Spatial prediction of soil organic carbon stocks across contrasting Andean basins, Peru

Carlos Carbajal, Merely Tumbalobos-Dextre, Tatiana Condori-Ataupillco, Nestor Cuellar-Condori, Carla Gavilan

PII: S2352-0094(25)00111-7

DOI: https://doi.org/10.1016/j.geodrs.2025.e01026

Reference: GEODRS 1026

To appear in: *Geoderma Regional*

Received date: 29 April 2025

Revised date: 30 October 2025

Accepted date: 5 November 2025

Please cite this article as: C. Carbajal, M. Tumbalobos-Dextre, T. Condori-Ataupillco, et al., Spatial prediction of soil organic carbon stocks across contrasting Andean basins, Peru, *Geoderma Regional* (2024), https://doi.org/10.1016/j.geodrs.2025.e01026

Spatial Prediction of Soil Organic Carbon Stocks Across Contrasting Andean Basins, Peru

Carlos Carbajal[a*], Merely Tumbalobos-Dextre[a], Tatiana Condori-Ataupillco[b], Nestor Cuellar-Condori[c], Carla Gavilan[d]

[a]Dirección de Servicios Estratégicos Agrarios, Instituto Nacional de Innovación Agraria (INIA), Av. La Molina 1981, Lima 15024, Perú; merely.tdextre@gmail.com (M.T.-D.)

[b]Dirección de Servicios Estratégicos Agrarios, Instituto Nacional de Innovación Agraria (INIA), Ayacucho 05002, Perú; tatiana.condori1423@gmail.com (T.C.-A)

[c]Dirección de Servicios Estratégicos Agrarios, Instituto Nacional de Innovación Agraria (INIA), Puno 21001, Perú; ncuellar07@gmail.com (N.C.-C)

[d]Department of Environmental Sciences, Rutgers, The State University of New Jersey, New Brunswick, NJ 08904, USA; cg1141@envsci.rutgers.edu (C.G.)

*Corresponding Author: cmcarbajal@gmail.com

Abstract

Soil organic carbon stocks (SOCS) are critical components of the global carbon cycling and play a central role in climate change mitigation. However, their dynamics in high-altitude Andean ecosystems remain poorly understood despite their importance for carbon sequestration. The significant spatial heterogeneity of SOCS in mountainous terrain makes accurate quantification and mapping challenging. This study evaluated the performance of geospatial regression and machine learning (ML) approaches for predicting SOCS in two Peruvian Andean basins: Torobamba and Coata. We compared Geographically Weighted Regression (GWR), GWR with collinearity analysis (GWRC), their kriging-adjusted variants, and ML models (Random Forest, Gradient Boosting). Models were built using key SOCS covariates for each basin and validated through 5-fold cross-validation with Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). In Torobamba, GWRC markedly improved performance, reducing the RMSE by 79–90% and achieving R^2 up to 0.99. In contrast, Coata, showed only modest improvements (RMSE reductions of 7.8–9.8%, R^2 = 0.30–0.39). ML models performed poorly (negative R^2), likely due to feature selection, parameter tuning, or limited sample size. Overall, locally weighted regression approaches (GWRK/GWRCK) outperformed conventional ML methods for SOCS prediction in complex mountain environments, particularly with small to medium sample sizes. These results highlight the importance of accounting for spatial non-

stationarity in SOCS and provide methodological guidance for SOCS mapping in Andean ecosystems.

Keywords: Digital soil mapping; Soil organic carbon stock; Geographically weighted regression; Machine learning regression algorithms; Andes.

1. Introduction

Digital soil mapping (DSM) has become an essential tool for understanding and predicting soil properties in diverse landscapes (Chen et al., 2022). In the Andean region, a complex terrain with significant ecological diversity, soil organic carbon (SOC) is critical for soil health, productivity, and ecosystem sustainability (Munoz et al., 2015). Traditional soil sampling methods, although effective at local scales, are often labor-intensive, time-consuming (Minasny and McBratney, 2016), and limited in their ability to provide comprehensive information of SOC distribution across vast, heterogeneous areas (Minasny et al., 2013). SOC is vital for maintaining soil structure, enhancing water retention, and supporting biodiversity, functions particularly important in steep, erosion-prone Andean landscapes. However, SOC levels are spatially variable and challenging to measure across mountainous topography (Román-Sánchez et al., 2018). DSM enables the generation of continuous SOC maps that capture the variability of soil properties driven by climate, vegetation, land use, and topography (Moura-Bueno et al., 2021). This approach is well suited for the Andes, where mountainous landscapes and diverse microclimates drive to distinct soil formation processes (Anderson et al., 2011).

To monitor and quantify C sequestration, SOC concentrations are expressed in terms of volume and are referred to as soil organic carbon stocks (SOCS). Several studies investigated the spatial distribution of SOCS under different land use and land cover (LULC) types in the montane ecosystem (Dorji et al., 2014) and forests (Ottoy et al., 2017; Vallejos-Torres et al., 2024). Other have assessed SOCS at various depths using environmental covariates, and compared them across soil types and land uses, reporting that SOCS increase under grasslands (Hengl et al., 2023). Understanding SOCS distribution in the high Andean regions of Peru is constrained by restricted access and logistical challenges due to the complex geography, further limiting the sample size; moreover, if observations at different depths are required.

Predicting the current or future spatial distribution of SOCS is possible using statistical and ML models (Yigini and Panagos, 2016). In the present study, we focused on the performance of Geographically Weighted Regression (GWR), which employs local models estimated from subsets of observations centered at a focal point (Szakács et al., 2011). GWR has consistently outperformed other models, such as ordinary least squares (OLS), when estimating the spatial distribution of SOCS in forests, effectively reducing data saturation and uncertainty (Wu et al., 2023). Compared to Ordinary Kriging (OK), inverse distance weighted (IDW), and multiple linear regression (MLR), GWR produced plausible results, revealing spatial patterns affected by environmental variables (Zhang et al., 2011). GWR is usually applied to address multicollinearity effects among auxiliary variables in regression prediction and to capture effectively spatially non-stationary structures and relationships between the target and response variables (Chen et al., 2021; Guo et al., 2018). Future studies should examine how GWR models address multicollinearity while systematically evaluating their predictive performance under varying bandwidth selection, kernel functions, and spatial autocorrelation patterns. For example, GWR Kriging (GWRK) proved more accurate than standard GWR when used to correct for residual autocorrelation (Kumar et al., 2012). Similarly, the potential benefits of integrating machine learning require further investigation (Zhang et al., 2025). For instance, Zeng et al. (2024) developed hybrid models integrating GWR and ML approaches, such as geographically weighted neural networks, to predict Cu distributions in urban topsoil. In other instances, the method was combined with RF to develop spatially weighted RF models, which were used to investigate the drivers of forest dynamics (Santos et al., 2019). For similar purposes, the strengths of Bayesian inference were leveraged through the Bayesian Geographically Weighted Regression model (Faisal et al., 2025).

Advancements in ML, particularly random forest (RF) and gradient boosting (GB) techniques, offer alternatives for the efficient and precise prediction of SOCS. RF is an ensemble learning method that constructs multiple decision trees during training and combines them to improve predictive accuracy and control over-fitting (Breiman, 2001). It is particularly effective for handling large datasets with complex interactions among variables (Genuer et al., 2017; Schonlau and Zou, 2020). RF leverages environmental, topographic, and climatic information from remote sensors and spectral indices to generate detailed SOC maps (Pouladi et al., 2023; Triantakonstantis and Karakostas,

2025), with good results in grasslands (Szakács et al., 2011). A notable application of RF is the creation of a gridded SOCS map across Australia, which enhanced the representation of uncertainties distribution (Wang et al., 2024). GB, another ensemble learning method, builds multiple decision trees sequentially, with each tree correcting the errors of the previous one (Friedman, 2001; Zhang and Jung, 2021). This approach results in a highly accurate predictive model particularly effective in capturing complex patterns and relationships within data (Habib et al., 2024; Tahmouresi et al., 2024), making GB suitable for SOCS prediction. In previous research, GB models were used environmental covariates, topographic information, and remote sensing data to estimate SOC with high precision (Chen et al., 2024; Kumar et al., 2023; Taghizadeh-Mehrjardi et al., 2020). ML model accuracy depends on input data quality and availability. While incomplete or erroneous data can compromise performance, the iterative nature of GB allows continuous model refinement. This characteristic makes GB a reliable approach for predicting SOC across diverse landscapes (Taghizadeh-Mehrjardi et al., 2020). Nevertheless, ensemble methods such as RF and GB face two main limitations: high computational demands and poor interpretability due to their complex models. These factors can hinder its adoption in practical applications (Halder et al., 2024). The effectiveness of ML depends on data quality, predictor selection and parameter tuning. Combining deep sampling with hybrid models can enhance interpretability and predictive power, while prioritizing the selection of appropriate predictors for the specific study area is also recommended.

This study aims to a) develop and evaluate advanced predictive frameworks, specifically focused on GWR models and ML algorithms, to predict SOCS across two contrasting basins in the Andean region of Southern Peru, b) investigate the influence of environmental and spatial patterns on the distribution of SOCS in mountain ecosystems, and c) provide guidance on land management to improve SOC storage.

2. Materials and methods

2.1. Study area

Two study areas were selected for the present study: the Torobamba and Coata basins (hereafter referred to as Torobamba and Coata), located in the southern Peruvian region (Fig. 1). Torobamba (-74°.16’ N, -13°.20’ E), with an area of approximately 1040.47 km², is located in the Peruvian Andean southern region (between 1500 and 4500 masl) and comprises extensive agricultural areas, shrublands, native grasslands, and seasonal

peatlands and bogs (Zanaga et al., 2022). The mean temperature ranges between 1 °C to 20 °C during the dry season (May to September) and between 4 °C to 20 °C during the wet season (October to April), with annual precipitation of 536 mm. Soils in this basin are classified mainly as Cambisols, and other soils reported in the area are Leptosols, Andosols, Regosols, and Vertisols (IUSS Working Group, 2007). Coata (-70°.29' N, -15°.62' E), with an area of approximately 460.16 km², is located in the southern Peruvian region between 3825 and 4250 masl. This basin comprises extensive agricultural areas and native grasslands presenting soils classified as Cambisols, Chernozems, and Phaeozems (IUSS Working Group, 2007). The historical average annual precipitation is 751 mm, average minimum temperatures of -5 °C, and average maximum temperatures of 17 °C. Both basins exhibit distinct biogeographic characteristics based on an ecological classification system proposed by Pulgar Vidal, (2014). Torobamba demonstrates high environmental heterogeneity, spanning four ecological zones: Yunga (warm-dry), Quechua (temperate, dry and pleasant), Suni (cold-dry), and Puna (very cold). However, Coata is only found in the Puna zone, which is known for being cold, semi-humid to humid environments and natural grassland. The Puna ecosystem plays a crucial role in soil carbon sequestration throughout the Southern Cone, particularly due to the presence of peatlands and organic soil systems (García Lino et al., 2024)

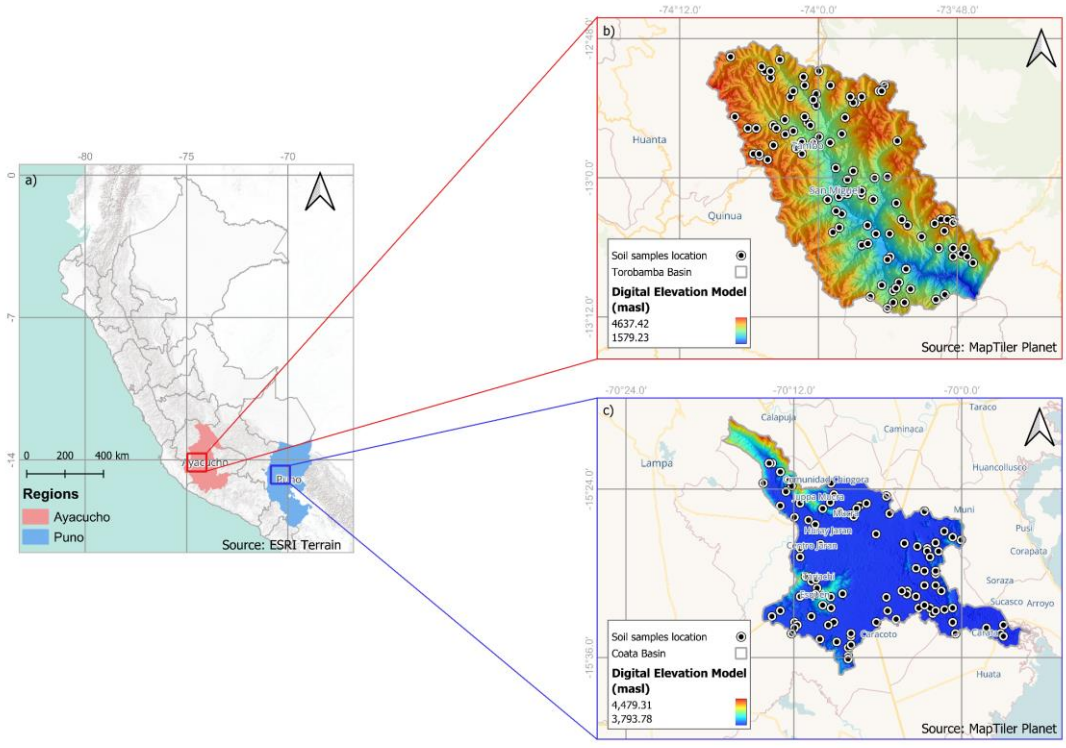

Fig. 1. Location of sampling points in a) Ayacucho and Puno Regions of Peru with the respective basins, b) Torobamba, and c) Coata.

