Discover Soil

Research

Predicting bulk density in Brazilian soils for carbon stocks calculation: a comparative study of multiple linear regression and Random Forest models using continuous and categorical variables

Wharley Pereira dos Santos[1] · Carlos Manoel Pedro Vaz[1] · Ladislau Martin-Neto[1] · Adriano Anselmi[2] · Javier Tomasella[3] · Falberni de Souza Costa[4] · Jackson Adriano Albuquerque[5] · Quirijn de Jong van Lier[6] · Rafael Galbieri[7] · Fabiano José Perina[8]

Received: 18 November 2024 / Accepted: 17 January 2025
Published online: 03 February 2025

Abstract
Soil bulk density (ρ_b) is vital for assessing soil organic carbon (SOC) and nutrient stocks, as well as for modeling soil processes. Although ρ_b can be measured using traditional methods, these are labor-intensive and time-consuming. Consequently, there is a growing interest in developing pedotransfer functions (PTFs) to predict ρ_b based on more easily measured soil properties. The vast Brazilian territory, with different climates and biomes, presents a challenge for development of national-scale PTFs, particularly for ρ_b. In this study, a comprehensive dataset was compiled from various sources to develop ρ_b PTFs across Brazil. The dataset includes particle size distribution (PSD), SOC, ρ_b, sampling depth, soil order, land use, and geographic coordinates, allowing for the incorporation of additional numerical and categorical variables. Rigorous data preprocessing ensured quality and reliability. PTFs were developed using multiple linear regression (MLR) and Random Forest (RF) models. Model accuracy was evaluated using mean absolute error, bias, root mean square error (RMSE), and coefficient of determination. Both MLR and RF models accurately predicted ρ_b, with log-transformed PSD and SOC emerging as key predictors. The RF model slightly outperformed the MLR model (RMSE = 0.12 vs. 0.13 g cm^{-3}) on the test dataset, underscoring the importance of environmental and categorical variables in predicting ρ_b. The developed PTFs, along with other PTFs for Brazil, were applied to estimate SOC stocks across different biomes and land uses. Best estimations were obtained with the RF model, with an R^2 of 0.97, emphasizing the value of categorical variables in improving SOC stock estimations.

Keywords PTFs · Soil formation · Clay content · SOC · Multiple linear regression · Random forest

✉ Wharley Pereira dos Santos, wharleypereira@gmail.com; Carlos Manoel Pedro Vaz, carlos.vaz@embrapa.br; Ladislau Martin-Neto, ladislau.martin@embrapa.br; Adriano Anselmi, adriano.anselmi1@bayer.com; Javier Tomasella, javier.tomasella@inpe.br; Falberni de Souza Costa, falberni.costa@embrapa.br; Jackson Adriano Albuquerque, jackson.irai@gmail.com; Quirijn de Jong van Lier, qdjvlier@usp.br; Rafael Galbieri, rafaelgalbieri@imamt.org.br; Fabiano José Perina, fabiano.perina@embrapa.br | [1]Embrapa Instrumentation, São Carlos, SP 13560-970, Brazil. [2]Bayer Crop Science, São Paulo, SP 04761-000, Brazil. [3]National Institute for Space Research (INPE), Cachoeira Paulista, SP 12630-000, Brazil. [4]Embrapa Acre, Rio Branco, AC 69900-970, Brazil. [5]Santa Catarina State University (UDESC), Lages, SC CEP 88520-000, Brazil. [6]DVECO/CENA, University of São Paulo, Piracicaba, SP 13400-970, Brazil. [7]Mato Grosso Cotton Institute (IMAmt), Primavera do Leste, MT 78850-000, Brazil. [8]Embrapa Cotton, Campina Grande, PB 58428-095, Brazil.

1 Introduction

Mitigating and adapting to climate change is an urgent challenge that requires immediate action for timely and effective solutions. Global surface temperature has risen, especially over land, largely due to the historical increase in CO_2 emissions from fossil fuel combustion, industrial processes, and land use changes [1]. Climate warming has a significant impact on soil organic carbon (SOC) stocks by altering the soil carbon dynamics. To reduce greenhouse gas (GHG) emissions and mitigate climate change, cropland soils are considered a significant potential carbon sink [2]. This potential can be enhanced through the use of cover crops and the adoption of crop rotation and diversification strategies [3–5].

Regarding SOC stocks, soil bulk density (ρ_b) is an essential soil physical property, as it allows the conversion of SOC mass concentration (g/kg) to SOC stock per unit of area (kg/ha) for a defined soil layer. Bulk density also significantly influences other soil properties, such as porosity, water retention, and hydraulic conductivity. These factors, in turn, affect plant root development and ultimately crop yield, as they are closely linked to soil compaction [6] and soil water holding capacity [7]. Soil bulk density shows considerable spatial variations across different landscapes and depths, requiring its assessment to account for spatial variability and soil layer depth [8]. This variability spans a wide range of soil types, environments, climatic regions, and land use [9].

Changes in soil physical properties under agricultural management are complex and influenced by the interaction of environmental and agronomic factors, such as cropping intensity, soil texture, and climate conditions [10]. Therefore, when managing SOC, it is essential to consider soil type and environmental factors in addition to land use and management practices, as these factors significantly contribute to SOC stock variations [11]. For instance, higher soil bulk density increases the need for organic matter inputs to promote long-term carbon accumulation. This is particularly important in compacted soils or sandy soils of tropical regions, where warm, humid climates with high rainfall and increased soil aeration stimulate microbial activity. This, in turn, accelerates the mineralization of organic matter, leading to increased CO_2 release from the soil. Addressing this issue requires a strategic approach that prioritizes enhancing plant functional diversity. Implementing cover crops in agricultural management systems can significantly increase soil carbon stocks and improve organic matter stability in these soils [12].

Mathematical functions or algorithms that link measured soil data to specific soil properties are essential for understanding and relating soil processes and functions. In soil science, these are known as pedotransfer function (PTFs), a concept first introduced by [13]. Pedotransfer functions embody the knowledge rules in a soil inference system, enabling the estimation of soil properties, such as ρ_b, through PTFs tailored to specific geomorphic regions or soil types [14]. Despite the importance of soil bulk density for evaluating and modeling soil processes and functions, comprehensive global datasets on ρ_b remain limited. This is because ρ_b measurements are labor-intensive, time-consuming, and expensive [8]. Consequently, several PTFs have been developed to estimate ρ_b using more easily measured and readily available soil parameters, such as clay, silt, and sand content (texture), soil organic matter content, and other soil attributes [15–20].

Brazil comprises a wide range of climates, including equatorial, humid tropical, subtropical, and semi-arid zones. This diversity poses a challenge to develop PTFs for the entire country, especially for bulk density, due to its large variation for different soils, biomes and land use [16, 18, 21–25]. Thus, developing PTFs for Brazil requires careful consideration of the wide range of soil properties, requiring strategic groupings that effectively address the country's pedogenetic and environmental variability [26]. It is also worth to highlight the unique physical-hydric properties of highly weathered tropical soils, which differ significantly from those in temperate climates [27]. This challenge is similarly relevant to other extensive regions with substantial variability in soil types and biomes [28–32].

The development of PTFs for the prediction of bulk density has advanced significantly with the integration of modern technologies, particularly through the widespread application of machine learning algorithms. However, traditional methods, such as stepwise multiple linear regression, remain widely used [19, 21, 33]. Indeed, the pursuit of improved PTF performance has driven the creation of more complex models for ρ_b prediction. In linear models, the assumptions of multivariate normality and homoscedasticity often require transformations of the explanatory variables and/or the ρ_b response variable to improve model accuracy [34–36].

The application of advanced modelling techniques, such as artificial neural network (ANN) and regression tree (RT)-based models, does not always result in more accurate bulk density predictions [37]. Additionally, these methods are often perceived as "black boxes" [38]. To address these challenges, efforts have been made to use approaches like the Cubist regression-tree method, which adapts the PTFs to local contexts by incorporating soil morphology,

land use, and quantitative soil properties as predictors [39]. In general, machine learning-based PTFs outperform conventional PTFs, making them promising candidates for further investigation. They offer an advantage over purely empirical methods, such as linear or nonlinear regression models, which often require determining predictors a priori. Additionally, the relationship between soil properties and predictors can vary across different parts of the database, introducing further complexity [40]. Recent studies using Random Forest have also shown highly promising results, delivering accurate predictions for soil properties [41–44].

Giving the growing need for detailed soil data on a global scale to assess the impacts of global warming, and considering the distinct behavior of tropical Oxisols compared to clay-rich soils in temperate climates [27], developing more accurate approaches for estimating soil physics parameters using Brazilian database is crucial for several applications in a global context.

The present study aims to develop pedotransfer functions for estimating ρ_b using stepwise multiple linear regression applied to soil properties from available databases, which are considered suitable predictors (PTF-1). Additionally, it seeks to enhance bulk density prediction by incorporating categorical variables, such as land uses, regions, and soil pedogenetic classifications within the stepwise multiple linear regression model (PTF-2). The study also explores the use of the Random Forest technique (PTF-3), which incorporates additional predictors such as climatic, topographic, and vegetation indices. These PTFs were designed to improve ρ_b prediction for all soil types and depths across various landscapes and geopolitical regions in Brazil. The approaches in PTF-2 and PTF-3 strategically group data to analyze the variability of ρ_b under different soil-landscape-land use conditions. The developed PTFs were compared with other pedotransfer functions for Brazil available in the literature. Their applicability was further assessed to estimate soil organic carbon (SOC) stocks using the predicted soil bulk densities across various Brazilian biomes and land uses.

2 Materials and methods

2.1 Dataset compilation

A comprehensive data acquisition effort was conducted to compile information on the physical and chemical soil properties across Brazil. This effort involved consulting a variety of sources, including open-source soil databases, published papers applied to Brazil, dissertations, theses, and other relevant documents. Additionally, well-structured and reliable datasets from various research project reports were incorporated. The selected data included soil physical properties (bulk density, sand, silt and clay contents) and organic carbon (SOC) content across various soil depths, soil classes and textures encompassing all Brazilian states and inherently covering the six Brazilian biomes (Amazon Rainforest, Atlantic Forest, Cerrado, Caatinga, Pampas, and Pantanal) as illustrated in Fig. 1. Supplementary information, such as geographic coordinates, soil class, and land use, was also collected to ensure comprehensive coverage and facilitate the creation of categorical variables. Literature searches were conducted across multiple sources, including peer-reviewed papers, theses and dissertations from Brazil. These searches were performed using various platforms, such as citation databases (e.g. Web of Science) and search engines like Google Scholar, among others.

The database comprises data from the PRO Carbono project, a collaborative research initiative between the Brazilian Agricultural Research Corporation (EMBRAPA) and Bayer Company. Additionally, it integrates data from the Hydrophysical Database for Brazilian Soils-HYBRAS [45]; a dataset provided by a project involving EMBRAPA, the Cotton Producer Association of Bahia State (ABAPA), Bahia Foundation (Research and Development Support Foundation of Western Bahia), and the Brazilian Cotton Institute (IBA) [46]; a dataset from a project between EMBRAPA, Mato Grosso Cotton Institute (IMAmt), Cotton Producer Association of Mato Grosso State (AMPA), Seed Producer Association of Mato Grosso State (Aprosmat), and IBA [47]; the RadamBrasil soil survey for the Brazilian Amazon region [48]; and other data sourced from the aforementioned literature review. This extensive compilation from multiple sources enhances the database's comprehensiveness compared to previous similar research efforts.

The continuous numerical variables used included sand content (wt%; particle-size between 2000 and 50 µm), silt (wt%; particle-size between 50 and 2 µm), clay (wt%; particle-size lower than 2 µm), which were determined using either the pipette or densimeter method [49]. Soil Organic Carbon (SOC) content [wt%] was determined either by dry combustion [50] or the Walkley–Black chromic acid wet oxidation method [51]. Soil bulk density (g cm^{-3}) was determined by collecting undisturbed samples in steel cylinders (5 cm in diameter and 5 cm in height) from trenches. These samples were submitted to weight determination on a balance in the laboratory after being oven-dried at 105 °C for 24 h.

