



Improving the spatial prediction of topsoil properties in a typical grazing area using multi-season PlanetScope spectral covariates and data mining techniques

Kwanele Phinzi¹ · László Bertalan² · Gashaw Gismu Chakilu³ · Szilárd Szabó²

Received: 13 October 2024 / Accepted: 25 January 2025
© The Author(s) 2025

Abstract

Understanding the spatial distribution of topsoil properties in grassland ecosystems is essential for improving soil ecosystem services, quality, and erosion resilience. The availability of free, high-resolution satellite imagery and advanced data mining techniques offers new opportunities for efficient soil property assessment. This study aimed to evaluate the potential utility of multi-season PlanetScope imagery to predict soil organic carbon (SOC), pH, and calcium carbonate (CaCO₃). Using random sampling, 121 topsoil samples (0–30 cm depth) were collected with an auger across grasslands, bare soil, and eroded areas within a typical grazing land use. Three data mining techniques: random forest (RF), extreme gradient boosting (XGB), and support vector machines (SVM), were applied and evaluated using a 10-fold cross-validation. The results indicated that multi-season spectral covariates considerably improved the accuracy of the target soil properties compared to single-season imagery. SVM was the most effective algorithm for predicting SOC, achieving a root mean square error (RMSE) of 0.52%, mean absolute error (MAE) of 0.24%, and R² of 0.92. RF was the best-performing algorithm for predicting soil pH (RMSE=0.22, MAE=0.17, and R² = 0.97) and CaCO₃ (RMSE=0.55%, MAE=0.42%, and R² = 0.96). While XGB failed to capture the variability in soil pH, the other models generated interpretable maps that accurately represented the distribution of soil properties across different land cover categories. The green-red vegetation index (GRVI) was the most critical covariate for predicting SOC, while elevation and the topographic wetness index (TWI) were key predictors for soil pH and CaCO₃, respectively. This study underscores the potential of multi-season PlanetScope imagery for accurately predicting soil properties and recommends conducting similar studies in diverse geographical settings to validate these findings and develop more generalizable models.

Keywords Digital soil mapping · Machine learning · Multispectral · Remote sensing · Soil property prediction

Introduction

Soil quality is influenced by a combination of various topsoil physical, chemical, and biological properties, including pH, calcium carbonate (CaCO₃), and soil organic carbon (SOC) (Loveland and Webb 2003; Xiao et al. 2023). Among these, SOC is considered the most critical factor influencing soil structure, nutrient retention, and moisture-holding capacity, all of which are essential for soil fertility and crop yield (Oelofse et al. 2015). Beyond its impact on soil productivity, SOC also plays an integral role in mitigating global climate change through carbon sequestration (Lal 2004; Bradford et al. 2021). The amount of carbon stored in the soil as CO₂ is estimated to be two to three times greater than that stored in the atmosphere (Zhang et al. 2023), making it

Communicated by Hassan Babaie.

✉ Kwanele Phinzi
phinzik@unizulu.ac.za

¹ Department of Geography and Environmental Studies,
University of Zululand, KwaDlangezwa 3886, South Africa

² Department of Physical Geography and Geoinformatics,
University of Debrecen, Debrecen 4032, Hungary

³ Department of Natural Resources Management, Debarq
University, Debarq, Ethiopia

a crucial component of the global carbon cycle (Post et al. 1982). CaCO_3 also plays a crucial role in soil biogeochemistry by increasing soil pH (i.e., neutralizing soil acidity) and extractable Ca, thereby enhancing SOC content (Rowley et al. 2018, 2020). It also acts as a cementing agent in soil, with higher soil strength being associated with increased CaCO_3 content (Fukue et al. 1999). Soil pH, a measure of soil acidity or alkalinity, is another critical component of productive soils. It regulates plant nutrient availability by altering the chemical forms and reactions of nutrients, thereby influencing plant growth and biomass production (Minasny et al. 2016; Oshunsanya 2018; Neina 2019). (Xiao et al. 2023) Maintaining plant nutrient availability is particularly crucial in livestock-grazing areas, like the study site, as it directly impacts forage quality and quantity.

These soil properties also influence the soil structure, which plays a key role in mitigating erosion (Teague and Kreuter 2020). For example, soils with stable structures are often more resistant to erosion. Thus, predicting the spatial distribution of these topsoil properties is essential for enhancing soil ecosystem services, improving soil quality, and increasing resilience against soil erosion (Blanco-Canqui et al. 2013; Nokhandan et al. 2023). Data mining techniques like machine learning (ML) have gained significant traction in soil science, offering promising solutions for predicting various soil properties, which in turn may support agricultural decision-making and soil quality management (Padarian et al. 2020). ML techniques analyze the spatial variation of specific soil properties using a set of covariates that represent different soil-forming factors, such as vegetation, topography, climate, and parent material (Were et al. 2015). By establishing statistical relationships between in-situ soil property observations and these covariates, ML models can infer and map soil properties at locations lacking direct observations (Heung et al. 2016; Xiao et al. 2023). ML methods such as random forest (RF), extreme gradient boosting (XGB), and support vector machines (SVM) have demonstrated high efficacy in predicting soil properties, excelling in capturing non-linear relationships within soil data (Padarian et al. 2020; Wang and Zhou 2022; Hateffard et al. 2024).

The increasing availability of free remote sensing data, from which most environmental covariates are derived, has significantly accelerated the adoption and implementation of ML in digital soil analysis (Padarian et al. 2020; Motia and Reddy 2021). Traditionally, mapping soil properties involved interpolating observations from sampled locations to predict properties at unsampled locations using geostatistical methods. One notable advantage of geostatistical methods like kriging is their ability to account for spatial autocorrelation in model residuals (Hengl et al. 2015). However, there are several reasons to consider remote

sensing and ML as preferable alternatives to geostatistical techniques. A frequently cited issue with geostatistical methods is the propagation of interpolation errors in soil property estimations, especially when sampling points are sparse (Were et al. 2015). Achieving accurate predictions thus requires a dense and representative set of observations, a process that is both labor-intensive and time-consuming due to the significant spatial variability of soil properties (Song et al. 2017). In contrast, the combination of remote sensing and data mining techniques provides a rapid and cost-effective approach for mapping the spatio-temporal distribution of soil properties with limited in-situ observations (Zhai 2019; Wang and Zhou 2022). When verified with field survey, this approach can generate accurate, continuous, and spatially detailed information on soil properties (Du et al. 2015; Naimi et al. 2022; Keshavarzi et al. 2023).

Satellite remote sensing offers a relatively high temporal resolution compared to ground-based and airborne sensors (Xiao et al. 2019). Additionally, it provides broader aerial coverage, making it highly suitable for rapid digital soil mapping. Recent studies have demonstrated that combining static terrain derivatives with dynamic covariates from multitemporal satellite imagery significantly enhances the prediction of soil properties such as calcium carbonate equivalent, SOC, and sand content (Fatholouloumi et al. 2020). Lamichhane et al. (2021) applied the random forest (RF) algorithm to predict SOC content using multi-season Sentinel-2 data, discovering that multi-season data better capture the dynamic conditions of surface soils compared to single-season images. Similarly, Wang and Zhou (2022) estimated soil organic matter (SOM) content using single and multi-season Sentinel-2 A images, finding that the latter provided superior results. The rationale for using multi-season imagery lies in the fact that topsoil properties also vary in time as influenced by land cover dynamics. Thus, incorporating images from all seasons can capture such variation, thereby enhancing the predictive capacity of ML models (Lamichhane et al. 2021).

The potential of the recently launched PlanetScope data, with its high spatial (3 m) and spectral (8 bands) resolutions, remains underexplored in the realm of digital soil mapping, despite its free availability for scholarly research and education purposes. Although some studies have compared Landsat-8 and Sentinel-2 with PlanetScope in mapping soil properties (Žižala et al. 2019; Silvero et al. 2021), the application of multi-season PlanetScope data is still lacking. Against this background, the objectives of this study are: (a) to evaluate the potential of multi-season PlanetScope imagery to improve the accuracy of digital soil mapping and, (b) select the best predictive model for each soil property. The underlying hypothesis was that multi-season imagery outperforms single-season imagery for predicting the

target soil properties. To achieve these objectives, three ML algorithms, including RF, XGB, and SVM, were applied and their predictive performance was evaluated through a repeated 10-fold cross-validation procedure.

Materials and methods

Study area

This study was conducted in a small semi-arid region of eastern South Africa, characterized primarily by subsistence farming, particularly livestock rearing. The study area covers 0.20 km², stretching between latitudes: 30° 42' 50" S, and 30° 43' 20" S and longitudes: 29° 26' 04" E and 29° 26' 20" E (Fig. 1). It features an undulating topography with varying elevations. The northern part, predominantly vegetated, is at a higher elevation, while the lower elevations in the footslope are mainly eroded. Grassland, which constitutes 37% of the surface cover in this area, is the dominant vegetation type. However, overgrazing and other

human activities are increasingly transforming these grasslands into bare soil and built-up areas. The region is entirely covered by Luvisols and receives an average annual rainfall of 671 mm, primarily during the summer months from November to February.