2.2. Soil sampling and analysis

The conditional Latin Hypercube Sampling (cLHS) design (Minasny and McBratney, 2006) was used to find the best distribution of sampling points across the study area. The cLHS was implemented using the 'clhs' R package (Roudier, 2012). A set of environmental variables was used as conditioning factors to generate sampling points using the cLHS method (Table S1). A total of 100 and 96 sampling sites were identified for Torobamba and Coata, respectively. The geographic coordinates were extracted for each of the sampling sites.

At each identified sampling site, five soil subsamples were collected within a radius of 5 meters at a depth of 30 cm. The subsamples were homogenized to obtain approximately 1 kg of composite sample per site. Undisturbed samples were collected at 30 cm depth using a 5 cm by 5 cm core ring sampler in order to assess soil bulk density (BD) at each identified sampling site.

The SOC (%) was determined by dry combustion (ISO, 1996) using an elementary analyzer LECO CN828 (Leco Corp., St. Joseph. MI, USA), and BD was obtained using the core method (ISO, 2017). Equation 1 (Peng, et al., 2024), SOCS at 30 cm were determined by calculating soil mass (BD × soil depth) per unit area and multiplying by SOC concentration (%).

$$\text{SOCS (Mg ha}^{-1}) = \text{SOC (\%)} * \text{BD (Mg m}^{-3}) * \text{Soil Depth (cm)} \quad (1)$$

2.3. Covariates data collection

Multiple environmental factors influence SOC distribution at a landscape level. To model the response of SOCS, it is essential to identify relevant predictor covariates. The foundation of this approach was laid out by Jenny, (1994) and the SCORPAN paradigm of soil forming factors (McBratney et al., 2003). A total of 21 covariates were considered in this study (Table S2). A set of topographic variables was derived from the ALOS PALSAR digital elevation model (DEM) at 30m resolution using SAGA tools (Conrad et al., 2015) and the 'elevatr' R package (Hollister et al., 2023). Climate variables, including annual mean temperature and precipitation, were extracted from the WorldClim v2.1 database at 1km resolution (Fick and Hijmans, 2017) using the

'geodata' R package (Hijmans et al., 2021). Land Surface Temperature (LST) was incorporated as a critical climatic indicator governing soil-atmosphere interactions and moisture dynamics (Ghaderpour et al., 2024). Following the method described in (Parastatidis et al., 2017), Google Earth Engine (GEE) (Gorelick et al., 2017) was used to retrieve LST by using Landsat 8 imagery, including the thermal band (ST_B10) and the NDVI (Normalized Difference Vegetation Index).

A suite of vegetation indices (Table 1) was produced from USGS Landsat 8 Level 2, Collection 2 imagery (30m resolution) acquired through GEE. Several indices, such as NDVI, TSAVI (Transformed Soil Adjusted Vegetation Index), SATVI (Soil Adjusted Total Vegetation Index), GNDVI (Green Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), SAVI (Soil Adjusted Vegetation Index), NDWI (Normalized Difference Water Index), BSI (Bare Soil Index), and BI2 (Second Brightness Indices), were obtained and included in the analysis.

Table 1

List of the indices derived from remote sensing used for spatial modeling.

Index	Acronym	Equation	Details	Reference
Normalized Difference Vegetation Index	NDVI	$\frac{(NIR - R)}{(NIR + R)}$		(Rouse et al., 1974)
Transformed Soil Adjusted Vegetation Index	TSAVI	$\frac{s(NIR - s * R - a)}{(a * NIR + R - a * s + X * (1 + s^2))}$	s = Soil line slope a = Soil line intercept X = Adjustment factor to minimize soil noise	(Baret and Guyot, 1991)
Soil Adjusted Total Vegetation Index	SATVI	$\frac{SWIR1 - R}{SWIR1 + R + L} * (1 + L) - \frac{SWIR2}{2}$	L = Soil brightness factor (values between 0 – 1)	(Marsett et al., 2006)
Green Normalized Difference Vegetation Index	GNDVI	$\frac{(NIR - G)}{(NIR + G)}$		(Gitelson et al., 1996)
Enhanced Vegetation Index	EVI	$2.5 * \frac{(NIR - R)}{(NIR + C1 * R - C2 * B + L)}$	L = Canopy Background Adjustment (values between 0 – 1) $C1$ = Aerosol Resistance Coefficient in band Red (values between 0 – 6) $C2$ = Aerosol Resistance Coefficient in band Blue (values between 0 – 7.5)	(Huete et al., 2002)

Soil Adjusted Vegetation Index	SAVI	$\frac{(NIR - R)}{(NIR + R + L)} * (1 + L)$	L = Soil brightness factor (values between 0 – 1)	(Huete, 1988)
Normalized Difference Water Index	NDWI	$\frac{(NIR - SWIR1)}{(NIR + SWIR1)}$		(Gao, 1996)
Bare Soil Index	BSI	$\frac{(SWIR1 + R) - (NIR + B)}{(SWIR1 + R) + (NIR + B)}$		(Rikimaru et al., 2002)
Second Brightness Index	BI2	$\frac{\sqrt{R^2 + G^2 + NIR^2)}}{3}$		(Escadafal, 1989)

2.4. Statistical analysis and data preprocessing

A structured feature selection process was implemented in R software 4.4.3 (R Core Team, 2023) to enhance model efficacy and mitigate overfitting. The process involved two sequential steps: first, address redundancy by removing variables exhibiting correlations exceeding 90%, and second, iteratively eliminate features with Variance Inflation Factors (VIFs) above a threshold of 10. Prior research (Naimi et al., 2021; Zhang et al., 2022) suggests discarding variables with high VIF (>10) to enhance model robustness and prediction accuracy. This threshold is considered a thumb rule, and should be treated with caution when used to assess signals of persistent multicollinearity (O'brien, 2007). The VIF was computed according to Eq. 2. The procedure retained predictors that met the VIF criteria, as illustrated in the first block of Fig. 2. Correlation matrix heatmaps are provided in the supplementary materials (Figures S1 and S2).

$$VIF_i = \frac{1}{1-R_i^2} \qquad (2)$$

where R_i^2 is the R-squared value obtained from the regression of X_i on the other predictors.

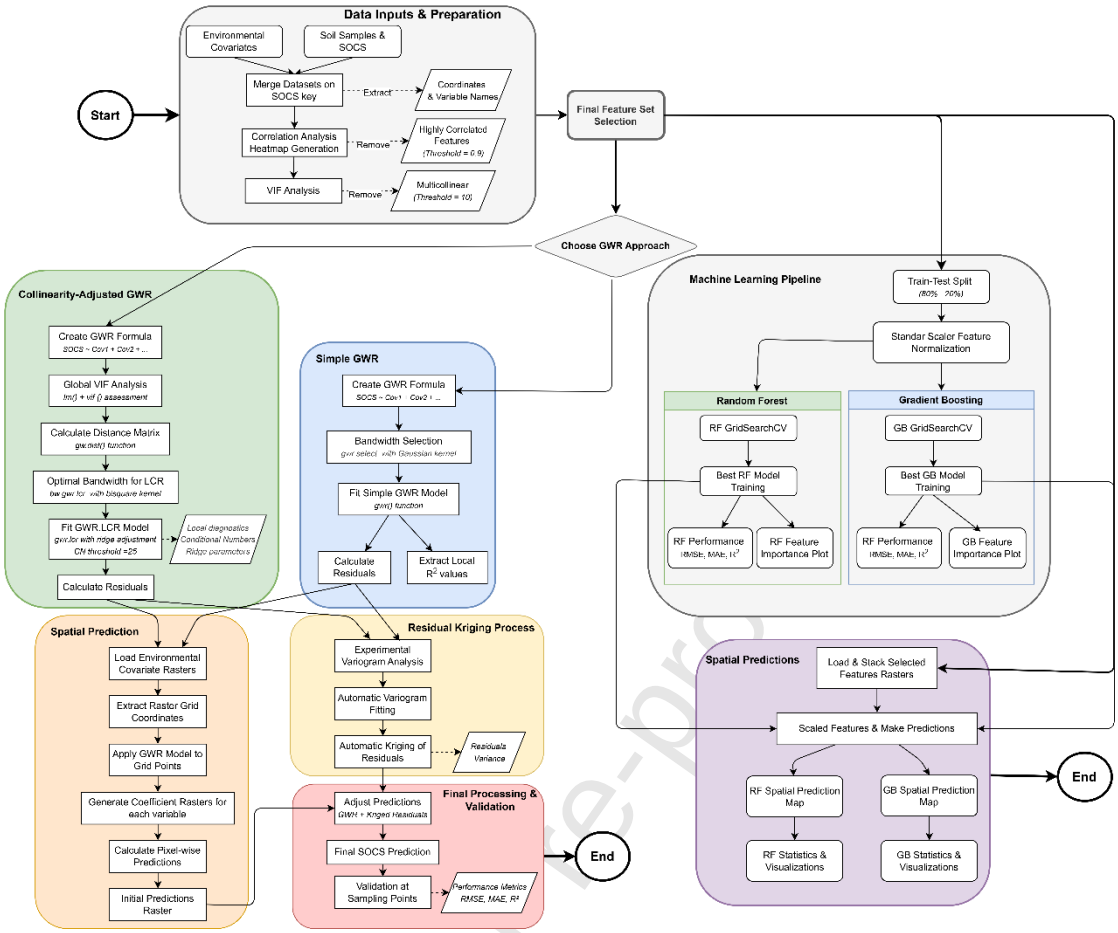

Fig. 2. Schematic representation of the modelling framework utilizing Geographically Weighted Regression and Machine Learning approaches.

2.5. Modeling SOCS

The present study uses Geographically Weighted Regression (GWR), Random Forest (RF), and Gradient Boosting (GB) models to predict SOCS. These models were trained using a dataset of environmental predictors selected during preprocessing, including surface soil properties, topographic and climatic variables, as well as vegetation spectral indices. The approaches assumed spatially variable relationships between SOCS and environmental covariates. The GWR model explicitly addressed this relationship in its localized regression framework (Zeng et al., 2016). To complement the GWR characteristics, ML models such as RF and GB were also implemented to predict SOCS by synthesizing and identifying diverse covariate groups, balancing complexity and generalizability (Hounkpatin et al., 2021; Ottoy et al., 2017).

2.5.1. Geographically Weighted Regression (GWR)

GWR is a local spatial analysis method that explores relationships between predictor variables, often addressing spatial heterogeneity or non-stationarity (Brunsdon et al., 1996). This technique generates localized regression results at any point in a region by using statistical curve-fitting and smoothing techniques (Thrift and Kitchin, 2009). Unlike global regression outputs, it provides mappable statistics that describe local relationships (Fotheringham et al., 1998). The following Eq. 3 presents an adaptation of the GWR model (Emamgholizadeh, 2017) that is used in this study to predict SOCS.

$$y(u) = \beta_0(\mathrm{u}) + \sum_{k=1}^{p} \beta_k(u)\, X_k\,(u) + \in (u) \qquad (3)$$

where $y(u)$ represents the dependent variable (SOCS) at location u, $\beta_0(\mathrm{u})$ is the geographically varying intercept, $\beta_k(u)$ are the unknown regression coefficients that are spatially variant, $X_k\,(u)$ are the kth independent variables with the value X_k at location u, p is the total number of samples of soils, and $\in (u)$ represents the random error at location u.

GWR analysis began by extracting coordinates to determine spatial relationships. The model incorporated all variables except SOCS (dependent variable). The optimal bandwidth selection was obtained by employing cross-validation with a Gaussian weighting kernel using the 'gwr.sel' function from the 'spgwr' R package (Bivand and Yu, 2006). The GWR model, fitted with optimal bandwidth, included hat matrix calculations for diagnostics and generated location-specific R^2 values to identify spatial variations in model performance. Prediction layers were built using the extracted coordinates from the selected covariates, including the coefficient and intercept values. Model refinement employed kriging interpolation of residuals using the 'autofitVariogram' and 'autoKrige' functions from the automap R package (Hiemstra et al., 2009), resulting in adjusted predictions.