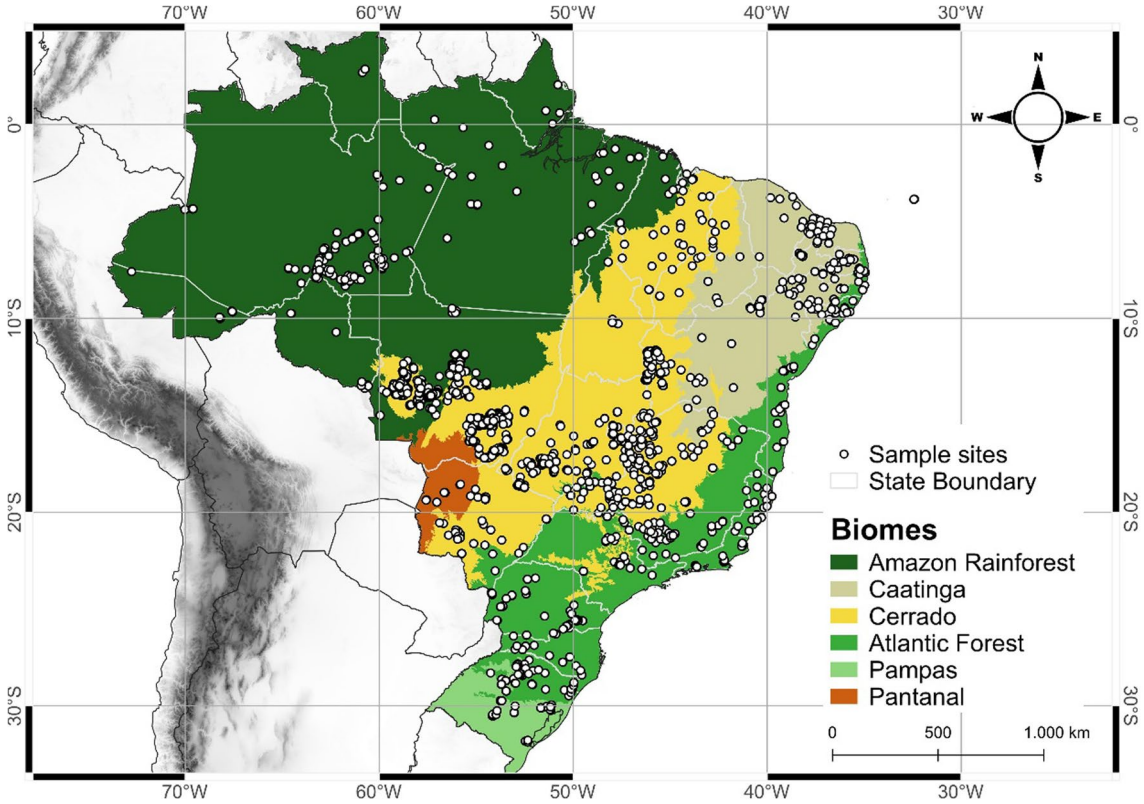

Fig. 1 Geographic distribution of sampling points for measured soil bulk density across the Brazilian territory

2.2 Data preprocessing

The original database contained 26,855 entries, representing all major Brazilian soil classes and soil textures, various soil layers, and different land uses, including native vegetation and croplands. To maintain data integrity, entries with incomplete or duplicated data were excluded. Additionally, some samples in this study may contain significant recording errors, particularly from legacy soil survey databases. These errors could result in outliers that do not accurately represent the underlying correlation structure among the measured soil properties, thereby potentially biasing the analysis. To prevent this, a quality control check was performed on the soil properties, identifying and excluding values outside the acceptable range, resulting in 16,505 samples available for analysis. The textural distribution of the soil dataset used in this study is illustrated in Fig. 2, encompassing the twelve soil texture classes defined by the United States Department of Agriculture (USDA) classification system.

2.3 Development, training and testing of the PTFs

The PTFs were developed using two predictive methods: Multiple Linear Regression (MLR) and Random Forest model (RF) models. Both methods have been widely applied in the development of PTFs for soil bulk density prediction [15, 29, 32, 52–54].

Different categories of model predictors were built using a selection of variables and covariates that have a relevant pedogenetic association with soil bulk density. These categories include soil properties and environmental factors related to terrain attributes, climate, geology, geomorphology, organisms, land use/land cover, and management practices. Terrain attributes primarily consisted of elevation, slope, and aspect, derived from USGS/NASA's Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM). The SRTM 90 m data, provided by CGIAR-CSI GeoPortal as SRTM 90 m data [55], was accessed through the Google Earth Engine platform [56]. The geomorphic domains were derived from datasets originally compiled during the RadamBrasil geological and geomorphological

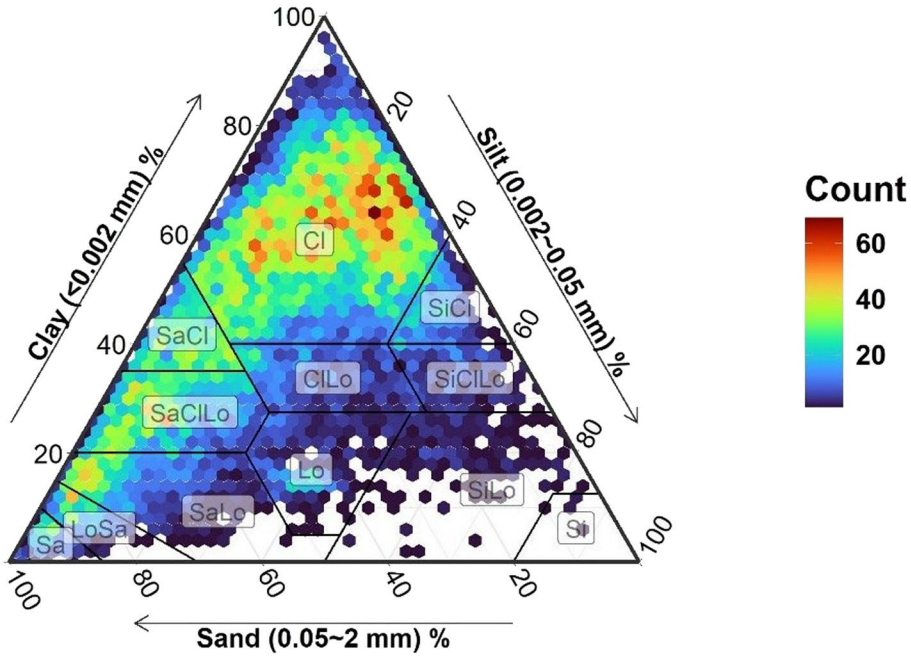

Fig. 2 Particle size distribution of soil samples used to develop and test the models, plotted on a texture triangle and categorized according to the United States Department of Agriculture (USDA) classification system. The legend indicates the number of soil samples within each hexagonal bin, with colors ranging from red to blue to represent high and low data density. *Cl* clay, *SiCl* silty clay, *SaCl* sandy clay, *ClLo* clay loam, *SiClLo* silty clay loam, *SaClLo* sandy clay loam, *Lo* loam, *SiLo* silty loam, *SaLo* sandy loam, *Si* silt, *LoSa* loamy sand, *Sa* sand

surveys, updated by the Brazilian Institute of Geography and Statistics – IBGE [57]. Climate variables included average temperature (°C) and precipitation (mm), sourced from WorldClim version 2.1 climate data covering the period from 1970 to 2000 [58], downloaded through the geodata package [59] in R (version 4.3.3). Additionally, Köppen's climate classification, as mapped by [60] for Brazil's climatic regions, was used as a climate-related predictor. Vegetation-related features were primarily represented by the normalized difference vegetation index (NDVI) derived from Moderate Resolution Imaging Spectroradiometer (MODIS) data using the MOD13Q1 product, which provides 16-day composite NDVI images [61]. The data were acquired and preprocessed on the Google Earth Engine platform with a spatial resolution of 250 m, and averaged over the period from 2001 to 2023. The conceptual framework used in this study was based on the SCORPAN method, as described by [62], which extends [63] factorial equation. This approach characterizes soil attributes by integrating climate, organisms, relief, parent material, time, and spatial factors, along with intrinsic soil properties and their relationships within the soil-landscape.

2.3.1 Multiple linear regression

Multiple Linear Regression (MLR) is a widely used and straightforward technique for developing analytical functions to predict soil physical and hydraulic properties. Initially, a stepwise procedure was employed to introduce or remove predictor variables, with the goal of optimizing the model's goodness-of-fit, measured by the coefficient of determination (R^2). This process was carried out using Stepwise Multiple Linear Regression (SMLR). Multiple linear regression modeling, using stepwise methods such as forward selection, backward elimination, and bi-directional procedures, is highly effective for testing the statistical significance of predictors and addressing multicollinearity issues, which can result from strong correlations between variables. The Akaike Information Criterion (AIC) is employed to identify the optimal set of predictors for predicting bulk density. During the stepwise regression process, AIC is computed at each step to assess model improvement when a new predictor is added. The model with the greatest reduction in AIC is considered superior. The process ends either when no further improvement is achieved by adding more predictors or when all potential predictors have been included in the model. Consequently, SMLR streamlines the selection of variables, identifying those most representative of bulk density. Tolerance and Variance Inflation Factor (VIF) were employed to assess multicollinearity among the predictor variables.

$$\mathrm{VIF} = \frac{1}{1 - R^2} = \frac{1}{\mathrm{Tolerance}} \quad (1)$$

The tolerance is the reciprocal of the VIF. A lower tolerance value indicates a higher probability of multicollinearity among the variables. A VIF value of 1 signifies that the independent variables are uncorrelated. VIF values between 1 and 5 suggest moderate correlation, while values between 5 and 10 indicate a significant level of correlation that could pose challenges. A VIF greater than 10 suggests pronounced multicollinearity among predictors, which can lead to poorly estimated regression coefficients [64]. In models incorporating categorical predictors, a variant known as the Generalized VIF (GVIF) was employed to ensure consistent multicollinearity assessments [65].

The MLR model can be expressed as follows:

$$f(x_0)=\beta_0+\sum_{i=1}^{n}\beta_i{}^{*}p_i+\varepsilon \tag{2}$$

where $f(x_0)$ represents the estimated dependent variable (soil bulk density), β_0 is the y-intercept (constant term), β_l corresponds to the regression coefficients, p_l denotes the independent explanatory variables, and ε is the model's error term or residuals.

Two sub-approaches were employed in the development of the multiple linear regression models. The first focused on ratios and log-transformations applied to soil particle-size distribution (PSD), including clay, silt, and sand contents, as well as the log-transformation of SOC. Additionally, it included the inverse of the clay ratio [clay_ratio = % clay/(% sand + % silt)] [66] and the silt/clay ratio [67] (model 1). The former ratio reflects the relative erodibility of soils, while the latter serves as a textural weathering index representing intrinsic pedogenetic processes. The second MLR approach involved the incorporation of categorical variables to improve the prediction of bulk density (model 2).

2.3.1.1 Stepwise multiple linear regression model 1 The first model for ρ_b estimation, referred to as PTF-1, relies exclusively on continuous numerical variables as demonstrated by the following relationships.

$$\rho_{b\,PTF-1} = \beta_0 + \beta_1 alr_1 + \beta_2 alr_2 + \beta_3 \ln(SOC+1) + \beta_4 clay\_ratio^{-1} + \beta_5(silt/clay) + \beta_6 Depth + \varepsilon \tag{3}$$

$$alr_1 = \ln\left(\frac{sand}{clay}\right); alr_2 = \ln\left(\frac{silt}{clay}\right) \tag{4}$$

In Eq. 3, alr denotes the additive log-ratio transformation applied to the PSD, as detailed in Eq. 4. This transformation is essential for compositional data, such as PSD or soil texture, to address potential biases arising from inherent negative correlations between sand, silt, and clay contents within the ternary compositional space [68]. The compositional data are projected from the simplex space S^d into the real space R^D through the application of log-ratio transformations, as described by [69]. The first and second vector components of the transformed PSD (Eq. 4) serve as predictors in the regression model. In Eq. 3, β_0, β_1, β_2, β_3, β_4, β_5 and β6 represent the estimated regression coefficients, while ε denotes the residuals to PTF-1.

After evaluating the goodness-of-fit and checking the model assumptions for SMLR, observations exerting significant influence on the estimated regression coefficients were identified within the dataset. These observations were subsequently considered genuine outliers. To mitigate the impact of these outliers, cases with excessive influence were removed, followed by a re-evaluation of the model´s performance. Various techniques were employed to identify and address potential weaknesses in the model, leveraging statistical methods such as Cook's distance, the Hat Diagonal statistic, Covariance Ratio (CovRatio) statistic, DFFITS, and DFBETAS [70]. Subsequently, 70% of the samples (10,677) were allocated to the training or calibration data subset, and the remaining 30% (4,573) were designated as the test data subset. This allocation employed random sampling within stratified splitting based on soil bulk density outcomes to develop and evaluate the PTFs. These subsets were used to determine the performance of the multiple regression models. The model parameters were optimized using k-fold cross-validation, with k set to 10 to prevent overfitting.

2.3.1.2 Stepwise multiple linear regression model 2 To investigate the potential benefits of incorporating categorical variables for enhancing bulk density prediction, a second multiple linear regression model, referred to as PTF-2, was developed. This model combines both numerical and categorical predictor variables, as outlined in Eq. 5.

$$\begin{aligned}\rho_{b\,PTF-2} = {} & \beta_0 + \beta_1 alr_1 + \beta_2 alr_2 + \beta_3 \ln(SOC+1) + \beta_4 clay\_ratio^{-1} + \beta_5(silt/clay) \\ & + \beta_6 Depth + Biome + LULC + Order + Texture + \varepsilon\end{aligned} \tag{5}$$

The PTF-2 model incorporates additional predictor variables beyond those derived from PSD transformations, SOC, and depth. These additional variables include categorical predictors that represent k-level factors, such as landscape patterns and geographic explanatory variables. Specifically, the data were grouped into the six Brazil's biomes: Caatinga, Amazon Rainforest, Pampas, Atlantic Forest, Cerrado, and Pantanal. Moreover, land use/land cover (LULC) categories were grouped into agricultural land (AL), native vegetation (NV), and other land uses (OL). The model also includes soil orders according to the Brazilian Soil Classification System (SiBCS) [71] (refer to Table 1). Additionally, a generic soil texture classification was included, categorizing soils into coarse-grained, medium-grained, and fine-grained soils, corresponding to sandy, loamy, and clayey soils, respectively, based on the USDA system [72].