Soil sampling and measurements

Sampling design is a critical first step in digital soil mapping, with the selection of an appropriate sampling method being a key component of this process (Brus 2019). In this study, simple random sampling was chosen due to budget limitations and the need to position samples in strategically accessible locations, as the study area is characterized by irregular terrain and large gullies. A total of 121 topsoil samples (0–30 cm depth) were randomly collected using a soil sampling auger. A hand-held Garmin eTrex® 20x GPS device, with a horizontal accuracy of 3.65 m, was used to record the coordinates of sampling locations, ensuring that each soil sample was taken at least 10 m apart. The samples were then transported to the Institute

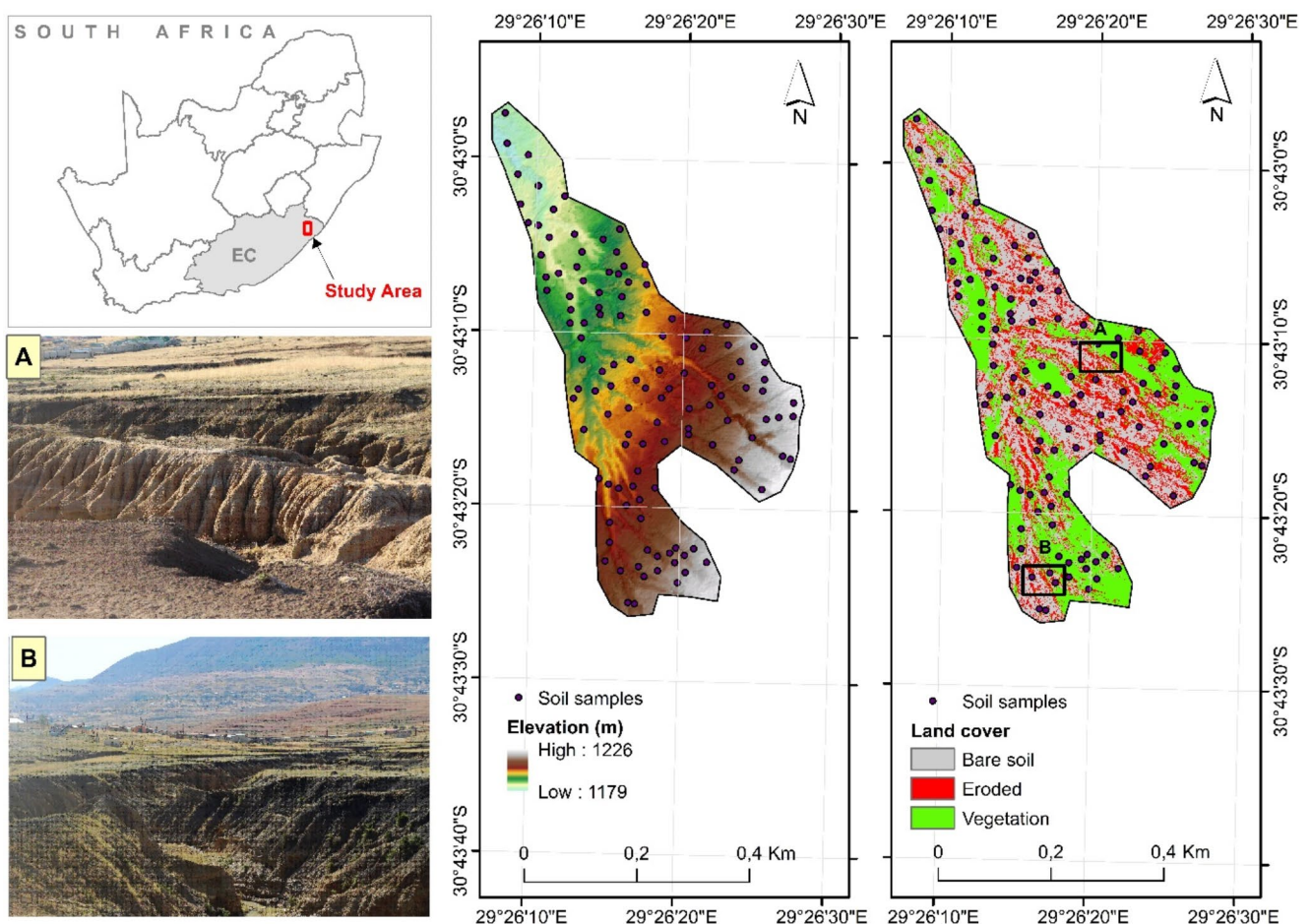


Fig. 1 Study area location in the Eastern Cape (EC) Province of South Africa. A and B are field photographs showing dominant land cover in the study area

of Geoscience laboratory, University of Debrecen, for further analysis. Before any analysis, the samples were oven-dried for a minimum of 48 h, crushed, and sieved through a 0.2 mm mesh to remove rock fragments and plant debris. Standard laboratory procedures were followed to analyze the soil properties. Soil pH was measured with a Consort C3010 Multiparameter Analyzer, manufactured by Consort bvba (<https://consort.be/>). CaCO_3 content was determined using the volumetric Scheibler calcimeter method, which is recognized internationally for determining carbonate content in accordance with ISO 10693:1995 standards. SOC was determined through the Walkley-Black wet oxidation method (Walkley and Black 1934). Each measurement was performed in triplicate to ensure accurate results. Descriptive statistics for each soil property were computed using R software and visualized with the ggplot2 package. To analyze the distribution of target soil properties across different land covers, a high-resolution (1.5 m) land cover map of the study area was generated. This map, with an overall accuracy of 78%, was produced from a pansharpened Satellite pour l'Observation de la Terre (SPOT-7) image using the random forest (RF) classifier. To examine the effects of slope steepness on the target soil properties, boxplots depicting the distribution of each soil property across various slope steepness classes were also produced.

Feature set generation

Six topographic covariates including elevation, topographic wetness index (TWI), curvature, slope, slope length and steepness factor (LS-factor), and aspect were generated from a 51.2 cm digital elevation model (DEM) using ArcGIS 10.4. This DEM was generated using Structure-from-Motion (SfM) processing of archive aerial imagery (Eltner and Sofia 2020), which were provided by the National Geospatial Information (NGI), a key entity in spatial information management in South Africa. The dataset comprised of nine high-resolution images (7740×13824 pixels, 107.00 MPixels) with a resolution of 600 dpi in TIF format, covering an area of 75 km. The SfM processing was performed using Agisoft Metashape Professional v2.0 software. Initially, the images were aligned using the SIFT algorithm,

which identified common keypoints between the overlapping images. Next, ground control points (GCPs) were utilized to georeference the model due to the absence of proper exterior orientation data (Bakker and Lane 2017). A total of 28 GCPs were selected at strategic locations with open surfaces, such as road crossings, where vegetation cover was minimal. The elevation values were extracted from the SRTM database and used to mark the evenly distributed GCPs. The overall root mean square error (RMSE) of the applied GCPs was 0.102 m. The next step involved the generation of a dense point cloud, comprising 311,003,382 points (3.82 points/sq m), which was later classified to remove surface objects and vegetation. Finally, the DEM was interpolated with a spatial resolution of 51.2 cm.

Additionally, four PlanetScope images representing different seasons: Summer (08 December 2023), Spring (23 March 2023), Winter (10 June 2023), and Autumn (21 September 2023) were downloaded from the Planet Labs website (<https://www.planet.com>). PlanetScope has a spatial resolution of 3 m and includes eight spectral bands: coastal blue, blue, green, yellow, red, red-edge, and near-infrared (NIR). From each image, nineteen covariates were computed, comprising eight raw spectral bands and five vegetation indices (VIs): normalized difference vegetation index (NDVI), transformed normalized difference vegetation index (TNDVI), atmospheric resistance vegetation index (ARVI), green-red vegetation index (GRVI), and salinity index (SI) (Table 1). These VIs were selected based on the PlanetScope bands but also because they are among the most widely used indices as covariates in previous digital soil mapping studies (Taghizadeh-Mehrjardi et al. 2021; Liu et al. 2022; Yuvaraj et al. 2022). All covariates were computed using spatial analyst tools in ArcMap 10.4. To ensure consistency between DEM-derived and satellite-derived covariates, the DEM-derived covariates were resampled to match the 3 m spatial resolution of the PlanetScope data.

Table 1 Spectral indices derived from the PlanetScope image

Name	Formula	References
Normalized difference vegetation index (NDVI)	$\frac{(\text{NIR} - \text{Red})}{(\text{NIR} + \text{Red})}$	(Rouse et al. 1974)
Transformed normalized difference vegetation index (TNDVI)	$\sqrt{\frac{(\text{NIR} - \text{Red})}{(\text{NIR} + \text{Red})}} + 0.5$	(Tucker 1979)
Salinity index (SI)	$\frac{(\text{Red} - \text{NIR})}{(\text{Green} + \text{NIR})}$	(Taghizadeh-Mehrjardi et al. 2021)
Atmospheric resistance vegetation index (ARVI)	$\frac{(\text{NIR} - (2 \times \text{Red}) + \text{Blue})}{(\text{NIR} + (2 \times \text{Red}) + \text{Blue})}$	(Kaufman and Tanre 1992)
Green-red vegetation index (GRVI)	$\frac{(\text{Green} - \text{Red})}{(\text{Green} + \text{Red})}$	(Tucker 1979)

Note Blue=band 2, Green=band 4, Red=band 6, NIR (near infra-red)=band 8

Prediction models

Random forest (RF)

Random Forest (RF), introduced by (Breiman 2001), is a widely used machine learning method that enhances decision tree performance by aggregating numerous randomly generated trees through bootstrap sampling. Each tree is trained on a subset of the data, while the out-of-bag samples, not included in training, are used for evaluation (Abdi 2020). Additionally, RF randomly selects a subset of features for each tree, reducing overfitting and enhancing accuracy (Irizarry 2019). Advantages of RF include its non-parametric nature, high classification accuracy, ability to assess variable importance, and flexibility for various data analyses (Rodriguez-Galiano et al. 2012). The “rf” function from the caret package (Kuhn 2008) was used for RF, with a grid search method to tune the mtry parameter.