2.5.2. Geographically Weighted Regression with Collinearity analysis (GWRC)

To enhance model robustness and provide a reference point for comparing local collinearity patterns, global multicollinearity was assessed using Variance Inflation Factors (VIF), followed by localized collinearity analysis incorporating coordinate extraction and distance matrix computation. The distance matrix in GWR serves as a crucial element, as it records the pairwise distances between all data points based on their coordinates. The optimal spatial bandwidth for the GWR framework was

determined via *bisquare* kernel and adaptive bandwidth selection to determine the number of nearest neighbors using the 'bw.gwr.lcr' function within the GWmodel package (Gollini et al., 2015). The GWRC model incorporated a locally compensated ridge term (*gwr.lcr*) to mitigate spatially varying multicollinearity. A Local Conditional Number (CN) threshold of 25 was used to identify locations requiring local ridge parameter adjustment. Tuning the Lambda parameter (lambda.adjust) further stabilized the regression coefficients at these sites. This approach reduces their variance and stabilizes the model (Wheeler, 2007). Diagnostic plots were generated to visualize CN, a measure of the local multicollinearity among the independent variables at each location, and ridge parameters. Also, predictions were extrapolated across the study areas, ensuring spatially explicit outputs.

2.5.3. Random Forest (RF)

Random Forests (RF), which leverage the principles of classification and regression trees, bagging, and added randomness, are robust prediction tools that enhance accuracy through the law of large numbers while mitigating overfitting (Breiman, 2001; Tyralis et al., 2019). With selected variables during the preprocessing, the data were partitioned into training (80%) and testing (20%) subsets, standardized to ensure consistent scaling. A grid search was implemented using the function 'GridSearchCV', systematically testing hyperparameter combinations on scaled training data to optimize the performance of the 'RandomForestRegressor' model with 5-fold cross-validation. The process culminated in identifying the best-performing model, with the optimal hyperparameters that maximized predictive accuracy and generalization capability. This approach ensures methodological rigor in balancing model complexity and validation robustness.

2.5.4. Gradient Boosting (GB)

Gradient Boosting (GB) is an ensemble ML method that builds predictive models by iteratively combining weak base learners (e.g., decision trees) to minimize residuals through a forward stage-wise optimization process (Belyadi and Haghighat, 2021). Designed for regression tasks, GB operates by sequentially fitting regression trees to the negative gradient of a differentiable loss function, incrementally improving predictions for continuous target variables (Pedregosa et al., 2011). This additive modeling approach balances bias-variance trade-offs, enabling robust performance in complex regression scenarios while maintaining computational efficiency.

After splitting and scaling out the data, the 'GradientBoostingRegressor' method applied to the train data, and the hyperparameter optimization of the model was performed using the function 'GridSearchCV', to evaluate various combinations of key parameters, such as the number of estimators (100, 200, 500), learning rate (0.01, 0.1, 0.2), maximum tree depth (3,4, 5), minimum samples required to split a node (2, 5, 10), and minimum samples required in a leaf node (1, 2, 4). This optimization procedure used a 5-fold cross-validation and employed a negative mean squared error as the performance metric, leveraging parallel processing to improve computational efficiency.

2.6. SOCS and Land Cover

To analyze the spatial distribution of SOCS across different land cover (LC) types in the study area, a supervised classification was performed using Sentinel-2 imagery within the GEE platform (Gorelick et al., 2017). Training points were incorporated for representative sampling. Prediction results from each model were intersected by LC type; the mean statistics of training points were calculated to show the SOCS variation. LC types in the legend are based on the ESA WorldCover 10m 2021 v200 thematic map (Zanaga et al., 2022). The LC maps for each basin are available in the supplementary material (Figure S3).

2.7. Model validation

The accuracy assessment of the preliminary models involved calculating residuals from predicted and observed values. Kriging via residual interpolation was used to adjust only the GWR predictions. We then used metrics such as the root square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) for a thorough model performance assessment, as described in Eqs. (4)–(6), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \left(y_i - \underline{y}\right)^2} \quad (6)$$

where y_i are the observed values, $\widehat{y_i}$ are the predicted values, $\underline{y}$ the mean of the observed values, and *n* is the number of samples.

3. Results

3.1. Exploratory Data Analysis

A descriptive statistical summary of topsoil (0-30 cm) SOC and BD for both study areas is provided in Table 2. The average SOC values for Torobamba and Coata were 2.52 % (±1.04) and 1.34% (±0.39), respectively. Torobamba exhibited significantly higher values of SOC and heterogeneity (0.83 – 5.16%). Average BD values were 1.24 Mg m-3 (±0.19) for Torobamba and 1.32 Mg m-3(±0.17) for Coata. Torobamba has the highest SOC values and the lowest BD values, unlike Coata, which has the lowest SOC values but the highest BD values.

Table 2

Descriptive statistics values of Soil Organic Carbon (SOC) and bulk density (BD) at the sampling sites in Torobamba and Coata.

Basin	Variable	Units	Mean	SD[1]	Min[2]	Max[3]
Torobamba	SOC	%	2.52	1.04	0.83	5.16
	BD	Mg m^{-3}	1.24	0.19	0.65	1.59
Coata	SOC	%	1.34	0.39	0.52	2.42
	BD	Mg m^{-3}	1.32	0.17	0.71	1.75

[1] Standard deviation, [2] Minimum, [3] Maximum

Two tables in the supplementary materials provide descriptive statistics for each basin based on the study covariates and the response variable SOCS (Tables S3 and S4).

3.2. Selected covariates

Table 3 shows the selected covariates for each basin based on the VIF criteria. Nine covariates were selected to model and predict SOCS. Multicollinearity among independent variables in the regression models, measured by the VIF, varied between 1.39 and 8.67 for Torobamba and 1.07 and 8.89 for Coata.

Table 3

Variance Inflation Factor (VIF) values for selected covariates in the two study areas

Torobamba		Coata	
Covariate	**VIF**	**Covariate**	**VIF**
Soil Adjusted Total Vegetation (SATVI)	8.67	Valley Depth (V_depth)	8.89
Normalized Height (N_height)	7.78	Soil Adjusted Total Vegetation Index (SATVI)	8.89
Second Brightness Index (BI2)	6.00	Topographic Wetness Index (TWI)	8.54
Valley Depth (V_depth)	5.61	Normalized Height (N_height)	6.68
Length Slope Factor (LS_factor)	5.26	Bare Soil Index (BSI)	6.17
Topographic Position Index (TPI)	2.88	Length Slope Factor (LS factor)	2.53
Plan Curvature (PLcurv)	2.71	Topographic Position Index (TPI)	2.21
Profile Curvature (PRcurv)	2.45	Profile Curvature (PRcurv)	1.51
Bare Soil Index (BSI)	1.39	Plan Curvature (PLcurv)	1.07

3.3. GWR Model analysis and predictions

The ample coefficient range displayed by the GWR models lead to significant spatial variation in the relationships between the selected covariates and the SOCS predictions (Table 4). The GWR model achieves an optimal bandwidth of 55,023.57 meters; however, GWRC model achieves an optimal bandwidth of 63 neighbors for Torobamba. For Coata, the GWR model has a fixed optimal bandwidth of 40,064.74 meters, while the GWRC has an optimal bandwidth of 95 neighbors.

Table 4

Summary of the coefficients used in geographically weighted regression model (GWR), and GWR with local collinearity adjustment (GWRC).

Covariates	GWR			GWRC		
	Minimum	**Maximum**	**Median**	**Minimum**	**Maximum**	**Median**
Torobamba						
Intercept	50.65	54.13	52.45	-19.41	109.8	45.95
Second Brightness Index (BI2)	-311.99	-242.2	-276.76	-868.29	394.62	-287.94
Bare Soil Index (BSI)	-6.89	1.13	-3.01	-61.02	84.29	-8.81
Length Slope Factor (LS_factor)	0.16	0.24	0.19	-0.3	0.61	-0.01
Normalized Height (N_height)	22.51	27.64	25.41	-18.53	73.35	33.92
Plan Curvature (PLcurv)	-583.62	690.22	20.59	-13224	10612.41	-6880.9
Profile Curvature (PRcurv)	-579.65	46.46	-238.7	-3959.2	1253.14	-1576.5
Soil Adjusted Total Vegetation Index (SATVI)	135.94	139.61	137.67	9.72	281.42	93.77
Topographic Position Index (TPI)	0.61	0.75	0.67	-1.92	3.34	1.01
Valley Depth (V_depth)	0.02	0.03	0.03	-0.06	0.12	0.04

	Coata					
Intercept	73.21	74.96	74.24	63.29	92.63	76.39
Normalized Height (N_height)	-27.43	-24.44	-26.05	-42.84	-6.27	-24.61
Bare Soil Index (BSI)	-19.51	-14.35	-17.6	-66.44	-1.3	-28.84
Length Slope Factor (LS_factor)	0.93	0.97	0.95	-0.02	0.79	0.56
Plan Curvature (PLcurv)	-257.75	-252.08	-255.28	-262.93	-172.64	-225.5
Profile Curvature (PRcurv)	-3260.3	-2556.6	-2937.1	-6246.7	-499.83	-3116.3
Valley Depth (V_depth)	-0.06	-0.05	-0.05	-0.1	-0.03	-0.05
Soil Adjusted Total Vegetation Index (SATVI)	44.66	52.15	49.38	5.44	72.84	50.02
Topographic Position Index (TPI)	1.15	1.27	1.23	-0.23	1.91	1.02
Topographic Wetness Index (TWI)	-1.34	-1.29	-1.31	-1.76	-0.97	-1.33

Spatial variation in the local R^2 values for the GWR models is minimal (around 1-2%). Across locations in Torobamba, the R^2 values fell within a range of 0.133 to 0.141. Coata shows similar results, with R^2 values between 0.164 and 0.189. In both basins, the higher R^2 values were located in the southeast region (Fig. 3). The model explains only 13-14% and 16-19% of SOCS variation, which means that the GWR model is consistent but has weak local prediction power across the study areas.

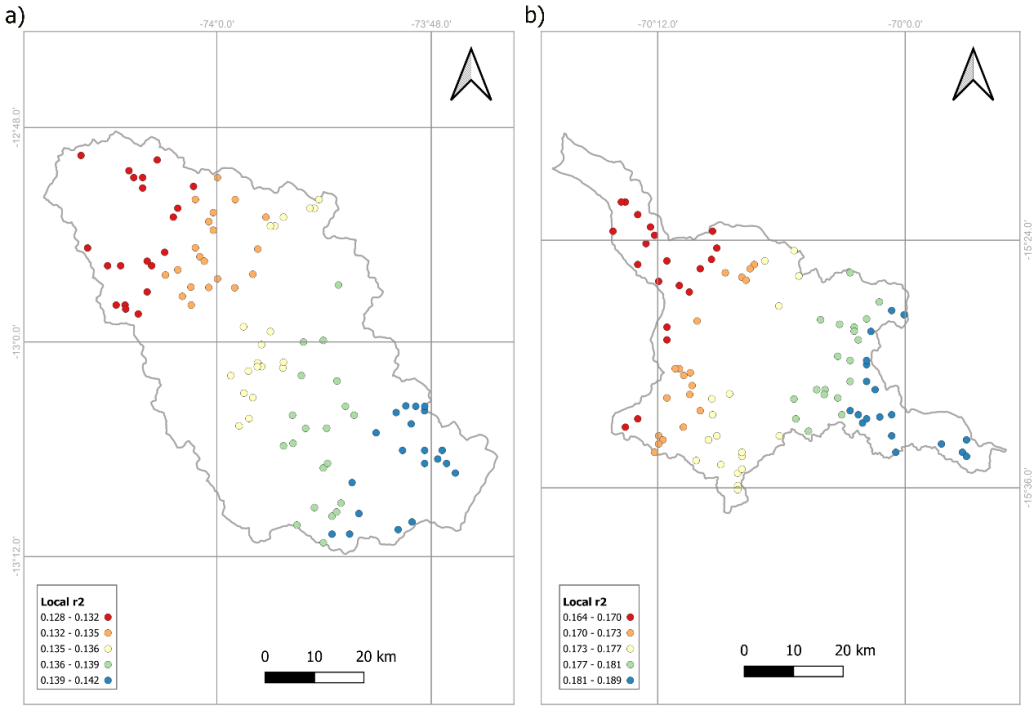

Fig. 3. Spatial distribution of local R^2 in a) Torobamba; b) Coata

When applying Geographically Weighted Ridge Regression (GWRR) in GWRC, we obtained the spatial distribution of CN, showing values between 17.77 and 27.33 for Torobamba and 20.11 and 26.70 for Coata (Fig. 4).