To encode categorical information into a numerical format suitable for linear regression analysis, a coding scheme was employed. This approach assigns a unique numerical value to each category within a variable. For example, in binary encoding using dummy variables (0 for absence and 1 for presence), the soil order category 'Latosol' might be assigned a value of 1, while all other soil orders would be coded as 0 (indicating 'not Latosol'). This procedure creates a new binary dummy variable that serves as a predictor in the linear regression model. For a categorical predictor with k levels, typically k-1 dummy variables are included in the model to avoid redundancy. This approach prevents perfect collinearity, which would occur if all k dummy variables were used, leading to a situation where each observation sums to 1 as every observation belongs to exactly one category. As a result, one dummy variable is consistently omitted, known as the reference level. Furthermore, the development of the soil bulk density model follows the methodology outlined for Regression model 1. A total of 14,336 samples were used, with 70% allocated to training data and the remaining 30% reserved for testing, following the same procedure described in the previous section. The pre-processing steps, data splitting, and MLR modelling were performed in R version 4.3.3 [73], using the *tidyverse* collection of packages [74], the *metan* package [75], the *soiltexture* package [76], and the *caret* package [77].

2.3.2 Random forest

In addition to the linear regression models (PTF-1 and PTF-2), an alternative approach utilizing advanced non-parametric techniques was explored for PTF development. Specifically, the Random Forest (RF) algorithm was employed to develop what is referred to as PTF-3. The Random Forest algorithm [78], was selected for its effectiveness in handling both numerical and categorical predictor variables within the dataset. Additionally, RF offers advantages in addressing potential imbalances in class distributions, capturing complex interactions between predictors, and providing insights into the importance of variables for predicting bulk density.

Table 1 Soil Order/suborder according to the Brazilian soil classification (SiBCS), with partial equivalence to the international soil classification system (WRB/FAO), for the soils examined in this study

SiBCS	Code	N[a]	WRB/FAO[b]
Argisols	P	2829	Acrisols; Lixisols; Alisols
Cambisols	C	1776	Cambisols
Chernosols	M	100	Kastanozems; Phaeozems; Chernozems (some)
Spodosols	E	124	Podzols
Gleysols	G	332	Gleysols; Stagnosols (some); Solonchaks
Latosols	L	15964	Ferralsols
Luvisols	T	208	Luvisols
Nitosols	N	1744	Nitisols; Lixisols; Alisols
Neosols			–
Fluvic Neosols	RY	181	Fluvisols
Litholic Neosols	RL	791	Leptsols
Quartzarenic Neosols	RQ	1349	Arenosols
Regolithic Neosols	RR	144	Regosols
Organosols	O	56	Histosols
Planosols	S	382	Planosols; Solonetz
Plinthosols	F	317	Plinthosols
Vertisols	V	122	Vertisols

[a]number of samples for each soil order/suborder

[b]World Reference Base for Soil Resources

The RF regression algorithm is an advanced ensemble-learning approach that integrates an extensive collection of regression trees. While it is designed to prevent overfitting, ensuring proper configuration of the input parameters in the RF model is essential to ensure optimal model performance. The algorithm generates robust predictors through bagging (bootstrap aggregation) and random input selection. Bagging creates multiple regression trees from random bootstrap samples of the dataset, while random input selection chooses subsets of variables for split decisions at each tree node. In this study, each Random Forest comprised 1,000 trees to ensure accuracy, with trees grown to maximum depths without pruning to minimize bias. One-third of the input variables were randomly sampled for split decisions, a strategy considered effective. Two key parameters in RF are the number of trees (ntree; default value of 500 trees) and the number of input variables sampled for splitting at each node (mtry; default value is one-third of the total number of variables). Optimizing these parameters significantly impacts the model's explained variance and error rate, with performance assessed through the prediction error using out-of-bag (OOB) cross-validation. In this study, the RF model was built using predictor variables from model 2 (Eq. 5). Additionally, the environmental and topography covariates, which represent soil-forming factors, as well as vegetation index, were incorporated as extra predictors beyond those used in PTF-1 and PTF-2 models. For the RF modeling, 16,087 samples were used, with 70% allocated to training and 30% used for testing, following the same procedure previously described. Entries with categorical predictors that had missing factor levels were excluded. The significance of predictor variables in the model was evaluated using the Variable Importance Metric (VIMP) as proposed by [78] and further developed by [79]. VIMP quantifies the influence of each predictor variable by comparing the performance of the original RF model with a version in which the variable is randomly permuted, a technique also described by [78]. The *RandomForestSRC* package [80] in R was used to develop the random forests. This package is effective for handling large datasets due to its parallel processing capabilities, which significantly enhance computational efficiency.

2.4 Models performance evaluation metrics

The performance of the PTFs was assessed using the mean absolute error (MAE), bias, and root mean square error (RMSE). Additionally, the coefficient of determination (R^2) was employed as a performance metric, derived from linear regression between the predicted and the observed values. All statistical computations were performed using the R programming language [73].

2.5 Applying the PTFs to estimate SOC stocks

The three PTFs developed in this study, along with other previously published PTFs for Brazilian soils [15, 18, 48], termed PTF–Bnt, PTF–Brn, and PTF–T&H, as shown in the Eqs. 6, 7, and 8 respectively, were employed to estimate SOC stocks.

$$\text{PTF - Bnt: } \rho_b\ \left[\text{g cm}^{-3}\right] = 1.5688\ -\ 0.0005\ \times \text{Clay}\ \left[\text{g kg}^{-1}\right]\ -\ 0.009\ \times \text{SOC}\ \left[\text{g kg}^{-1}\right]; R^2\ =\ 0.63 \tag{6}$$

$$\text{PTF - Brn: } \rho_b\ \left[\text{g cm}^{-3}\right] = 1.398\ -\ 0.0047\ \times \text{Clay}\ \left[\text{dag kg}^{-1}\right]\ -\ 0.042\ \times \text{SOC}\ \left[\text{dag kg}^{-1}\right]; R^2\ =\ 0.50 \tag{7}$$

$$\text{PTF - T\&H: } \rho_b\ \left[\text{g cm}^{-3}\right] = 1.578\ -\ 0.054\ \times \text{SOC}\ \left[\text{dag kg}^{-1}\right]\ -\ 0.006\ \times \text{Silt}\ \left[\text{dag kg}^{-1}\right]\ -\ 0.004\ \times \text{Clay}\ \left[\text{dag kg}^{-1}\right]; R^2\ =\ 0.60 \tag{8}$$

The estimations were based on holdout sampling of soil profile data from the dataset used to develop the PTFs, which represents the common dataset across the three PTFs, excluding points without profile sampling. Thus, 2,100 complete soil profiles with their corresponding geographical coordinates were selected from the total data set to obtain the carbon stocks for each soil profile. SOC stocks were calculated using Eq. 9.

$$\text{SOC stock}\ \left[\text{Mg C ha}^{-1}\right] = \rho_b\ \left[\text{g cm}^{-3}\right]\ \times \text{SOC}\ \left[\text{dag kg}^{-1}\right]\ \times \text{Depth}\ \ [\text{cm}] \tag{9}$$

Estimated SOC stocks were calculated using bulk density values predicted by the PTFs, combined with measured SOC. These estimates were then compared to measured SOC stocks determined using both measured ρ_b and SOC.

Due to the diverse origins of the dataset sources, soil profile depths varied widely, ranging from 0.2 m to more than 1 m. These variations significantly impact SOC stocks, as they are dependent on profile depth. Initial comparisons between measured and estimated SOC stocks were made irrespective of profile depth, resulting in considerable variability in SOC

stocks as a result of the differing depths. To improve the comparison of SOC stocks across different land uses and biomes, soil profile depth was standardized to 30 cm (i.e. 0–30 cm topsoil layer at each site) using the method developed by [81], as applied by [82] and [83] (Eqs. 10 and 11).

$$Y_d = 1 - \beta^d \tag{10}$$

$$SOC_{30} = \frac{Y_d}{Y_{d0}} \times SOC_{d0} = \frac{1 - \beta^{30}}{1 - \beta^{d0}} \times SOC_{d0} \tag{11}$$

where Y_d represents the cumulative proportion of soil carbon stock from the soil surface down to a depth of d (cm), SOC_{30} is the SOC (Mg C ha^{-1}) pool in the first 30 cm of the soil profile [81], β is the relative decrease rate of the soil carbon pool with soil depth (β = 0.9786 as the global average), d_0 is the original soil depth available in each soil sampled profile (cm), and SOC_{d0} is the original SOC stocks.

The performance of the different ρ_b PTFs in predicting SOC stocks across biomes and two major land use/land cover (viz. Agricultural Lands and Native Vegetation) was evaluated comparing both measured and estimated SOC stock averaged by biomes and land uses, analyzing the distribution of the deviance residuals.

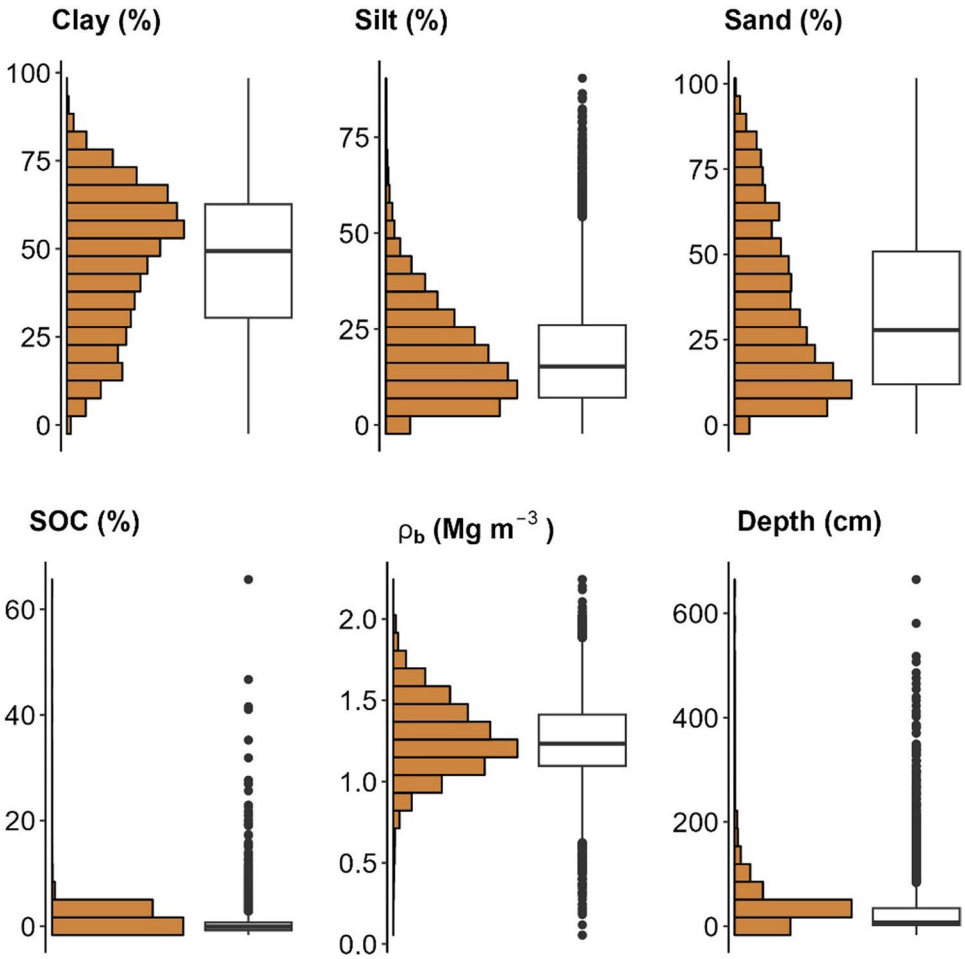

Fig. 3 The distribution of clay, silt, sand, and soil organic carbon contents, soil bulk density, and soil depth in the compiled dataset

3 Results and discussion

3.1 Descriptive statistics of soil properties

Before performing variable transformations and calculating indices for developing PTFs 1 and 2, the soil properties were submitted to statistical analysis. Figure 3 summarizes the variability of these soil properties, highlighting the broad ranges observed across Brazil's tropical and subtropical regions. Tropical environments are typically characterized by highly weathered soils, rich in kaolinite clay and sesquioxides of iron and aluminum within the finer soil fractions. These components significantly influence carbon storage throughout the soil profile [84].

The mean bulk density was 1.27 g cm^{-3}, ranging from 0.13 to 2.21 g cm^{-3}. Clay content varied between 0.1 and 96%, with an average of 47%. Soil samples with clay content higher than 31% comprised 75% of the dataset. Sand content ranged from 0.1 to 99%. The SOC content ranged from near zero to 64%, with an average of 1.52%.