Extreme gradient boosting (XGB)

Extreme Gradient Boosting (XGB) is an advanced version of the gradient boosting decision tree algorithm, designed to improve prediction performance and computational efficiency (Chen and Guestrin 2016). Unlike classical gradient boosting machines (GBM), which build additive models of weak learners by optimizing a loss function (Friedman 2001), XGB incorporates an objective function combining a loss function and a regularization term. This dual approach controls model complexity and prevents overfitting, thereby enhancing predictive accuracy (Ahmetoglu and Das 2022). XGB has numerous hyperparameters that can be adjusted to make the model more robust. These include eta (learning rate), max_depth (maximum depth of the trees), subsample (fraction of observations used for each tree), colsample_bytree (fraction of features used for each tree), gamma (minimum loss reduction required to partition a leaf node), and min_child_weight (minimum sum of instance weight needed in a child). A grid search was performed to find the

optimal combination of these parameters. The model was implemented using the “xgbTree” function from the xgboost package (Chen et al. 2015).

Support vector machines

Support Vector Machine (SVM), a binary classifier based on statistical learning theory (Vapnik 1999), works by identifying a hyperplane that optimally separates data points. Using the kernel trick, SVM expands the feature space to achieve perfect class separability, employing various kernel functions for flexibility in finding complex non-linear boundaries (Boehmke and Greenwell 2019). Key hyperparameters include the cost (C) parameter, which penalizes misclassifications, and the sigma parameter, which influences the decision boundary. The “svmRadial” function from the caret package (Kuhn 2008) was selected as it provided the highest accuracy among kernel functions. A grid search was conducted for finding an optimal combination of the C and sigma parameters.

Experimental procedure for soil property prediction

The prediction of each soil property was carried out under five different scenarios, each representing a set of covariates from various seasons. This was done to examine the impact of both individual seasons and a combination of all seasons on predicting the target soil property (Table 2). Throughout these scenarios, topographic covariates were held constant to focus on the influence of multi-season PlanetScope data on prediction accuracy. The RF, XGB, and SVM models were calibrated and validated using a repeated 10-fold cross-validation (CV) method. The root mean square error (RMSE) was utilized to optimize the hyperparameters of each machine learning model during training. This approach did not necessitate a train-test split of the dataset (Phinzi et al. 2020). Instead, the entire dataset of 121 observations was randomly divided into ten subsamples. In each iteration, nine subsamples were used to train the models while the remaining one was used for testing. This process was repeated until each subsample had been used as a test set against all others. The entire procedure was repeated three times. To assess the performance of the models for each target soil property, the RMSE, mean absolute error (MAE), and coefficient of determination (R^2) of 30 models (i.e., 10-fold CV with three repetitions) were calculated. These accuracy metrics were computed within the 10-fold CV procedure as follows:

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

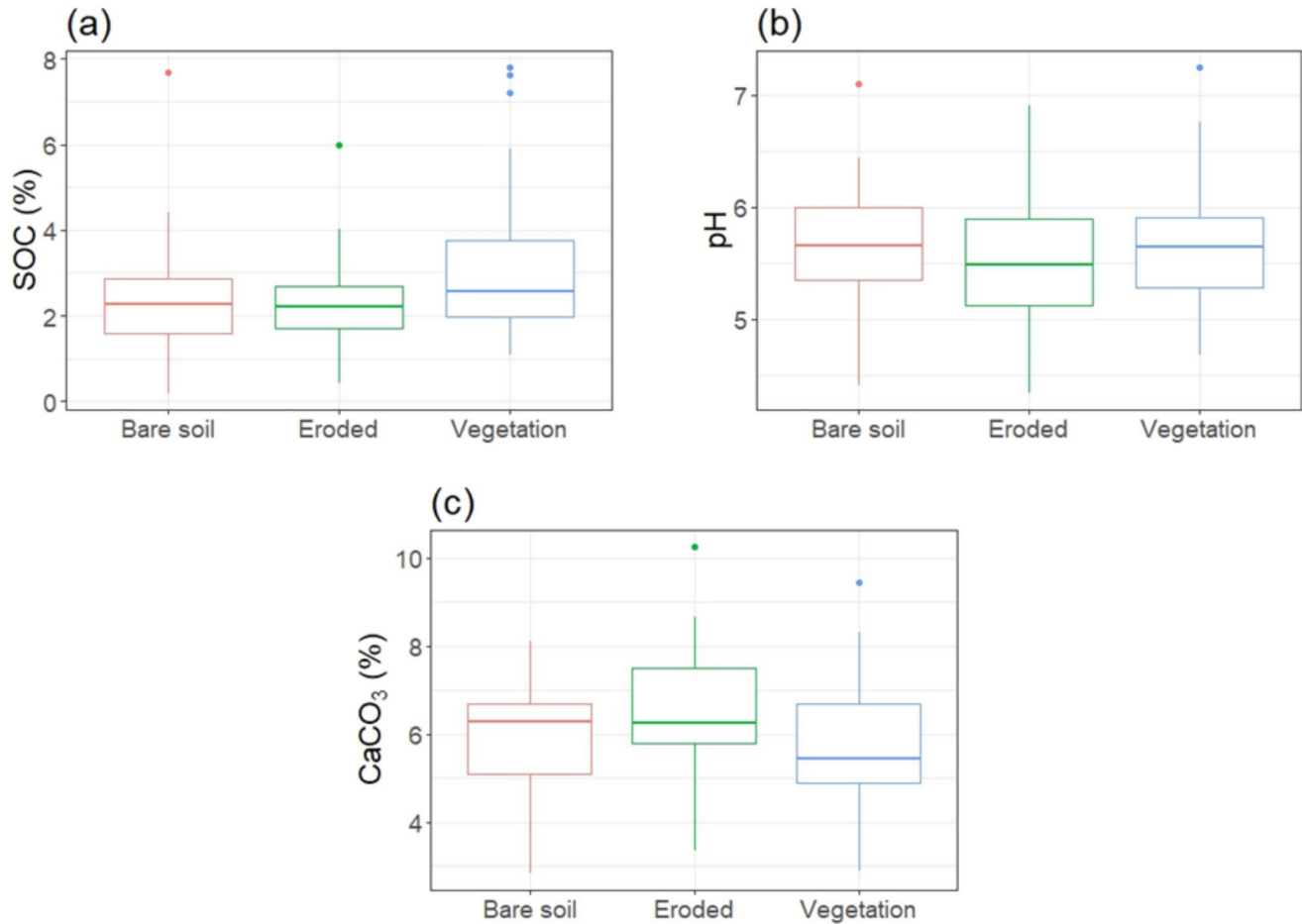
Table 2 Different combinations of the covariates involved in each prediction scenario

Scenario	Covariates	Number of covariates
Spring (March)	Six topographic + thirteen Spring spectral covariates	19
Winter (June)	Six topographic + thirteen Winter spectral covariates	19
Autumn (September)	Six topographic + thirteen Autumn spectral covariates	19
Summer (December)	Six topographic + thirteen Summer spectral covariates	19
All seasons	Six topographic + spectral covariates from all seasons	58

Table 3 Summary statistics of the soil properties (SOC=soil organic carbon, pH, and CaCO₃=calcium carbonate)

Soil property	Min	Max	Range	Mean	Median	SD
SOC (%)	0.19	7.81	7.62	2.67	2.46	1.40
pH	4.34	7.25	2.92	5.61	5.63	0.54
CaCO ₃ (%)	2.84	10.25	7.41	5.98	6.07	1.36

Note Min=minimum, Max=maximum, SD=standard deviation

**Fig. 2** Target soil properties across different land covers: **(a)** SOC, **(b)** pH, and **(c)** CaCO₃

$$MAE = \frac{1}{n} \sum_{i=1}^n (P_i - O_i)^2 \quad (2)$$

$$R^2 = \frac{\frac{1}{n} \sum_{i=1}^n (P_i - \hat{O}_i)^2}{\frac{1}{n} \sum_{i=1}^n (O_i - \hat{O}_i)^2} \quad (3)$$

Where n represents the entire dataset of 121 observations, P_i and O_i are the predicted and observed values for a target soil property, respectively. $\hat{P}_i$ and $\hat{O}_i$ denote the mean values of the predicted and observed soil properties, respectively. Lower values of RMSE and MAE, along with higher R^2 values, indicate better predictive performance of the model.