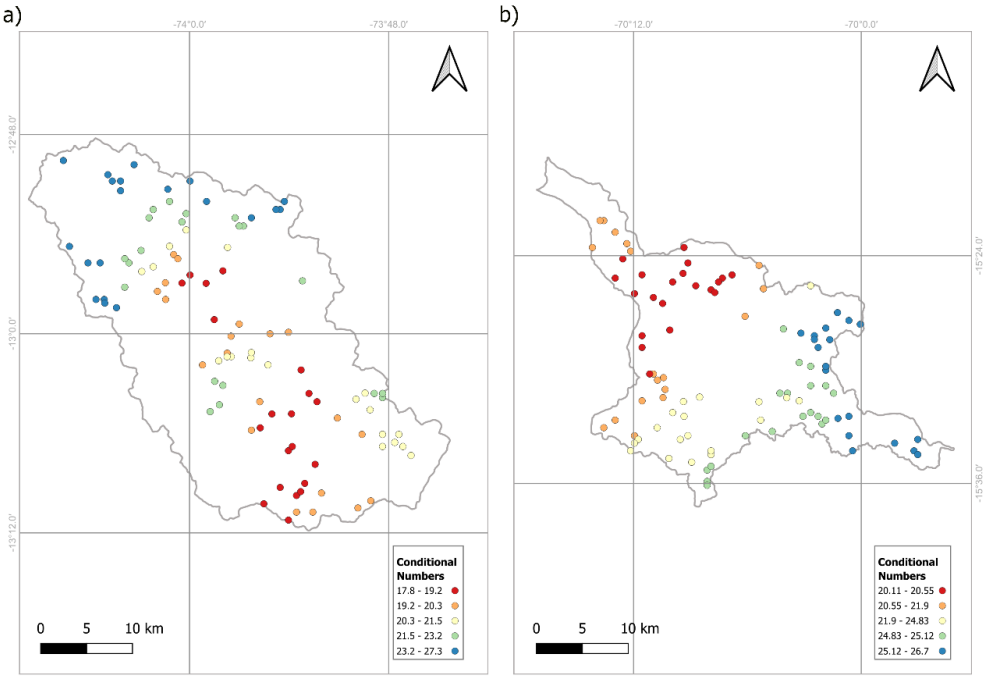

Fig. 4. Spatial distribution of local condition numbers (CN) values in a) Torobamba; b) Coata

The predicted SOCS using GWR with adjusted residuals ranged from 37.31 to 173.97 (Mg ha^{-1}) in Torobamba and from 10.99 to 100.21 (Mg ha^{-1}) in Coata. The kriged-interpolated residuals in the GWRC model presented values ranging from 27.84 to 161.01 (Mg ha^{-1}) in Torobamba and from 13.95 to 95.35 (Mg ha^{-1}) in Coata (Fig. 5).

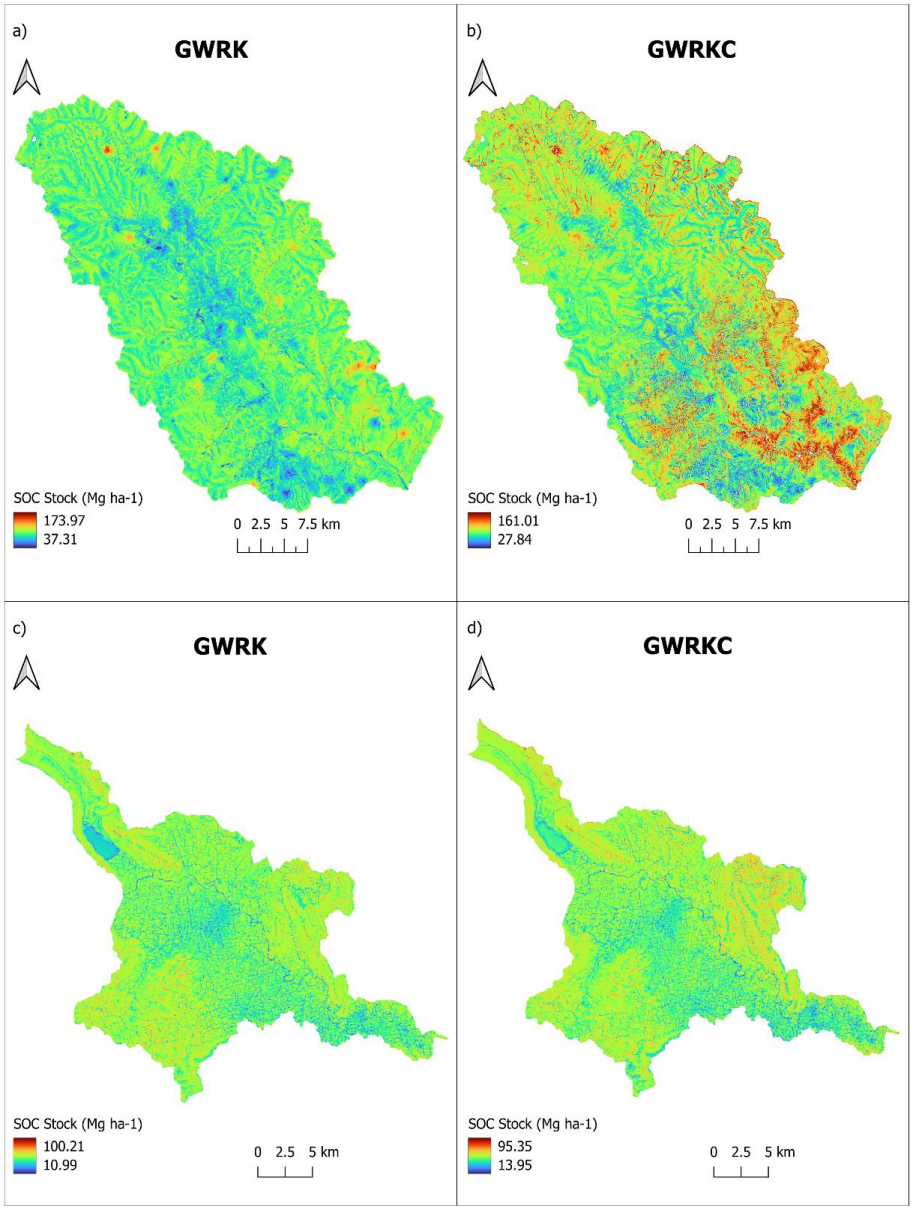

Fig. 5. Spatial distribution of SOCS predicted values in a) Torobamba using Geographically Weighted Regression Kriging (GWRK); b) Torobamba employing Geographically Weighted Regression with local collinearity adjustment Kriging (GWRCK); c) Coata using GWRK; and d) Coata employing GWRCK.

3.4. Machine learning models analysis

Variable importance (VI) plots show the contribution of covariates to the accuracy of the prediction model (Fisher et al., 2019). For the RF models, variables such as V_depth, SATVI, BI2, TPI, and BSI were identified as the most important in Torobamba, while V_depth, BSI, SATVI, N_height, and PRcurv were the most important in Coata (Fig. 6).

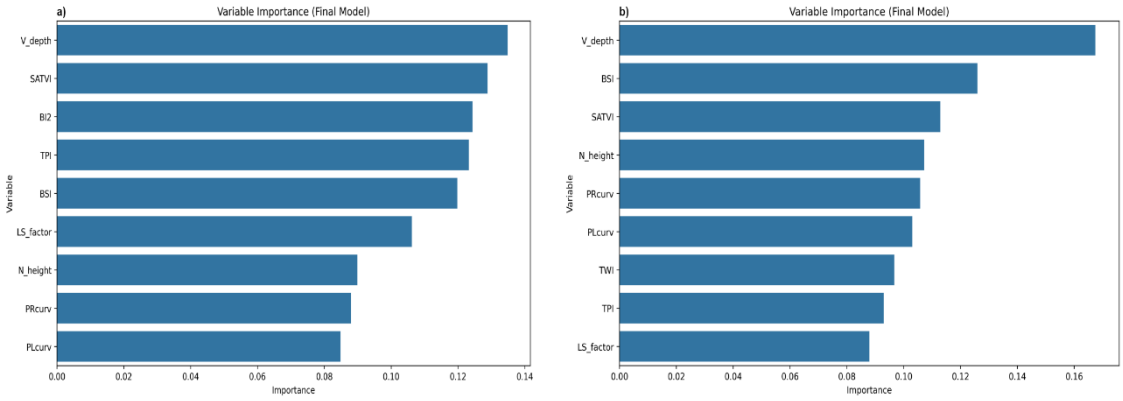

Fig. 6. Feature of importance plots for the Random Forest model in a) Torobamba; b) Coata.

For the GB models, topographic variables such as TPI, LS_factor, and V_depth, and vegetation-related variables such as BSI and SATVI, were the most important in Torobamba. In Coata, V_depth, TWI, N_height, and PLcurve were the most important topographic variables, while only SATVI represents the vegetation-related covariates (**Fig. 7**).

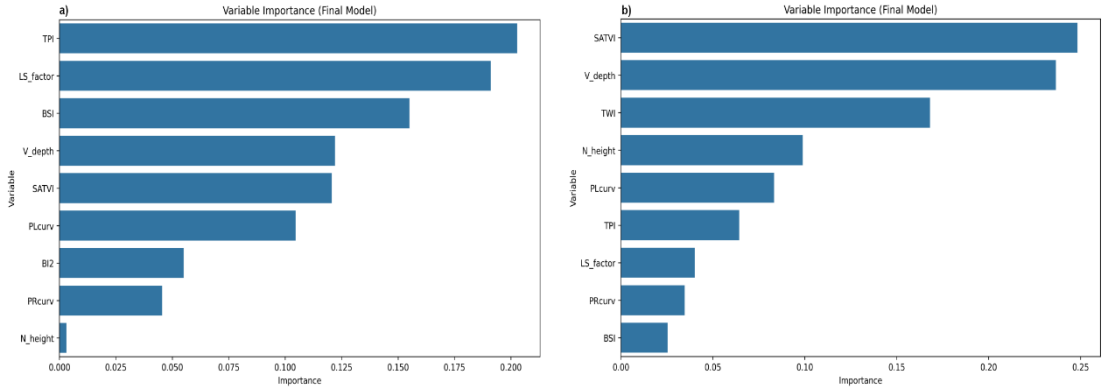

Fig. 7. Feature of importance plots for the Gradient Boosting model in a) Torobamba; b) Coata.

The spatial distribution of SOCS predicted by RF and GB models is illustrated in **Fig. 8**. Torobamba exhibited high SOCS values, ranging from 71.8 to 122.95 Mg ha^{-1} (RF) and 68.3 to 124.72 Mg ha^{-1} (GB), compared to Coata, which showed values between 40.02 to 69.9 Mg ha^{-1} (RF) and 37.29 to 75.02 Mg ha^{-1} (GB). While both models yielded comparable predictions, GB demonstrated a more uniform spatial distribution pattern, whereas RF showed more localized values.