Pearson correlation analysis (Fig. 4) showed that soil bulk density is strongly influenced by both clay and sand contents, with opposite effects: it is positively correlated with sand and negatively correlated with clay content. Additionally, soil bulk density exhibits a significant inverse relationship with organic carbon content. The absolute Pearson correlation coefficient between sand and clay content exceeds 0.8, suggesting a high likelihood of collinearity and supporting the use of the additive log-ratio (alr) transformation. Figure 5 presents the frequency distributions of the transformed PSD using alr, the logarithm transformation of the soil organic carbon (ln(SOC + 1)), soil bulk density, and the corresponding Pearson´s correlations. This logarithmic transformation improves the relationship between these variables and soil bulk density.

3.2 Multiple linear regression model structure

Stepwise linear regression analysis identified soil properties, specifically SOC and soil PSD as the primary factors influencing ρ_b estimation. Both the first vector component of the logarithmically transformed PSD (alr_1) and SOC content showed a highly significant relationship with the soil bulk density. This significance was confirmed by the t-statistic test of the estimated regression coefficients for PTF-1 (Table 2) and PTF-2 (Table 3). The second PTF based on SMLR improved model accuracy by incorporating categorical variables (Table 3). The Variance inflation factor (VIF) was used to assess the multicollinearity among variables, with each regression coefficient β_i in the SMLR models having a VIF below 10, indicating acceptable levels of multicollinearity (Tables 2 and 3). The F-values for the selected linear regression models were highly significant. Low VIF values associated with the variables suggest the absence of collinearity issues.

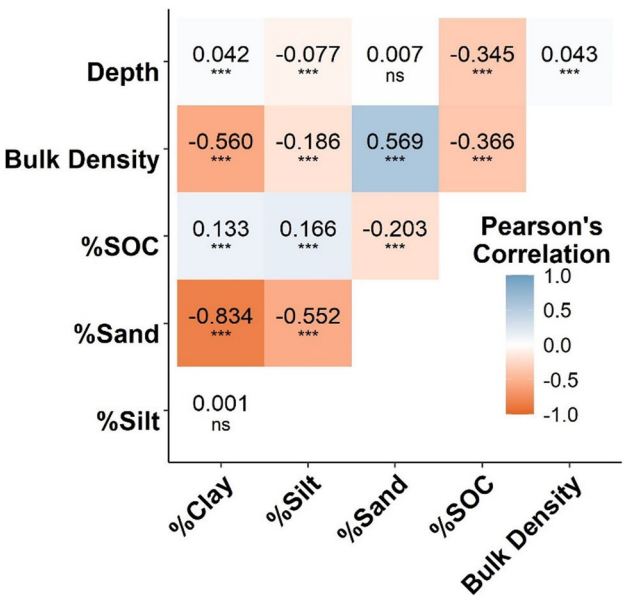

Fig. 4 Pearson correlation coefficient (r) between soil properties. Significance levels: ns (not significant) for $p \geq 0.05$; * for $p < 0.05$; ** for $p < 0.01$; and *** for $p < 0.001$

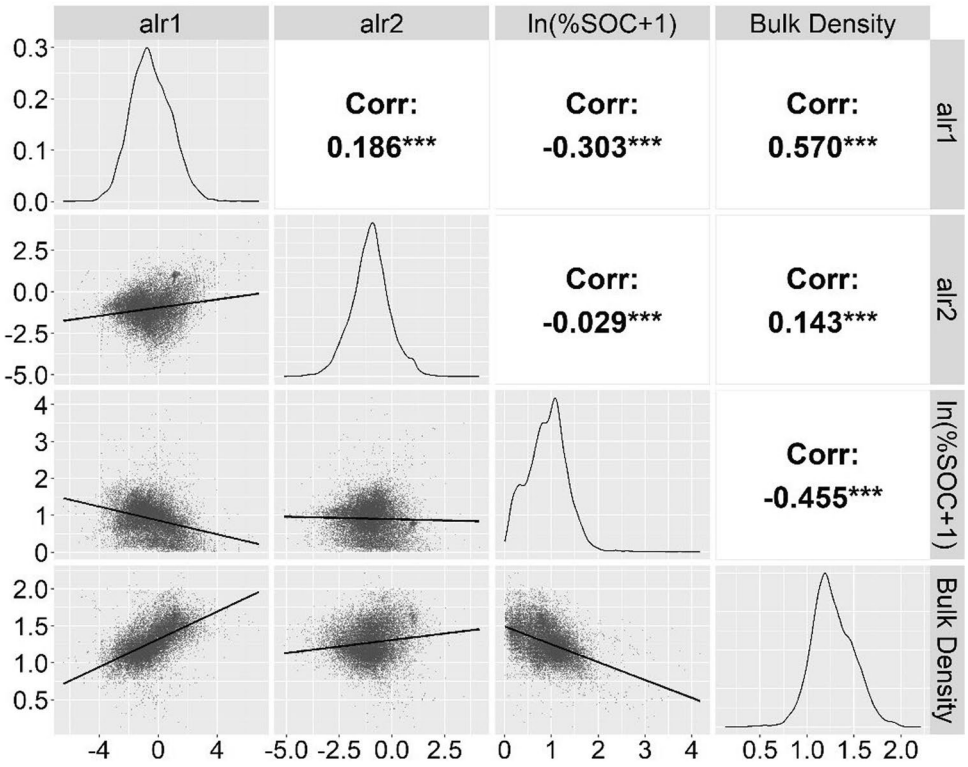

Fig. 5 Frequency distributions of transformed variables (central diagonal graphs) and Pearson correlation coefficients (r) between soil bulk density and transformed variables. Significance levels: ns (not significant) for $p \geq 0.05$; * for $p < 0.05$; ** for $p < 0.01$; and *** for $p < 0.001$

Table 2 Regression coefficient (β) estimates, significance statistics, and variance inflation factor (VIF) for multiple regression analysis of model 1 (PTF-1) during training

Coefficient	Estimate	Std.Error	t value	Pr(>|t|)	VIF	1/VIF
$\beta_{0\ (Intercept)}$	1.547	0.008	184.549	< 0.001	–	–
$\beta_{1\ (alr1)}$	0.058	0.002	28.066	< 0.001	4.071	0.246
$\beta_{2\ (alr2)}$	−0.011	0.003	−3.433	< 0.001	4.472	0.224
$\beta_{3\ (ln(SOC+1))}$	−0.224	0.005	−44.681	< 0.001	2.122	0.471
$\beta_{4\ (clay\_ratio^{-1})}$	−0.030	0.003	−9.477	< 0.001	4.237	0.236
$\beta_{5\ (sampling\ depth)}$	−0.001	0.00006	−20.705	< 0. 0.001	1.851	0.540
$\beta_{6\ (silt/clay)}$	0.031	0.005	5.971	< 0. 0.001	3.849	0.260
Adjusted R^2	0.529					
Residual standard error	0.139					
F statistic	2003					
Probability of F statistic	P < 0.0001					

3.3 Performance evaluation of multiple linear regression models

Given the fact that soil texture and organic carbon data are readily available at various depths across soil profiles in Brazilian soil surveys, a simplified pedogenetic model was developed using only continuous numerical variables to predict soil bulk density (PTF-1). The large dataset used in this study enabled the simplification of the model development through a stepwise procedure. Table 2 shows the β parameters for PTF-1. Internal validation on the training set revealed an adjusted R^2 of 0.53 and a residual standard error of 0.139 g cm^{-3}. In comparison, PTF-2 (Table 3) demonstrated better performance during internal validation on the training set, with an adjusted R^2 of 0.62 and a residual standard error of 0.126 g cm^{-3}.

The two SMLR PTF models yielded slightly different accuracies in predicting soil bulk density. Figure 6 presents scatterplots of predicted versus measured ρ_b along with descriptive regression statistics. The inclusion of categorical variables, such as soil types, Brazilian biomes, land use/land cover types, and soil texture, improved prediction accuracy, resulting in an R^2 of 0.61, a bias of -0.01 g cm^{-3}, an RMSE of 0.125 g cm^{-3}, and a MAE of 0.10 g cm^{-3}. Incorporating relevant categorical

Table 3 Regression coefficient (β), significance statistics, and generalized variance inflation factor (GVIF) for multiple regression analysis of model 2 (PTF-2) during training

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)	GVIF
$\beta_{0\ (Intercept)}$	1.512	0.009	177.248	< 0.001	-
$\beta_{1\ (alr1)}$	0.061	0.002	29.230	< 0.001	6.954
$\beta_{2\ (alr2)}$	−0.028	0.003	−9.303	< 0.01	4.548
$\beta_{3\ (ln(SOC+1))}$	−0.169	0.005	−36.599	< 0.001	2.124
$\beta_{4\ (clay\_ratio^{-1})}$	−0.028	0.003	−10.874	< 0.001	4.414
$\beta_{5\ (sampling\ depth)}$	−0.001	0.00005	−16.985	< 0.001	1.858
$\beta_{6\ (silt/clay)}$	−0.008	0.005	−1.759	< 0.1	3.766
$\beta_{7\ (Soil\ Order)}$					3.068
Cambisol	−0.041	0.005	−7.762	< 0.001	
Gleysol	−0.079	0.016	−4.965	< 0.001	
Latosol	−0.040	0.004	−11.143	< 0.001	
Quartzarenic Neosol	−0.044	0.007	−6.710	< 0.001	
Planosol	0.085	0.011	7.726	< 0.001	
Plintosol	−0.052	0.021	−2.522	< 0.05	
$\beta_{8\ (Texture\ Group)}$					3.492
Loamy soil	0.049	0.004	12.332	< 0.001	
$\beta_{9\ (Biome)}$					3.182
Cerrado	−0.055	0.003	−17.151	< 0.001	
Pampas	0.129	0.005	25.715	< 0.001	
$\beta_{10\ (LULC)}$					1.655
Agricultural land	0.023	0.005	4.876		
Native vegetation	−0.036	0.005	−6.538	< 0.001	
Adjusted R^2	0.615				
Residual standard error	0.126				
F statistic	944.2				
Probability of F statistic	P < 0.0001				

data has proven essential for enhancing model predictability, as it better captures the relationship between bulk density and the prevailing conditions within landscape unit (e.g., current land use/land cover in each biome). However, for some applications, a linear regression model using only numerical variables, like PTF-1, may be preferable due to its simplicity and ease of implementation. By selecting only a few key predictors that account for most of the variance of the soil bulk density, unnecessary complexity can be avoided.

Considering that soil texture and organic carbon are commonly available in soil evaluations and investigations, a simplified PTF regression model can be developed with log transformations of these variables to predict soil bulk density in Brazil, as demonstrated by PTF-1. Incorporating soil indices such as clay ratio into the model highlights the aggregation effect of clay particles, which affects soil structure and, consequently, the bulk density of both surface and deeper soil layers. Most predictive models for estimating soil bulk density in Brazil rely on linear combinations of soil organic matter and clay, silt and sand contents [15, 18, 48, 85].

3.4 Performance of the RF regression model (PTF-3)

The results showed a direct relationship between the number of predictors and model performance, with performance improving as additional predictors were incorporated beyond the soil propriety-related variables used in the PTF-1 and PTF-2. The variance explained by the model increased with the inclusion of categorical, climatic, and vegetation index predictors. This was demonstrated by the higher accuracy of the RF model in predicting soil bulk density, reflected in its higher coefficient of determination and lower RMSE, as compared to SMLR. The Random Forest model achieved an RMSE of 0.114 g/cm^3, with superior R^2 values of 0.75 in the Out-of-Bag (OOB) evaluation and 0.74 in the test set, indicating that RF outperformed SMLR (Fig. 6e, f). An important step in the OOB evaluation was determining the optimal number of