Results

Statistical analysis

SOC ranged from 0.19 to 7.81%, with a mean value of 2.67% (Table 3). Samples from vegetated locations recorded relatively high SOC values compared to bare soil and eroded areas, which exhibited similar median SOC values. The mean pH value of all sampled locations was 5.63, with a relatively narrow range of 2.92, indicating minimal variation in pH values across different sampling locations. Indeed, the pH boxplots for different land covers appear similar, although eroded areas had slightly lower median pH values (Fig. 2). CaCO₃ values ranged from 2.84 to 10.25%,

showing considerable variation across different land covers. Eroded and bare soil sampling locations recorded the lowest and highest CaCO_3 values, respectively. Samples from vegetated locations had a lower median CaCO_3 value compared to the overall median value of 6.07%.

Higher SOC values were associated with flat and gently sloping areas and decreased as slope steepness increased (Fig. 3). The highest SOC value (7.81%) was observed within a slope steepness range of 0–3%, while the lowest value (0.19%) was found in the 12–20% slope steepness range. The median SOC values across various slope steepness categories showed little variation, generally hovering around the overall median value of 2.46%. Similarly, pH values followed a comparable pattern, with higher pH values occurring in less steep areas. Areas with gentle slopes of 0–3%, 3–12%, and 12–20% exhibited distinctively higher median pH values ($\text{pH} > 5.5$) compared to much steeper areas (25–35% and >35%). Regarding CaCO_3 , the median values also varied across different slope steepness classes, with less steep areas having higher median values.

Model performance

Models demonstrated varying performance in predicting target soil properties under different prediction scenarios (Table 4). Generally, models utilizing the entire feature set (all seasons) outperformed those based on a single season's feature set. For SOC, the SVM model using all seasons achieved the lowest RMSE (0.52) and MAE (0.24), though it had a slightly lower R^2 (0.92) compared to the Autumn RF model ($R^2=0.93$). In single-season scenarios, the SVM model performed poorly, recording the highest RMSE (1.26–1.31) and MAE (0.83–0.87) values, and explaining only about 20–30% of the total variation in SOC (i.e., $R^2 < 0.30$). RF and XGB models performed exceptionally well in predicting SOC under the Autumn scenario compared to other scenarios.

The RF model consistently outperformed other models in predicting pH across all scenarios. Notably, when utilizing data from all seasons, the RF model achieved the lowest RMSE (0.22) and MAE (0.17), accounting for 97% of the variation in pH. A similar trend was observed for

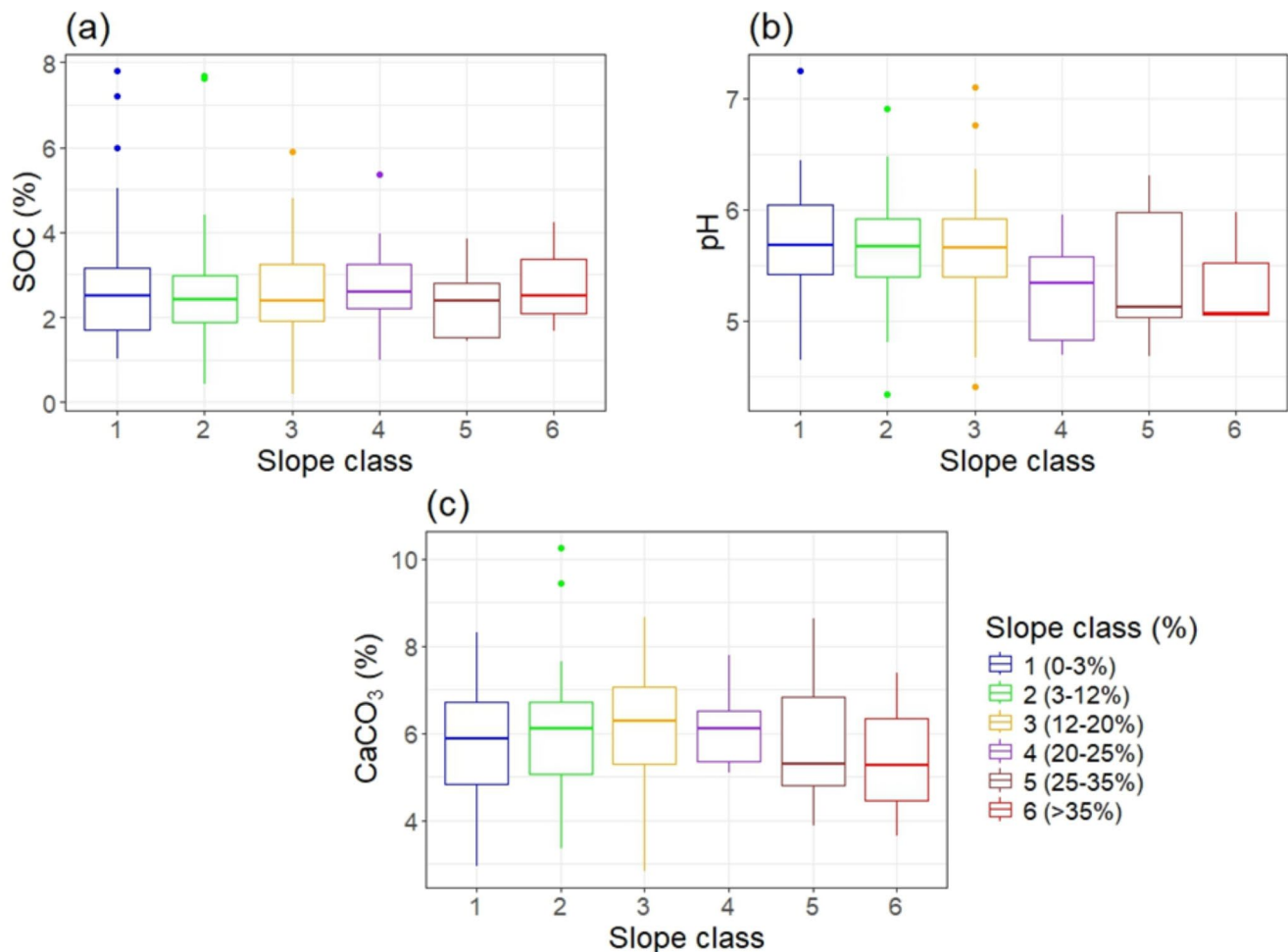


Fig. 3 Target soil properties across different slope classes: (a) SOC, (b) pH, and (c) CaCO_3

Table 4 Performance of machine learning models for predicting soil properties under different scenarios, with a color gradient from green (indicating high performance) to red (indicating low performance)

Soil property	Scenario	XGB			RF			SVM		
		RMSE	MAE	R ²	RMSE	MAE	R ²	RMSE	MAE	R ²
SOC	Spring	0.99	0.72	0.68	0.71	0.49	0.92	1.32	0.87	0.19
	Winter	0.94	0.69	0.72	0.72	0.50	0.92	1.29	0.86	0.22
	Autumn	0.86	0.63	0.78	0.70	0.48	0.93	1.26	0.83	0.27
	Summer	0.93	0.68	0.73	0.72	0.51	0.91	1.31	0.85	0.23
	All	0.94	0.67	0.76	0.71	0.49	0.92	0.52	0.24	0.92
pH	Spring	0.54	0.42	0.00	0.22	0.17	0.95	0.45	0.31	0.37
	Winter	0.54	0.42	0.00	0.23	0.17	0.94	0.31	0.19	0.73
	Autumn	0.54	0.42	0.00	0.26	0.20	0.96	0.48	0.33	0.23
	Summer	0.54	0.42	0.00	0.23	0.18	0.95	0.35	0.21	0.63
	All	0.54	0.42	0.00	0.22	0.17	0.97	0.33	0.18	0.67
CaCO ₃	Spring	1.02	0.78	0.70	0.66	0.51	0.94	1.25	0.92	0.17
	Winter	0.95	0.73	0.72	0.65	0.51	0.94	0.78	0.47	0.7
	Autumn	1.07	0.83	0.63	0.68	0.53	0.94	1.25	0.92	0.15
	Summer	1.05	0.83	0.61	0.66	0.51	0.93	1.25	0.96	0.16
	All	0.90	0.70	0.77	0.55	0.42	0.96	0.78	0.44	0.71

RMSE = root mean square error, MAE = mean absolute error, R² = coefficient of determination, XGB = extreme gradient boosting, RF = random forest, SVM = support vector machines, SOC = soil organic carbon, CaCO₃ = calcium carbonate, All = all seasons.

predicting CaCO₃, where the RF model also excelled with the all-season feature set, recording RMSE, MAE, and R² values of 0.55, 0.42, and 0.96, respectively. Conversely, the XGB model exhibited poor performance in predicting pH across all scenarios, with the highest RMSE (0.54) and MAE (0.42) values, and an R² of 0, indicating it could not explain the variation in soil pH. However, the XGB model performed relatively well for CaCO₃ prediction across all scenarios, with RMSE values ranging from 0.90 to 1.07, MAE from 0.70 to 0.83, and R² from 0.61 to 0.77, achieving the best results with the entire feature set. The SVM model showed variable performance, returning the best results for pH prediction with the Winter feature set (RMSE=0.31, MAE=0.19, and R²=0.73) and for CaCO₃ prediction with the all-season feature set (RMSE=0.78, MAE=0.44, and R²=0.71).