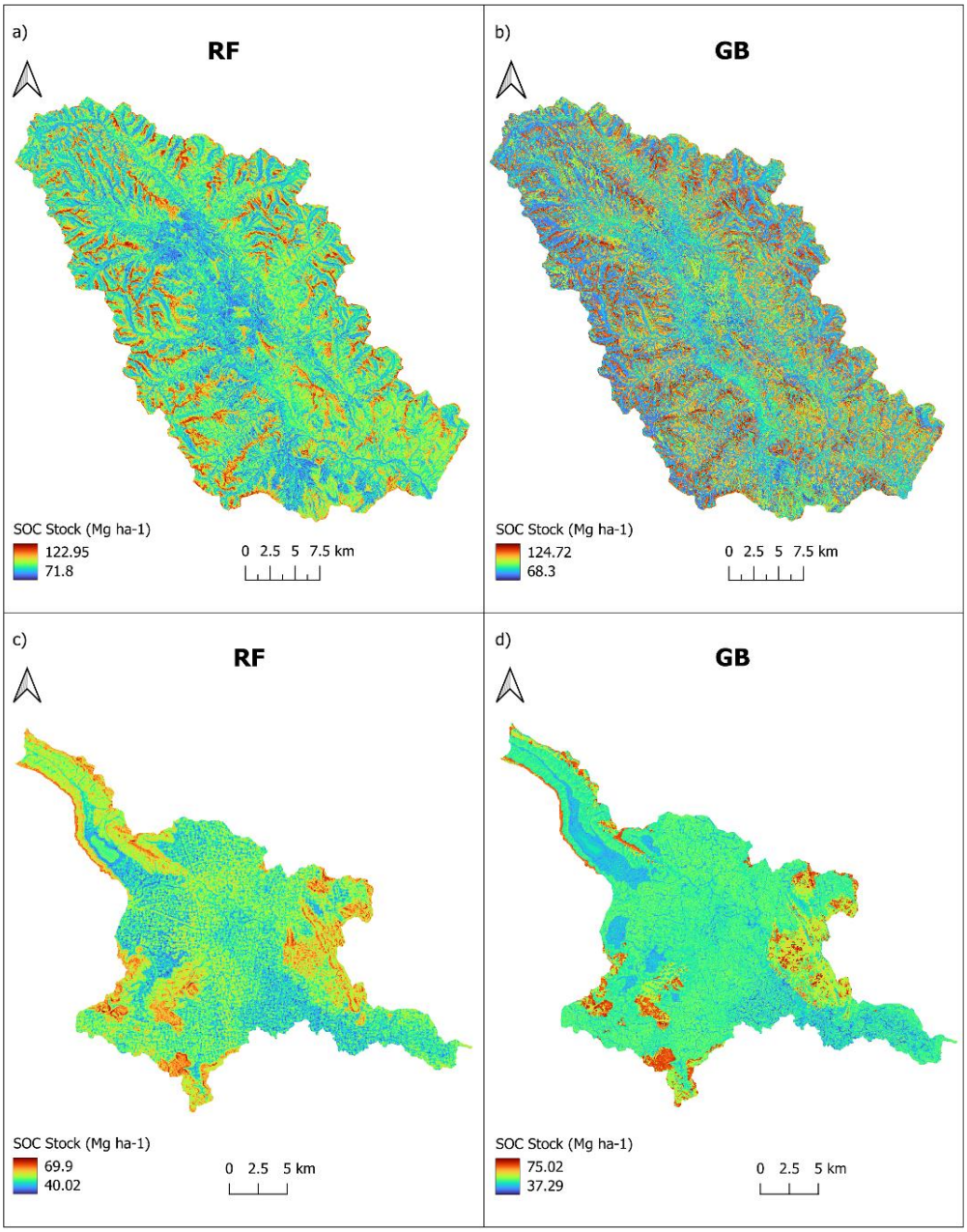

Fig. 8. Maps of predicted values of Soil Organic Carbon Stocks (SOCS) in a) Torobamba using Random Forest (RF), b) Torobamba using Gradient Boosting (GB), c) Coata using RF, and d) Coata using GB.

3.5. Change in SOCS by Land Cover

The predicted SOCS grouped by their corresponding LC are shown in Fig. 9 and Fig. 10. Torobamba exhibits the highest values of SOCS in the Shrubland cover (95.79 Mg ha^{-1}) predicted by the GWRKC model, followed by the Tree cover (94.17 Mg ha^{-1}) using

the GWRK model. Cropland has the lowest values when using GWRKC (85.88 Mg ha^{-1}) and GWRK (86.38 Mg ha^{-1}) while ML models predicted high values in the same LC. In Coata, SOCS predictions are similar for all LC types. As in Torobamba, the higher values were observed in the Tree (59.60 Mg ha^{-1}) and Shrubland (57.49 Mg ha^{-1}) areas, with the lowest values observed in Cropland (51.89 Mg ha^{-1}).

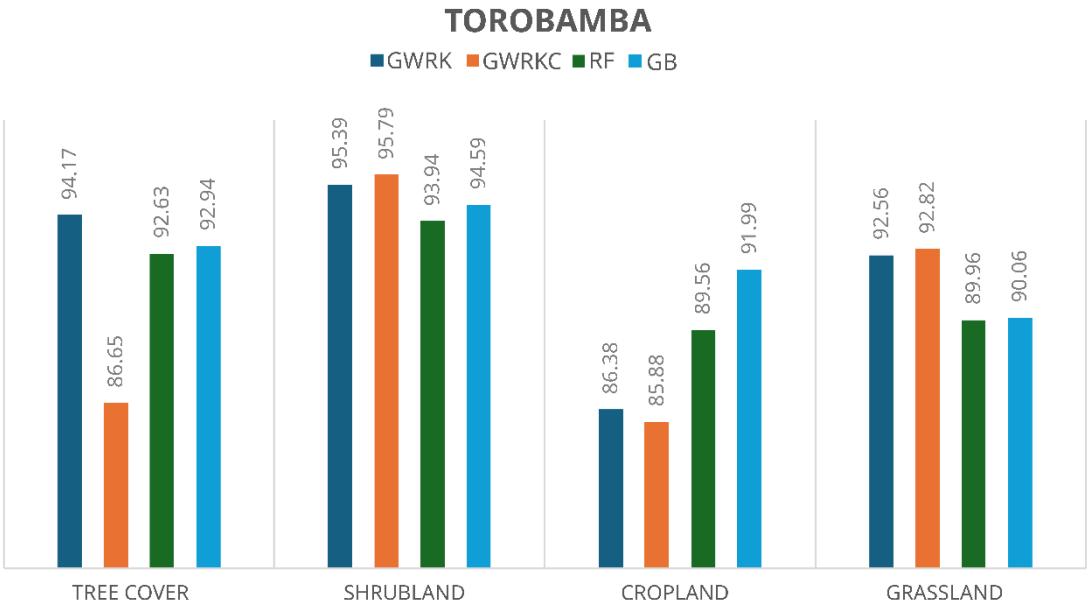

Fig. 9. Distribution of predicted Soil Organic Carbon Stocks in Mg ha^{-1} on Land Cover Types using all models in Torobamba.

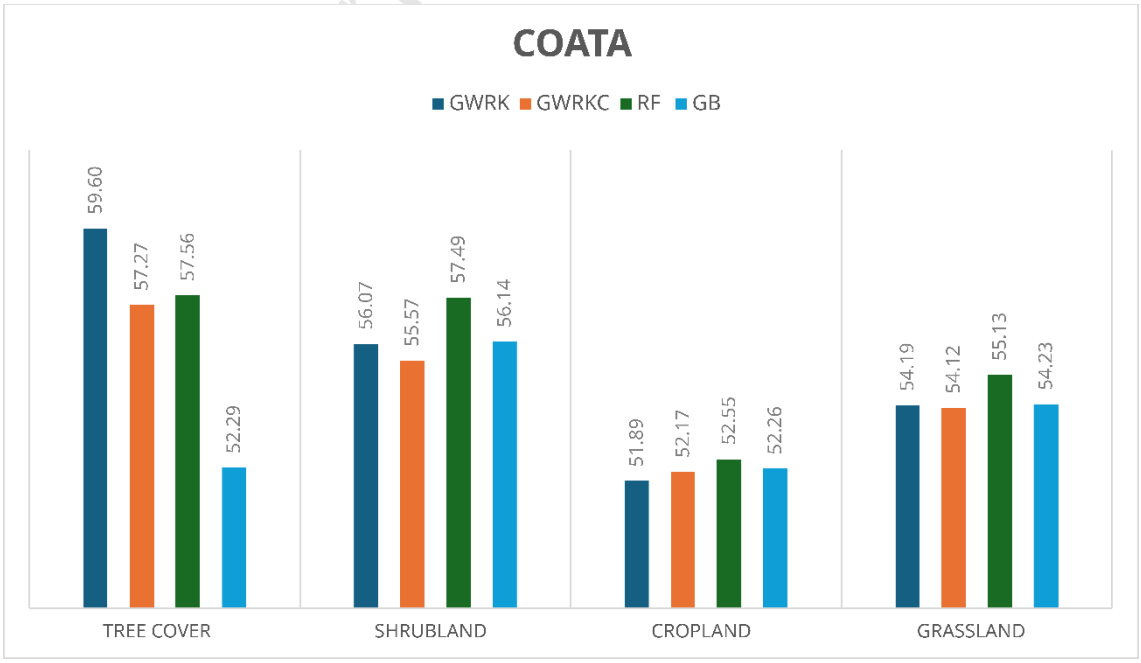

Fig. 10. Distribution of predicted Soil Organic Carbon Stocks in Mg ha^{-1} on Land Cover Types using all models in Coata.

3.6. Validation and model performance

Table 5 shows the model performance metrics for each location. These metrics help assess the accuracy and reliability of the SOCS predictions made by the models. The data presents models with high accuracy and discusses their potential application in regions that share similar characteristics. The GWRCK was the best model when assessing the datasets based on RMSE, MAE, and R^2 for both study areas.

Table 5

Summary of performance metrics for Geographically Weighted Regression (GWR) and their kriging adjusted (GWRK), GWR with collinearity analysis (GWRC) and their kriging adjusted (GWRCK), Random Forest (RF), and Gradient Boosting (GB) models.

Basin	Metrics	GWR	GWRK	Change[1] (%)	GWRC	GWRCK	Change[1] (%)	RF	GB
	RMSE	31.28	6.52	79.14	25.26	2.48	90.17	29.29	30.67
Torobamba	MAE	26.76	5.57	79.18	20.70	2.02	90.22	26.11	27.57
	R^2	0.14	0.96	609.36	0.44	0.99	127.77	0.11	0.03
	RMSE	15.60	14.37	7.84	14.89	13.43	9.80	14.19	16.01
Coata	MAE	12.74	11.74	7.88	12.21	11.01	9.87	13.15	14.75
	R^2	0.17	0.30	71.48	0.25	0.39	56.81	-0.01	-0.28

[1] Percentage improvement observed with the application of Kriging Interpolation

4. Discussion

Values of SOC and BD

Soils in Torobamba exhibited significantly higher SOC concentrations (2.52%) compared to Coata (1.34%). The SOC values in Torobamba surpassed those reported for Ferralsols, Acrisols, Planosols, and Podzols in the Amazon (2.18%) (Castañeda-Martín and Montes-Pulido, 2017), yet remained lower than levels reported in cultivated pastures and grasslands in the Peruvian Andes (3.94%, 3.83%) (Carbajal et al., 2024). The BD values in this study (1.24-1.32 Mgm^{-3}) were notably higher than those reported for Andean soils under diverse land uses (0.48-0.53 gcm^{-3}) (Bravo-Medina et al., 2023). Similarly, BD values exceeded those documented in watersheds of the Ecuadorian and Colombian Andes (0.5 to 1.0 g cm-3) (Andrade et al., 2022; Beltrán-Dávalos et al.,

2022), suggesting higher levels of soil compaction. These differences likely reflect interactions among climate, vegetation, land use, and topography.

Influence of Explanatory Variables on SOCS Predictions with GWR and GWRC models

Geographically Weighted Regression (GWR) models confirmed spatial non-stationarity in the relationships between predictor variables and SOCS across the study areas (Table 4). In Torobamba, terrain curvature (PLcurv, PRCurv) and vegetation indices (particularly SATVI) dominated, while in Coata PR_curv exerted the strongest influence, followed by N_height and BSI. Coefficient ranges demonstrated strong spatial variability, often switching between positive and negative influences depending on location. Kernel function modifications and bandwidth parameters selection were found to directly affect GWR coefficient estimation (Wang et al., 2012).

Comparisons between standard GWR and GWRC revealed key differences in coefficient distributions. GWRC produced broader coefficient ranges (e.g., PL_curv in Torobamba: -583.62 to 690.22 in GWR vs. to -13224.00 to 10612.41 in GWRC). This underscores the effectiveness of GWRC in addressing local collinearity, improving interpretability of local regression coefficients despite minimal global collinearity (Wheeler, 2007). While GWRC enhances coefficient reliability under multicollinearity (Bárcena et al., 2014), the increased variability requires careful interpretation. Bandwidth optimization was particularly critical, with adaptive bandwidths proving more effective under uneven sample densities (Kiani et al., 2024).

Despite accounting for spatial variation, local R^2 values remained low (mean R^2: 0.136 in Torobamba, 0.177 in Coata) (Fig. 3), suggesting selected predictors only partially explained SOCS variability. Possible causes include missing explanatory variables, unaddressed non-linearities, or inherent system randomness. Strong spatial heterogeneity, unrecorded local effects, or locations where the underlying spatial process is inherently noisy or less predictable likely contribute to these results. In regions with complex geomorphology and spatial heterogeneity, finding a clear pattern can be difficult (Costa et al., 2018). Sparse sampling further limited local reliability (Devkota et al., 2013) while residual spatial autocorrelation pointed to model limitations (Gaspard et al., 2019). Bandwidth sensitivity further influenced outcomes: overly broad bandwidths risked over-smoothing and convergence toward a standard regression model

(Weku et al., 2022), which can lead to. Multicollinearity (CN values ~ 20-28) (Fig.4) can produce unstable coefficient estimates highly sensitive to minor changes, impacting the interpretation of local predictor effects (Brunsdon et al., 2012).