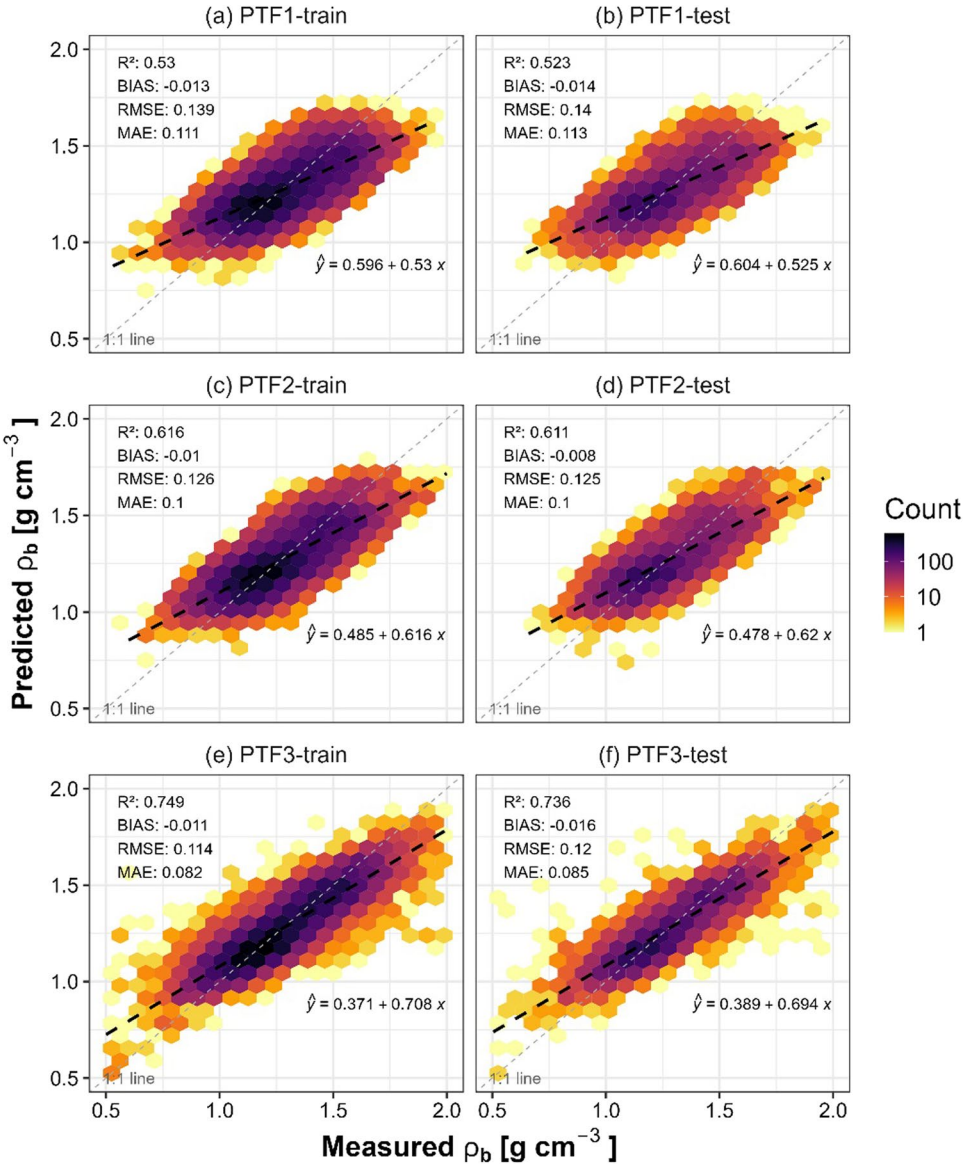

Fig. 6 Relationship between measured and predicted ρ_b for the training set (**a**, **c**, **e**) and testing set (**b**, **d**, **f**) for SMLR-PTF-1, SMLR-PTF-2 (without outliers), and RF-PTF-3. The count indicates the number of instances in each hexagon

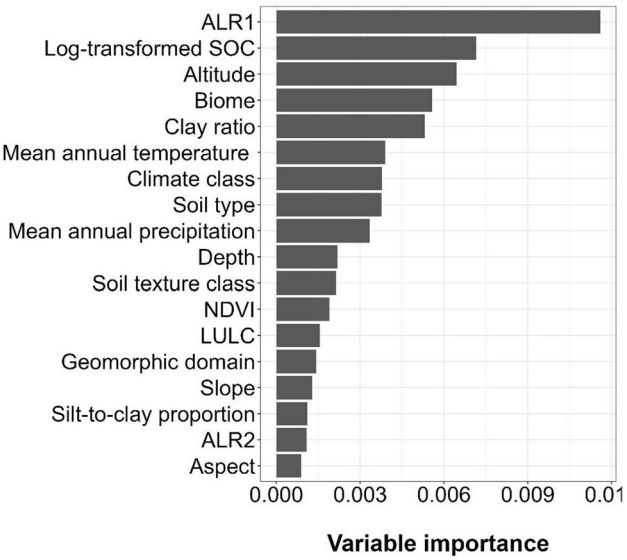

Fig. 7 Variable importance (VIMP) of the RF-PTF-3 model in predicting soil bulk density

trees for prediction. Based on the relationship between OOB error and the number of trees, 1,000 trees were identified as optimal for effective model prediction.

3.4.1 Importance of each predictor

The variable importance (VIMP), calculated using the Breiman-Cutler permutation method [78], is presented in Fig. 7. The VIMP ranking of baseline variables for bulk density ranges from the highest (alr1) to the lowest (Aspect). VIMP values are visually represented with bar charts, showing the increase in prediction error caused by the random permutation of each variable, thus highlighting their relative contributions to model accuracy. A higher VIMP value indicates a greater predictive capability for the corresponding variable.

The inclusion of the first alr transformation component as a predictor indicates a strong positive relationship between log-transformed sand/clay content and soil bulk density. This result underscores the importance of soil mineral particles in shaping the structure of highly weathered tropical soils. In deeper layers of these soil profiles, the reduction in stable SOC content is controlled by climate, vegetation type, and management practices that do not include deeper-rooted vegetation, leading to greater undersaturation of mineral-associated carbon [86]. This condition is common in soils of tropical and subtropical climates in Brazil, where iron and aluminum sesquioxides in the clay fraction play a key role in soil aggregation. These stabilizing agents contribute to favorable physical conditions in deeper soil layers and are strongly correlated with bulk density. Additionally, iron sesquioxides such as goethite and hematite play a significant role in SOC adsorption in tropical soils [87]. The clay fraction's mineralogy also significantly influences bulk density predictions, as it varies with climate and parent material across Brazil. Among minerals in the clay fraction, kaolinite is likely the most abundant in Brazilian soils, especially in Latosols, with exceptions for the most weathered gibbsitic types [88].

In particular, due to the prevalence of highly weathered soils in Brazil, Latosols with prevalence of gibbsite exhibit relatively low bulk density values, while Latosols prevalent in kaolinite show high values, especially in their B horizon [89]. The former refers to the soils that are the result of long-term, intense weathering and complete leaching [88], which occur with great prominence in the Cerrado Region of the Central Plateau of Brazil. Gibbsite, a highly stable end product of weathering, is also found in various tropical regions globally, such as the predominantly gibbsitic soils in Hawaii [90]. Thus, the alr1 and the clay ratio highlight the importance of the clay-size fraction, as a physical soil property that is strongly influenced by iron/aluminum oxides and clay mineralogy, for ρ_b prediction. This is reflected in its relationship with other soil particle fractions, indicating characteristics of well-weathered soils in the humid tropics of Brazil. The greater significance of the alr1 predictor suggests a stronger direct relationship with soil bulk density at lower sand/clay ratios, a trend emphasized by the log-ratio transformation. Numerous studies have already highlighted the importance of PSD (sand, silt, clay) and SOC in predicting ρ_b [15, 21, 29, 53, 54, 85, 91–94]. However, this study demonstrates that transformations of these variables, and the incorporation of environmental and categorical predictors, can further improve the accuracy of bulk density predictions.

Land use and management practices significantly influence soil aggregation, structure, and consequently, soil bulk density. Soil organic carbon plays a key role in the formation and stabilization of larger aggregates [95], while clay mineral composition governs the formation of smaller aggregates. To promote the distribution of SOC at depth and stabilize it as recalcitrant carbon, thereby prolonging its turnover time [96], it is essential to enhance both aboveground and belowground inputs of organic matter. These efforts must consider the specific characteristics of soils and agroecosystems. SOC emerged as the second most important factor in estimating soil bulk density, a significance underscored by the high Variable Importance value following the logarithmic transformation of SOC. This transformation highlights small SOC values, which are common in Brazil's tropical climate conditions.

In addition to predictors based on particle size distribution (sand, silt, clay) and SOC, other key factors in the development of PTF-3 included variables related to climate conditions and topography. These factors, which vary widely across Brazilian biomes, had a significant influence on ρ_b. Notably, variables such as altitude, biome, and climate (mean annual temperature, mean annual precipitation, and climate classification) emerged as important predictors, enhancing the accuracy of soil bulk density estimates. The great importance of altitude and mean annual temperature variables is likely due to their effect on SOC retention and soil weathering processes. High altitudes in tropical regions can enhance SOC retention, as cooler temperatures at higher elevations slow the decomposition of organic matter, promoting SOC stabilization [97, 98]. Studies by [99] further support this association, reporting higher SOC content and correspondingly lower soil bulk density values at altitudes above 1,000 m a.s.l. in the semiarid highlands of the Caatinga biome. These

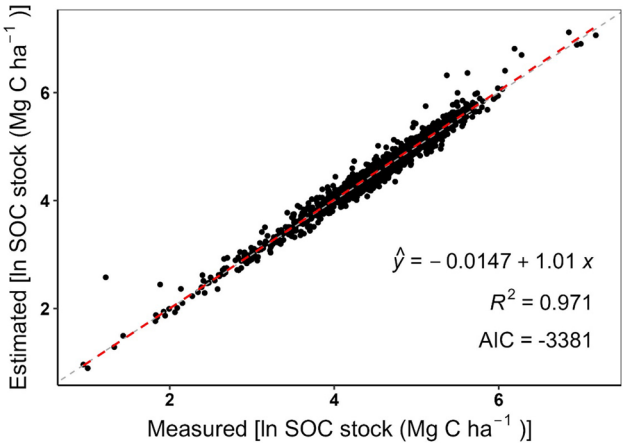

Fig. 8 Estimated soil organic carbon stocks using the developed bulk density PTF-3. Red dashed line is a linear fit, the grey dashed line along the fitted line represents the 1:1 line. AIC is Akaike information criterion

Table 4 Statistical metrics for SOC stock predictions versus measured values, using different models for bulk density estimation

Model	R^2	AIC	MAE Mg ha^{-1}	RMSE	RMSLE
PTF-1	0.958	–2735	10.9	33.2	0.12
PTF-2	0.966	–3173	9.3	24.1	0.11
PTF-3	0.971	–3381	9.8	22.1	0.11
PTF-Bnt[a]	0.949	–2368	12.7	32.6	0.16
PTF-Brn[b]	0.957	–2626	16.5	30.4	0.18
PTF-T&H[c]	0.957	–2611	12.2	24.2	0.15

[a] [15]; [b] [18]; [c][48]

high-altitude soils, formed under paleoclimates and paleoweathering, serve as important proxies for understanding climate change.

The results presented here demonstrate that RF regression modeling achieved the highest prediction accuracy, while MLR also yielded a satisfactory level of accuracy. In regions with limited soil data, such as the tropical rainforest area [48, 100], locally trained PTFs using MLR can provide suitable estimations. This effectiveness arises from the selection of appropriate statistical methods and soil predictors according to specific pedo-environmental conditions. The study by de [24] explored the predictive capabilities of RF and MLR-derived PTFs for estimating soil bulk density, considering both soil properties and environmental factors. They found that MLR and RF models were more accurate when combining soil properties with environmental variables, while RF outperformed MLR when solely focused on soil properties. Our findings are consistent with these conclusions and align with those of [101], who reported that RF and MLR models performed better in estimating ρ_b when both soil and environmental data were used as predictors.

The performance of the PTFs developed in this study was compared with tropical-climate PTFs from Central Africa, as reported by [29], who utilized a dataset of highly weathered soils. Despite differences in dataset size and input variables, our PTFs performed better overall, even when accounting for the challenges of predicting soil bulk density due to short-term variability caused by natural or anthropogenic influences on land use.

Nevertheless, it is important to highlight that the approach outlined here could be applied to other regions with highly weathered tropical soils, provided that the soil formation factors and their related biotic and abiotic properties are adequately characterized and include relevant indicators for quantifying soil bulk density at a similar study scale. Additionally, certain tropical soil types with significance in other tropical regions lack direct counterparts in the Brazilian Soil Classification System. This underscores the need for tailored approaches and region-specific data to enhance the applicability and accuracy of PTFs in tropical environments worldwide.

3.5 Estimation of SOC stocks using PTF-derived ρ_b

The soil bulk density prediction methods developed in this study, along with existing pedotransfer functions (PTFs) for general Brazilian soils, were applied to estimate SOC stocks at all measured points, while accounting for the original depth

of the collected soil profiles. Figure 8 presents a comparison between the estimated and measured values. Due to the highly skewed distribution of SOC data, a natural logarithmic (ln) transformation was applied. The considerable variation in SOC stocks likely reflects differences in SOC content, bulk density, land use, biomes, soil classes, textures, environmental conditions, and other relevant factors, as well as the significant variability in soil depth across the measured points.

The evaluation of the prediction models for estimating SOC stocks demonstrated that PTF-3 performed the best, with a slightly higher R^2 value of 0.97 and the largest reduction in AIC, reaching -3381. The AIC is particularly useful for comparing the relative fit of different PTF models applied to the same dataset. A lower AIC value indicates a better-fitting model for estimating SOC stocks among the models compared, as shown in Table 4. Alongside the error metrics, MAE, RMSE, and RMSLE were also employed to evaluate the model performance. RMSLE is an error metric that applies a logarithmic transformation to the values before calculating the errors, in contrast to RMSE, which directly computes the square root of the mean squared differences between observed and predicted values. RMSLE is particularly useful in situations involving skewed data, such as the positively skewed distribution of SOC Stock data across the Brazilian territory. Most metrics confirmed the superiority of PTF-2 and PTF-3 in estimating SOC stock.

The assessment of entire soil profiles for calculating total carbon stocks resulted in higher error metrics, such as increased MAE and RMSE, compared to metrics reported by other studies that used ρ_b PTFs to estimate carbon stocks in individual soil layers [52, 102]. Depending on the soil type, significantly higher SOC stocks and variability in SOC content may be observed, as seen in Brazilian tropical podzols (Spodosols) in the Amazon. These tropical podzols are characterized by a SOC-rich topsoil horizon, a deep, thick SOC-rich spodic horizon (Bh), overlying a C horizon with progressively lower carbon content extending deep into the soil profile. For these soils [103], reported RMSE values of 13.6 kg C m^{-2}, 15.9 kg C m^{-2}, and 15.0 kg C m^{-2} when estimating SOC stocks using soil bulk density PTFs.

Mean soil organic carbon (SOC) stocks, both measured and estimated across different biomes and land uses, for a standardized soil layer (0–30 cm) at the selected soil profiles, are shown in Table 5. Differences between measured and estimated SOC stocks are illustrated in the box plot graphs in Fig. 9.

The results revealed differences between bulk density PTFs and their impacts on SOC stocks in agricultural lands and native vegetation across Brazilian biomes. PTF-2 and PTF-3 demonstrated the smallest deviations when predicting SOC stocks across soil layers sampled from various biomes, including both agricultural and native vegetation areas. When estimating SOC stocks within standardized soil profiles to a depth of 30 cm, the RMSE for SOC stock predictions was 16.3, 13.0, and 12.6 Mg ha^{-1} for PTF-1, PTF-2, and PTF-3, respectively. In comparison, the RMSE values for PTFs from [15, 18], and [48] were 14.1, 19.7, and 18.1 Mg ha^{-1}, respectively. The errors were more uniformly distributed around the median SOC stock values for PTF-2 and PTF-3, highlighting the importance of incorporating categorical variables in MLR (PTF-2) and Random Forest (PTF-3) models. This approach improves the accuracy of bulk density predictions across diverse landscapes and pedo-environmental conditions.

Table 5 Mean SOC stocks (Mg ha^{-1}) in the 0–30 cm soil layer for native vegetation and agricultural land cover in Brazilian biomes, calculated from this study, and comparison with other SOC stocks estimated from bulk density PTFs available for general Brazilian soils

Biome	Land Use	N	SOC stock (Mg ha^{-1})						
			Measured	PTF-1	PTF-2	PTF-3	Bnt	Brn	T&H
Amazon rainforest	NV	30	50.2 (±5.0)	53.7 (±5.2)	51.8 (±5.2)	52.7 (±5.5)	48.5 (±4.4)	46.7 (±4.6)	47.5 (±4.8)
	AL	53	84.3 (±16.7)	74.4 (±10.6)	78.3 (±12.4)	83.5 (±15.3)	46.8 (±6.3)	60.8 (±7.1)	57.8 (±6.3)
Atlantic forest	NV	97	90.4 (±11.6)	100.1 (±10.9)	100 (±12.6)	104.8 (±15.6)	31.8 (±36.9)	67.7 (±10.3)	56.9 (±15.9)
	AL	468	77.4 (±1.2)	74.6 (±1.3)	76.7 (±1.4)	77.6 (±1.3)	68.0 (±1.0)	66.8 (±1.2)	67.6 (±1.1)
Caatinga	NV	59	42.7 (±3.8)	41.9 (±3.8)	41.1 (±3.8)	41.6 (±3.8)	37.9 (±3.4)	35.9 (±3.4)	38.3 (±3.6)
	AL	79	40.9 (±3.0)	40.1 (±3.0)	41.0 (±3.1)	39.8 (±3.0)	36.7 (±2.7)	34.4 (±2.6)	36.1 (±2.8)
Cerrado	NV	158	68.0 (±2.1)	72.9 (±2.1)	68.4 (±2.0)	73.0 (±2.1)	65.4 (±1.8)	63.6 (±1.9)	67.8 (±2.0)
	AL	884	76.6 (±0.8)	76.5 (±0.7)	75.2 (±0.7)	77.8 (±0.8)	68.6 (±0.6)	66.7 (±0.7)	71.5 (±0.7)
Pampas	NV	24	66.6 (±3.4)	63.5 (±2.9)	68.2 (±3.0)	64.6 (±3.0)	58.8 (±2.6)	55.5 (±2.6)	58.0 (±2.6)
	AL	182	67.2 (±1.5)	60.1 (±1.2)	67.3 (±1.4)	61.3 (±1.3)	55.8 (±1.1)	52.6 (±1.1)	54.4 (±1.1)
Pantanal	NV	7	39.4 (±5.7)	49.7 (±10.3)	47.5 (±10.1)	46.1 (±9.5)	45.6 (±9.4)	42.8 (±9.3)	45.2 (±9.1)
	AL	3	62.4 (±10.3)	66.1 (±7.5)	69.4 (±7.9)	68.5 (±8.7)	60.4 (±7.0)	57.2 (±6.9)	62.4 (±6.8)

NV Native Forest, *AL* Agricultural Lands; Bnt: [15]; Brn: [18]; T&H: [48]; N: number of points

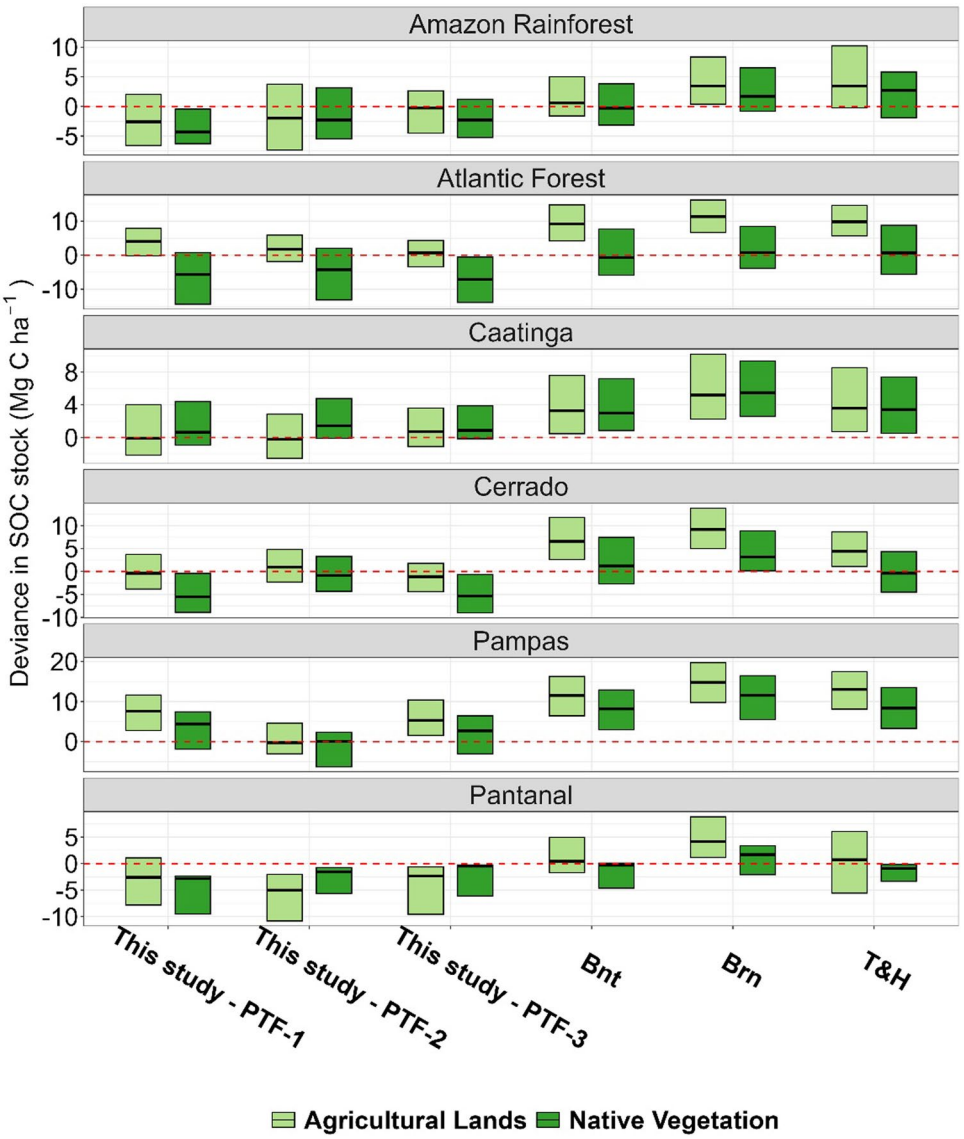

Fig. 9 Performance of SOC stock prediction in two different land uses (Native Vegetation and Agricultural Lands) using the PTFs from this study, compared to three reference PTFs available for Brazilian soils. The 0.25, and 0.75 quantiles (box), and median (solid horizontal line) 0–30 cm soil organic carbon storage deviance estimates were drafted in the plot. Reference PTFs:Bnt: [15]; Brn: [18], and T&H: [48]

The Random Forest model (PTF-3) proved particularly effective for evaluating SOC stocks, leveraging an indicator- or proxy-based approach, which is especially useful when analyzing larger regions. This is a critical consideration when assessing soil organic carbon storage as a key soil function across different spatial scales [104], especially when using terrain features as predictors in the model. Depending on the landscape scale, terrain attributes can effectively explain the variability of SOC stocks by influencing patterns of precipitation, temperature, and soil water-holding capacity. In this study, elevation above sea level emerged as a more effective predictor of bulk density, and consequently a better indicator of SOC storage than climate variables. This is because elevation integrates the effects of temperature and precipitation on net primary productivity and decomposition, while also reflecting the processes that shape soil type distribution across the landscape. Therefore, depending on terrain conditions, elevation can be considered a reliable proxy for climate across Brazil's vast and diverse territory. Advanced PTF development techniques improve soil property estimation by reducing geographical biases, especially as soil properties change in response to tillage and climate, impacting the spatial distribution of predicted variables [105].

According to Table 5, in the Amazon biome, PTF-3 demonstrated the best performance for agricultural land use, with estimates closely aligning with measured SOC stock values in areas of natural vegetation. In the Cerrado biome, the developed PTFs produced estimates with minimal deviations from measured SOC stock values for both agricultural and natural areas, while also exhibiting lower standard errors, indicating reduced variability in SOC stocks predictions. Overall, the PTFs developed in this study outperformed those from existing literature for agricultural soils in the Cerrado biome.

A similar trend was observed in the Caatinga biome, where slight improvements in SOC stock estimation were noted for areas of native vegetation areas. In the Atlantic Forest, the PTFs showed better performance in estimating SOC stock for native vegetation compared to the three PTFs from previous studies. Finally, in the Pampas biome, PTF-2 slightly outperformed the others for both native vegetation and agricultural land, with estimates closely matching measured values.

The finding of higher SOC stocks in agricultural lands compared to native vegetation within the Amazon biome has been explored in meta-analyses. This variation may result from a higher dominance of pasture-derived SOC relative to cropland-derived SOC in the sampled regions. Croplands often have lower carbon restitution and lose forest carbon during cultivation, while pastures benefit from grass root systems that enhance SOC storage and offset carbon losses from native forests [106]. However, the ongoing conversion of tropical forests to pastures and croplands in the Amazon raises concerns regarding future biogeochemical cycles in the Amazon Basin. Declines in SOC stocks in the Amazon have been projected by [107] under different land-use change scenarios. This study also reports that soils adjacent to the Amazon River have the highest SOC stocks due to their hydric conditions.

The lowest carbon stock values were found in the two Brazilian biomes located at the extremes of the dry Northeast-Southwest (NE-SW) axis in Brazil. Both the Caatinga and Pantanal biomes are highly vulnerable to climate change due to their predominantly dry and seasonal climates and their lack of influence from the higher altitudes of the Central Plateau, which would otherwise contribute to greater carbon accumulation, as shown by [52]. In the Pantanal, agricultural areas primarily result from the conversion of native forest to cultivated pasture, along with continuous grazing of native pasture. Interestingly, agricultural lands in this biome exhibited higher SOC stock values than native vegetation, which contrasts with the findings of [108]. However, the limited number of soil profile samples from the Pantanal biome may have compromised the representativeness of its land use patterns (Table 5). Additionally, the Pantanal is highly susceptible to fire due to changes in land cover and climate, particularly shifts in precipitation patterns [109]. These factors influence its annual SOC contributions, which are largely dependent on river flooding, leading to soil inundation and sediment deposition. Therefore, future studies should focus on SOC stock losses, especially in areas vulnerable to erosion or sediment deposition, such the Pantanal, Pampas and Caatinga biomes on a large scale, and in smaller areas prone to colluvial and alluvial deposition and localized erosion.

4 Conclusions

This study compiled a comprehensive database encompassing particle size distribution (sand, silt and clay contents), soil organic carbon content, sampling depth, soil bulk density, soil type, soil texture group, biome, land use and geographic coordinates. This dataset allowed for the inclusion of additional covariates, such as climatic and topographic variables and vegetation index as abiotic and biotic indicators of the soil formation, from other open-source databases. The study reveals significant variability in soil properties across Brazil's tropical and subtropical regions, influenced by factors such as climate, topography, parent material, age, and spatial position. The highly weathered soils in these areas, which are rich in kaolinite clay and iron/aluminum sesquioxides, play a significant role in carbon storage due to its influence on soil bulk density in Brazil. Using Multiple Linear Regression and Random Forest modeling, the study developed predictive models for soil bulk density, emphasizing the importance of soil PSD and organic carbon content. While both MLR and RF exhibited similar performance, RF demonstrated slightly superior accuracy, especially with the inclusion of environmental covariates such as topographical, climatic, and categorical predictors. RF's advantages include its ability to handle high dimensionality, capture non-linear relationships, and its robustness to outliers. The accuracy of the MLR models improved with the use of log-transformed PSD and SOC data, as well as the inclusion of categorical variables such as biome, land use, soil type and textural class.

The slight differences between MLR and RF modeling emphasize MLR's practicality and ease of application. However, machine learning approaches like RF-PTF-3 are increasingly valued for their ability to capture the spatial drivers of soil properties (e.g., climatic variation and topographic factors at scales relevant to SOC storage). This poses a challenge for agricultural extension services, which must adapt to new technologies that enhance land use in the context of climate change and inform long-term strategies that balance immediate and medium-term priorities.

This study utilized soil bulk density PTFs to estimate SOC stocks in Brazilian biomes. A log transformation addressed skewed SOC data, revealing significant variability by land use and environment, according to the data scatter ranging from the lowest to the highest values of SOC Stock. The Random Forest model (PTF-3) outperformed others, with an R^2 of 0.97, emphasizing the value of categorical variables in improving SOC stock estimations in the Amazon and Atlantic

Forests biomes. The PTFs developed herein outperformed existing soil bulk density PTFs from the literature used for comparison in estimating SOC stocks.

The research underscores the importance of incorporating environmental continuous and categorical variables (RF model) to improve prediction accuracy, thereby improving the understanding of soil behavior across diverse landscapes. Moreover, the study emphasizes the need for comprehensive soil datasets to advance soil management strategies and contribute to policy decisions. Accurate monitoring of soil bulk density is essential for assessing soil health and implementing effective conservation practices. Prioritizing agricultural sustainable management practices in Brazilian policies, such as promoting cover crops and minimal tillage, is essential for mitigating compaction and preserving soil structure and function.

Acknowledgements This research was financially supported by a joint project between Embrapa and Bayer Corporation (Brazil branch) titled "Balance and carbon footprint in the production of soybean, maize, and sugarcane: metrics, tools, and protocols for tropical and subtropical areas of Brazil" (proc. 20.22.00.118.00.02).

Author contributions W. P. S.: writing—review & editing, writing—original draft, visualization, investigation, conceptualization, formal analysis, methodology, software, validation. C. M. P. V.: writing—review & editing, writing—original draft, investigation, formal analysis, methodology, validation, supervision, data curation, conceptualization. L. M.: writing—review & editing, supervision, resources, project administration, funding acquisition. A. A.: writing—review & editing, funding acquisition. J. T.: writing—review & editing, formal analysis, conceptualization, data curation. F. S. C.: writing—review & editing, formal analysis, data curation. J. A. A.: writing—review & editing, conceptualization. Q. J. V. L.: writing—review & editing, conceptualization. R. G.: WRITING—REVIEW & EDITING, formal analysis, data curation. F. J. P.: writing—review & editing, formal analysis, data curation.

Data availability The dataset generated during this study is available from the corresponding author upon reasonable request.

Declarations

Competing interests The authors declare no competing interests.

References

1. IPCC: Sections. In: Climate Change 2023: Synthesis Report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland; 2023. p. 35–115. https://doi.org/10.59327/IPCC/AR6-9789291691647.
2. Amelung W, Bossio D, de Vries W, Kögel-Knabner I, Lehmann J, Amundson R, et al. Towards a global-scale soil climate mitigation strategy. Nat Commun. 2020;11(1):5427.
3. Poeplau C, Don A. Carbon sequestration in agricultural soils via cultivation of cover crops – a meta-analysis. Agric Ecosyst Environ. 2015;200:33–41.
4. Jian J, Du X, Reiter MS, Stewart RD. A meta-analysis of global cropland soil carbon changes due to cover cropping. Soil Biol Biochem. 2020;143:107735.
5. Beillouin D, Ben-Ari T, Malézieux E, Seufert V, Makowski D. Positive but variable effects of crop diversification on biodiversity and ecosystem services. Glob Chang Biol. 2021;27(19):4697–710.
6. Reichert JM, Suzuki LEAS, Reinert DJ, Horn R, Håkansson I. Reference bulk density and critical degree-of-compactness for no-till crop production in subtropical highly weathered soils. Soil Tillage Res. 2009;102(2):242–54.
7. Assouline S. Modeling the relationship between soil bulk density and the water retention curve. Vadose Zo J. 2006;5(2):554–63. https://doi.org/10.2136/vzj2005.0083.
8. Sequeira CH, Wills SA, Seybold CA, West LT. Predicting soil bulk density for incomplete databases. Geoderma. 2014;213:64–73.
9. Brahim N, Bernoux M, Gallali T. Pedotransfer functions to estimate soil bulk density for Northern Africa: Tunisia case. J Arid Environ. 2012;81:77–83. https://doi.org/10.1016/j.jaridenv.2012.01.012.
10. Li Y, Li Z, Cui S, Zhang Q. Trade-off between soil pH, bulk density and other soil physical properties under global no-tillage agriculture. Geoderma. 2020;361:114099.

11. Rabbi SMF, Tighe M, Delgado-Baquerizo M, Cowie A, Robertson F, Dalal R, et al. Climate and soil properties limit the positive effects of land use reversion on carbon storage in Eastern Australia. Sci Rep. 2015;5(1):17866.
12. dos Cordeiro CFS, Rodrigues DR, da Silva GF, Echer FR, Calonego JC. Soil organic carbon stock is improved by cover crops in a tropical sandy soil. Agron J. 2022;114(2):1546–56. https://doi.org/10.1002/agj2.21019.
13. Bouma J. Using soil survey data for quantitative land evaluation. In: Stewart BA, editor. Advances in soil science. New York: Springer Verlag; 1989. p. 177–213.
14. McBratney AB, Minasny B, Cattle SR, Vervoort RW. From pedotransfer functions to soil inference systems. Geoderma. 2002;109:41–73.
15. Benites VM, Machado PLOA, Fidalgo ECC, Coelho MR, Madari BE. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. Geoderma. 2007;139:90–7.
16. Moreira T, Brandao DN, Haddad DB, Ceddia MB, Pinheiro EFM, Oliveira RF. A first approach using neural network to estimating soil bulk density of Urucu basin in Central Amazon-Brazil. In: 2017 International Joint Conference on Neural Networks (IJCNN) [Internet]. IEEE; 2017. p. 3236–3239. https://doi.org/10.1109/IJCNN.2017.7966260.
17. Kaur R, Kumar S, Gurung HP. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs. Aust J Soil Res. 2002;40(5):847–58.
18. Bernoux M, Cerri C, Arrouays D, Jolivet C, Volkoff B. Bulk densities of Brazilian Amazon soils related to other soil properties. Soil Sci Soc Am J. 1998;62(3):743–9. https://doi.org/10.2136/sssaj1998.03615995006200030029x.
19. Heuscher SA, Brandt CC, Jardine PM. Using soil physical and chemical properties to estimate bulk density. Soil Sci Soc Am J. 2005;69(1):51–6. https://doi.org/10.2136/sssaj2005.0051a.
20. Périé C, Ouimet R. Organic carbon, organic matter and bulk density relationships in boreal forest soils. Can J Soil Sci. 2008;88(3):315–25.
21. Beutler SJ, Pereira MG, Tassinari WDS, de Menezes MD, Valladares GS, dos Anjos LHC. Bulk density prediction for histosols and soil horizons with high organic matter content. Rev Bras Ciência do Solo. 2017;41:1–13.
22. Barros HS, Fearnside PM. Pedo-transfer functions for estimating soil bulk density in Central Amazonia. Rev Bras Ciência do Solo. 2015;39(2):397–407.
23. de Carvalho Junior W, Calderano Filho B, da Chagas CS, Bhering SB, Pereira NR, Pinheiro HSK. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas. Pesqui Agropecuária Bras. 2016;51(9):1428–37.
24. de Souza E, Fernandes Filho EI, Schaefer CEGR, Batjes NH, dos Santos GR, Pontes LM. Pedotransfer functions to estimate bulk density from soil properties and environmental covariates: Rio Doce basin. Sci Agric. 2016;73(6):525–34.
25. Ramos HMM, Valladares GS, Andrade Junior AS, Matos Filho CHA, De Sousa RS, Mousinho FEP. Pedotransference functions for prediction of density in soils of Piauí, Brazil. Brazilian J Dev. 2022;8(7):50832–50.
26. Barros AHC, de Jong Van Lier Q. Pedotransfer functions for Brazilian soils. In: Teixeira WG, Ceddia MB, Ottoni MV, Donnagema GK, editors. Application of soil physics in environmental analyses: measuring, modelling and data integration progress in soil science. 1st ed. Cham: Springer; 2014. p. 131–62.
27. Tomasella J, Hodnett MG, Rossato L. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. Soil Sci Soc Am J. 2000;64(1):327–38. https://doi.org/10.2136/sssaj2000.641327x.
28. Abdelbaki AM. Evaluation of pedotransfer functions for predicting soil bulk density for U.S. soils. Ain Shams Eng J. 2018;9(4):1611–9.
29. Botula YD, Nemes A, Van Ranst E, Mafuka P, De Pue J, Cornelis WM. Hierarchical pedotransfer functions to predict bulk density of highly weathered soils in Central Africa. Soil Sci Soc Am J. 2015;79(2):476–86.
30. Minasny B, Hartemink AE. Predicting soil properties in the tropics. Earth-Science Rev. 2011;106:52–62.
31. Panagos P, De Rosa D, Liakos L, Labouyrie M, Borrelli P, Ballabio C. Soil bulk density assessment in Europe. Agric Ecosyst Environ. 2024;364:108907.
32. Han GZ, Zhang GL, Gong ZT, Wang GF. Pedotransfer functions for estimating soil bulk density in China. Soil Sci. 2012;177(3):158–64.
33. Perreault S, El Alem A, Chokmani K, Cambouris AN. Development of pedotransfer functions to predict soil physical properties in Southern Quebec (Canada). Agronomy. 2022;12(2):526.
34. Huntington TG, Johnson CE, Johnson AH, Siccama TG, Ryan DF. Carbon, organic matter, and bulk density relationships in a forested Spodosol. Soil Sci. 1989;148(5):380–6.
35. Lark RM, Rawlins BG, Robinson DA, Lebron I, Tye AM. Implications of short-range spatial variation of soil bulk density for adequate field-sampling protocols: methodology and results from two contrasting soils. Eur J Soil Sci. 2014;65(6):803–14. https://doi.org/10.1111/ejss.12178.
36. Premrov A, Cummins T, Byrne KA. Bulk-density modelling using optimal power-transformation of measured physical and chemical soil parameters. Geoderma. 2018;314:205–20.
37. Tranter G, Minasny B, Mcbratney AB, Murphy B, Mckenzie NJ, Grundy M, et al. Building and testing conceptual and empirical models for predicting soil bulk density. Soil Use Manag. 2007;23(4):437–43. https://doi.org/10.1111/j.1475-2743.2007.00092.x.
38. Palladino M, Romano N, Pasolli E, Nasta P. Developing pedotransfer functions for predicting soil bulk density in Campania. Geoderma. 2022;412:115726.
39. Richard C, Lemercier B, Michot D, Pichelin P, Rémy A, Berthier L, et al. End-user-oriented pedotransfer functions to estimate soil bulk density and available water capacity at horizon and profile scales. Soil Use Manag. 2023;39(1):270–85. https://doi.org/10.1111/sum.12851.
40. Van Looy K, Bouma J, Herbst M, Koestel J, Minasny B, Mishra U, et al. Pedotransfer functions in earth system science: challenges and perspectives. Rev Geophys. 2017;55(4):1199–256. https://doi.org/10.1002/2017RG000581.
41. Chen S, Chen Z, Zhang X, Luo Z, Schillaci C, Arrouays D, et al. European topsoil bulk density and organic carbon stock database (0–20 cm) using machine-learning-based pedotransfer functions. Earth Syst Sci Data. 2024;16(5):2367–83.
42. Ferreira ACS, Pinheiro ÉFM, Costa EM, Ceddia MB. Predicting soil carbon stock in remote areas of the Central Amazon region using machine learning techniques. Geoderma Reg. 2023;32:e00614.
43. Hateffard F, Szatmári G, Novák TJ. Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions. Soil Sci Annu. 2023;74(1):1–11.

44. Nabiollahi K, Eskandari S, Taghizadeh-Mehrjardi R, Kerry R, Triantafilis J. Assessing soil organic carbon stocks under land-use change scenarios using random forest models. Carbon Manag. 2019;10(1):63–77. https://doi.org/10.1080/17583004.2018.1553434.
45. Ottoni MV, Ottoni Filho TB, Schaap MG, Lopes-Assad MLRC, Rotunno Filho OC. Hydrophysical database for Brazilian Soils (HYBRAS) and pedotransfer functions for water retention. Vadose Zo J. 2018;17(1):1–17. https://doi.org/10.2136/vzj2017.05.0095.
46. Perina FJ, Bogiani JC, Ribeiro GC, Breda CE, Fabris A, dos Santos IA, et al. Levantamento e Manejo de Fitonematoides em Algodoeiro no Oeste da Bahia, resultados safra 2016/17. Circular Técnica. Luís Eduardo Magalhães/BA; 2017; 1–8.
47. Galbieri R, Asmus GLA, Vaz CMP, Lamas FM, Crestana S, Torres ED, et al. Áreas de produção de algodão em Mato Grosso: nematoides, murcha de fusarium, sistemas de cultivo, fertilidade e física de solo. Circular Técnica. Primavera do Leste-MT: Instituto Mato-grossense do Algodão (IMAmt); 2014.
48. Tomasella J, Hodnett MG. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Sci. 1998;163(3):190–202.
49. Donagemma GK, Viana JHM, de Almeida BG, Ruiz HA, Klein VA, Dechen SCF, et al. Análise granulométrica. In: Teixeira PC, Donagemma GK, Fontana A, Teixeira WG, editors., et al., Manual de Métodos de Análise de Solo 3 ed rev e ampl. 3rd ed. Brasília: Embrapa; 2017. p. 95–116.
50. Nelson DW, Sommers LE. Total carbon, organic carbon, and organic matter. In: Sparks L, Page AL, Helmke PA, Loeppert RH, Soltanpour PN, Tabatabai MA, Johnston CT, Sumner ME, editors. Methods of soil analysis: part 3 chemical methods. Madison: American Society of Agronomy; 1996. p. 961–1010.
51. Fontana A, Campos DVB. Carbono orgânico. In: Teixeira PC, Donagemma GK, Fontana A, Teixeira WG, editors. Manual de Métodos de Análise de Solo 3 ed rev e ampl. 3rd ed. Brasília: Embrapa; 2017. p. 360–7.
52. Gomes LC, Faria RM, de Souza E, Veloso GV, Schaefer CEGR, Filho EIF. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma. 2019;340:337–50.
53. Katuwal S, Knadel M, Norgaard T, Moldrup P, Greve MH, de Jonge LW. Predicting the dry bulk density of soils across Denmark: comparison of single-parameter, multi-parameter, and vis–NIR based models. Geoderma. 2020;361:114080.
54. Ramcharan A, Hengl T, Beaudette D, Wills S. A soil bulk density pedotransfer function based on machine learning: a case study with the NCSS soil characterization database. Soil Sci Soc Am J. 2017;81(6):1279–87. https://doi.org/10.2136/sssaj2016.12.0421.
55. Jarvis A, Reuter HI, Nelson A, Guevara E. Hole-filled seamless SRTM data V4. International Centre for Tropical Agriculture (CIAT). 2008. https://srtm.csi.cgiar.org
56. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ. 2017;202:18–27.
57. IBGE. BDIA–Banco de Dados de Informações Ambientais. 2023. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2102042.
58. Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. 2017;37(12):4302–15. https://doi.org/10.1002/joc.5086.
59. Hijmans RJ, Barbosa M, Ghosh A, Mandel A. Package "geodata": Download geographic data. 2024.
60. Alvares CA, Stape JL, Sentelhas PC, de Moraes Gonçalves JL, Sparovek G. Köppen's climate classification map for Brazil. Meteorol Zeitschrift. 2013;22(6):711–28.
61. Huete A, Didan K, Miura T, Rodriguez E, Gao X, Ferreira L. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ. 2002;83(1–2):195–213.
62. McBratney A, Mendonça SM, Minasny B. On digital soil mapping. Geoderma. 2003;117(1–2):3–52.
63. Jenny H. Factors of soil formation - a system of quantitative pedology. Vol. 35, geographical review. New York: McGraw-Hill; 1941.
64. Belsley DA. Conditioning diagnostics: collinearity and weak data in regression. 1st ed. Chichester: Wiley-Interscience; 1991. p. 396.
65. Fox J, Monette G. Generalized collinearity diagnostics. J Am Stat Assoc. 1992;87(417):178–83. https://doi.org/10.1080/01621459.1992.10475190.
66. Wischmeier WH, Mannering JV. Relation of soil properties to its erodibility. Soil Sci Soc Am J. 1969;33(1):131–7. https://doi.org/10.2136/sssaj1969.03615995003300010035x.
67. Van WAR. Criteria for classifying tropical soils by age. J Soil Sci. 1962;13(1):124–32. https://doi.org/10.1111/j.1365-2389.1962.tb00689.x.
68. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B. 1982;44(2):139–77.
69. Aitchison J. The statistical analysis of compositional data. London: Chapman and Hall; 1986.
70. Belsley DA, Kuh E, Welsch RE. Regression diagnostics — identifying influential data and sources of collinearity. New York, NY: John Wiley & Sons; 1980.
71. de Santos HG, Jacomine PKT, de Anjos LHC, de Oliveira VÁ, Lumbreras JF, Coelho MR, et al. Sistema Brasileiro de Classificação de Solos. Embrapa Solos. 5th ed. Brasília: Embrapa; 2018.
72. Soil Survey Staff. Keys to soil taxonomy. 13th ed. Washington, DC: USDA Natural Resources Conservation Service; 2022.
73. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 2024. https://www.r-project.org/
74. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4(43):1686. https://doi.org/10.21105/joss.01686.
75. Olivoto T, Lúcio AD. metan: an R package for multi-environment trial analysis. Methods Ecol Evol. 2020;11(6):783–9. https://doi.org/10.1111/2041-210X.13384.
76. Moeys J, Shangguan W, Petzold R, Minasny B, Rosca B, Jelinski N, et al. Package "soiltexture": functions for soil texture plot, classification and transformation. 2024. https://doi.org/10.32614/CRAN.package.soiltexture.
77. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008. https://doi.org/10.18637/jss.v028.i05.
78. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
79. Ishwaran H, Lu M, Kogalur UB. randomForestSRC: variable importance (VIMP) with subsampling inference vignette. 2021.
80. Ishwaran H, Kogalur UB. Package 'randomForestSRC'. 2023; 1–128. https://www.randomforestsrc.org/ https://ishwaran.org/

81. Jobbagy EG, Jackson RB. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol Appl. 2000;10(2):423–36.
82. Siddique IA, Grados D, Chen J, Lærke PE, Jørgensen U. Soil organic carbon stock change following perennialization: a meta-analysis. Agron Sustain Dev. 2023;43(5):58. https://doi.org/10.1007/s13593-023-00912-w.
83. Abdalla M, Hastings A, Cheng K, Yue Q, Chadwick D, Espenberg M, et al. A critical review of the impacts of cover crops on nitrogen leaching, net greenhouse gas balance and crop productivity. Glob Chang Biol. 2019;25(8):2530–43. https://doi.org/10.1111/gcb.14644.
84. Rodríguez-Albarracín HS, Demattê JAM, Rosin NA, Contreras AED, Silvero NEQ, Cerri CEP, et al. Potential of soil minerals to sequester soil organic carbon. Geoderma. 2023;436:116549.
85. dos Reis AMH, Teixeira WG, Fontana A, Barros AHC, de Victoria DC, Vasques GM, et al. Hierarchical pedotransfer functions for predicting bulk density in Brazilian soils. Sci Agric. 2024. https://doi.org/10.1590/1678-992x-2022-0255.
86. Georgiou K, Jackson RB, Vindušková O, Abramoff RZ, Ahlström A, Feng W, et al. Global stocks and capacity of mineral-associated soil organic carbon. Nat Commun. 2022;13(1):3797.
87. Benke M, Mermut A, Shariatmadari H. Retention of dissolved organic carbon from vinasse by a tropical soil, kaolinite, and Fe-oxides. Geoderma. 1999;91(1–2):47–63.
88. Schaefer CEGR, Fabris JD, Ker JC. Minerals in the clay fraction of Brazilian Latosols (Oxisols): a review. Clay Miner. 2008;43(1):137–54.
89. Ferreira MM, Fernandes B, Curi N. Influência da mineralogia da fração argila nas propriedades físicas de latossolos da região sudeste do Brasil. Rev Bras Ciência do Solo. 1999;23(3):515–24.
90. Uehara G, Ikawa H, Sherman GD. Desilication of halloysite and its relation to gibbsite formation. Pacific Sci. 1966;20(1):119–24.
91. Al-Qinna MI, Jaber SM. Predicting soil bulk density using advanced pedotransfer functions in an arid environment. Trans ASABE. 2013;56(3):963–76.
92. Choudhury BU, Santra P, Singh N, Chakraborty P. Development of land-use-specific pedotransfer functions for predicting bulk density of acidic topsoil in eastern Himalayas (India). Geoderma Reg. 2023;34:e00671.
93. Rodríguez-Lado L, Rial M, Taboada T, Cortizas AM. A pedotransfer function to map soil bulk density from limited data. Procedia Environ Sci. 2015;27:45–8.
94. Zheng G, Jiao C, Xie X, Cui X, Shang G, Zhao C, et al. Pedotransfer functions for predicting bulk density of coastal soils in East China. Pedosphere. 2023;33(6):849–56. https://doi.org/10.1016/j.pedsph.2023.01.014.
95. Pessoa TN, Bovi RC, Nunes MR, Cooper M, Uteau D, Peth S, et al. Clay mineral composition drives soil structure behavior and the associated physical properties in Brazilian Oxisols. Geoderma Reg. 2024;38:e00837.
96. Lorenz K, Lal R. The depth distribution of soil organic carbon in relation to land use and management and the potential of carbon sequestration in subsoil horizons. Adv Agron. 2005;88:35–66.
97. Araujo MA, Zinn YL, Lal R. Organic carbon retention in tropical highlands. Geoderma. 2017;300:1–10. https://doi.org/10.1016/j.geoderma.2017.04.006.
98. Dalmolin RSD, Gonçalves CN, Dick DP, Knicker H, Klamt E, Kögel-Knabner I. Organic matter characteristics and distribution in Ferralsol profiles of a climosequence in southern Brazil. Eur J Soil Sci. 2006;57(5):644–54. https://doi.org/10.1111/j.1365-2389.2005.00755.x.
99. de Souza JJLL, Souza BI, Xavier RA, Cardoso ECM, de Medeiros JR, da Fonseca CF, et al. Organic carbon rich-soils in the brazilian semiarid region and paleoenvironmental implications. CATENA. 2022;212:106101.
100. Obidike-Ugwu EO, Ogunwole JO, Eze PN. Derivation and validation of a pedotransfer function for estimating the bulk density of tropical forest soils. Model Earth Syst Environ. 2023;9(1):801–9. https://doi.org/10.1007/s40808-022-01531-2.
101. Akpa SIC, Ugbaje SU, Bishop TFA, Odeh IOA. Enhancing pedotransfer functions with environmental data for estimating bulk density and effective cation exchange capacity in a data-sparse situation. Soil Use Manag. 2016;32(4):644–58. https://doi.org/10.1111/sum.12310.
102. Moquedace CM, Baldi CGO, Siqueira RG, Cardoso IM, de Souza EFM, Fontes RLF, et al. High-resolution mapping of soil carbon stocks in the western Amazon. Geoderma Reg. 2024;36:e00773.
103. Pereira OJR, Montes CR, Lucas Y, Melfi AJ. Evaluation of pedotransfer equations to predict deep soil carbon stock in tropical podzols compared to other soils of the Brazilian amazon forest. In: Hartemink AE, Minasny B, editors. Digital soil morphometrics. Cham: Springer International; 2016. p. 331–49.
104. Wiesmeier M, Urbanski L, Hobley E, Lang B, von Lützow M, Marin-Spiotta E, et al. Soil organic carbon storage as a key function of soils - a review of drivers and indicators at various scales. Geoderma. 2019;333:149–62.
105. Nemes A, Timlin DJ, Pachepsky YA, Rawls WJ. Evaluation of the Rawls et al. (1982) pedotransfer functions for their applicability at the U.S. national scale. Soil Sci Soc Am J. 2009;73(5):1638–45. https://doi.org/10.2136/sssaj2008.0298.
106. Fujisaki K, Perrin A, Desjardins T, Bernoux M, Balbino LC, Brossard M. From forest to cropland and pasture systems: a critical review of soil organic carbon stocks changes in Amazonia. Glob Chang Biol. 2015;21(7):2773–86. https://doi.org/10.1111/gcb.12906.
107. Cerri CEP, Easter M, Paustian K, Killian K, Coleman K, Bernoux M, et al. Predicted soil organic carbon stocks and changes in the Brazilian Amazon between 2000 and 2030. Agric Ecosyst Environ. 2007;122(1):58–72.
108. Cardoso EL, Silva MLN, Silva CA, Curi N, de Freitas DAF. Estoques de carbono e nitrogênio em solo sob florestas nativas e pastagens no bioma Pantanal. Pesqui Agropecuária Bras. 2010;45(9):1028–35.
109. Kumar S, Getirana A, Libonati R, Hain C, Mahanama S, Andela N. Changes in land use enhance the sensitivity of tropical ecosystems to fire-climate extremes. Sci Rep. 2022;12(1):1–11. https://doi.org/10.1038/s41598-022-05130-0.