Spatial distribution of predicted soil properties

The spatial distribution maps of the target soil properties, generated using the best predictive models with the lowest RMSE and MAE values, are illustrated in Fig. 4. For comparative purposes, a land cover map was included to understand how each predicted soil property varies across different land cover categories. While the CaCO₃ map

displayed a distinct visual pattern compared to the SOC map, both exhibited relatively consistent pattern across bare soil and eroded areas. In contrast, the pH map did not exhibit a clear generalizable pattern of variation in pH with land cover.

We also compared the observed and predicted soil properties within each land cover category. The predicted soil property values closely matched the observed values across different land cover categories, as evidenced by the boxplots and scatterplots (Fig. 5). Specifically, the median values of both observed and predicted SOC were identical for each land cover category. Similarly, the predicted and observed median values for pH were the same. However, the predicted and observed median values for CaCO₃ showed slight differences. The predicted minimum and maximum values deviated from the observed values for all soil properties, suggesting that the models underestimated the target properties. Additionally, the boxplots for all predicted soil properties were slightly narrower compared to those for the observed soil properties.

Variable importance

In the analysis, only the variable importance of the best predictive model for each soil property was considered (Fig. 6).

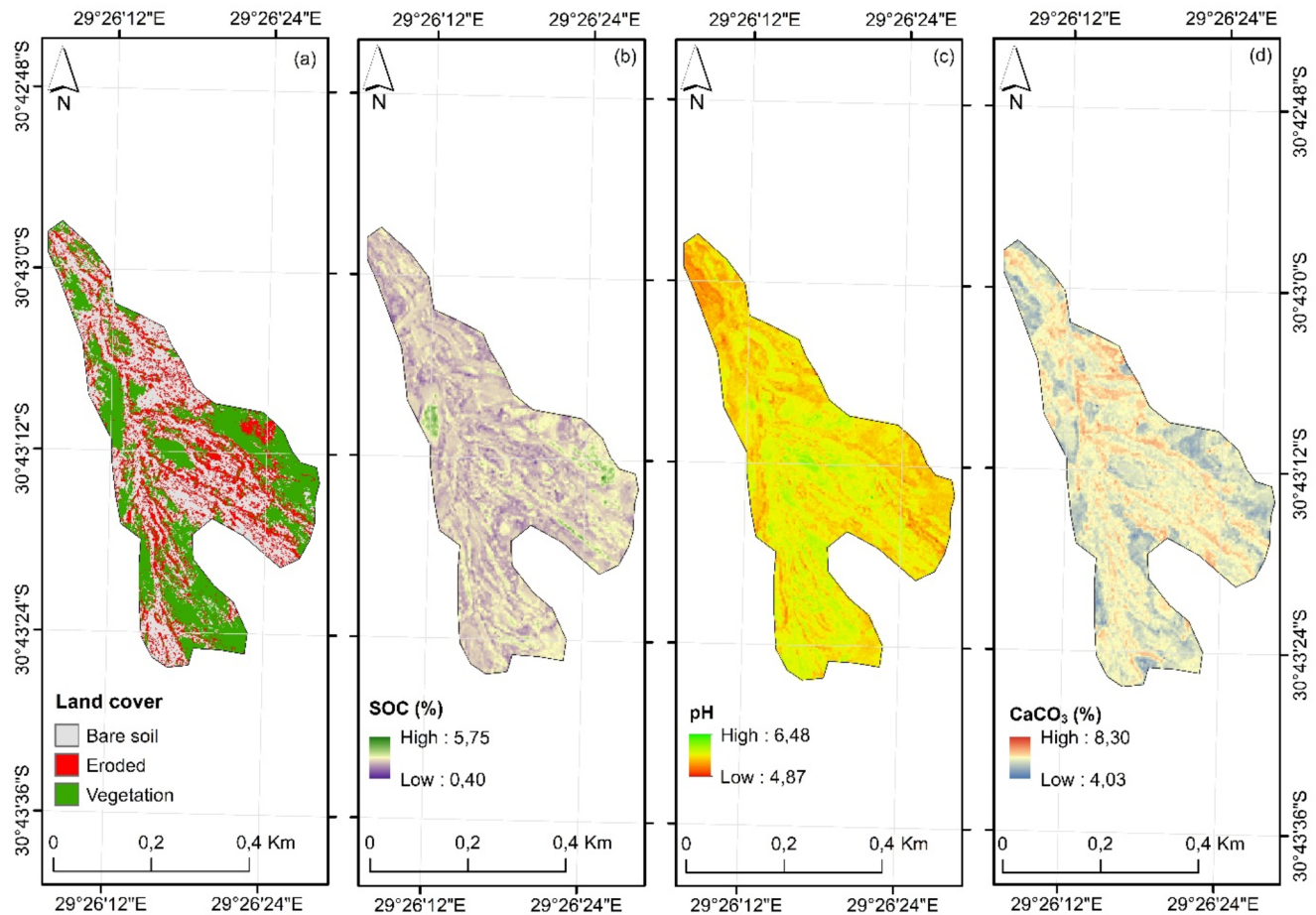


Fig. 4 Spatial distribution maps for (a) land cover and predicted soil properties: (b) SOC, (c) pH, and (d) CaCO_3

Spectral covariates (PlanetScope bands and VIs) from multiple seasons were more critical than topographic covariates for predicting SOC. Notably, the top six most important spectral covariates were GRVI, Green I band, two NIR bands, and Red Edge band. These variables were derived from the images of various seasons, with Winter and Spring variables each appearing twice in the top six. The LS factor was the least significant variable for predicting SOC. In contrast, topographic variables such as elevation, slope, LS factor, and aspect were the most critical covariates for predicting soil pH, followed by two Winter spectral covariates (GRVI and NIR band). TNDVI and NDVI from the Summer season were among the least important variables for predicting pH. For CaCO_3 , the most important variables were TWI, aspect, Coastal Blue band, GRVI, and slope, while the Green I band and SI from the Autumn and Spring seasons, respectively, had a minimal contribution.

Discussion

Our study revealed that utilizing multi-season spectral covariates was more effective for mapping the target soil properties compared to using covariates from a single-season image. This finding aligns with the results of Wang and Zhou (2022) and Lamichhane et al. (2021). The improved efficiency can be attributed to the unique spectral characteristics provided by different seasons, influenced by varying environmental factors such as soil moisture. These seasonal variations may have enhanced the detection of soil properties by offering a broader range of spectral data. As soil moisture levels fluctuate seasonally, they affect the soil's reflectance in various spectral bands (Lobell and Asner 2002; Lesaignoux et al. 2013). Additionally, certain soil properties may be more detectable or pronounced during specific seasons. For example, spectral features such as soil darkness which indicates the presence of SOC may be more visible in rainy season (i.e. Summer, in our case) than in others (Mandal 2016). Multi-season imagery captured these seasonal changes, leading to a more accurate and dynamic representation of soil properties.

