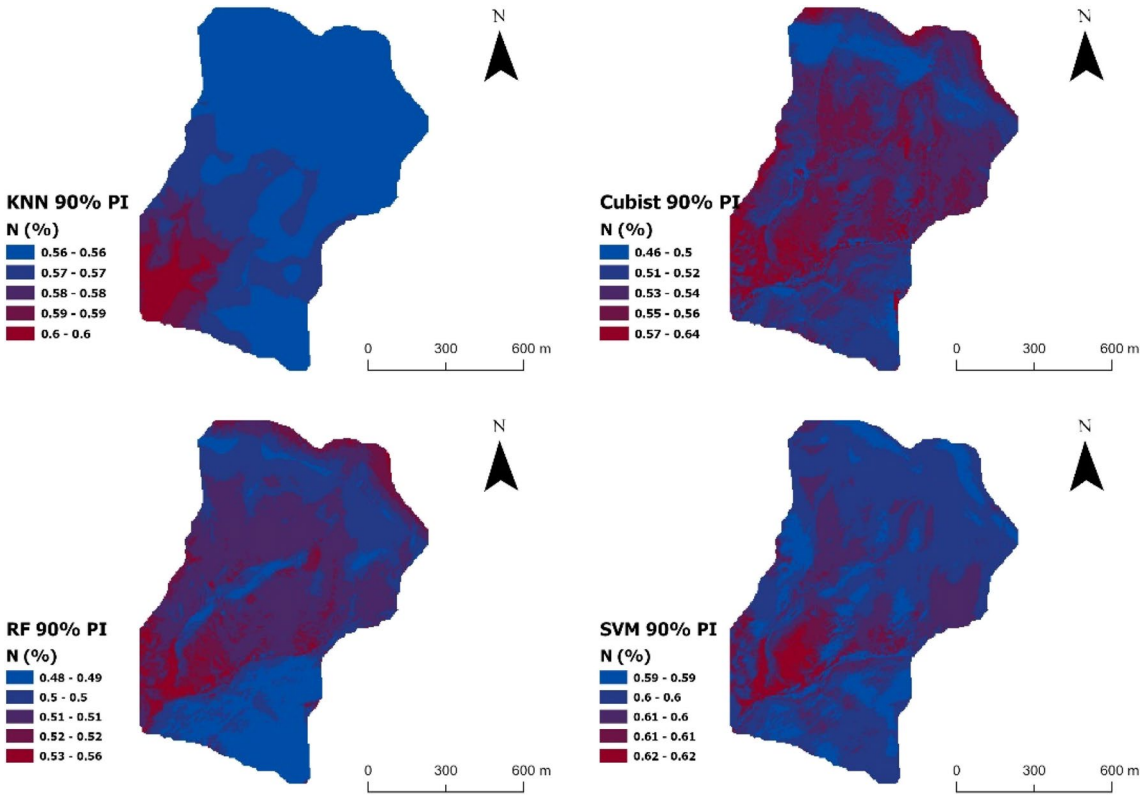

Fig. 22 Uncertainty maps of soil TN (%) content showing the 90% prediction interval generated using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