GWR outperformed OLS by reducing AICc (Sartika and Suryani, 2020), explicitly capturing spatial heterogeneity and local variation often ignored by conventional ML models (Yan et al., 2022; Mishra et al., 2010). GWRK improved further by combining local regression with kriging of residuals (Kumar et al., 2012; Imran et al., 2015). While standalone ML offers predictive accuracy, it lacks spatial interpretability and risks bias without adaptations such as Random Forest Spatial Interpolation (Sun et al., 2021; Kmoch et al., 2025). Integration of GWR and kriging remains a robust approach for SOCS mapping (Adeniyi et al., 2024; Faisal et al., 2025).

GWRC provided more stable and interpretable local coefficient by mitigating multicollinearity effects (Czarnota et al., 2015), though at the expense of wider coefficient ranges or altered weighting schemes. Even though these changes are designed to address collinearity, local models may still use different neighbor sets or weights, which may account for the differences in the GWRC coefficient estimates (Leong and Yue, 2017). Multi-scale GWR (MGWR) may further improve performance by capturing spatial heterogeneity and residual structure more effectively than OLS and supporting accurate, spatially explicit carbon stock estimates (Song et al., 2024; Kuang and Chen, 2025).

Variable Importance (VI) in Machine Learning Models

Analyzing variable importance provides crucial insights into model functionality and is vital for improving performance (Chamma et al., 2024). High VI scores indicate strong predictive power and a close relationship between a predictor and SOC variation in the study area. However, establishing causality is different, as it requires evaluating whether randomizing the predictor reduces prediction accuracy (Lamsaf et al., 2025; Wang et al., 2024). Numerous studies have shown topographic and vegetation variables to be critical predictors of SOC. Topography can directly influence SOC accretion in valley bottoms (Schwanghart and Jarmer, 2011), or indirectly reflect soil moisture conditions regulating vegetation growth and residue decomposition (Rasel et al., 2017). In the present study, V_depth strongly influenced both basins (Fig. 4). Topographic variables derived from DEMs, such as TWI and V_depth, affect SOC erosion and deposition processes (Dorji et al., 2014; Wang et al., 2020). Previous research found V_depth to

have moderate influence in SOC stock prediction compared to elevation and LS_factor (Adhikari et al., 2019), while other models such as RF and BRT identified standardized height and altitude as best predictors (Ließ et al., 2016). In Coata, the relative homogeneous terrain limited topographic influence, with TWI emerging as the primary variable (John et al., 2020).

Land use/land cover and vegetation indices are also critical SOC predictors in the Peruvian Andes (Carbajal et al., 2024). Precipitation and NDVI have been found to outperform terrain attributes (Emadi et al., 2020). In this study, SATVI was a crucial predictor in Torobamba, while BSI was more important in Coata. BSI combined with alongside geological units, soil taxonomy, precipitation, elevation, and the LS_factor, has improved SOC mapping accuracy (Ayala Izurieta et al., 2021). Recent findings show that combining spectral indices with topographical features reduces validation errors and optimizes SOC predictions (Cutting et al., 2024).

SOCS prediction using GWR and ML models

All four modeling approaches generated similar within-basin predictions, though inter-basin differences were substantial. In Torobamba, GWRK predicted the highest SOCS (173.97 $Mgha^{-1}$), followed by GWRKC (161.01 $Mgha^{-1}$), while GB estimates were much lower (68.3 Mg ha^{-1}). In Coata, GWRK also produced the highest SOCS (100.21 $Mgha^{-1}$), while GB again predicted the lowest (37.29 Mg ha^{-1}). These differences underscore the need for ensemble methods when estimating SOCS (Mishra et al., 2020). While GWR models proved useful in areas of high spatial SOCS variability, ML produced narrower SOCS ranges, failing to fully capture the inherent variability, potentially due to sampling limitations.

SOCS under different land cover types

Distinct SOCS patterns emerged across land cover types in both basins. In Torobamba (Fig. 9), shrublands showed the highest SOC stocks (mean: 94.93 $Mgha^{-1}$), followed by grasslands (91.60 $Mgha^{-1}$), tree cover (91.60 $Mgha^{-1}$), and croplands (88.95 $Mgha^{-1}$). The SOCS in shrublands may be attributed to the combination of slower decomposition rates and substantial biomass input from shrub species, which contribute to organic matter accumulation in the soil (Liu et al., 2023). In Coata (Fig. 10), tree cover exhibited the highest SOCS (mean: 56.71 Mg ha^{-1}), followed by shrublands (56.32 $Mgha^{-1}$), with grasslands (54.42 $Mgha^{-1}$) and croplands (52.22 $Mgha^{-1}$) showing lower values. These

results align with global findings that forests typically maintain higher SOCS due to continuous litter input and reduced soil disturbance compared to managed agricultural lands (Lal, 2004). The consistently lower cropland SOCS across both basins reflects the impact of agricultural activities on soil carbon dynamics. These findings corroborate that significant SOC losses occur after converting natural ecosystems to cropland due to reduced carbon input, increased decomposition rates, and soil disturbance from tillage practices (Deng et al., 2016; Wei et al., 2014).

Environmental differences, such as precipitation and temperature, may partly explain greater biomass production and slower decomposition rates in Torobamba, contributing to higher SOCS relative to Coata. Such regional variations in SOCS have been reported, emphasizing the importance of climate gradients and geomorphological features in determining regional SOC distribution patterns (Wiesmeier et al., 2019). Despite numerical differences among models, consistency in relative patterns among land cover types increases confidence in the overall findings.

Performance of Models for Predicting and Mapping SOCS

GWR and its variants, RF, and GB models to reveal significant differences in performance to predict SOCS under Andean conditions (Table 5). In Torobamba, GWR variants (GWRK and GWRCK) achieved the highest predictive performance, with RMSE reductions of 79–90% and R^2 values up to 0.99. Kriging interpolation enhanced predictions by addressing residual spatial autocorrelation. These findings are consistent with prior research indicating the effectiveness of GWR models in addressing spatial non-stationarity (Cao et al., 2022; Li et al., 2010). This attribute is critical in heterogeneous Andean landscapes with variable topography and microclimates. In Coata, improvements were smaller (RMSE reduction 7.8–9.8%) though R^2 values (0.3–0.39) still outperformed other models, suggesting less pronounced spatial variability or limitations in capturing local effects.

RF and GB performed poorly in both basins. In Torobamba, R^2 values remained below 0.11, while in Coata, RF (RMSE 14.19) and GB (RMSE 16.01) showed moderate performance but still produced negative R^2 values, suggesting poor fit. These outcomes may stem from inadequate feature selection, suboptimal hyperparameter tuning, insufficient training data, or lack of spatial covariates. While RF often excels in SOC prediction (Hengl et al., 2018) (Ayala Izurieta et al., 2021; Ließ et al., 2016), its

performance here indicates challenges unique to Andean conditions. One reason might be that non-spatial ML models may not adequately capture the complex spatial dependencies in complex landscapes. Another one might be that the model performance is highly context dependent. In contrast to the relatively uniform landscape of Coata, the rugged terrain of Torobamba may enhance the usefulness of GWR. Conversely, the environmental and land-use patterns in Coata could either mask spatial variations or introduce confounding effects.

Several studies have shown that ML and GWR models developed for SOCS prediction in one region may be transferable to others when similarities in climate, soil types, land use, and covariate datasets exist. However, differences in local environmental drivers and data availability often require model recalibration or local adaptation. Cross-regional transfer is most effective when models rely on harmonized datasets, such as the Harmonized World Soil Database (HWSD) or WISE30se (Batjes, 2016). Recent advances, such as domain-adaptive ML for crop yield prediction (Priyatikanto et al., 2023), pasture growth transfer modeling across diverse climates to predict nitrogen response rates (Pylianidis et al., 2023), and SOC prediction using Convolutional Neural Network- based transfer learning– (Han et al., 2025) highlight promising future directions.

5. Conclusions

Digital SOCS mapping provides critical insights for characterizing high Andean soils in Peru, supporting the development of regulatory frameworks to prevent degradation and contributing to global SOC assessments. These soils play critical roles in regulating hydrological processes, sequestering atmospheric carbon and providing ecosystem services, underscoring their regional and global significance. The high SOCS values obtained in the present study highlight the importance of characterizing SOC reserves in Andean landscapes.

Our results demonstrated that GWR and its variants outperformed conventional ML for SOCS prediction in complex Andean landscapes, especially under small to medium-sized sampling GWRK, in particular, accurately captured spatial heterogeneity, enhancing predictive reliability. By contrast, RF and GB showed limited performance without explicit spatial adaptation mechanisms.

Although SOCS modeling techniques such as GWR and ML have broad applicability, they require thorough validation and context-specific adjustments to ensure reliable predictions. Future research should explore hybrid modeling frameworks that integrate spatial weighting and ML to optimize predictions in heterogeneous mountain environments.

Funding

This research was funded by the INIA project "Mejoramiento de los servicios de investi-gación y transferencia tecnológica en el manejo y recuperación de suelos agrícolas degradados y aguas para riego en la pequeña y mediana agricultura en los departamentos de Lima, Áncash, San Martín, Cajamarca, Lambayeque, Junín, Ayacucho, Arequipa, Puno y Ucayal", CUI 2487112.

Credit authorship contribution statement

Carlos Carbajal: Writing – original draft, Methodology, Formal analysis, Investigation, Software, Data Curation. **Merely Tumbalobos-Dextre**: Methodology, Software, Visualization. **Tatiana Condori-Ataupillco**: Investigation, Resources, **Nestor Cuellar-Condori**: Investigation, Resources, **Carla Gavilan**: Conceptualization, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

To the Soil, Water, and Foliar Laboratory (LABSAF) network technicians, especially of La Molina, Canaán, and Illpa Experimental Agrarian Stations headquarters. Special thanks go to Marilia Coila Mamani and Fredy Flores Galindo for their help collecting soil samples.

References

Adeniyi, O.D., Brenning, A., Maerker, M., 2024. Spatial prediction of soil organic carbon: Combining machine learning with residual kriging in an agricultural lowland area (Lombardy region, Italy). Geoderma 448, 116953. https://doi.org/10.1016/j.geoderma.2024.116953

Adhikari, K., Owens, P.R., Libohova, Z., Miller, D.M., Wills, S.A., Nemecek, J., 2019. Assessing soil organic carbon stock of Wisconsin, USA and its fate under future land use and climate change. Science of The Total Environment 667, 833–845. https://doi.org/10.1016/j.scitotenv.2019.02.420

Anderson, E., Marengo, J., Villalba, R., Halloy, S., Young, B., Cordero, D., Gast, F., Jaimes, E., Ruiz Carrascal, D., 2011. Consequences of climate change for ecosystems and ecosystem services in the tropical andes, in: Climate Change and Biodiversity in the Tropical Andes. MacArthur Foundation Inter American Institute For Global Change Research.IAI The Scientific Committee on Problems of the Environment.SCOPE, pp. 1–18.

Andrade, H.J., Segura, M.A., Canal-Daza, D.S., 2022. Conservation of Soil Organic Carbon in the National Park Santuario de Fauna y Flora Iguaque, Boyacá-Colombia. Forests 13, 1275. https://doi.org/10.3390/f13081275

Ayala Izurieta, J.E., Márquez, C.O., García, V.J., Jara Santillán, C.A., Sisti, J.M., Pasqualotto, N., Van Wittenberghe, S., Delegido, J., 2021. Multi-predictor mapping of soil organic carbon in the alpine tundra: a case study for the central Ecuadorian páramo. Carbon Balance and Management 16, 32. https://doi.org/10.1186/s13021-021-00195-2

Bárcena, M.J., Menéndez, P., Palacios, M.B., Tusell, F., 2014. Alleviating the effect of collinearity in geographically weighted regression. J Geogr Syst 16, 441–466. https://doi.org/10.1007/s10109-014-0199-6

Baret, F., Guyot, G., 1991. Potentials and limits of vegetation indices for LAI and APAR assessment. Remote Sensing of Environment 35, 161–173. https://doi.org/10.1016/0034-4257(91)90009-U

Batjes, N.H., 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. Geoderma 269, 61–68. https://doi.org/10.1016/j.geoderma.2016.01.034

Beltrán-Dávalos, A.A., Ayala Izurieta, J.E., Echeverria Guadalupe, M.M., Van Wittenberghe, S., Delegido, J., Otero Pérez, X.L., Merino, A., 2022. Evaluation of Soil Organic Carbon Storage of Atillo in the Ecuadorian Andean Wetlands. Soil Systems 6, 92. https://doi.org/10.3390/soilsystems6040092

Belyadi, H., Haghighat, A., 2021. Chapter 5 - Supervised learning, in: Belyadi, H., Haghighat, A. (Eds.), Machine Learning Guide for Oil and Gas Using Python. Gulf Professional Publishing, pp. 169–295. https://doi.org/10.1016/B978-0-12-821929-4.00004-4