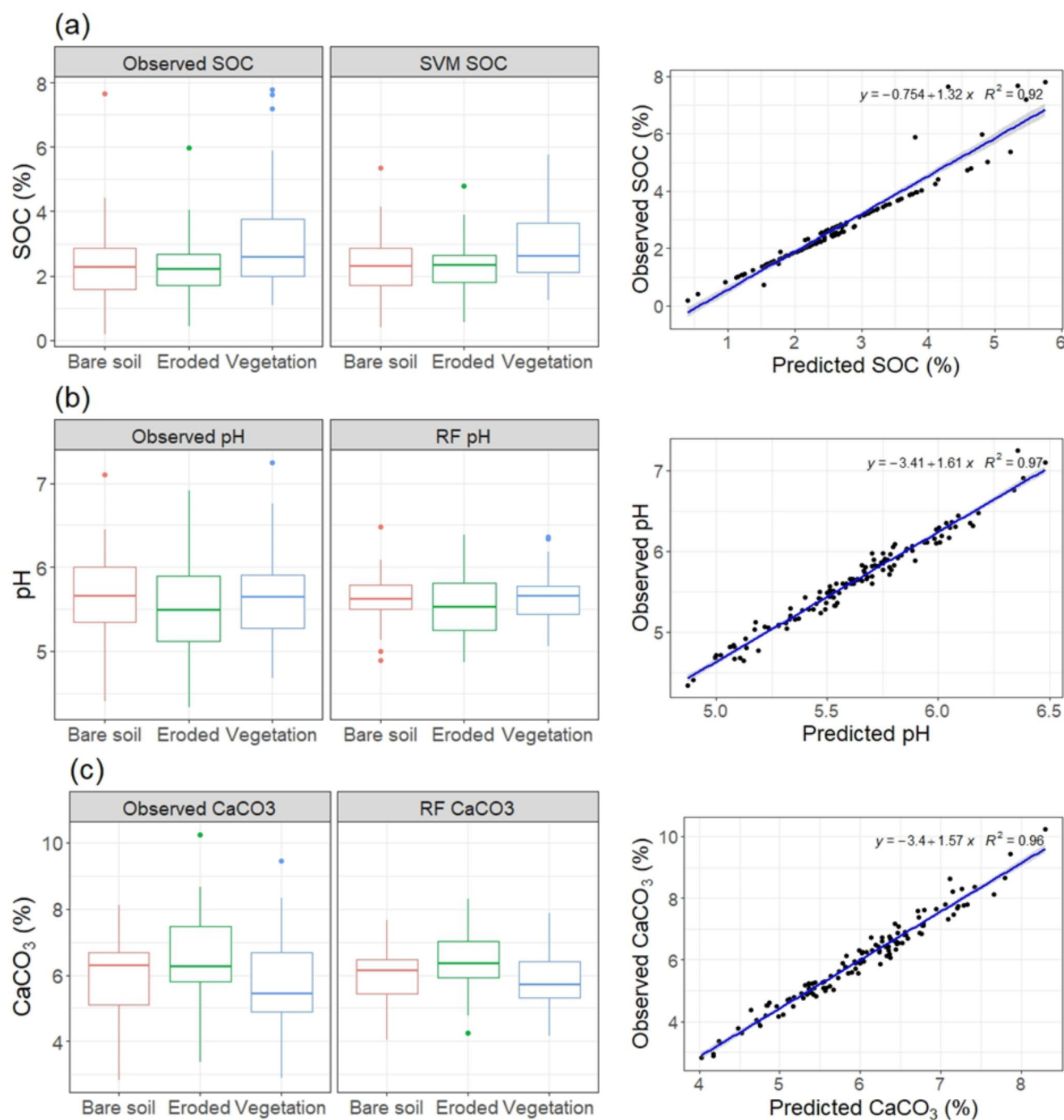


Fig. 5 Boxplots and corresponding scatterplots comparing observed and predicted SOC (a) using the SVM model, and pH (c) and CaCO₃ (d) using the RF model

Different data mining techniques can significantly impact the prediction accuracy of digital soil mapping models. In our study, RF consistently produced impressive results ($R^2 > 0.90$) across all scenarios for each soil property, with only one instance where SVM slightly outperformed it for SOC. RF outperformed both SVM and XGB in predicting soil pH and CaCO₃. This consistent high performance,

even with single-season datasets, underscores the stability and robustness of RF compared to other algorithms, which can be attributed to its functioning mechanism. RF constructs an ensemble model of decision trees from random subsets of features (Nguyen et al. 2015). Using a larger feature set for each decision tree enhances the robustness of individual trees, thereby reducing the risk of overfitting

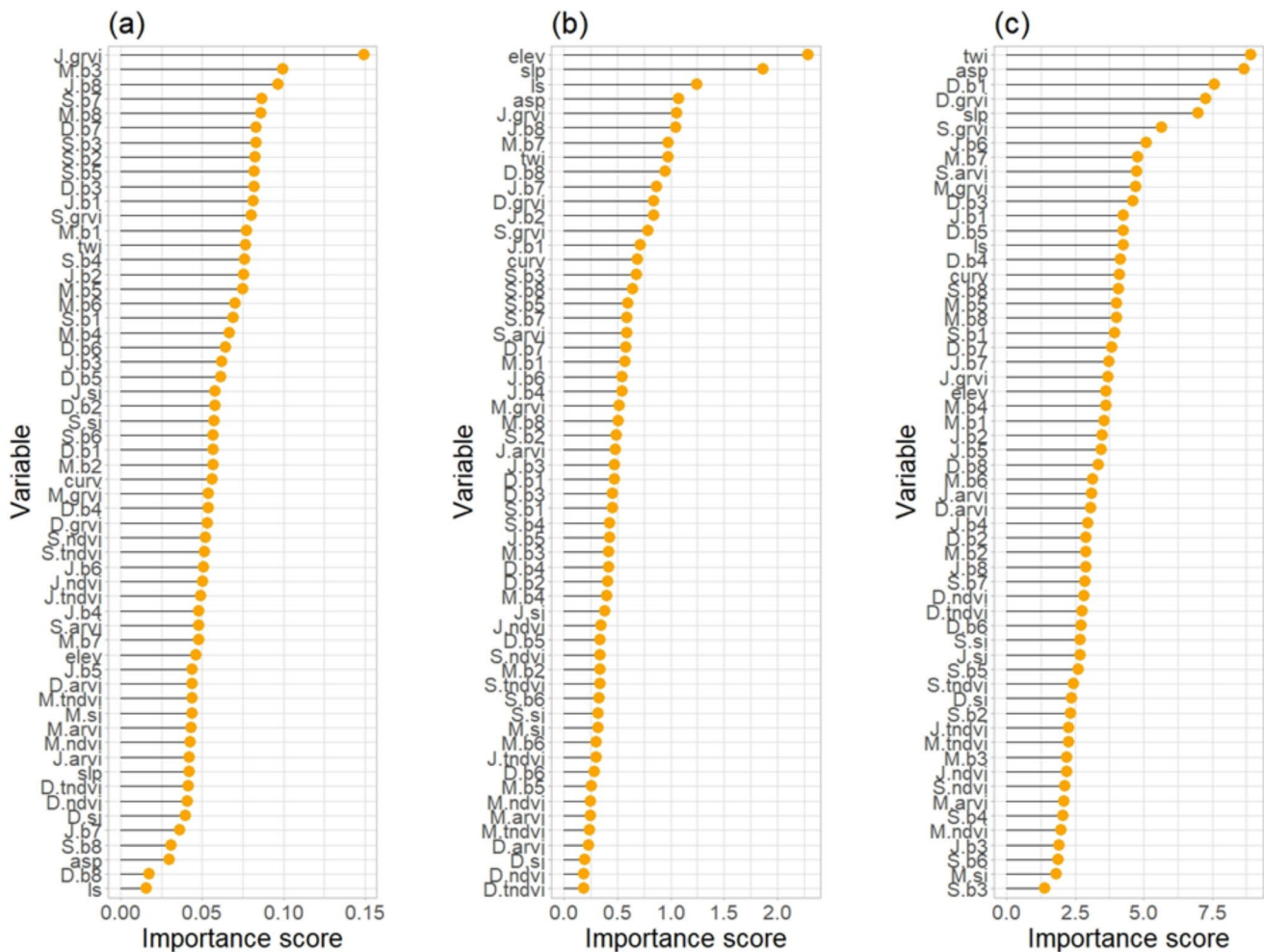


Fig. 6 Variable importance for predicting SOC (a) using the SVM model, and pH (b) and CaCO₃ (c) using the RF model. Note: M=March, J=June, S=September, D=December, grvi=green red vegetation index, ndvi=normalized difference vegetation index.

arvi=atmospheric resistance, vegetation index, tndiv=transformed normalized difference vegetation index, si=salinity index, elev=elevation, slp=slope, ls=slope length-steepness, asp=aspect, twi=topographic wetness index, curv=curvature, bxx=band numbers

(Phinzi and Szabó 2024), which explains the improved performance when using the entire feature set (covariates from all seasons).

Surprisingly, XGB's performance for soil pH was consistently inferior to that of RF and SVM. Despite being a tree-based model like RF, XGB was ineffective at predicting soil pH and failed to explain any variability in the soil pH data ($R^2 = 0$). This discrepancy can be attributed to the lower variability in the soil pH values, which ranged from 4.34 to 7.25, compared to SOC (0.19–7.81%) and CaCO_3 (2.84–10.25%), both of which had a much wider range. The narrow range and limited variability in pH values likely hindered XGB's ability to capture subtle differences and patterns, leading to poor performance. This observation is supported by XGB's good performance for SOC and CaCO_3 , which exhibited greater variability than pH. In a study by Taghizadeh-Mehrjardi et al. (2021), XGB demonstrated

better predictive performance for CaCO_3 which had a much wider range (10–45%) compared to the narrower ranges of SOC (0–3.07%) and pH (7.50–10).

The best predictive models for each soil property also generated interpretable soil maps. For example, the SOC map produced by the SVM model demonstrated natural variability across different land cover categories. Predicted SOC values were lower in degraded and bare soil areas and higher in vegetated areas, which is expected in regions experiencing severe erosion. According to Celestina et al. (2019), degraded areas typically have low soil organic matter. These areas often have steeper slopes that facilitate erosion, explaining the predicted lower SOC values. Similarly, the RF model produced an accurate CaCO_3 map with a generalizable pattern. CaCO_3 values were higher in eroded and bare soil surfaces, which is logical because the study area is semi-arid and lacks protective vegetation cover.

predisposing it to erosion. The removal of topsoil by erosion exposes calcium carbonate-rich subsoil, leading to higher detection of CaCO_3 values in these areas. In contrast, the pH map generated by the RF model did not exhibit a clear, generalizable pattern across different land cover types, as observed with other soil properties, despite the impressive performance in terms of accuracy metrics. This lack of pattern can be mainly attributed to the small variation in the observed pH values, which affected the predicted pH.

The importance of covariates varied depending on the target soil properties. GRVI derived from Winter season imagery emerged as the most crucial covariate for predicting SOC. This is unsurprising because, during Winter, the reduced biomass of many plants results in less vegetation cover. This exposes more soil surface, allowing soil properties like SOC to be more detectable through remote sensing indices like GRVI. This finding is consistent with previous studies, where satellite-based indices like NDVI (Taghizadeh-Mehrjardi et al. 2021) and Leaf Area Index (Lamichhane et al. 2021) were identified as key covariates for predicting SOC. The reduced vegetative cover during Winter likely enabled the sensor to capture soil characteristics more directly, thereby enhancing SOC prediction accuracy. However, covariates from other seasons, such as the Green I and NIR bands from Spring and the Red Edge band from Autumn, also proved to be critical, underscoring the importance of incorporating multi-season covariates in predictive models.

Topographic covariates, specifically elevation, and TWI, were identified as the most significant predictors for soil pH and CaCO_3 , respectively, underscoring the critical role of topography as a soil-forming factor. This aligns with findings from Xiao et al. (2023), who also identified elevation, which affects local microclimate and leaching processes, as a key predictor for soil pH. Elevated areas with steeper slopes typically exhibit better drainage and reduced water accumulation, which impacts the leaching of minerals and organic matter, thereby influencing soil pH (Schoonover and Crim 2015). In our study, we observed higher pH values in areas with gentler slopes, indicating the combined impact of elevation and slope. On the contrary, soil pH was lower in eroded areas compared to other land cover types. The observed lower pH values were associated with SOM depletion and nutrient loss, potentially due to site-specific processes such as leaching or erosion-induced SOM removal. While higher pH (>6) has been linked to unstable soil aggregates and increased erosion risk in other studies (Matsumoto et al. 2018), the relationship between pH, SOM, and erosion risk is likely soil-specific. Future studies investigating predominant soil types and management practices could provide more definitive conclusions. The eroded sites also exhibited minimal biological activity, such

as reduced presence of microorganisms and vegetation, which can be partially attributed to increased soil acidity, as reflected in the lower pH values. For CaCO_3 , TWI, which indicates the potential for water accumulation, was ranked as the foremost predictor. Areas with high TWI values tend to retain water for longer periods, facilitating distinct chemical weathering processes that affect CaCO_3 levels. Despite these findings, spectral covariates also proved relevant for predicting these soil properties. The GRVI from Winter imagery ranked fifth in importance for predicting pH, while the Coastal Blue band and GRVI from Summer imagery ranked third and fourth, respectively, for predicting CaCO_3 .

The findings of this study enhance understanding of the variability in soil property distribution, which is crucial for targeted management strategies. The study identified a spatial correlation between reduced SOC levels and eroded areas. Livestock management can play a pivotal role in promoting greater permanent soil cover through plant litter and vegetation, effectively mitigating soil erosion and enhancing SOC accumulation (Teague and Kreuter 2020). However, for these benefits to be realized, it is essential to implement rotational grazing practices, which are currently absent in the region, to stabilize soils and prevent further degradation caused by overgrazing. Although this study successfully predicted soil properties and highlighted the potential of multi-season covariates to enhance soil property prediction, several limitations should be addressed in future research. The influence of varying environmental factors, such as soil moisture, on spectral characteristics was acknowledged, but these factors were not directly measured. Additionally, other soil-forming factors such as climate and parent material were excluded due to data limitations, which likely contributed to the underestimation of soil properties even by the best models. Including factors such as soil moisture, temperature, precipitation, and parent material could refine predictive models. Furthermore, the study observed that different algorithms yielded varying performances across soil properties. Future research should explore hybrid modeling approaches that combine the strengths of different algorithms to improve predictive accuracy. Such attempts should also incorporate spatial uncertainty to enhance the reliability of the findings.

Conclusions

This study assessed the potential of multi-season PlanetScope imagery to enhance the accuracy of digital soil mapping and identified the most effective predictive models for each soil property. The key conclusions were as follows:

- The use of multi-season spectral covariates greatly improved the accuracy of digital soil mapping compared to single-season covariates. This improvement is attributed to the diverse spectral characteristics provided by different seasons, which more effectively captured variations in soil properties.
- SVM was the best-performing algorithm for predicting soil organic carbon (SOC) with an RMSE of 0.52%, MAE of 0.24%, and R^2 of 0.92. RF proved to be the most effective algorithm for predicting soil pH (RMSE=0.22%, MAE=0.17%, and $R^2 = 0.97$) and CaCO_3 (RMSE=0.55%, MAE=0.42%, and $R^2 = 0.96$). These models also produced interpretable maps that realistically represented the distribution of soil properties across different land cover categories.
- The GRVI from the Winter season image emerged as the most important covariate for predicting SOC due to reduced vegetation cover, which allowed for better detection of SOC. Topographic variables, such as elevation and TWI, were identified as key predictors for soil pH and CaCO_3 , respectively, underscoring the significant role of topography in soil formation and properties.
- To further improve predictive models, future studies should incorporate additional environmental factors such as soil moisture, temperature, precipitation, and parent material. Exploring hybrid modeling approaches could also enhance predictive accuracy.

Acknowledgements Sincere appreciation goes to Mr. Thabiso Shinga, Sluleko Gwala, and Mzikayise Phinzi for their assistance in soil sample collection. A special thanks to Dr. Tamás Mester and the Soil Lab team at the University of Debrecen, including Mrs. Krisztina Sósne Mező and Ms. Csilla Tóth, for their invaluable expertise and support in processing and measuring soil samples. We also extend our gratitude to Planet Labs Inc. for providing PlanetScope imagery through their Education and Research Program, and to the National Geo-spatial Information (NGI) of South Africa for the aerial imagery archive used to generate the DEM for this study. Finally, the first author is deeply thankful to the Tempus Public Foundation for funding his studies through the Stipendium Hungaricum Scholarship Program, and to the Department of Higher Education and Training of South Africa for their additional support.

Author contributions K. P conceptualized, designed the study, performed analysis, wrote the first draft, and revised the manuscript. L. B produced the DEM and edited the manuscript. S. S supervised the research, reviewed, and edited the manuscript. G. G. C also reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by University of Zululand. This research was funded by the Department of Higher Education and Training (DHET) of South Africa.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdi AM (2020) Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *ISci Remote Sens* 57:1–20. <https://doi.org/10.1080/15481603.2019.1650447>
- Ahmetoglu H, Das R (2022) A comprehensive review on detection of cyber-attacks: data sets, methods, challenges, and future research directions. *Internet Things* 20:100615. <https://doi.org/10.1016/j.iot.2022.100615>
- Bakker M, Lane SN (2017) Archival photogrammetric analysis of river–floodplain systems using structure from Motion (SfM) methods. *Earth Surf Process Landf* 42:1274–1286. <https://doi.org/10.1002/esp.4085>
- Blanco-Canqui H, Shapiro CA, Wortmann CS et al (2013) Soil organic carbon: the value to soil properties. *J Soil Water Conserv* 68. <http://doi.org/10.2489/jswc.68.5.129A>. 129A–134A
- Boehmke B, Greenwell BM (2019) Hands-on machine learning with R. CRC, New York
- Bradford MA, Wood SA, Addicott ET et al (2021) Quantifying microbial control of soil organic matter dynamics at macrosystem scales. *Biogeochemistry* 156:19–40. <https://doi.org/10.1007/s10533-021-00789-5>
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Brus DJ (2019) Sampling for digital soil mapping: a tutorial supported by R scripts. *Geoderma* 338:464–480. <https://doi.org/10.1016/j.geoderma.2018.07.036>
- Celestina C, Hunt JR, Sale PWG, Franks AE (2019) Attribution of crop yield responses to application of organic amendments: a critical review. *Soil Tillage Res* 186:135–145. <https://doi.org/10.1016/j.still.2018.10.002>
- Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. pp 785–794
- Chen T, He T, Benesty M et al (2015) Xgboost: extreme gradient boosting. R package version 0.4-2 1:1–4
- Du F, Zhu A, Band L, Liu J (2015) Soil property variation mapping through data mining of soil category maps. *Hydrol Process* 29:2491–2503. <https://doi.org/10.1002/hyp.10383>
- Eltner A, Sofia G (2020) Structure from motion photogrammetric technique. *Developments in Earth surface processes*. Elsevier, pp 1–24
- Fatholouloumi S, Vaezi AR, Alavipanah SK et al (2020) Improved digital soil mapping with multitemporal remotely sensed satellite data

- fusion: a case study in Iran. *Sci Total Environ* 721:137703. <https://doi.org/10.1016/j.scitotenv.2020.137703>
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 1189–1232
- Fukue M, Nakamura T, Kato Y (1999) Cementation of soils due to calcium carbonate. *Soils Found* 39:55–64. https://doi.org/10.3208/sandf.39.6_55
- Hateffard F, Steinbuch L, Heuvelink GBM (2024) Evaluating the extrapolation potential of random forest digital soil mapping. *Geoderma* 441:116740. <https://doi.org/10.1016/j.geoderma.2023.116740>
- Hengl T, Heuvelink GBM, Kempen B et al (2015) Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLoS ONE* 10:e0125814. <https://doi.org/10.1371/journal.pone.0125814>
- Heung B, Ho HC, Zhang J et al (2016) An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* 265:62–77. <https://doi.org/10.1016/j.geoderma.2015.11.014>
- Irizarry RA (2019) Introduction to data science: data analysis and prediction algorithms with R. CRC, New York
- Kaufman YJ, Tanre D (1992) Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. *IEEE Trans Geosci Remote Sens* 30:261–270. <https://doi.org/10.1109/36.134076>
- Keshavarzi A, Kaya F, Başayığit L et al (2023) Spatial prediction of soil micronutrients using machine learning algorithms integrated with multiple digital covariates. *Nutr Cycl Agroecosyst* 127:137–153. <https://doi.org/10.1007/s10705-023-10303-y>
- Kuhn M (2008) Building predictive models in R using the caret package. *J Stat Softw* 28:1–26. <https://doi.org/10.18637/jss.v028.i05>
- Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. *Sci* (1979) 304:1623–1627. <https://doi.org/10.1126/science.1097396>
- Lamichhane S, Adhikari K, Kumar L (2021) Use of multi-seasonal satellite images to predict SOC from cultivated lands in a montane ecosystem. *Remote Sens (Basel)* 13:4772. <https://doi.org/10.3390/rs13234772>
- Lesaignoux A, Fabre S, Briottet X (2013) Influence of soil moisture content on spectral reflectance of bare soils in the 0.4–14 μm domain. *Int J Remote Sens* 34:2268–2285. <https://doi.org/10.1080/01431161.2012.743693>
- Liu Q, He L, Guo L et al (2022) Digital mapping of soil organic carbon density using newly developed bare soil spectral indices and deep neural network. *Catena (Amst)* 219:106603. <https://doi.org/10.1016/j.catena.2022.106603>
- Lobell DB, Asner GP (2002) Moisture effects on soil reflectance. *Soil Sci Soc Am J* 66:722–727. <https://doi.org/10.2136/sssaj2002.72.20>
- Loveland P, Webb J (2003) Is there a critical level of organic matter in the agricultural soils of temperate regions: a review. *Soil Tillage Res* 70:1–18. [https://doi.org/10.1016/S0167-1987\(02\)00139-3](https://doi.org/10.1016/S0167-1987(02)00139-3)
- Mandal UK (2016) Spectral color indices based geospatial modeling of soil organic matter in Chitwan District, Nepal. *The International archives of the Photogrammetry. Remote Sens Spat Inform Sci* 41:43–48. <https://doi.org/10.5194/isprs-archives-XLI-B2-4.3-2016>
- Matsumoto S, Ogata S, Shimada H et al (2018) Effects of pH-induced changes in soil physical characteristics on the development of soil water erosion. *Geosci (Basel)* 8:134. <https://doi.org/10.3390/geosciences8040134>
- Minasny B, Hong SY, Hartemink AE et al (2016) Soil pH increase under paddy in South Korea between 2000 and 2012. *Agric Ecosyst Environ* 221:205–213. <https://doi.org/10.1016/j.agee.2016.01.042>
- Motia S, Reddy SRN (2021) Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation. In: *Journal of Physics: Conference Series*. IOP Publishing, p 012037
- Naimi S, Ayoubi S, Demattè JAM et al (2022) Spatial prediction of soil surface properties in an arid region using synthetic soil image and machine learning. *Geocarto Int* 37:8230–8253. <https://doi.org/10.1080/10106049.2021.1996639>
- Neina D (2019) The role of soil pH in plant nutrition and soil remediation. *Appl Environ Soil Sci* 2019:5794869. <https://doi.org/10.1155/2019/5794869>
- Nguyen T-T, Huang JZ, Nguyen TT (2015) Unbiased feature selection in learning random forests for high-dimensional data. *The Scientific World Journal* 2015
- Nokhandan FN, Ghahraman K, Horváth E (2023) Erosion susceptibility mapping of a loess-covered region using Analytic Hierarchy Process—A case study: Kalat-e-Naderi, northeast Iran. *Hung Geographical Bull* 72:339–364. <https://doi.org/10.15201/hungeobull.72.4.2>
- Oelofse M, Markussen B, Knudsen L et al (2015) Do soil organic carbon levels affect potential yields and nitrogen use efficiency? An analysis of winter wheat and spring barley field trials. *Eur J Agron* 66:62–73. <https://doi.org/10.1016/j.eja.2015.02.009>
- Oshunsanya SO (2018) Introductory chapter: Relevance of soil pH to agriculture. In: Oshunsanya SO (ed) *Soil PH for nutrient availability and crop performance*. IntechOpen, pp 1–82
- Padarian J, Minasny B, McBratney AB (2020) Machine learning and soil sciences: a review aided by machine learning tools. *SOIL* 6:35–52
- Phinzi K, Szabó S (2024) Predictive machine learning for gully susceptibility modeling with geo-environmental covariates: main drivers, model performance, and computational efficiency. *Nat Hazards* 120:7211–7244. <https://doi.org/10.1007/s11069-024-06481-9>
- Phinzi K, Abriha D, Bertalan L et al (2020) Machine learning for gully feature extraction based on a pan-sharpened multispectral image: Multiclass vs. binary approach. *ISPRS Int J Geoinf* 9. <https://doi.org/10.3390/ijgi9040252>
- Post WM, Emanuel WR, Zinke PJ, Stangenberger AG (1982) Soil carbon pools and world life zones. *Nature* 298:156–159. <https://doi.org/10.1038/298156a0>
- Rodriguez-Galiano VF, Ghimire B, Rogan J et al (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J Photogrammetry Remote Sens* 67:93–104. <https://doi.org/10.1016/j.isprsjprs.2011.11.002>
- Rouse JW, Haas RH, Schell JA, Deering DW (1974) Monitoring vegetation systems in the Great Plains with ERTS. *NASA Spec Publ* 351:309
- Rowley MC, Grand S, Verrecchia ÉP (2018) Calcium-mediated stabilisation of soil organic carbon. *Biogeochemistry* 137:27–49. <https://doi.org/10.1007/s10533-017-0410-1>
- Rowley MC, Grand S, Adatte T, Verrecchia EP (2020) A cascading influence of calcium carbonate on the biogeochemistry and pedogenic trajectories of subalpine soils. *Switz Geoderma* 361:114065. <https://doi.org/10.1016/j.geoderma.2019.114065>
- Schoonover JE, Crim JF (2015) An introduction to soil concepts and the role of soils in watershed management. *J Contemp Water Res Educ* 154:21–47. <https://doi.org/10.1111/j.1936-704X.2015.03186.x>
- Silvero NEQ, Demattè JAM, de Souza Vieira J et al (2021) Soil property maps with satellite images at multiple scales and its impact on management and classification. *Geoderma* 397:115089. <https://doi.org/10.1016/j.geoderma.2021.115089>
- Song Y-Q, Yang L-A, Li B et al (2017) Spatial prediction of soil organic matter using a hybrid geostatistical model of an extreme learning machine and ordinary kriging. *Sustainability* 9:754. <https://doi.org/10.3390/su9050754>

- Taghizadeh-Mehrjardi R, Hamzehpour N, Hassanzadeh M et al (2021) Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. *Geoderma* 399:115108. <https://doi.org/10.1016/j.geoderma.2021.115108>
- Teague R, Kreuter U (2020) Managing grazing to restore soil health, ecosystem function, and ecosystem services. *Front Sustain Food Syst* 4:534187. <https://doi.org/10.3389/fsufs.2020.534187>
- Tucker CJ (1979) Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens Environ* 8:127–150. [https://doi.org/10.1016/0034-4257\(79\)90013-0](https://doi.org/10.1016/0034-4257(79)90013-0)
- Vapnik V (1999) *The nature of statistical learning theory*. Springer science & business media
- Walkley A, Black IA (1934) An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Sci* 37:29–38
- Wang L, Zhou Y (2022) Combining Multitemporal Sentinel-2A spectral imaging and Random Forest to improve the Accuracy of Soil Organic Matter estimates in the Plough Layer for Cultivated Land. *Agriculture* 13:8. <https://doi.org/10.3390/agriculture13010008>
- Were K, Bui DT, Dick ØB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an afro-montane landscape. *Ecol Indic* 52:394–403. <https://doi.org/10.1016/j.ecolind.2014.12.028>
- Xiao J, Chevallier F, Gomez C et al (2019) Remote sensing of the terrestrial carbon cycle: a review of advances over 50 years. *Remote Sens Environ* 233:111383. <https://doi.org/10.1016/j.rse.2019.111383>
- 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* 39:900–916. <https://doi.org/10.1111/sum.12874>
- Yuvaraj D, Jayachandran K, Ashokkumar L (2022) Performance of spectral indices for soil properties: a case study from Redland farm, south Florida. *Model Earth Syst Environ* 8:4829–4841. <https://doi.org/10.1007/s40808-022-01371-0>
- Zhai M (2019) Inversion of organic matter content in wetland soil based on Landsat 8 remote sensing image. *J Vis Commun Image Represent* 64:102645. <https://doi.org/10.1016/j.jvcir.2019.102645>
- Zhang S, Tian J, Lu X, Tian Q (2023) Temporal and spatial dynamics distribution of organic carbon content of surface soil in coastal wetlands of Yancheng, China from 2000 to 2022 based on landsat images. *Catena (Amst)* 223:106961. <https://doi.org/10.1016/j.catena.2023.106961>
- Žížala D, Minařík R, Zádorová T (2019) Soil organic carbon mapping using multispectral remote sensing data: Prediction ability of data with different spatial and spectral resolutions. *Remote Sens (Basel)* 11:2947. <https://doi.org/10.3390/rs11242947>

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