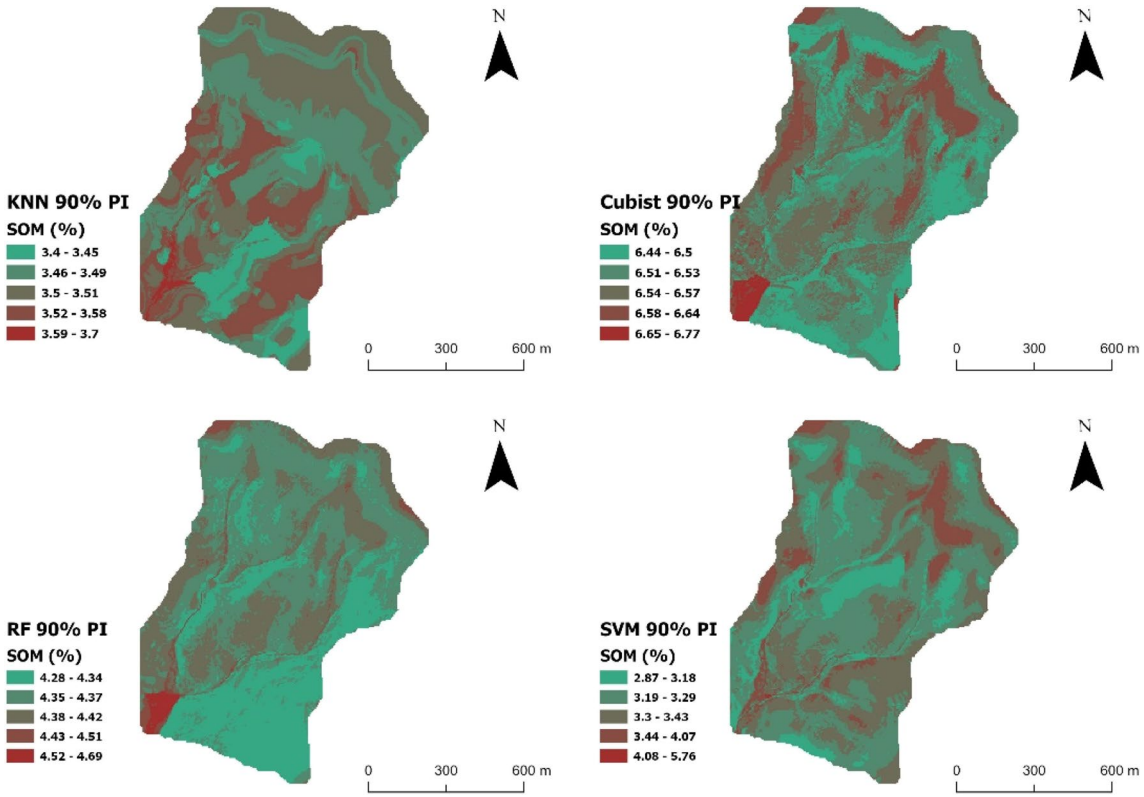

Fig. 23 Uncertainty maps of SOM (%) content showing the 90% prediction interval generated using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

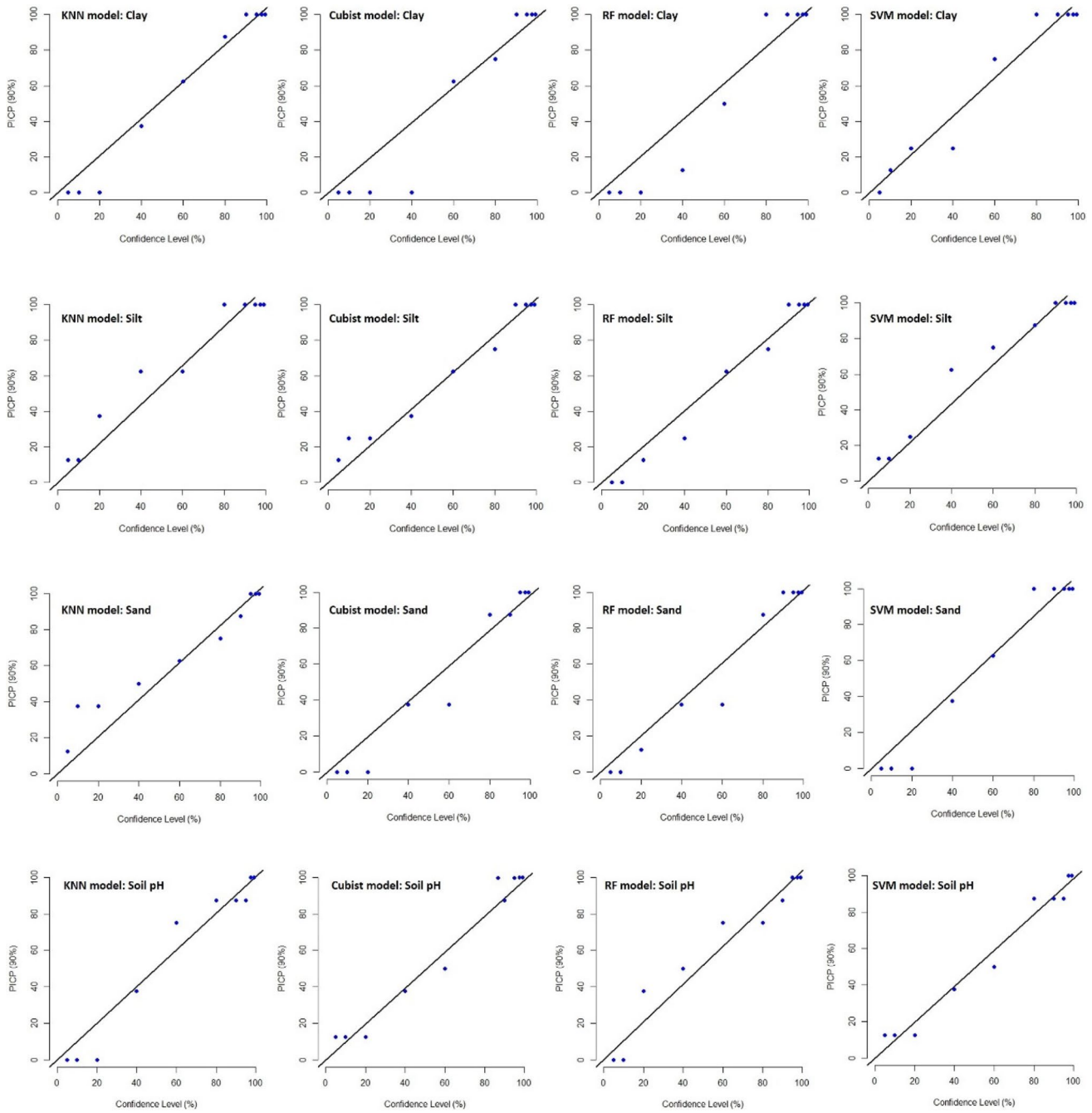

Fig. 24 Prediction interval coverage probability (PICP) plots of clay, silt, sand, and soil pH using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

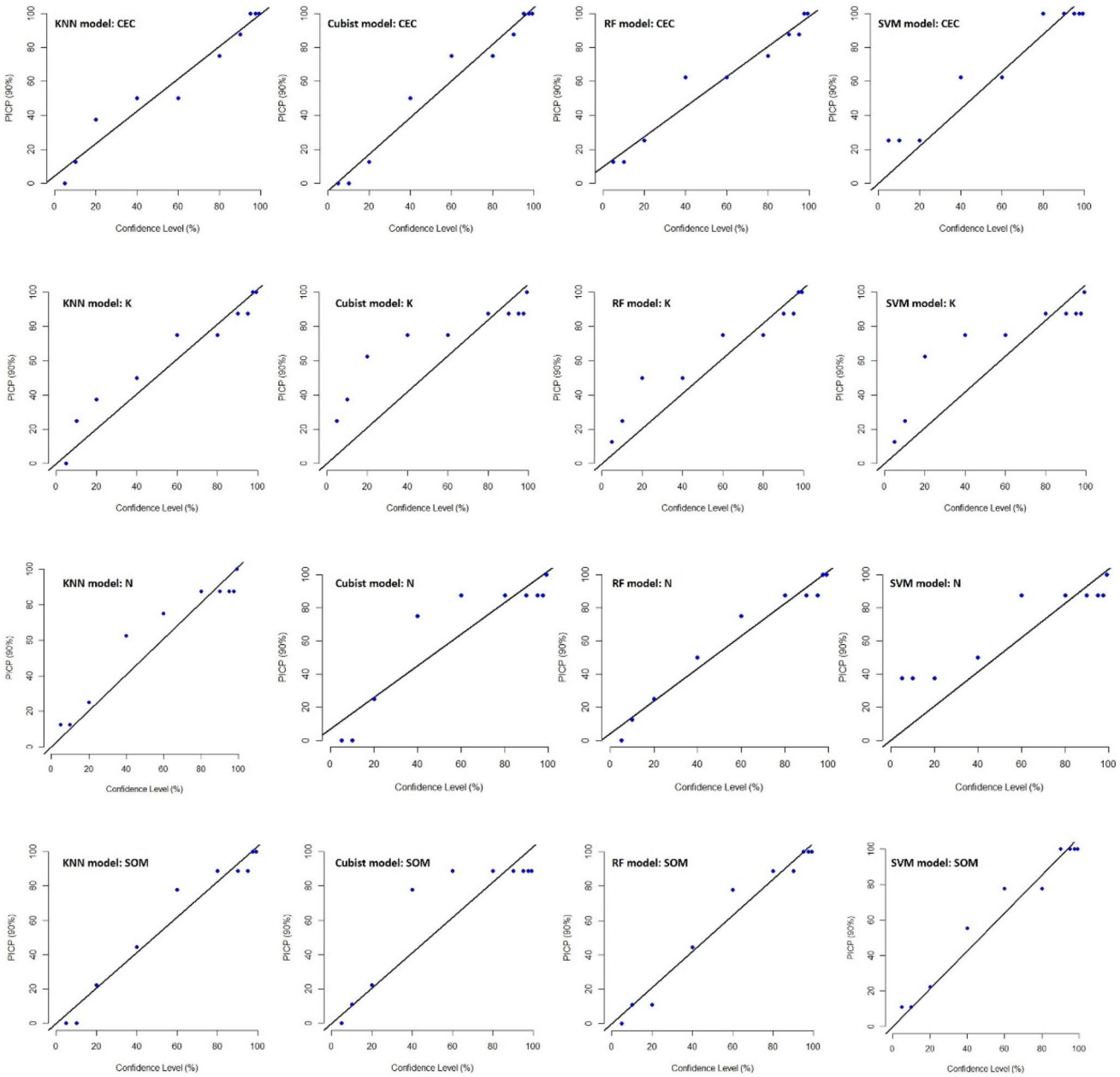

Fig. 25 Prediction interval coverage probability (PICP) plots of Cation Exchange Capacity (CEC), AK (cmol kg^{-1}), TN (%), and SOM (%) content using kNN (k-nearest Neighbor), Cubist, Random Forest (RF), and Support Vector Machines (SVM)

Acknowledgements This work was made possible thanks to funding from the Consejería de Economía, Ciencia y Agenda Digital de la Junta de Extremadura and from the European Regional Development Fund of the European Union through reference grant IB16052. We would also like to thank the European Social Fund and the Junta de Extremadura for funding PhD student Jesús Barrena González (PD18016) and ANID PIA/BASAL FB0002 for funding to Jorge Mora. We extend our sincere gratitude to the European Union's Horizon 2020 Marie Skłodowska-Curie Actions (MSCA) Research and Innovation Staff Exchange (RISE) programme under Grant Agreement number: 872384 for funding the project "Creating knowledge for UNDERsTanding ecosystem seRvicEs of agroforEStry systems through a holistic methodological framework" (H2020 MSCA-RISE "UNDERTREES").

Author contributions The study was collaboratively conceptualized by Jesús Barrena-González, Joaquín Francisco Lavado Contador, and Manuel Pulido Fernández, who established the research objectives. Victor Anthony Gabourel-Landaverde and Jorge Mora contributed to the methodology design. The validation process involved Jesús Barrena-González, Joaquín Francisco Lavado Contador, Manuel Pulido Fernández, and Victor Anthony Gabourel-Landaverde. Jesús Barrena-González conducted the formal analysis using statistical techniques to interpret the data. Field investigations and data collection were performed by Jesús Barrena-González, Joaquín Francisco Lavado Contador, and Manuel Pulido Fernández. The initial draft of the manuscript was collectively written by Jesús Barrena-González, Victor Anthony Gabourel-Landaverde, and Jorge Mora, with revisions by Joaquín Francisco Lavado Contador and Manuel Pulido Fernández. Manuel Pulido Fernández provided supervision and guidance throughout the research process. All authors reviewed and approved the final version of the manuscript prior to submission for publication.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability The data utilized in this study can be made available by the corresponding author upon reasonable request.

Declarations

Competing interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

Adeniyi OD, Brenning A, Bernini A, Brenna S, Maerker M (2023) Digital Mapping of Soil Properties Using Ensemble Machine Learning Approaches in an Agricultural Lowland Area of Lombardy. Italy Land 12(2):494

Adhikari K, Owens PR, Ashworth AJ, Sauer TJ, Libohova Z, Richter JL, Miller DM (2018) Topographic controls on soil nutrient variations in a silvopasture system. Agrosystems, Geosciences & Environment 1(1):1–15

Agam N, Kustas WP, Anderson MC, Li F, Colaizzi PD (2007) Utility of thermal sharpening over Texas high plains irrigated agricultural fields. J Geophys Res-Atmos 112(D19):110

Alfonso-Torreño A, Gómez-Gutiérrez Á, Schnabel S (2021) Dynamics of erosion and deposition in a partially restored valley-bottom gully. Land 10(1):62

Altaf S, Meraj G, Romshoo SA (2014) Morphometry and land cover based multi-criteria analysis for assessing the soil erosion susceptibility of the western Himalayan watershed. Environmental Monitoring Assessment 186:8391–8412

Andivia E, Fernández M, Alejano R, Vázquez-Piqué J (2015) Tree patch distribution drives spatial heterogeneity of soil traits in cork oak woodlands. Ann for Sci 72:549–559

Aqdam KK, Mahabadi NY, Ramezanpour H, Rezapour S, Mosleh Z, Zare E (2022) Comparison of the uncertainty of soil organic carbon stocks in different land uses. J Arid Environ 205:104805

Arabameri A, Cerda A, Rodrigo-Comino J, Pradhan B, Sohrabi M, Blaschke T, Tien Bui D (2019) Proposing a novel predictive technique for gully erosion susceptibility mapping in arid and semi-arid regions (Iran). Remote Sensing 11(21):2577

Araya SN, Ghezzehei TA (2019) Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. Water Resour Res 55(7):5715–5737

Bailey NJ, Motavalli PP, Udawatta RP, Nelson KA (2009) Soil CO 2 emissions in agricultural watersheds with agroforestry and grass contour buffer strips. Agrofor Syst 77:143–158

Beguin J, Fuglstad G-A, Mansuy N, Paré D (2017) Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma 306:195–205

Behrens T, Schmidt K, MacMillan RA, ViscarraRossel R (2018) Multi-scale digital soil mapping with deep learning. Scientific Reports 8(1):15244. https://doi.org/10.1038/s41598-018-33516-6

Bouslihim Y, Rochdi A, Paaza NEA (2021) Machine learning approaches for the prediction of soil aggregate stability. Heliyon 7(3):e06480

Bui DT, Moayedi H, Kalantar B, Osouli A, Gör M, Pradhan B, Rashid ASA (2019) Harris hawks optimization: A novel swarm intelligence technique for spatial assessment of landslide susceptibility. Sensors 19(16):3590

Ceballos A, Schnabel S (1998) Hydrological behaviour of a small catchment in the dehesa landuse system (Extremadura, SW Spain). J Hydrol 210:146–160. https://doi.org/10.1016/S0022-1694(98)00180-2

Cresto Aleina F, Runkle BR, Kleinen T, Kutzbach L, Schneider J, Brovkin V (2015) Modeling micro-topographic controls on boreal peatland hydrology and methane fluxes. Biogeosciences 12(19):5689–5704

Feng Q, Zhao W, Qiu Y, Zhao M, Zhong L (2013) Spatial heterogeneity of soil moisture and the scale variability of its influencing factors: A case study in the Loess Plateau of China. Water 5(3):1226–1242

Fitria AD, Kurniawan S (2021) Land-use changes and slope positions impact on the degradation of soil functions in nutrient stock within the Kalikungkuk micro watershed, East Java, Indonesia. Journal of Degraded Mining Lands Management 8(2):2689–2702

Forkuor G, Hounkpatin OK, Welp G, Thiel M (2017) High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. PLoS ONE 12(1):e0170478

Gallant JC, Dowling TI (2003) A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour Res 39(12):1347

Gazol A, Hereş A-M, Yuste JCJA, Meteorology F (2021) Land-use practices (coppices and dehesas) and management intensity modulate responses of Holm oak growth to drought. Agricultural Forest Meteorology 297:108235

Gerstoft P (2001) SAGA User Manual 4.1: An inversion software package. SACLANT Undersea Research Centre, La Spezia, Italy and Marine Physical Laboratory, Scripps Institution of Oceanography, University of California at San Diego, USA

Gómez Gutiérrez Á, Schnabel S, Lavado Contador JF, Pulido Fernández M (2009) Factors controlling gully erosion at different spatial and temporal scales in rangelands of SW Spain. Geophys Res Abstr 11(EGU2009):7635

Guo Z, Adhikari K, Chellasamy M, Greve MB, Owens PR, Greve MH (2019) Selection of terrain attributes and its scale dependency on soil organic carbon prediction. Geoderma 340:303–312

Hassan-Vásquez JA, Maroto-Molina F, Guerrero-Ginel JE (2022) GPS tracking to monitor the spatiotemporal dynamics of cattle behavior and their relationship with feces distribution. Animals 12(18):2383

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer, New York, pp 1–758

Hawthorne S, Miniat CF (2018) Topography may mitigate drought effects on vegetation along a hillslope gradient. Ecohydrology 11(1):e1825

Isermann M (2005) Soil pH and species diversity in coastal dunes. Plant Ecol 178:111–120

IUSS Working Group WRB (2015) World Reference Base for Soil Resources 2014, update 2015. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106. Roma, FAO

Kasraei B, Heung B, Saurette DD, Schmidt MG, Bulmer CE, Bethel W (2021) Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environmental Modelling Software 144:105139

Khaledian Y, Miller B (2020) Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling 81:401–418. https://doi.org/10.1016/j.apm.2019.12.016

Khaledian Y, Brevik EC, Pereira P, Cerdà A, Fattah MA, Tazikeh H (2017) Modeling soil cation exchange capacity in multiple countries. CATENA 158:194–200

Khlosi M, Alhamdoosh M, Douaik A, Gabriels D, Cornelis W (2016) Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. Eur J Soil Sci 67(3):276–284

Khosravi Aqdam K, Asadzadeh F, Momtaz HR, Miran N, Zare E (2022) Digital mapping of soil erodibility factor in northwestern Iran using machine learning models. Environ Monit Assess 194(5):387

Kursa MB, Rudnicki WR (2010) Feature selection with the Boruta package. J Stat Softw 36:1–13

Lassaletta L, Sanz-Cobena A, Aguilera E, Quemada M, Billen G, Bondeau A, Garnier J (2021) Nitrogen dynamics in cropping systems under Mediterranean climate: a systemic analysis. Environmental Research Letters 16(7):073002

Lavado Contador JF, Maneta M, Schnabel S (2006) Prediction of near-surface soil moisture at large scale by digital terrain modeling and neural networks [Científico]. Environ Monit Assess 121:213–232

Li H, Liang Y, Xu Q (2009) Support vector machines and its applications in chemistry. Chemometrics Intelligent Laboratory Systems 95(2):188–198

Li A, Tan X, Wu W, Liu H, Zhu J (2017) Predicting active-layer soil thickness using topographic variables at a small watershed scale. PLoS ONE 12(9):e0183742

Luizão RC, Luizão FJ, Paiva RQ, Monteiro TF, Sousa LS, Kruijt B (2004) Variation of carbon and nitrogen cycling processes along a topographic gradient in a central Amazonian forest. Glob Change Biol 10(5):592–600

Mahmoudzadeh H, Matinfar HR, Taghizadeh-Mehrjardi R, Kerry R (2020) Spatial prediction of soil organic carbon using machine learning techniques in western Iran. Geoderma Reg 21:e00260

Malone BP, Minasny B, McBratney AB (2017) Using R for digital soil mapping, vol 35. Springer, pp 1–262

Mishra U, Lal R, Slater B, Calhoun F, Liu D, Van Meirvenne M (2009) Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. Soil Sci Soc Am J 73(2):614–621

Mishra U, Gautam S, Riley WJ, Hoffman FM (2020) Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. Frontiers in Big Data 3:528441

Montanarella L, Pennock DJ, McKenzie N, Badraoui M, Chude V, Baptista I, Yagi K (2016) World's soils are under threat. SOIL 2(1):79–82

Morellos A, Pantazi X-E, Moshou D, Alexandridis T, Whetton R, Tziotzios G, Mouazen AM (2016) Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosystems engineering 152:104–116

Mosleh Z, Salehi MH, Jafari A, Borujeni IE, Mehnatkesh A (2016) The effectiveness of digital soil mapping to predict soil properties over low-relief areas. Environmental monitoring assessment 188:1–13. https://doi.org/10.1007/s10661-016-5204-8

Padarian J, Minasny B, McBratney AB (2019) Machine learning and soil sciences: A review aided by machine learning tools. SOIL 6:35–52. https://doi.org/10.5194/soil-6-35-2020

Parsaie F, Farrokhian Firouzi A, Mousavi SR, Rahmani A, Sedri MH, Homaee M (2021) Large-scale digital mapping of topsoil total nitrogen using machine learning models and associated uncertainty map. Environmental Monitoring Assessment 193:1–15

Peel MC, Finlayson BL, McMahon TA (2007) Updated world map of the Köppen-Geiger climate classification. Hydrol Earth Syst Sci 11(5):1633–1644

Plieninger T, Pulido FJ, Konold W (2003) Effects of land-use history on size structure of holm oak stands in Spanish dehesas: implications for conservation and restoration. Environ Conserv 30:61–70

Poggio L, De Sousa LM, Batjes NH, Heuvelink G, Kempen B, Ribeiro E, Rossiter D (2021) SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7(1):217–240

Pulido M, Schnabel S, LavadoContador JF, Lozano-Parra J, González F (2018) The impact of heavy grazing on soil quality and pasture production in rangelands of SW Spain. Land Degradation & Development 29(2):219–230. https://doi.org/10.1002/ldr.2501

Pulido M, Schnabel S, Contador JFL, Lozano-Parra J, Gómez-Gutiérrez Á (2017) Selecting indicators for assessing soil quality and degradation in rangelands of Extremadura (SW Spain). Ecol Indic 74:49–61. http://www.sciencedirect.com/science/article/pii/S1470160X16306537

Pulido-Fernández M, Schnabel S, Lavado-Contador JF, Miralles Mellado I, Ortega Pérez R (2013) Soil organic matter of Iberian open woodland rangelands as influenced by vegetation cover and land management. CATENA 109(2013):13–24. https://doi.org/10.1016/j.catena.2013.05.002

Pulleman M, Bouma J, Van Essen E, Meijles E (2000) Soil organic matter content as a function of different land use history. Soil Sci Soc Am J 64(2):689–693

Quinlan JR (1992, November) Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence, vol 92, pp 343–348

Ramcharan A, Hengl T, Nauman T, Brungard C, Waltman S, Wills S, Thompson J (2018) Soil property and class maps of the conterminous United States at 100-meter spatial resolution. Soil Sci Soc Am J 82(1):186–201

Reyna-Bowen L, Fernandez-Rebollo P, Fernández-Habas J, Gómez JA (2020) The influence of tree and soil management on soil organic carbon stock and pools in dehesa systems. CATENA 190:104511

Saidi S, Ayoubi S, Shirvani M, Azizi K, Zeraatpisheh M (2022) Comparison of Different Machine Learning Methods for Predicting Cation Exchange Capacity Using Environmental and Remote Sensing Data. Sensors 22(18):6890

Schnabel S, Dahlgren RA, Moreno-Marcos G (2013) Soil and water dynamics. In: Campos P, Hutsinger L, Oviedo JL, Starrs PF, Díaz M, Standiford R, Montero G (eds) Mediterranean Oak Woodland Working Landscapes: Dehesas of Spain and Ranchlands of California, Landscape Series 16. Springer-Verlag, pp 91–121

Seybold C, Grossman R, Reinsch T (2005) Predicting cation exchange capacity for soil survey using linear models. Soil Sci Soc Am J 69(3):856–863

Sharififar A (2022) Accuracy and uncertainty of geostatistical models versus machine learning for digital mapping of soil calcium and potassium. Environmental Monitoring Assessment 194(10):760

Simón N, Montes F, Díaz-Pinés E, Benavides R, Roig S, Rubio A (2013) Spatial distribution of the soil organic carbon pool in a Holm oak dehesa in Spain. Plant Soil 366(1–2):537–549. https://doi.org/10.1007/s11104-012-1443-9

Szatmári G, Pásztor L (2019) Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337:1329–1340

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

Tang Q, Xu Y, Bennett SJ, Li Y (2015) Assessment of soil erosion using RUSLE and GIS: a case study of the Yangou watershed in the Loess Plateau, China. Environmental Earth Sciences 73:1715–1724

Terefe H, Argaw M, Tamene L, Mekonnen K, Recha J, Solomon D (2020) Effects of sustainable land management interventions on selected soil properties in Geda watershed, central highlands of Ethiopia. Ecol Process 9:1–11

Tesfahunegn GB, Tamene L, Vlek PL (2011) Catchment-scale spatial variability of soil properties and implications on site-specific soil management in northern Ethiopia. Soil Tillage Research 117:124–139

Wadoux AM-C, Minasny B, McBratney AB (2020) Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth Sci Rev 210:103359

Wang Y, Shao M, Gao L (2010) Spatial variability of soil particle size distribution and fractal features in water-wind erosion crisscross region on the Loess Plateau of China. Soil Sci 175(12):579–585

Wang J, Lu P, Valente D, Petrosillo I, Babu S, Xu S, Li C, Huang D, Liu M (2022) Analysis of soil erosion characteristics in small watershed of the loess tableland Plateau of China. Ecol Indic 137:108765

Wei J-B, Xiao D-N, Zeng H, Fu Y-K (2008) Spatial variability of soil properties in relation to land use and topography in a typical small watershed of the black soil region, northeastern China. Environ Geol 53:1663–1672

Xiao S, Ou M, Geng Y, Zhou T (2023) Mapping soil pH levels across Europe: An analysis of LUCAS topsoil data using random forest kriging (RFK). Soil Use Manag 39(2):673–987

Zeraatpisheh M, Ayoubi S, Jafari A, Tajik S, Finke P (2019) Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338:445–452. https://doi.org/10.1016/j.geoderma.2018.09.006