Bivand, R., Yu, D., 2006. spgwr: Geographically Weighted Regression. https://doi.org/10.32614/CRAN.package.spgwr

Bravo-Medina, C., Torres-Navarrete, B., Arteaga-Crespo, Y., Garcia-Quintana, Y., Reyes-Morán, H., Changoluisa-Vargas, D., Paguay-Sayay, D., 2023. Soil properties variation in a small-scale altitudinal gradient of an evergreen foothills forest, Ecuadorian Amazon region. European Journal of Forest Research 142, 1325–1339. https://doi.org/10.1007/s10342-023-01593-6

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brunsdon, C., Charlton, M., Harris, P., 2012. Living with Collinearity in Local Regression Models. Spatial Accuracy.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis 28, 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Cao, Z., Zhu, W., Luo, P., Wang, S., Tang, Z., Zhang, Y., Guo, B., 2022. Spatially Non-Stationary Relationships between Changing Environment and Water Yield Services in Watersheds of China's Climate Transition Zones. Remote Sensing 14, 5078. https://doi.org/10.3390/rs14205078

Carbajal, M., Ramírez, D.A., Turin, C., Schaeffer, S.M., Konkel, J., Ninanya, J., Rinza, J., De Mendiburu, F., Zorogastua, P., Villaorduña, L., Quiroz, R., 2024. From Rangelands to Cropland, Land-Use Change and Its Impact on Soil Organic Carbon Variables in a Peruvian Andean Highlands: A Machine Learning Modeling Approach. Ecosystems. https://doi.org/10.1007/s10021-024-00928-7

Castañeda-Martín, A.E., Montes-Pulido, C.R., 2017. Carbono almacenado en páramo andino. Entramado 13, 210–221. https://doi.org/10.18041/entramado.2017v13n1.25112

Chamma, A., Thirion, B., Engemann, D., 2024. Variable Importance in High-Dimensional Settings Requires Grouping. AAAI 38, 11195–11203. https://doi.org/10.1609/aaai.v38i10.28997

Chen, J., Qu, M., Zhang, J., Xie, E., Zhao, Y., Huang, B., 2021. Improving the spatial prediction accuracy of soil alkaline hydrolyzable nitrogen using GWPCA-GWRK. Soil Science Soc of Amer J 85, 879–892. https://doi.org/10.1002/saj2.20189

Chen, Q., Wang, Y., Zhu, X., 2024. Soil organic carbon estimation using remote sensing data-driven machine learning. PeerJ 12, e17836. https://doi.org/10.7717/peerj.17836

Chen, S., Arrouays, D., Mulder, V.L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A.C., Walter, C., 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma 409, 115567. https://doi.org/10.1016/j.geoderma.2021.115567

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8, 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015

Costa, E.M., Tassinari, W. de S., Pinheiro, H.S.K., Beutler, S.J., dos Anjos, L.H.C., 2018. Mapping Soil Organic Carbon and Organic Matter Fractions by Geographically Weighted Regression. J. Environ. Qual. 47, 718–725. https://doi.org/10.2134/jeq2017.04.0178

Cutting, B.J., Atzberger, C., Gholizadeh, A., Robinson, D.A., Mendoza-Ulloa, J., Marti-Cardona, B., 2024. Remote Quantification of Soil Organic Carbon: Role of Topography in the Intra-Field Distribution. Remote Sensing 16, 1510. https://doi.org/10.3390/rs16091510

Czarnota, J., Wheeler, D.C., Gennings, C., 2015. Evaluating Geographically Weighted Regression Models for Environmental Chemical Risk Analysis. Cancer Inform 14, 117–127. https://doi.org/10.4137/CIN.S17296

Deng, L., Zhu, G., Tang, Z., Shangguan, Z., 2016. Global patterns of the effects of land-use changes on soil carbon stocks. Global Ecology and Conservation 5, 127–138. https://doi.org/10.1016/j.gecco.2015.12.004

Devkota, M., Hatfield, G., Chintala, R., 2014. Effect of Sample Size on the Performance of Ordinary Least Squares and Geographically Weighted Regression. Br. J. Math. Comput. Sci. 4, 1–21. https://doi.org/10.9734/BJMCS/2014/6050

Dorji, T., Odeh, I.O.A., Field, D.J., Baillie, I.C., 2014. Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems, Eastern Himalayas. Forest Ecology and Management 318, 91–102. https://doi.org/10.1016/j.foreco.2014.01.003

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., Scholten, T., 2020. Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. Remote Sensing 12, 2234. https://doi.org/10.3390/rs12142234

Emamgholizadeh, S., Shahsavani, S., Eslami, M.A., 2017. Comparison of artificial neural networks, geographically weighted regression and Cokriging methods for predicting the spatial distribution of soil macronutrients (N, P, and K). Chin. Geogr. Sci. 27, 747–759. https://doi.org/10.1007/s11769-017-0906-6

Escadafal, R., 1989. Remote sensing of arid soil surface color with Landsat thematic mapper. Advances in Space Research 9, 159–163. https://doi.org/10.1016/0273-1177(89)90481-X

Faisal, F., Pramoedyo, H., Astutik, S., Efendi, A., 2025. Bayesian Geographically Weighted Regression with Kriging for Enhanced Spatial Prediction: A Comparison of Jeffreys' and Conjugate Priors. Math. Model. Eng. Probl. 12. https://doi.org/10.18280/mmep.120515

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Intl Journal of Climatology 37, 4302–4315. https://doi.org/10.1002/joc.5086

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20, 1–81.

Fotheringham, A.S., Charlton, M.E., Brunsdon, C., 1998. Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis. Environ Plan A 30, 1905–1927. https://doi.org/10.1068/a301905

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29, 1189–1232. https://doi.org/10.1214/aos/1013203451

Gao, B., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 58, 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

García Lino, M.C., Pfanzelt, S., Domic, A.I., Hensen, I., Schittek, K., Meneses, R.I., Bader, M.Y., 2024. Carbon dynamics in high-Andean tropical cushion peatlands: A review of geographic patterns and potential drivers. Ecol. Monogr. 94, e1614. https://doi.org/10.1002/ecm.1614

Gaspard, G., Kim, D., Chun, Y., 2019. Residual spatial autocorrelation in macroecological and biogeographical modeling: a review. J. Ecol. Environ. 43, 19. https://doi.org/10.1186/s41610-019-0118-3

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., Villa-Vialaneix, N., 2017. Random Forests for Big Data. Big Data Research 9, 28–46. https://doi.org/10.1016/j.bdr.2017.07.003

Ghaderpour, E., Mazzanti, P., Bozzano, F., Scarascia Mugnozza, G., 2024. Trend Analysis of MODIS Land Surface Temperature and Land Cover in Central Italy. Land 13, 796. https://doi.org/10.3390/land13060796

Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sensing of Environment 58, 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P., 2015. GWmodel : An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. J. Stat. Soft. 63. https://doi.org/10.18637/jss.v063.i17

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, Big Remotely Sensed Data: tools, applications and experiences 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Guo, L., Luo, M., Zhangyang, C., Zeng, C., Wang, S., Zhang, H., 2018. Spatial modelling of soil organic carbon stocks with combined principal component analysis and geographically weighted regression. J. Agric. Sci. 156, 774–784. https://doi.org/10.1017/S0021859618000709

Habib, M., Habib, A., Alibrahim, B., 2024. Prediction and parametric assessment of soil one-dimensional vertical free swelling potential using ensemble machine learning models. Advanced Modeling and Simulation in Engineering Sciences 11, 26. https://doi.org/10.1186/s40323-024-00277-z

Halder, K., Srivastava, A.K., Ghosh, A., Nabik, R., Pan, S., Chatterjee, U., Bisai, D., Pal, S.C., Zeng, W., Ewert, F., Gaiser, T., Pande, C.B., Islam, A.R.Md.T., Alam, E., Islam, M.K., 2024. Application of bagging and boosting ensemble machine learning techniques for groundwater potential mapping in a drought-prone agriculture region of eastern India. Environmental Sciences Europe 36, 155. https://doi.org/10.1186/s12302-024-00981-y

Han, J., Wu, M., Qi, Y., Li, X., Chen, X., Wang, J., Zhu, J., Li, Q., 2025. A soil organic carbon mapping method based on transfer learning without the use of exogenous data. Front. Environ. Sci. 13, 1580085. https://doi.org/10.3389/fenvs.2025.1580085

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518. https://doi.org/10.7717/peerj.5518

Hengl, T., Sorenson, P., Parente, L., Cornish, K., Battigelli, J., Bonannella, C., Gorzelak, M., Nichols, K., 2023. Assessment of soil organic carbon stocks in Alberta using 2-scale sampling and 3D predictive soil mapping. FACETS 8, 1–17. https://doi.org/10.1139/facets-2023-0040

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J.W., Heuvelink, G.B.M., 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Computers & Geosciences 35, 1711–1721. https://doi.org/10.1016/j.cageo.2008.10.011

Hijmans, R.J., Barbosa, M., Ghosh, A., Mandel, A., 2021. geodata: Download Geographic Data. https://doi.org/10.32614/CRAN.package.geodata

Hollister, J., Shah, T., Nowosad, J., Robitaille, A.L., Beck, M.W., Johnson, M., 2023. elevatr: Access elevation data from various apis (manual). https://doi.org/10.5281/zenodo.8335450

Hounkpatin, K.O.L., Stendahl, J., Lundblad, M., Karltun, E., 2021. Predicting the spatial distribution of soil organic carbon stock in Swedish forests using a group of covariates and site-specific data. SOIL 7, 377–398. https://doi.org/10.5194/soil-7-377-2021

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment, The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring 83, 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment 25, 295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Imran, M., Stein, A., Zurita-Milla, R., 2015. Using geographically weighted regression kriging for crop yield mapping in West Africa. Int. J. Geogr. Inf. Sci. 29, 234–257. https://doi.org/10.1080/13658816.2014.959522

IUSS Working Group, 2007. World reference base for soil resources 2006, first update 2007, World Soil Resources Reports No. 103. ed. FAO, Rome.

Jenny, H., 1994. Factors of soil formation: a system of quantitative pedology, Unabridged, unaltered republ., new foreword. ed, Dover books on earth sciences. Dover Publ, New York.

Kiani, B., Sartorius, B., Lau, C.L., Bergquist, R., 2024. Mastering geographically weighted regression: key considerations for building a robust model. Geospatial Health 19. https://doi.org/10.4081/gh.2024.1271

Kmoch, A., Harrison, C.T., Choi, J., Uuemaa, E., 2025. Spatial autocorrelation in machine learning for modelling soil organic carbon. Ecol. Inform. 86, 103057. https://doi.org/10.1016/j.ecoinf.2025.103057

Kuang, Y., Chen, X., 2025. Spatial heterogeneity of forest carbon stocks in the Xiangjiang river Basin urban agglomeration: analysis and assessment based on the multiscale geographically weighted regression (MGWR) model. Front. Environ. Sci. 13, 1573438. https://doi.org/10.3389/fenvs.2025.1573438

Kumar, A., Moharana, P.C., Jena, R.K., Malyan, S.K., Sharma, G.K., Fagodiya, R.K., Shabnam, A.A., Jigyasu, D.K., Kumari, K.M.V., Doss, S.G., 2023. Digital Mapping of Soil Organic Carbon Using Machine Learning Algorithms in the Upper Brahmaputra Valley of Northeastern India. Land 12, 1841. https://doi.org/10.3390/land12101841

Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189–190, 627–634. https://doi.org/10.1016/j.geoderma.2012.05.022

Lal, R., 2004. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. Science 304, 1623–1627. https://doi.org/10.1126/science.1097396

Lamsaf, A., Carrilho, R., Neves, J.C., Proença, H., 2025. Causality, Machine Learning, and Feature Selection: A Survey. Sensors 25, 2373. https://doi.org/10.3390/s25082373

Leong, Y.-Y., Yue, J.C., 2017. A modification to geographically weighted regression. International Journal of Health Geographics 16, 11. https://doi.org/10.1186/s12942-017-0085-9

Li, S., Zhao, Z., Miaomiao, X., Wang, Y., 2010. Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression. Environmental Modelling & Software 25, 1789–1800. https://doi.org/10.1016/j.envsoft.2010.06.011

Ließ, M., Schmidt, J., Glaser, B., 2016. Improving the Spatial Prediction of Soil Organic Carbon Stocks in a Complex Tropical Mountain Landscape by Methodological Specifications in Machine Learning Approaches. PLoS ONE 11, e0153673. https://doi.org/10.1371/journal.pone.0153673

Liu, Y., Wang, K., Dong, L., Li, J., Wang, X., Shangguan, Z., Qu, B., Deng, L., 2023. Dynamics of litter decomposition rate and soil organic carbon sequestration following vegetation succession on the Loess Plateau, China. CATENA 229, 107225. https://doi.org/10.1016/j.catena.2023.107225

Marsett, R.C., Qi, J., Heilman, P., Society for Range Management, 2006. Remote Sensing for Grassland Management in the Arid Southwest. REM 59. https://doi.org/10.2458/azu_jrm_v59i5_marsett

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32, 1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009

Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital mapping of soil carbon, in: Advances in Agronomy. Elsevier, pp. 1–47. https://doi.org/10.1016/b978-0-12-405942-9.00001-3

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

Mishra, U., Lal, R., Liu, D., Van Meirvenne, M., 2010. Predicting the Spatial Variation of the Soil Organic Carbon Pool at a Regional Scale. Soil Sci. Soc. Am. J. 74, 906–914. https://doi.org/10.2136/sssaj2009.0158

Mishra, U., Gautam, S., Riley, W.J., Hoffman, F.M., 2020. Ensemble Machine Learning Approach Improves Predicted Spatial Variation of Surface Soil Organic Carbon Stocks in Data-Limited Northern Circumpolar Region. Front. Big Data 3. https://doi.org/10.3389/fdata.2020.528441

Moura-Bueno, J.M., Dalmolin, R.S.D., Horst-Heinen, T.Z., Grunwald, S., Ten Caten, A., 2021. Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. Geoderma 393, 114981. https://doi.org/10.1016/j.geoderma.2021.114981

Munoz, M.A., Faz, A., Mermut, A.R., 2015. Soil Carbon Reservoirs at High-Altitude Ecosystems in the Andean Plateau, in: Öztürk, M., Hakeem, K.R., Faridah-Hanum, I., Efe, R. (Eds.), Climate Change Impacts on High-Altitude Ecosystems. Springer International Publishing, Cham, pp. 135–153. https://doi.org/10.1007/978-3-319-12859-7_4

Naimi, S., Ayoubi, S., Zeraatpisheh, M., Dematte, J.A.M., 2021. Ground Observations and Environmental Covariates Integration for Mapping of Soil Salinity: A Machine Learning-Based Approach. Remote Sensing 13, 4825. https://doi.org/10.3390/rs13234825

O'brien, R.M., 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Qual Quant 41, 673–690. https://doi.org/10.1007/s11135-006-9018-6

Ottoy, S., De Vos, B., Sindayihebura, A., Hermy, M., Van Orshoven, J., 2017. Assessing soil organic carbon stocks under current and potential forest cover using digital soil mapping and spatial generalisation. Ecological Indicators 77, 139–150. https://doi.org/10.1016/j.ecolind.2017.02.010

Parastatidis, D., Mitraka, Z., Chrysoulakis, N., Abrams, M., 2017. Online Global Land Surface Temperature Estimation from Landsat. Remote Sensing 9, 1208. https://doi.org/10.3390/rs9121208

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others, 2011. Scikit-learn: Machine learning in python. Journal of machine learning research 12, 2825–2830.

Peng, Y., Chahal, I., Hooker, D.C., Van Eerd, L.L., 2024. Comparison of equivalent soil mass approaches to estimate soil organic carbon stocks under long-term tillage. Soil Tillage Res. 238, 106021. https://doi.org/10.1016/j.still.2024.106021

Pouladi, N., Gholizadeh, A., Khosravi, V., Borůvka, L., 2023. Digital mapping of soil organic carbon using remote sensing data: A systematic review. CATENA 232, 107409. https://doi.org/10.1016/j.catena.2023.107409

Priyatikanto, R., Lu, Y., Dash, J., Sheffield, J., 2023. Improving generalisability and transferability of machine-learning-based maize yield prediction model through domain adaptation. Agric. For. Meteorol. 341, 109652. https://doi.org/10.1016/j.agrformet.2023.109652

Pulgar Vidal, J., 2014. Las ocho regiones naturales del Perú. Terra Bras. https://doi.org/10.4000/terrabrasilis.1027

Pylianidis, C., Kallenberg, M.G.J., Athanasiadis, I.N., 2024. Domain adaptation with transfer learning for pasture digital twins. Environ. Data Sci. 3, e8. https://doi.org/10.1017/eds.2024.6

R Core Team, 2023. R: a language and environment for statistical computing (manual). R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Rasel, S.M.M., Groen, T.A., Hussin, Y.A., Diti, I.J., 2017. Proxies for soil organic carbon derived from remote sensing. International Journal of Applied Earth Observation and Geoinformation 59, 157–166. https://doi.org/10.1016/j.jag.2017.03.004

Rikimaru, A., Roy, P.S., Miyatake, S., 2002. Tropical forest cover density mapping. Tropical Ecology 43, 39–47.

Román-Sánchez, A., Vanwalleghem, T., Peña, A., Laguna, A., Giráldez, J.V., 2018. Controls on soil carbon storage from topography and vegetation in a rocky, semi-arid landscapes. Geoderma 311, 159–166. https://doi.org/10.1016/j.geoderma.2016.10.013

Roudier, P., 2012. clhs: Conditioned Latin Hypercube Sampling. https://doi.org/10.32614/CRAN.package.clhs

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the Great Plains with ERTS, in: NASA. Goddard Space Flight Center 3d ERTS-1 Symp., Vol. 1, Sect. A.

Santos, F., Graw, V., Bonilla, S., 2019. A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. PLOS ONE 14, e0226224. https://doi.org/10.1371/journal.pone.0226224

Sartika, E., Suryani, A., 2020. Comparison of geographically weighted regression analysis and global regression on modeling the unemployment rate in west java, in: Proceedings of the International Seminar of Science and Applied Technology (ISSAT 2020). Atlantis Press, pp. 472–478. https://doi.org/10.2991/aer.k.201221.078

Schonlau, M., Zou, R.Y., 2020. The random forest algorithm for statistical learning. The Stata Journal 20, 3–29. https://doi.org/10.1177/1536867X20909688

Schwanghart, W., Jarmer, T., 2011. Linking spatial patterns of soil organic carbon to topography — A case study from south-eastern Spain. Geomorphology 126, 252–263. https://doi.org/10.1016/j.geomorph.2010.11.008

Song, M., Huang, Z., Chen, C., Li, X., Mao, F., Huang, L., Zhao, Y., Lv, L., Yu, J., Du, H., 2024. Multi-scale geographically weighted regression estimation of carbon storage on coniferous forests considering residual distribution using remote sensing data. Ecol. Indic. 166, 112495. https://doi.org/10.1016/j.ecolind.2024.112495

Sun, Y., Ao, Z., Jia, W., Chen, Y., Xu, K., 2021. A geographically weighted deep neural network model for research on the spatial distribution of the down dead wood volume in Liangshui National Nature Reserve (China). IForest - Biogeosciences For. 14, 353–361. https://doi.org/10.3832/ifor3705-014

Szakács, G.G.J., Cerri, C.C., Herpin, U., Bernoux, M., 2011. Assessing soil carbon stocks under pastures through orbital remote sensing. Sci. agric. (Piracicaba, Braz.) 68, 574–581. https://doi.org/10.1590/S0103-90162011000500010

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020. Improving the Spatial Prediction of Soil Organic Carbon Content in Two Contrasting Climatic Regions by Stacking Machine Learning Models and Rescanning Covariate Space. Remote Sensing 12, 1095. https://doi.org/10.3390/rs12071095

Tahmouresi, M.S., Niksokhan, M.H., Ehsani, A.H., 2024. Enhancing spatial resolution of satellite soil moisture data through stacking ensemble learning techniques. Scientific Reports 14, 25454. https://doi.org/10.1038/s41598-024-77050-0

Thrift, N.J., Kitchin, R., 2009. International encyclopedia of human geography. Elsevier, Amterdam Boston.

Triantakonstantis, D., Karakostas, A., 2025. Soil Organic Carbon Monitoring and Modelling via Machine Learning Methods Using Soil and Remote Sensing Data. Agriculture 15, 910. https://doi.org/10.3390/agriculture15090910

Tyralis, H., Papacharalampous, G., Langousis, A., 2019. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. Water 11, 910. https://doi.org/10.3390/w11050910

Vallejos-Torres, G., Gaona-Jimenez, N., Pichis-García, R., Ordoñez, L., García-Gonzales, P., Quinteros, A., Lozano, A., Saavedra-Ramírez, J., Tuesta-Hidalgo, J.C., Reategui, K., Macedo-Córdova, W., Baselly-Villanueva, J.R., Marín, C., 2024. Carbon reserves in coffee agroforestry in the Peruvian Amazon. Front. Plant Sci. 15. https://doi.org/10.3389/fpls.2024.1410418

Wang, D., Li, X., Zou, D., Wu, T., Xu, H., Hu, G., Li, R., Ding, Y., Zhao, L., Li, W., Wu, X., 2020. Modeling soil organic carbon spatial distribution for a complex terrain based on geographically weighted regression in the eastern Qinghai-Tibetan Plateau. CATENA 187, 104399. https://doi.org/10.1016/j.catena.2019.104399

Wang, K., Zhang, C., Li, W., 2012. Comparison of Geographically Weighted Regression and Regression Kriging for Estimating the Spatial Distribution of Soil Organic Matter. GIScience & Remote Sensing 49, 915–932. https://doi.org/10.2747/1548-1603.49.6.915

Wang, L., Abramowitz, G., Wang, Y.-P., Pitman, A., Viscarra Rossel, R.A., 2024. An ensemble estimate of Australian soil organic carbon using machine learning and process-based modelling. SOIL 10, 619–636. https://doi.org/10.5194/soil-10-619-2024

Wei, X., Shao, M., Gale, W., Li, L., 2014. Global pattern of soil carbon losses due to the conversion of forests to agricultural land. Scientific Reports 4, 4062. https://doi.org/10.1038/srep04062

Weku, W., Pramoedyo, H., Widodo, A., Fitriani, R., 2022. Optimal Bandwidth for Geographically Weighted Regression to Model the Spatial Dependency of Land Prices in Manado, North Sulawesi Province, Indonesia. Geogr. Environ. Sustain. 15, 84–90. https://doi.org/10.24057/2071-9388-2019-154

Wheeler, D.C., 2007. Diagnostic Tools and a Remedial Method for Collinearity in Geographically Weighted Regression. Environ Plan A 39, 2464–2481. https://doi.org/10.1068/a38325

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.-J., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. Geoderma 333, 149–162. https://doi.org/10.1016/j.geoderma.2018.07.026

Wu, S., Jia, W., Wang, F., Sun, Y., Zhao, H., Lu, S., 2023. Estimation of Above-Ground Carbon Storage and Light Saturation Value in Northeastern China's Natural Forests Using Different Spatial Regression Models. Forests 14. https://doi.org/10.3390/f14101970

Yang, W., Deng, M., Tang, J., Luo, L., 2022. Geographically weighted regression with the integration of machine learning for spatial prediction. J. Geogr. Syst. 25, 213–236. https://doi.org/10.1007/s10109-022-00387-5

Yigini, Y., Panagos, P., 2016. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. Science of The Total Environment 557–558, 838–850. https://doi.org/10.1016/j.scitotenv.2016.03.085

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., Arino, O., 2022. ESA WorldCover 10 m 2021 v200. https://doi.org/10.5281/ZENODO.7254221

Zeng, C., Yang, L., Zhu, A.-X., Rossiter, D.G., Liu, Jing, Liu, Junzhi, Qin, C., Wang, D., 2016. Mapping soil organic matter concentration at different scales using a mixed geographically weighted regression method. Geoderma 281, 69–82. https://doi.org/10.1016/j.geoderma.2016.06.033

Zeng, Y., Shi, T., Liu, Q., Yang, C., Zhang, Z., Wang, R., 2024. A geographically weighted neural network model for digital soil mapping of heavy metal copper in coastal cities. J. Hazard. Mater. 480, 136285. https://doi.org/10.1016/j.jhazmat.2024.136285

Zhang, C., Tang, Y., Xu, X., Kiely, G., 2011. Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Applied Geochemistry 26, 1239–1248. https://doi.org/10.1016/j.apgeochem.2011.04.014

Zhang, W., Ji, J., Li, B., Deng, X., Xu, M., 2025. Integrating Genetic Algorithm and Geographically Weighted Approaches into Machine Learning Improves Soil pH Prediction in China. Remote Sensing 17, 1086. https://doi.org/10.3390/rs17061086

Zhang, W., Wan, H., Zhou, M., Wu, W., Liu, H., 2022. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. Ecological Indicators 143, 109420. https://doi.org/10.1016/j.ecolind.2022.109420

Zhang, Z., Jung, C., 2021. GBDT-MO: Gradient-Boosted Decision Trees for Multiple Outputs. IEEE Trans. Neural Netw. Learning Syst. 32, 3156–3167. https://doi.org/10.1109/TNNLS.2020.3009776

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: