MAPBIOMAS
[SOIL]

Algorithm Theoretical Basis Document (ATBD)

MapBiomas Soil

Collection 3

General Coordinator
Ferreira Jr., Laerte Guimarães

Scientific and Technical Coordinators
Horst, Taciara Zborowski
Samuel-Rosa, Alessandro

Support and Management
Azevedo, Tasso
Shimbo, Julia
Rosa, Marcos

May, 2026

Team

Training Field Soil Data

Anjos, Marcos Alexandre dos
Barth, Pedro Estevan
Capoane, Viviane
Cassamo, Delara Mariyamo Ibraimo
Dalzotto, Ana Vitoria
Huf dos Reis, Aline Mari
Kerber, Débora Liriel
Peruzzi, Vitor
Rossi, Gabriela Divino
Samuel-Rosa, Alessandro
Sobucki, Lisiane
Vergani, Felipe Brun

Space-Time Soil Modelling

Cardoso, Marcos Vinicius Souza
Dias, Mariana
Guelere, Rafael Ribeiro
Horst, Taciara Zborowski
Luz, Thalita da
Neto, Eduardo Carvalho da Silva
Oliveira, Fabrício Correia de
Oliveira, Monalisa
Parizzi, Talita Nogueira Terra
Penteado, Any Beatriz Moreira
Pretto, Ana Caroline
Rocha, Anderson Sandro da
Rosa, Eduardo Reis
Rosin, Luiz Felipe
Santos, Erli Pinto dos
Sárkány, Guilherme Attila Ribeiro
Silva, Barbara Costa
Silva, Luciana da Luz
Silva, Wallace Vieira da
Souza, Deorgia Tayane Mendes de
Vasconcelos, Vinícius
Weber, Eliseu

May, 2026

SUMMARY

To cite this document in your publications, please use the following format:

> Samuel-Rosa, Alessandro; Horst, Taciara Zborowski; Anjos, Marcos Alexandre dos; Santos, Erli Pinto dos; Silva, Barbara Costa da; Silva, Wallace Vieira da; Rocha, Anderson Sandro da; Huf dos Reis, Aline Mari; Sobucki, Lisiane; Cardoso, Marcos Vinicius Souza; Dias, Mariana; Pretto, Ana Caroline; Rosa, Eduardo Reis; Oliveira, Fabrício Correia de; Peruzzi, Vitor; Sárkány, Guilherme Attila Ribeiro; Vergani, Felipe Brun; Silva, Luciana da Luz; Weber, Eliseu; Capoane, Viviane; Souza, Deorgia Tayane Mendes de; Azevedo, Tasso; Shimbo, Julia; Rosa, Marcos; Ferreira Jr., Laerte Guimarães, 2025, "**MapBiomas Soil Brazil - Algorithm Theoretical Basis Document (ATBD) - Collection 3 (Beta)**", https://doi.org/10.58053/MapBiomas/OUQCLO, MapBiomas Data, V1

To cite the data collections in your publications, please use one of the following format:

> MapBiomas, 2025, **"Series of Annual Maps of Soil Organic Carbon Stock in Brazil 0-30 cm (1985-2024) – MapBiomas Soil Collection 3 (beta)"**, https://doi.org/10.58053/MapBiomas/2LUSVQ, MapBiomas Data, V2

> MapBiomas, 2025, **"Soil Particle Size and Texture Maps for Brazil 0-100 cm – MapBiomas Soil Collection 3 (beta)"**, https://doi.org/10.58053/MapBiomas/9ORUPF, MapBiomas Data, V2

> MapBiomas, 2025, **"Soil Depth Maps to Stoniness Thresholds for Brazil 0-100 cm – MapBiomas Soil Collection 3 (beta)"**, https://doi.org/10.58053/MapBiomas/1JGPIU, MapBiomas Data, V2

> MapBiomas, 2025, **"Training Field Soil Data for Mapping of Soil Particle Size Distribution (0-100 cm) in Brazil (MapBiomas Soil Collection 3, Beta Version)"**, https://doi.org/10.60502/SoilData/OXSR2N, SoilData, V3

> MapBiomas, 2025, **"Training Field Soil Data for Annual Mapping of Soil Organic Carbon Stock (0–30 cm) in Brazil, 1985–2024 (MapBiomas Soil Collection 3, Beta Version)"**, https://doi.org/10.60502/SoilData/IUZOAK, SoilData, V1

Executive Summary

Launched in 2021, the MapBiomas Soil project aims to explore the spatiotemporal dynamics of soil properties, focusing on soil organic carbon (SOC) stocks and their interplay with land cover and land use changes across Brazil. In 2023, the project delivered its first public release, a collection consisting of annual SOC stock maps for the top 30 cm of Brazilian soils. These maps, in tons per hectare (t/ha), spanned the period from 1985 to 2021. In Collection 2, MapBiomas Soil updated the SOC stock maps to span the period 1985-2023, and expanded its scope, presenting new maps of soil particle-size distribution and texture for three layers down to 30 cm.

For Collection 3, MapBiomas Soil updated the series of annual maps of SOC stocks, adding the year of 2023, as well as the particle-size distribution and texture maps. The key innovation is that soil texture data is now available for 10-cm thick layers from the topsoil down to 100 cm depth. Collection 3 also introduces maps of depth to stoniness thresholds, defined as the depth where the volume of the coarse fraction (> 2 mm) becomes dominant (>50%) or extreme (>90%). These products were generated from an enhanced dataset that integrates additional field samples and updated environmental covariates. To assess map quality, the Collection employs resampling methods and presents 'area-of-applicability' maps based on the geographic distance to the nearest training samples.

The maps in the MapBiomas Soil project were generated using regression models that establish correlations between target soil properties, derived from field measurements, and spatially explicit environmental covariates. These relationships were computed using the Random Forest and Gradient Boosted Trees machine learning algorithms, implemented within the Google Earth Engine platform. The resulting datasets and maps are openly accessible for download and visualization through the MapBiomas Soil platform (https://plataforma.brasil.mapbiomas.org). Users can explore mass and stocks of SOC as well as soil texture (sand, silt, and clay content, expressed in percentages) across various territorial divisions and land cover/use classifications.

This Algorithm Theoretical Basis Document (ATBD) outlines the methodology behind MapBiomas Soil Collection 3 and describes the datasets and processes involved in its creation. The datasets and processing codes are available on the project's GitHub repository (https://github.com/mapbiomas/brazil-soil), providing transparent access to the scientific community and the public.

1. Introduction

1.1 Scope

This document provides a comprehensive overview of the methodologies and processes used in the development of MapBiomas Soil Collection 3. It details the steps involved in processing field soil data, integrating environmental data, training models, and generating spatial, vertical, and temporal predictions. Additionally, the document includes historical context, background information, descriptions of the datasets and predictor variables used, and the quality assessment methods employed.

Each step of the methodological workflow is described to ensure transparency and facilitate the reproducibility of the processes and results. By documenting these procedures, the document serves as a critical resource for understanding the scientific and technical foundations of the MapBiomas Soil project.

1.2 Overview

The MapBiomas project was launched in July 2015 with the goal of enhancing the understanding of land cover and land use (LCLU) dynamics in Brazil. Since its inception, the MapBiomas network has produced annual LCLU maps for the country, utilizing Landsat satellite imagery at a spatial resolution of 30 meters. These maps provide a detailed and comprehensive record of Brazil's land use patterns, offering insights into the historical changes in native vegetation and anthropogenic activities such as agriculture, pasture, and urban expansion.

Each year, MapBiomas releases an updated collection of annual maps, spanning from 1985 to the present. These collections incorporate new classes, methodological improvements, and the latest annual data. This ongoing effort has enabled the network to deliver a quantitative history of land use changes, supporting research and policy-making related to environmental conservation and sustainable development.

In addition to LCLU data, MapBiomas has expanded its portfolio to include other critical products, such as maps of fire scars, water surfaces, deforestation alerts, and drivers of native vegetation degradation. Among these initiatives is the MapBiomas Soil project, launched in 2021, which aims to annually produce updated map collections that reveal the spatiotemporal dynamics of soil properties, particularly soil organic carbon (SOC) stocks, and their interactions with land cover and land use changes across Brazil.

MapBiomas Soil Collection 2 represented the second iteration of the project, building upon the beta collection released in 2023. The collection introduced new products, including maps of soil particle size distribution and texture classification, alongside updated annual maps of SOC stocks. The SOC stock maps covered the period from 1985 to 2023, providing a time series of estimates for the top 30 cm of soil, expressed in tons per hectare (t/ha). The particle size distribution maps were static, with a nominal reference year of 1990, and detailed the percentage of clay, sand, and silt across three soil layers (0–10 cm, 10–20 cm, and 20–30 cm). These distributions were further aggregated into soil texture classes using three classification schemes with 5, 8, and 13 classes.

Table 1. Evolution of the MapBiomas Soil Collections.

Collection* (date of launch)	Soil property	Measurement unit	Soil layer	Temporal reference	Version
C1 (21 July 2023)	SOC stock	t/ha	0-30 cm	Annual series between 1985 and 2021	Beta
C2 (06 December 2024)	SOC stock	t/ha	0-30 cm	Annual series between 1985 and 2023	Beta
	Soil particle size distribution (clay, silt and sand content)	%	0-10 cm 10-20 cm 20-30 cm 0-20 cm 0-30 cm	Nominal reference year of 1990	Beta
	Soil texture classification (5, 8, and 13 classes)	-	0-10 cm 0-20 cm 0-30 cm	Nominal reference year of 1990	Beta
C3 (05 December 2025)	SOC stock	t/ha	0-30 cm	Annual series between 1985 and 2024	Beta
	Soil particle size distribution (clay, silt and sand content)	%	0-10 cm 10-20 cm 20-30 cm 30-40 cm 40-50 cm 50-60 cm 60-70 cm 70-80 cm 80-90 cm 90-100 cm	Nominal reference year of 1990	Beta
	Soil texture classification (5, 8, and 13 classes)	-	0-10 cm 0-20 cm 0-30 cm 20-40 cm 30-60 cm 60-100 cm	Nominal reference year of 1990	Beta
	(depth-to-) Stoniness (sensored at 100 cm)	cm	50% threshold 90% threshold	Nominal reference year of 1990	Beta

* C1: Collection 1; C2: Collection 2; C3: Collection 3; SOC: soil organic carbon stocks.

For Collection 3, the annual series of SOC stock maps and the particle size distribution and texture maps were refined and updated. Notably, the project's scope expanded to include soil texture data for 10-cm thick layers, reaching a depth of 100 cm. Furthermore, Collection 3 introduces maps for depth-to-stoniness thresholds, specifically for the depths where the volume of the coarse fraction (> 2 mm) is greater than 50% (dominant) or 90% (extreme). These new products were made possible by an enhanced dataset, which incorporated extra field samples and updated environmental covariates. The modeling approach for the new particle size distribution products utilizes the Gradient Boosted Trees (GBT) machine learning algorithm, a methodological evolution from the Random Forest (RF) algorithm used in the previous collection. To assess quality, the collection provides bootstrapped error statistics and 'area-of-applicability' maps that illustrate the spatial reliability of the model predictions based on the proximity to the training samples (Table 1).

The static and annual soil property maps are generated using regression models based on machine learning algorithms. These models establish numerical relationships between field-measured soil properties and environmental covariates that cover the entire Brazilian territory over the mapping period. Field soil samples, obtained from the SoilData repository (https://soildata.mapbiomas.org/), span the period from 1960 to 2024 and undergo rigorous consistency checks, standardization, and harmonization before being used for model training. The processed data, including harmonized field samples and derived soil properties, are also released back into the SoilData repository, ensuring transparency and enabling further research and validation by the scientific community.

Environmental covariates represent soil-forming factors and drivers of change, serving as predictor variables in the machine learning models. Static covariates include soil properties, climate classification, phytophysiognomy, structural provinces, morphometric variables, and biome classifications. Dynamic covariates incorporate land use and land cover data, as well as spectral indices derived from satellite imagery. All covariates are sourced from open-access spatial databases and standardized to a consistent geographic framework.

The processing of covariates, model training, and soil property predictions are conducted using Google Earth Engine, supplemented by R and Python for less computationally intensive tasks. All final products are stored on the Google Cloud Storage Platform and made available under a Creative Commons Attribution (CC-BY) license, ensuring open access.

1.3 Region of Interest

MapBiomas Soil generates static and dynamic maps of soil properties for the entire Brazilian territory, encompassing the land areas of the country’s six official biomes: Amazon, Atlantic Forest, Caatinga, Cerrado, Pampa, and Pantanal (Figure 1).

The region of interest is the emerged soil area, explicitly excluding permanently submerged areas. For Collection 3, this area is delineated using a refined methodology that utilizes the project's own coarse fraction data (stoniness maps) to identify and exclude areas classified as non-soil. This mask is applied both spatially (e.g., to identify rocky outcrops on the surface based on the coarse fragment threshold) and vertically in depth, where estimates of particle size distribution, which extend up to 100 cm, are discontinued when the volume of the coarse fraction reaches 95%. Additionally, soil properties are not mapped in areas classified

as permanently submerged—which fall under the Water class (Class 26 and subclasses 31 and 33) of MapBiomas Collection 10.

Figure 1. Brazilian biomes used in the MapBiomas project to generate the Collection 2 of static and dynamic soil property maps (IBGE, 2019a).

1.4 Key Applications

Understanding SOC stocks, particle size distribution, soil texture, and stoniness—along with their spatial, vertical, and temporal dimensions—is essential for implementing public policies, effective conservation programs, and sustainable natural resource management. These soil properties play a critical role in mitigating climate change impacts, improving agricultural productivity, and supporting ecosystem health. The annual maps of SOC stocks and static maps of particle size distribution, soil texture, and stoniness produced by MapBiomas Soil have broad applications across public and private sectors, including:

SOC Stock Maps:

- **Preventing SOC Loss**: Identifying areas and factors (natural or anthropogenic) responsible for reducing SOC stocks in Brazil.
- **Guiding Conservation Policies**: Highlighting soils with the highest SOC stocks, particularly organic soils, which are vulnerable to land-use changes and require targeted conservation efforts to prevent carbon loss to the atmosphere.
- **Supporting Land Management**: Providing high-quality spatial-temporal information to help land users implement and maintain soil management practices that protect and increase SOC under local conditions.
- **Evaluating Public Policies**: Supplying data to assess the effectiveness of policies aimed at restoring degraded areas, a key goal for 2030.
- **Monitoring Rural Lands**: Enabling the evaluation of rural properties and their compliance with conservation and sustainability standards.
- **Assessing Climate Policies**: Supporting the evaluation of climate change mitigation and adaptation programs.
- **Carbon Credit Markets**: Identifying areas with high potential for carbon credit market entry.
- **Funding and Credit Evaluations**: Monitoring the outcomes of credit line applications and evaluating funding requests for climate change mitigation and adaptation projects.
- **Advancing Research**: Facilitating new and comprehensive studies on SOC dynamics and their interactions with land use and climate.

Particle Size Distribution and Soil Texture Maps:

- **Agricultural Planning**: Providing critical information on soil texture (clay, silt, and sand content) to guide crop selection, irrigation strategies, and soil management practices tailored to local conditions.
- **Erosion Risk Assessment**: Identifying areas with high erosion risk based on soil texture, enabling the implementation of targeted soil conservation measures.
- **Water Management**: Supporting water retention and drainage planning by understanding soil texture and its influence on water infiltration and storage.
- **Land Use Suitability**: Informing land use decisions by identifying areas suitable for specific agricultural, forestry, or conservation activities based on soil texture and particle size distribution.
- **Research and Education**: Providing a valuable resource for researchers and educators studying soil properties, land use, and ecosystem dynamics.

Soil Stoniness Maps:

- **Effective Soil Depth and Volume**: These maps serve as proxies for effective soil depth, defining the volume of fine earth, which is crucial for root development and accurate resource estimation.
- **Water Management & Hydrology**: They allow for a better understanding of soil processes, such as water retention and infiltration, and can be used to feed hydrological models to refine regional water storage capacity estimates.
- **Climate Risk Assessment**: The maps provide fundamental data for climate risk assessment concerning the soil's water storage capacity.

- **Land Use Suitability**: They inform land-use planning and conservation decisions by considering stoniness as a significant physical limitation for certain uses and an indicator of environmental fragility.

Enabling these applications is a key goal of MapBiomas Soil. User feedback, suggestions, and contributions of additional data to address any inconsistencies in this Collection 3 are essential to ensure the quality and reliability of the products. By meeting these objectives, MapBiomas Soil will contribute to climate change mitigation, food security, soil health preservation, and the reduction of soil degradation, while supporting sustainable land use and management practices across Brazil.

2. Background

This section addresses complementary contextual and critical information relevant to understanding the MapBiomas Soil products and methods used in the map collections.

2.1 Context

Soil plays a critical role in the global climate system, storing more carbon in its top 100 cm than the atmosphere (Lal, 2013). As such, soil can act as either a threat or a solution to climate change. Land-use changes and inadequate soil management practices have turned soil into a net source of greenhouse gases. However, conservation practices, such as restoring degraded areas and improving soil management, can reverse this trend by sequestering atmospheric carbon dioxide and storing it as soil organic carbon (SOC). These practices not only mitigate climate change but also enhance soil quality, as SOC positively influences soil properties such as water retention, nutrient availability, and erosion resistance (IPCC, 2014).

Understanding the location and temporal dynamics of SOC stocks across Brazilian biomes is essential for producing accurate estimates of the country's greenhouse gas emissions and removals (SEEG, 2022). Such estimates are critical for guiding public policies and programs aimed at achieving Brazil's emission reduction targets under the Paris Agreement (Brasil, 2015). Brazil's economy is heavily dependent on land use, yet the country lacks detailed spatial-temporal information on soil properties. Previous efforts to map SOC stocks have largely overlooked the temporal dynamics driven by land-use and climate changes, often presenting estimates as static maps or absolute values (Poggio et al., 2021; Vasques, Coelho, Dart, Baca, et al., 2021). This limitation stems primarily from the scarcity of field data required for training predictive models. Despite the large volume of soil data produced in Brazil (Camargo et al., 2010), much of it remains difficult to locate, access, and reuse (Hanson et al., 2011), efforts to address these challenges began with the establishment of the Brazilian Soil Data Repository (Samuel-Rosa et al., 2020). Over the following years, soil data from thousands of observations were rescued, curated, and made accessible through contributions from public and private organizations. This foundational work enabled the creation of MapBiomas Soil in 2021, which launched the first series of annual SOC stock maps in 2023, marking a significant milestone in the annual mapping of SOC for all Brazilian biomes.

MapBiomas Soil aims to deepen our understanding of how land-use and land-cover changes impact the spatial and temporal variation of SOC stocks across Brazil's biomes. The initial production of the beta collection was made possible through collaboration with public and private organizations, which expanded SOC data availability. The project's objective is to deliver annual updates to this historical series of SOC stock maps, including measures of local and global uncertainty, to further enhance data quality and utility. By providing high-resolution, dynamic maps of SOC stocks, MapBiomas Soil supports informed decision-making for sustainable land management and climate change mitigation in Brazil.

2.2 Historical Perspective

2.2.1 Legacy Soil Data

Brazilian scientists have generated one of the largest volumes of soil data in tropical regions (Camargo et al., 2010). However, inefficient storage and sharing practices have limited the discoverability and reusability of these datasets. Many are available only in printed or poorly scanned documents (Kämpf, 1971), while others are stored in inaccessible formats or lack critical metadata (Chagas et al., 2004; Ottoni et al., 2014). These challenges hinder research reproducibility and represent an inefficient use of limited investments in soil science.

Globally, efforts to compile and share legacy soil data have gained momentum. Initiatives like the Africa Soil Profiles Database, the Soil and Landscape Grid of Australia, and the World Soil Information Service (WoSIS) have demonstrated the value of centralized soil data repositories. WoSIS, managed by ISRIC, has enabled the creation of global soil maps through machine learning, such as those on the SoilGrids platform, widely used by organizations like the Intergovernmental Panel on Climate Change (IPCC) (Batjes et al., 2020; Hengl et al., 2017; Poggio et al., 2021).

In Brazil, the Brazilian Agricultural Research Corporation (Embrapa) and the Brazilian Institute of Geography and Statistics (IBGE) have led efforts to compile and organize soil data, particularly from the Radambrasil Project. Embrapa's Brazilian Soil Information System (BDSolos) and IBGE's updates to soil profile classifications have safeguarded an estimated 10,000 soil profiles (Samuel-Rosa & Vasques, 2017). Initiatives by individuals and groups of researchers, such as the recovery of Radambrasil data by ESALQ researchers, the creation of the Brazilian Soil Spectral Library (BESB), and the compilation of the 1075 soil samples of Hydrophysical Database for Brazilian Soils (HYBRAS), have further contributed to soil data availability (Cooper et al., 2005; Ottoni et al., 2018; Sato, 2015).

To address these challenges, SoilData (https://soildata.mapbiomas.org/), a data repository of the MapBiomas Soil initiative, provides a centralized platform for compiling, organizing, and sharing soil data. SoilData accepts voluntary data deposits from soil data producers free of charge and proactively digitizes soil data from old reports, articles, surveys, and theses, making them available for reuse. This ongoing effort ensures continuous growth and accessibility of soil data. With over 300 datasets, hundreds of users, and more than a thousand downloads, SoilData represents a significant advancement in soil data management in Brazil.

SoilData not only supports the development of MapBiomas Soil mapping products but also enables the creation of specialized databases that can be used to enhance the Brazilian Soil

Classification System and contribute to initiatives like the Brazilian National Soil Survey and Interpretation Program (PronaSolos). By increasing data findability and accessibility, SoilData fosters advancements in soil research, sustainable land management, and climate change mitigation, aligning with the goals of the MapBiomas Soil project.

2.2.2 SOC Stock Mapping

Global and national initiatives have produced SOC stock maps for Brazil using various techniques and approaches. These maps rely on field-measured SOC data and environmental covariates: auxiliary data representing soil-forming factors such as topography, hydrology, vegetation, and parent material. These covariates are critical for explaining the spatial distribution of SOC and modeling its stocks.

The first SOC stock map for Brazilian soils was developed two decades ago (Bernoux et al., 2002). This map, at a 1:1,000,000 scale, linked past vegetation and soil type data to estimate potential SOC stocks at 0–30 cm depth. It identified 75 soil-vegetation association categories, revealing regional variations in SOC stocks. For example, high SOC stocks in the Pantanal and northwest Amazon Basin were attributed to wet soils, while cooler climates in the south and semi-arid conditions in the northeast influenced SOC values. Dense forests in the Amazon showed higher SOC stocks than open forests, and soil type played a significant role in the Cerrado. The map estimated an average SOC stock ranging from 15.1 to 417.8 t/ha, totaling approximately 36.4 Gt across Brazil.

Recent national initiatives, such as those by Gomes et al. (2019) and Embrapa under the PronaSolos program (Vasques, Coelho, Dart, Baca, et al., 2021; Vasques et al., 2017), have advanced SOC stock mapping using digital soil mapping techniques. These efforts prioritize large-scale trends but often combine data from different periods due to limited temporal coverage. For instance, Gomes et al. (2019) used 8,227 soil profiles and 37,693 samples from the RADAMBRASIL project (1970s–1980s) and 74 environmental covariates to produce a 1 km resolution map. This map estimated 36 Gt of SOC in the top 30 cm, with the Amazon storing the highest stocks (36.1 Gt at 0–100 cm), while the Pantanal and Caatinga had the lowest stocks (0.77 Gt and 4.88 Gt, respectively). Embrapa's 2017 and 2021 maps, also at 1 km resolution, estimated 36 Gt of SOC in the 0–30 cm layer using less than 10,000 soil data points and about 40 covariates. These maps are publicly accessible through PronaSolos and Embrapa's Spatial Data Infrastructure.

At the global level, SoilGrids (https://soilgrids.org/), developed by ISRIC, produces SOC stock maps using machine learning models. Combining approximately 150,000 soil profiles and 158 covariates, SoilGrids provides SOC predictions at seven standard depths (0–200 cm). The 2016 version, with a 250 m resolution, included data from 5,086 Brazilian profiles (Poggio et al., 2021). However, uneven global data distribution limits its ability to capture fine-scale variations. Another global initiative, Soils Revealed (https://soilsrevealed.org/), visualizes the impact of soil management on SOC stocks. It highlights Brazil as a major SOC holder, with an estimated 31 Gt of SOC, but notes an average loss of 0.122 t/ha in the 0–30 cm layer between 2000 and 2018 due to land-use and land-cover changes.

The Global Soil Organic Carbon Map (GSOC), developed by the FAO under the Global Soil Partnership, represents the first global SOC stock map created through a participatory

process. Covering the 0–30 cm layer at 1 km resolution, the GSOC integrates national inventory data, published studies, and simulation models. Embrapa's SOC map was officially included as Brazil's contribution to the GSOC. Building on this, the FAO developed the Global Soil Organic Carbon Sequestration Potential Map (GSOCSeq), which estimates SOC sequestration potential using the RothC model. Brazil was identified as having the highest SOC sequestration potential under conservation practices, underscoring its critical role in climate change mitigation (FAO, 2020; Peralta et al., 2022).

While foundational, these national and global mapping initiatives were largely static or lacked the commitment to yearly updates with the granularity needed for local and regional analysis. This highlighted the critical need for a dynamic approach capable of monitoring annual changes in SOC stocks across Brazil. MapBiomas Soil was specifically designed to fill this gap. The project introduces annual maps of SOC stock (t/ha) for the top 30 cm of soil at a 30 m resolution, covering the period from 1985 to 2023. These maps are updated annually, providing a dynamic and detailed view of Brazil's SOC stocks. By accounting for the impacts of LULC changes, this approach offers an unprecedented tool for monitoring SOC dynamics and supporting sustainable land management practices across Brazil.

2.2.3 Soil Texture Mapping

Soil texture, defined by the relative proportions of sand, silt, and clay, is a key property influencing water retention, nutrient availability, erosion susceptibility, and agricultural productivity, as well as greenhouse gases emitted from soil (Adhikari et al., 2013; Bronick & Lal, 2005; Lal & Shukla, 2004). High-resolution soil texture mapping is essential for understanding soil behavior, supporting sustainable land management, and informing climate change mitigation strategies (Adhikari et al., 2013; Chagas et al., 2016; Mitran et al., 2024).

Historically, soil texture mapping in Brazil relied on conventional maps produced by Embrapa and IBGE at small cartographic scales (1:1,000,000 and, more recently, 1:250,000). These maps, based on field surveys, laboratory analyses, and landscape interpretation, integrate data from diverse sources and time periods. While valuable for regional and national planning, their coarse resolution limits their ability to capture fine-scale variations.

Under the PronaSolos program, Embrapa produced 90-m resolution maps of sand, silt, and clay content (g/kg) for Brazil at six depth intervals (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm), along with uncertainty maps (Vasques, Coelho, Dart, Baca, et al., 2021). These maps were generated using digital soil mapping techniques, integrating legacy soil data, freely available geospatial covariates, and open-source software. The dataset included 8,950 soil profiles for model training. However, the initiative faces two key limitations: it lacks a commitment to regular updates as new field data become available, and while the depth intervals align with international standards such as GlobalSoilMap (2015), they do not correspond to the 10 cm increments commonly required for practical applications.

Addressing the limitations of existing products in resolution, data density, and depth standardization was the primary motivation for developing the soil texture component of MapBiomas Soil. Building upon previous iterations, Collection 3 was developed using

machine learning algorithms and an enhanced dataset of field observations to generate high-resolution (30 m) maps of soil particle size distribution (clay, silt, and sand content) for ten 10-cm thick standardized depth layers, reaching a depth of 100 cm. Additionally, these maps are classified using three schemes (with five, eight, and thirteen textural classes) from the Brazilian System of Soil Classification, providing a valuable resource for agricultural planning, erosion risk assessment, and water management.

2.2.4 Soil Stoniness Mapping

Soil stoniness refers to the presence and abundance of coarse fragments (> 2 mm), including gravel, cobbles, and stones, within the soil profile. This attribute directly controls effective soil volume, influencing porosity, water retention capacity, hydraulic conductivity, erosion susceptibility, and mechanical resistance to root penetration and agricultural mechanization. Because coarse fragments reduce the volume of fine earth available for water and nutrient storage, stoniness is also a critical parameter for estimating SOC stocks. High-resolution stoniness mapping is therefore essential for land suitability assessment, hydrological and erosion modeling, territorial planning, and climate-related soil analyses (Cousin et al., 2003; Poesen & Lavee, 1994; Tetegan et al., 2015).

Historically in Brazil, information on soil stoniness has been restricted to conventional pedological surveys conducted by Embrapa, IBGE, and RADAMBRASIL. In these surveys, stoniness is described qualitatively in soil profile reports (IBGE, 2015; Santos et al., 2015) and associated with mapping units at medium or small cartographic scales. Although these datasets are fundamental for soil classification and regional planning, they do not provide a continuous, quantitative, and spatially explicit representation of coarse fragment distribution across the national territory. Stoniness has typically been recorded as a descriptive attribute or diagnostic feature of specific soil classes, which limits its direct integration into quantitative spatial modeling and large-scale environmental analyses.

The soil stoniness component of MapBiomas Soil Collection 3 introduces a national high-resolution (30 m) product representing the estimated depth (cm) to layers where coarse fragments exceed volumetric thresholds of 50% (dominant) and 90% (extreme) within the upper 100 cm of the soil profile. Rather than expressing stoniness solely as a surface percentage or qualitative class, this approach explicitly models the vertical occurrence of restrictive coarse-fragment layers. By estimating the depth to dominant and extreme stoniness conditions, the product provides a spatially explicit representation of physical rooting limitations and effective soil volume constraints for water storage and plant development.

3. Methodological Description

The collection of static and dynamic (annual) maps of soil properties for the Brazilian territory was generated using regression models. These models predict soil property values at unsampled locations based on the statistical relationships between field soil data (sampling points) and spatially exhaustive environmental covariates (raster maps). Model prediction performance is assessed using resampling techniques, which provide insights into the reliability of the predictions. Map algebra is then applied to generate soil texture and

(depth-to-) stoniness maps. The following sections describe each step involved in producing these static and dynamic maps of soil properties (Figure 2).

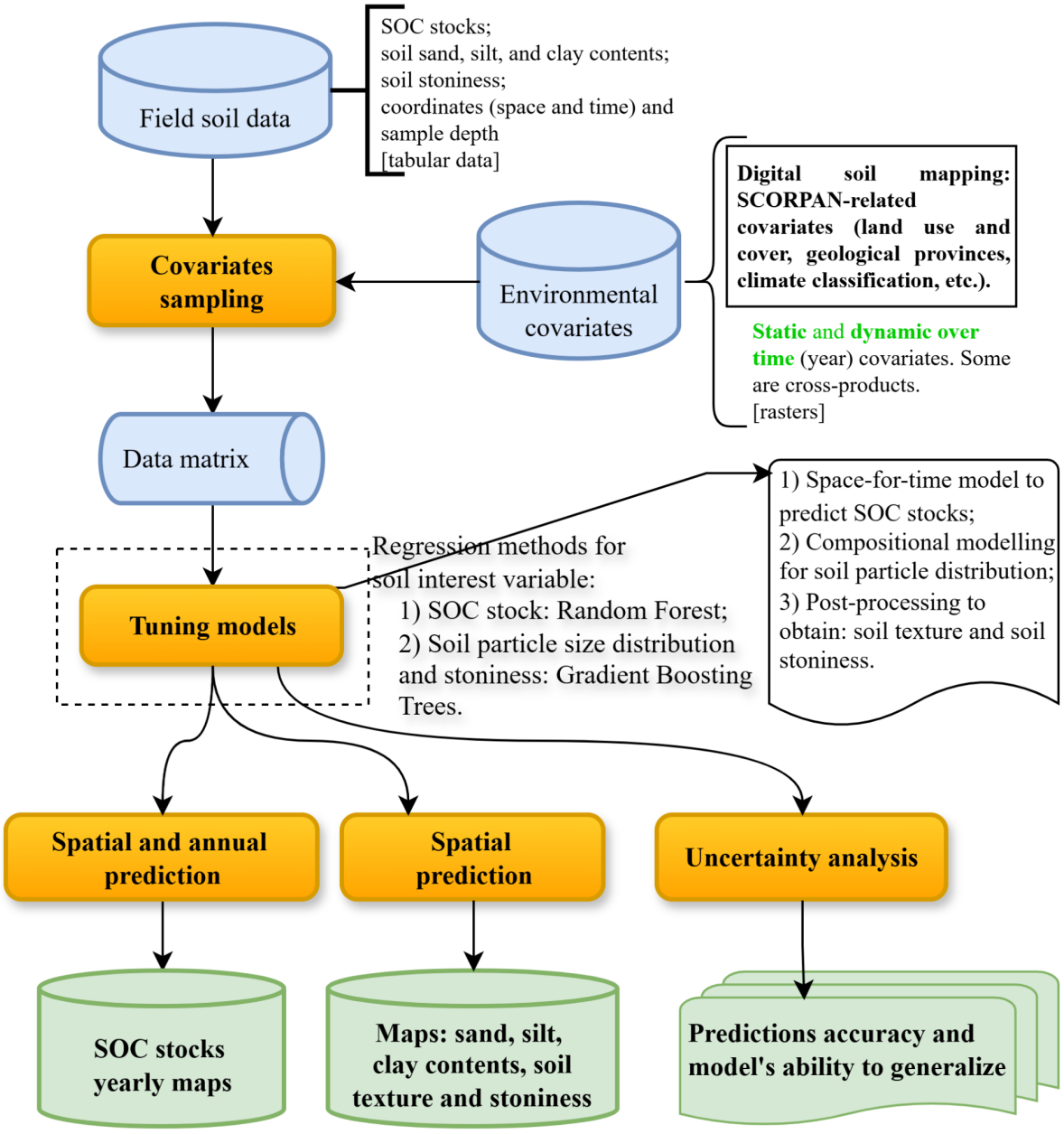

Figure 2. Flowchart of the method for generating annual soil organic carbon (SOC) stock maps and static soil particle size distribution (PSD), texture, and stoniness maps for Brazil.

3.1 Key Methodological Evolutions

Collection 3 incorporates more robust data processing, an expanded set of environmental predictors, and more sophisticated modeling techniques than Collection 2. These enhancements improve the accuracy of the SOC stock and texture maps and support the development of the entirely new maps of soil (depth-to-) stoniness. The key methodological evolutions are summarized below.

- **Enhanced Data Foundation and Processing:**
    - **Expanded Training Data:** The training dataset was substantially expanded to 14,983 unique points for modeling soil PSD (Collection 2 was based on 11,522 points). Similarly, for SOC modeling, the training dataset was increased to 16,068 unique points (Collection 2 used 12,666 unique points).
    - **Data Augmentation:** The training data was enhanced using tacit samples from dunes, beaches, sand banks, and from rock outcrops to improve model performance in areas of very high sand content and very low SOC stock.
- **Expanded Feature Space:**
    - **More Covariates:** The feature space increased from 106 covariates in Collection 2 to 185 covariates (158 static and 27 dynamic) in Collection 3. From this pool, 168 covariates were used for carbon modeling.
    - **New Static Variables:** Distance-to-class (rock outcrops and beach/dune/sandy areas) were derived from the MapBiomas Collection 10 to incorporate spatial gradients in the models. Categorical variables (pedology, lithology, and biome) were cross-referenced to create co-occurrence bands to capture more specific environmental interactions.
- **Advanced Modeling and Evaluation Strategies:**
    - **New Soil Texture Workflow:** A new workflow was developed for the soil texture product, featuring independent Gradient Boosted Trees (GBT) models that predict PSD for each 10-cm layer down to 100 cm depth.
    - **Back to a Global SOC Model:** The SOC modeling strategy went back from three distinct regional models to using a single, nationwide model such as in Collection 1.
    - **Dealing with Censored Data.** The modeling of SOC stocks was improved to use cumulative SOC stocks from soil layers up to 30 cm as training data, considering the censored nature of the soil data.
    - **Feature Selection and Sample Filtering.** Feature selection was optimized by minimizing collinearity among predictors, while training samples were filtered to exclude pedologically implausible feature space and soil property combinations.

3.2 Field Soil Data

The foundation for the mapping products of Collection 3 are two comprehensive, harmonized, analysis-ready field soil datasets, which are key project deliverables in their own right. These datasets were developed by sourcing data from datasets published in the SoilData repository (Table 2). Those datasets were deposited voluntarily by their authors or rescued/digitized by the SoilData team.

Table 2. Soil data sourced from the SoilData repository to prepare the training samples used across MapBiomas Soil collections.

Collection	Datasets	Points	Georeferenced points	Samples
1.0	258	12729	12139	44158
2.1	262	18415	14990	60862
3.0	266	19421	16515	61139

Each dataset was individually submitted to quality checks, both automated and manual, targeting key features such as the spatial coordinates, year of sampling, soil classification, sampling depth (cm), layer name, carbon content (g/kg), particle size distribution (g/kg), quantity of coarse fragments (g/kg), bulk density (g/cm^3), pH, and cation exchange capacity ($cmol_c/kg$). Manual corrections, adjustments, and the imputation of missing values using spline interpolation, downward extrapolation, and regression models were implemented when feasible, based on expert decisions informed by literature review and data analysis.

Random forest was used as a regression model for imputing missing data using soil properties and environmental covariates as predictors (Wright & Ziegler, 2017). Missing values in the covariates were handled using the missingness-in-attributes (MIA) approach (Twala et al., 2008). The best set of hyperparameters for each random regression forest was selected via grid search. The hyperparameters optimized were: the number of trees (num.trees), the number of covariates tried at each node split (mtry), the minimum size of a node (min.node.size), and the maximum depth of a tree (max.depth). This optimization used resampling methods based on the out-of-bag error statistics (Hastie et al., 2009).

All corrections implemented were mapped to and documented in the source dataset published in the SoilData repository. All procedures were conducted using R and Python open source software packages and libraries. The resulting source code is openly available at https://github.com/mapbiomas/brazil-soil.

3.2.1 Soil Particle Size Fractions

Samples from 0 to 100 cm depth, referencing the mid-layer depth, were selected to model the particle size distribution of the soil. These samples contained data on the four particle size fractions: clay, silt, sand, and coarse fragments (i.e., >2 mm diameter), which were either original measurements or imputed values. Imputation of missing values was performed using spline interpolation and downward extrapolation (for all fractions) and regression models (for coarse fragments only). To account for the compositional nature of the particle size fraction data, the mass percentages were converted for each layer to additive log-ratios: log(sand/clay), log(silt/clay), and log(gravel/clay) data (Samuel-Rosa et al., 2013). Samples without spatial coordinates were discarded.

3.2.1 Soil Organic Carbon Stock

Samples spanning from 0 to 100 cm depth were selected to model the soil organic carbon stock, with properties referenced to the mid-layer depth. These samples contained data on soil organic carbon content (g/kg), bulk density (g/cm^3), and mass percentage of coarse

fragments (%), all of which were either original measurements or imputed values. Imputation of missing values was performed using spline interpolation and downward extrapolation for all variables, and using regression models specifically for bulk density. The mass percentage of coarse fragments (%) was converted to volume percentage using the soil bulk density (g/cm³) and estimates of rock bulk density based on the literature (Sharma, 1997) (Table 3). Rock density was assigned to each training point by intersecting it with a map of simplified geological units ("subprovinces"), derived from the structural provinces data (IBGE, 2019b). Details on these simplified geological classes are described in the Environmental Covariates section.

Table 3. Approximate densities of major geological units (or approximate dry bulk density) based on literature values (Sharma, 1997).

Code of the Main Geological Unit	Rock type	Average density (or dry bulk density) (g/cm³)
1	Sediments	2.00
2	Sedimentary rocks	2.30
3	Volcanic igneous rocks	2.74
4	Plutonic igneous rocks	2.66
5	Metamorphic rocks	2.70
6	Sedimentary + igneous + metamorphic rocks	2.60
7	Sedimentary + igneous rocks	2.57
8	Igneous + metamorphic rocks	2.70

Georeferenced and timestamped samples were selected, and their SOC density (g/cm^3) was computed as: SOCdensity = carbon_content × (1 - coarse_fragments) × bulk_density. Bias correction was applied to National Forest Inventory samples using per-biome quantile mapping to match non-IFN sample distribution (Gudmundsson, 2025). The SOC stock (g/m²) was then computed as: SOCstock = SOCdensity × layer_thickness. Cumulative SOC stocks downward were then computed for all consecutive layers starting from the topsoil down to 30 cm depth.

3.3 Environmental Covariates

A comprehensive dataset of environmental covariates was assembled from open-access spatial databases to train regression models and predict soil properties at unsampled locations and times. The initial pool of covariates was selected based on pedological knowledge to represent the key factors influencing soil formation: climate, organisms, topography, and parent material (Table 4).

The covariates were incorporated into the modeling framework according to their temporal coverage, which could be either static (a single dataset for the entire series) or dynamic

(annual resolution). Static covariates, such as topography and geological classification, capture long-term influences on soil properties, while dynamic covariates, such as land-use and land-cover changes or vegetation indices, reflect temporal variations that drive soil change.

Table 4. Number of static and dynamic covariates employed across MapBiomas Soil collections.

Collection	Static covariates	Dynamic covariates	Total
1.0	62	15	77
2.1	85	21	106
3.0	137	23	160

3.3.1 Static Covariates

Static environmental covariates representing soil-forming factors were compiled from various sources. Existing maps of soil properties and classes were sourced from SoilGrids and FAO GBSmap, with multi-depth layers merged to match target prediction depths. National-scale datasets, including soil class maps, climate classification, and geological information from IBGE, complement this set of predictors. Probability maps of World Reference Base (WRB) soil classes support both individual and composite representations. To improve data quality, layers were smoothed, blended, and standardized. Particle size distribution maps generated within the project are part of this covariate set.

Land surface variables derived from Geomorpho90m and a digital elevation model (MERIT DEM) provide terrain attributes such as elevation, slope, curvature, and hydrological indices. Long-term water dynamics are represented through a water surface recurrence layer derived from MapBiomas Water Collection 3, expressing the frequency of water occurrence over the historical series.

Climate conditions are represented by the Koppen classification, while Brazilian biomes, phyto ecological regions, and structural provinces describe broad environmental patterns. A refined version of the geological provinces dataset, referred to as sub-provinces, captures the dominant lithological composition (e.g., sedimentary, metamorphic, vulcanic). Combinations between categorical layers (e.g., pedology, lithology, and biome) generate co-occurrence bands that represent specific ecological–geological contexts.

Distance-based covariates represent proximity to relevant landscape features, such as rock outcrops and sandy environments (e.g., coastal and dune systems), derived from MapBiomas Collection 10 and limited to a maximum distance of 7,000 meters, providing continuous spatial gradients. Stable land use and land cover areas, as defined by the MapBiomas Collection 10 legend, describe long-term cover/use persistence. Spatial coordinates (latitude and longitude) capture broad-scale spatial trends.

All covariates were processed to a consistent 30 m spatial resolution, and categorical variables were converted to binary (0 or 1) representations. All data sources and associated processing codes are open-access and are detailed in Appendix 1.

3.3.2 Dynamic Covariates

Dynamic covariates, representing transient environmental conditions and land use history, were prepared for the 1985-2024 period. These covariates were derived from various MapBiomas collections and the Landsat satellite archive.

The primary dynamic component is Land Use and Land Cover (LULC) persistence, derived from MapBiomas Collection 10. This multiclass variable is disaggregated into a set of individual, class-wise covariates. For any given year, each of these layers express the number of consecutive years a pixel remains in the same class, capturing landscape stability and transition dynamics. A uniform persistence period is assumed for the initial year of the time series.

Additional temporal variables describe the recurrence of key environmental events. Annual water surface maps and fire scar data from the most recent versions of the MapBiomas Water and MapBiomas Fire initiatives support the derivation of cumulative recurrence metrics, representing the frequency of occurrence of these events through time. Fire-related dynamics are further characterized by the number of years since the last recorded fire event. Proximity to anthropic land use, derived from the most recent version of the MapBiomas Degradation initiative, represents a proxy for disturbance and degradation processes.

Vegetation dynamics are represented by spectral indices derived from Landsat surface reflectance data (such as NDVI, SAVI, and EVI) derived from the Landsat 5, 7, 8, and 9 surface reflectance archives. To account for the legacy effect of past vegetation on soil carbon, the raw annual index values were transformed. The final covariates represent a temporally-weighted average of the six preceding years, calculated using an exponential decay function (alpha = 0.7). This procedure models the antecedent influence of vegetation conditions on soil organic carbon dynamics while also smoothing out short-term climatic variations that affect vegetation greenness but not soil properties.

All dynamic covariates were processed to a consistent 30 m spatial resolution for each year in the 1985-2024 time series. All data sources and associated processing codes are open-access and are detailed in Appendix 1.

3.4 Predictive Model

3.4.1 Soil Particle Size Fractions

We modeled the four soil particle size fractions (clay, silt, sand, and gravel) using three additive log-ratios: ALR1 = log(sand/clay), ALR2 = log(silt/clay), and ALR3 = log(gravel/clay). The complete dataset comprised 15,573 unique soil sampling points, a total of 60,261 soil layers. These points were intersected with a set of 129 static environmental covariates detailed in the Appendix. The mid-layer depth (in centimeters) was also included as a covariate in the model.

The resulting data matrix was sliced into ten submatrices, one for each of the 10-cm thick layers. Each submatrix was used to train three random regression forest models in R, one for each additive log-ratio. Hyperparameter tuning for each model was performed via a grid search, evaluating 256 combinations of num.trees, mtry, min.node.size, and max.depth using

out-of-bag error statistics (Hastie et al., 2009; Wright & Ziegler, 2017). The overall best set of hyperparameters was selected to be used across all models.

Using the overall optimized hyperparameters, three corresponding random regression forest models were then trained on each submatrix in Google Earth Engine with the ee.Classifier.smileGradientTreeBoost function (output mode: "Regression"). Spatial predictions of the additive log-ratios were generated for the midpoint of each 10-cm thick soil layer (5, 15, ..., 95 cm). These predictions were then back-transformed to mass percentages of clay, silt, sand, and gravel using the following equations (Samuel-Rosa et al., 2013):

$$\text{Sand (\%)} = e^{ALR1} / [e^{ALR1} + e^{ALR2} + e^{ALR3} + 1] \times 100$$

$$\text{Silt (\%)} = e^{ALR2} / [e^{ALR1} + e^{ALR2} + e^{ALR3} + 1] \times 100$$

$$\text{Gravel (\%)} = e^{ALR3} / [e^{ALR1} + e^{ALR2} + e^{ALR3} + 1] \times 100$$

$$\text{Clay (\%)} = 1 / [e^{ALR1} + e^{ALR2} + e^{ALR3} + 1] \times 100$$

Finally, the three fine fractions (clay, silt, and sand) were rescaled to sum the percent mass of fine earth material (< 2 mm diameter).

3.4.2 Soil Organic Carbon Stock

To predict the SOC stock for the 0-30 cm depth layer, a single, global machine learning model was trained using a dataset constructed from 12,574 georeferenced field data points and 130 environmental covariates (114 static and 17 dynamic). The dataset was augmented by creating two temporal replicas for samples located in stable natural land covers; this was done by assigning the original sample's data to the 10th and 20th year before the actual sampling year. Stable areas were defined as regions with no land use or cover change for at least 20 years and away from edges, i.e. neighboring anthropic disturbances.

A quality control procedure was implemented to identify and remove samples that conflict with ancillary environmental covariates. The filtering routine discarded samples that contradict expected physical, pedological, or ecological realities, specifically masking points if their target class conflicted with established thresholds for spectral indices, soil texture, ecological zones, or historical water recurrence. The final dataset was used to sample static and dynamic environmental covariates to form the training data matrix.

The model was trained using the Random Forest algorithm implemented in the ranger R package (Wright & Ziegler, 2017). Hyperparameter tuning was performed via a grid search, evaluating 256 combinations of num.trees, mtry, min.node.size, and max.depth using out-of-bag (OOB) error statistics (Hastie et al., 2009).

Using the optimized hyperparameters, the final Random Forest model was trained in Google Earth Engine with the ee.Classifier.smileRandomForest function. For prediction, the value for each pixel was derived by calculating the mean of the outputs from all individual trees in the forest ensemble. The final output of these models is a 30 m resolution time series of SOC stock maps for the 0-30 cm soil layer, covering all of Brazil from 1985 to 2024.

Table 5. Model and hyperparameters used to model SOC stocks in space and time across MapBiomas Soil Collections.

Collection	Model	n	p	num.trees	mtry	min.node.size	max.depth
1.0	Random Forest	9650	77	1000	25	5	30
2.1	Random Forest	12666	106	400	16	1-2[2]	40
3.0	Random Forest	27425 (14704)[1]	129	300	24	2	40

[1] Collection 3.0 employed layer-wise cumulative SOC stocks as target variable, thus largely increasing the number of training samples (n) compared to previous collections. The value between parenthesis is the number of points.

[2] Varying among models since Collection 2.1 employed three models (stratified by biome) to predict SOC stocks. Models 1 and 2 used a min.node.size value of 1, and Model 3 used 2.

3.5 Post-Processing

3.5.1 Soil Texture Maps

The maps of all four soil particle size fractions were masked using the MapBiomas land use and cover data (Collection 10). The mask applied excluded areas classified as water bodies (Class 26 and subclasses 31 and 33) as they do not contain soil. The resulting maps are left empty (labeled as 'no soil') in these regions. Other masks applied in previous collections (Rocky Outcrop - Class 29, and Dunes, Sands, and Beaches - Class 23) were discarded in Collection 3 in favor of the models predictions for these surfaces.

The maps of clay, silt, and sand content for the ten individual depth layers (0-10, 10-20, ..., 90-100 cm) were used to create maps for five wider depth intervals (0-20, 0-30, 20-40, 30-60, and 60-100 cm) using the arithmetic mean of the values from their constituent 10-cm thick layers.

The clay, silt, and sand percent mass for six depth intervals (0-10, 0-20, 0-30, 20-40, 40-60, and 60-100 cm) were classified using three textural classification schemes (Figure 3). The first scheme, derived from the Brazilian manuals of soil description in the field, categorizes soil into 13 textural classes (IBGE, 2015; Santos et al., 2015). The remaining two schemes are based on the Brazilian Systems of Soil Classification, one delineating eight textural subgroupings and the other defining five broader textural groupings (Santos et al., 2018).

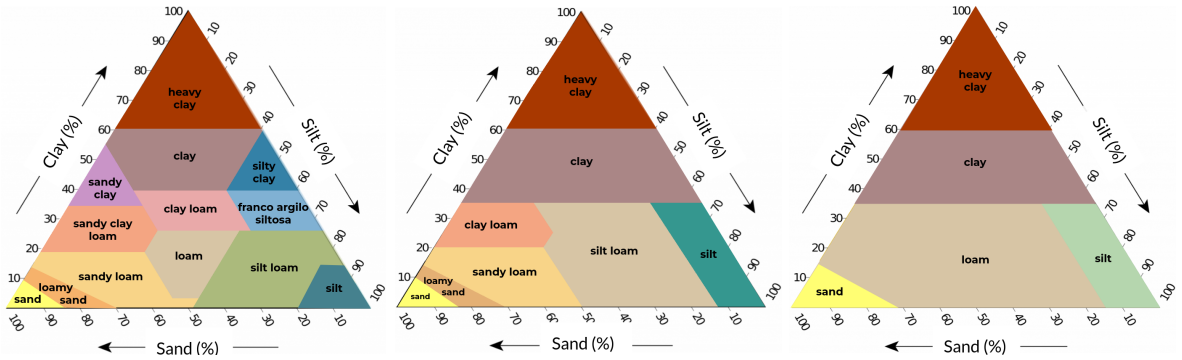

Figure 3. Ternary diagrams illustrating the three soil texture classification schemes used: (left) 13 textural classes, (center) eight textural subgroupings, and (right) five textural groupings.

3.5.2 Soil stoniness Maps

The depth-to-stoniness maps estimate the vertical distance from the soil surface to layers characterized by a large volume of coarse fragments (> 2 mm). Two distinct levels of stoniness are mapped: dominant, which identifies the onset of a fragment-dominated soil (more than 50% volume), and extreme, which marks the transition to a layer almost entirely composed of stones (more than 90% volume). The volume of coarse fragments (V_{rf}) in each 10-cm thick soil layer was calculated using the following equation:

$$\mathrm{V}_{rf} = \frac{W_{rf}/\rho_{rf}}{W_{rf}/\rho_{rf} + (1 - W_{rf})/\rho_b}$$

In which W_{rf} is the mass fraction of rock fragments (g/kg), ρ_{rf} is the density of the rock fragments (g/cm^3), and ρ_b is the bulk density of the fine-earth fraction (g/cm^3).

The values for ρ_{rf} were derived from an adaptation of the Geological Provinces of Brazil (IBGE, 2019b). Each pixel is assigned a constant density value based on the dominant lithology of the subprovince (Table 3), ranging from 2.00 g/cm³ for sediments to 2.74 g/cm³ for volcanic rocks. For the fine-earth bulk density (ρ_b), the model employs hierarchical pedotransfer functions (Huf dos Reis et al., 2024), which estimate density across ten depth intervals (0–100 cm) using the predicted sand and clay contents:

$\rho_b = 1.286 + 3.208 \times 10^{-3} \times \text{sand} - 2.013 \times 10^{-3} \times \text{clay}$

After calculating the volume of rock fragments for each depth interval, a critical threshold of 95% volume was applied. This value is used to define the transition to non-soil (rock or consolidated material), as it signifies the practical absence of fine earth. Above this 95% limit, the volumetric fragment content is considered 100%.

To generate the final depth maps, a vertical search algorithm was applied from the surface to the maximum depth (10 layers). This filter identifies the first occurrence of two specific volumetric thresholds: 50% (dominant) and 90% (extreme). The depth of the upper limit of the layer meeting these criteria is then recorded as the resulting pixel value. That is, once a threshold is met, the depth corresponding to the upper limit (top) of that layer is recorded (e.g., if the 20–30 cm layer meets the criteria, the recorded depth is 20 cm). If a threshold is

not reached within the 0–100 cm profile, the pixel is assigned a value of 100 cm. The final maps are masked for water surfaces (MapBiomas Water 2023) and exported at a 30-meter spatial resolution.

3.5.2 Soil Organic Carbon Stock Maps

The annual SOC stock maps were refined by integrating MapBiomas Collection 10 land use and land cover (LULC) data with specific soil thematic layers. The prediction domain and post-processing routines were updated to ensure physical consistency and temporal stability, following these criteria:

- **Masking:** The water bodies were masked. The mask was based on a 40-year water recurrence metric; pixels with a history of water recurrence and low photosynthetic activity were assigned a value of 0 t/ha.
- **Zero SOC Stock Assignment:** A SOC stock value of zero was assigned to the following land use and cover classes: Urban Infrastructure (Class 24) and Mining (Class 30), and Beach, Dune and Sand Spot (Class 23). For the latter, areas identified by the historical presence of sand bodies at any point across the 40-year MapBiomas series (1985–2024) were assigned a baseline SOC stock value of 10 t/ha. This acknowledges that stabilized sandy formations may have accumulated a minimal threshold of organic matter over decades. However, if a pixel is classified as Class 23 in the current prediction year, this baseline is overwritten with a value of 0 t/ha to reflect active or unstable sand bodies.
- **Temporal Filter:** A stability routine was implemented for the beginning of the series to minimize the effect of spectral noise and artifacts typical of earlier satellite sensors. In pixels where the land use and cover remained stable during the first five years (1985–1989), the predicted SOC stock values for the initial three years (1985–1987) were replaced by the 5-year median stock..

The final SOC stock predictions were generated in tons per hectare (t/ha). The product consists of a multi-band temporal series (1985–2024) with a spatial resolution of 30 meters.

3.6 Point and Zonal Statistics

Zonal statistics for soil properties were calculated for several areas of interest (AOI) across Brazil, including biomes, states, municipalities, conservation units, and LULC classes. The metrics computed were the SOC stock (t/ha) and mass (t) and the area (ha) of soil textural classes.

For soil texture, statistics were generated based on the spatial distribution of textural classes across three hierarchical levels: textural group, subgroup, and class. These analyses were replicated across six depth intervals (0-10, 0-20, 0-30, 20-40, 30-60, and 60-100 cm). The extent of each textural class within different AOI categories was determined by summing the area of pixels belonging to each category.

The mass of SOC within an AOI was calculated by summing the SOC mass (t) of all internal pixels, while the SOC stock (t/ha) was computed as an area-weighted average. In both cases, pixel area was determined using the `ee.Image.pixelArea()` function in Google Earth Engine to account for geographic distortions. This approach ensured that the value of each

pixel was weighted by its relative area within the AOI, correctly accounting for the variation in pixel size across Brazil's territory.

4 Evaluation Strategies

The evaluation strategy for Collection 3 retains the rigorous assessment methodology introduced in Collection 2, which aims to provide a realistic assessment of model performance, particularly in the absence of a dedicated independent validation dataset. This strategy relies on three components: leave-group-out cross-validation (LGO-CV) to prevent data leakage, quali-quantitative comparisons with existing maps, and area of applicability maps to communicate the spatial reliability of the predictions.

4.1 Generalization Error Assessment

The generalization error of the predictive models was estimated using a five-fold, leave-group-out cross-validation. A group was defined as the collection of layers in a soil profile and their spatial and temporal replicas. Five model performance statistics were computed: mean error (ME), mean absolute error (MAE), root mean squared error (RMSE), and model efficiency coefficient (MEC), and the regression slope (SLOPE).

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

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

$$\mathrm{MEC} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\mathrm{SLOPE} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $\hat{y}_i$ is the predicted value, y_i is the observed value, n is the sample size, and $\bar{y}_i$ is the mean of the observed values. These metrics were calculated both at the national level and for the six biomes (Amazon, Atlantic Forest, Caatinga, Cerrado, Pampa, and Pantanal) and the Coastal Zone ignoring tacit-samples of rock outcrops and beaches, dunes and sandspots.

4.2 Comparison with Existing Maps

The SOC stock maps are compared with existing static maps from other initiatives, which typically represent a synthesis of a historical period under a single nominal year. These benchmark products include the potential stock 1:5M (0-30 cm) (Bernoux et al., 2002), PronaSolos 90 m (0-30 cm) (Vasques, Coelho, Dart, Cintra, et al., 2021), SoilGrids 250 m (0-30 cm) (Poggio et al., 2021), and the GSOC 1 km (0-30 cm) map (FAO, 2022a). Because the MapBiomas Soil collection is a dynamic annual series, the comparison is limited to a

quali-quantitative assessment of agreement on large-scale spatial patterns and the mass of SOC estimated by each initiative. The MapBiomas Soil map for the year 2000 is used for this comparison, as this aligns with the nominal reference year for most of the other products.

For the particle size fractions, we make comparisons with ProNaSolos (90-m spatial resolution) (Vasques, Coelho, Dart, Cintra, et al., 2021), and SoilGrids (250-m spatial resolution) (Poggio et al., 2021). The particle size fractions (clay, silt, and sand) from these initiatives were first classified into textural groups to allow for a direct pixel-by-pixel comparison and compute confusion matrices. This cross-tabulation identifies where the primary spatial agreements and systematic discrepancies occur across the different products. These comparisons are made only for the 0-30 cm surface layer.

The maps of SOC stocks and textural classes of Collection 3 were also compared with maps of previous collections (1 and 2.1) using the same methods.

4.3 Spatial Reliability Assessment

A metric based on the geographic representativeness of the sampling network was calculated to communicate the spatial reliability of the model predictions. This metric evaluates the degree of spatial support for any given prediction location by calculating the Euclidean distance to the nearest training sample.

The metric calculated was the straight geographic distance (in meters) from each pixel to the closest soil observation used during model calibration. For PSD models, this distance was calculated each of the 10-cm depth intervals, and a single reliability map was produced by calculating the mean geographic distance across the 10 resulting layers. For the SOC models, the reliability was assessed by calculating a single layer over the entire (1985–2024).

5. Data Collections and Analysis

5.1 Field Soil Data

The two foundational datasets for Collection 3 were significantly expanded and improved compared with the two previous collections, enhancing the reliability and spatial representativeness of the final map products. This section details the evolution of the PSD and SOC stock datasets, which together form the basis for the Collection 3 maps. The two resulting dataset are openly available at https://doi.org/10.60502/SoilData/OXSR2N (soil particle size distribution) and https://doi.org/10.60502/SoilData/IUZOAK (soil organic carbon stock).

5.1.1 Soil Particle Size Distribution Dataset

An important advancement in Collection 3 is the inclusion of the soil gravel variable (coarse fragments > 2 mm) and the extension of soil mapping to a depth of 100 cm. In Collection 2, soil texture modeling was limited to 30 cm. Additionally, pseudo-samples derived from rocky outcrops, dunes, beaches, and sandy environments were incorporated, broadening the representation of extreme environments within the model. The number of unique points used for PSD modeling increased by 23.10%, rising from 11,522 in Collection 2 to 14,983 in Collection 3. This expansion substantially improved data representativeness, particularly in

biomes that had been previously under-sampled. The Pantanal exhibited a nearly sevenfold increase in sample size, while the Caatinga and Cerrado increased by 27.67% and 44.24%, respectively. Although the Amazon and Atlantic Forest remain the biomes with the largest number of samples, the newly incorporated data provide a more balanced overall distribution across biomes (Table 6 and Figure 4).

Table 6. Comparison of PSD sample distribution by biome (C2 vs. C3).

Biome	C2 # of Samples	C3 # of Samples	C3 % of Total	Change (%)
Amazônia	4,643	5,200	34.70	12.00
Mata Atlântica	3,497	4,037	26.90	15.45
Cerrado	1,554	2,787	18.60	79.37
Caatinga	703	972	8.92	38.27
Pampa	1,029	1,337	6.49	29.90
Pantanal	96	650	4.34	577.08
Total	11,522	14,983	100	30.15

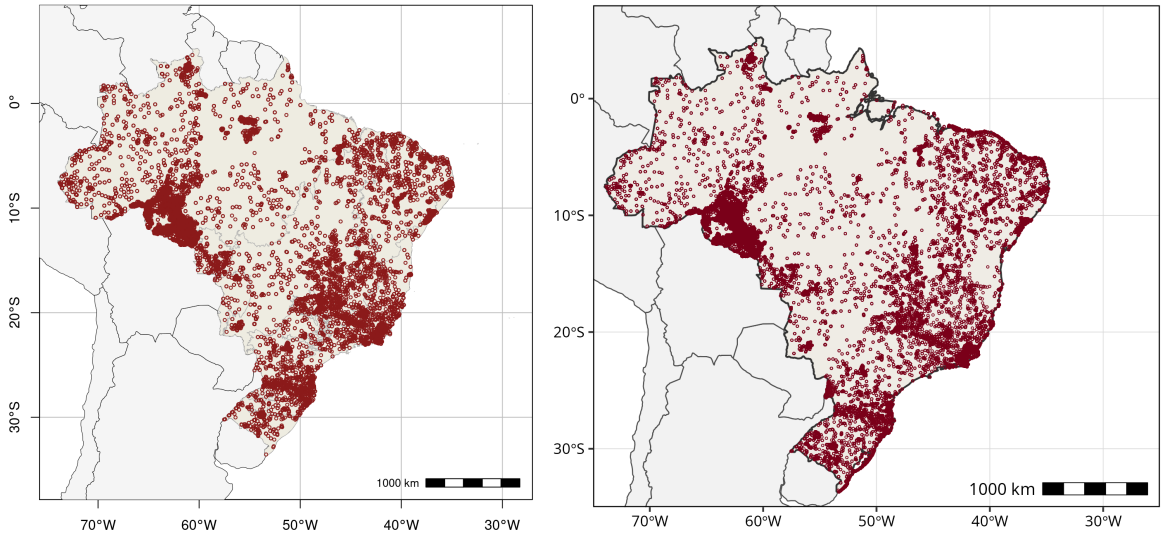

Figure 4. Spatial distribution of the particle size fractions in Collection 2 (left; n = 11,522) and Collection 3 (right; n = 14,983) data.

The frequency distributions show that clay and silt exhibit strong right-skewed patterns, with most samples concentrated at low values and a long tail toward higher contents. In contrast, sand displays a more uniform distribution across its entire range, reflecting greater diversity of sandy environments in the dataset. Rock fragments are predominantly clustered near 0%, but a secondary peak at high values indicates the presence of environments with substantial coarse material. Depth is most frequently represented in the upper soil layers, gradually

decreasing toward 100 cm, which reflects the typical sampling strategy in soil surveys and environmental databases (Figure 5).

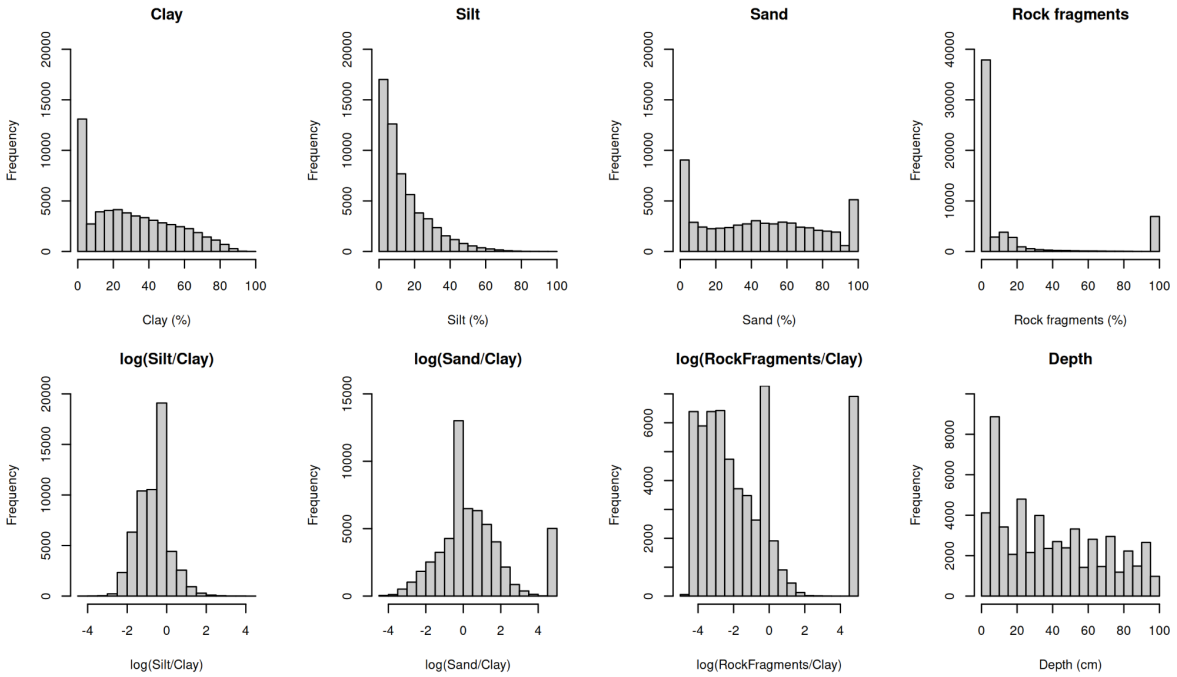

Figure 5. Frequency distributions of clay, silt, sand, rock fragments, their additive log-ratios, and mid-layer sampling depth (n = 57,321).

5.1.2 Soil Organic Carbon Stock Dataset

The number of unique georeferenced points used for soil organic carbon (SOC) stock modelling increased by 21.17%, rising from 12,666 in Collection 2 to 16,068 in Collection 3. This expansion substantially improved the representativeness of the data, particularly in biomes that were previously under-sampled. The Pantanal showed the largest relative increase, with a significant rise of 84.46% in sample count. The Cerrado also expanded considerably (by 47.16%), followed by the Pampa (by 20.82%) and the Caatinga (by 17.33%). The Amazon and the Atlantic Forest grew too, albeit to a lesser extent (8.84% and 9.55% respectively). While the Amazon and the Atlantic Forest remain the most sampled biomes, Collection 3 provides a more balanced overall distribution. Nevertheless, despite showing the greatest proportional increase, the Pantanal remains comparatively underrepresented in the dataset (Table 7 and Figure 6).

Table 7. Comparison of SOC sample distribution by biome (C2 vs. C3).

Biome	C2 # of Samples	C3 # of Samples	C3 % of Total	Change (%)
Amazônia	4,650	5,101	31.80	9.70
Mata Atlântica	4,143	4,575	28.50	10.42
Cerrado	1,564	2,960	18.40	89.29
Caatinga	1,197	1,448	9.01	20.96
Pampa	1,000	1,263	7.86	26.30
Pantanal	112	721	4.49	543.75
Total	12,666	16,068	100.00	26.87

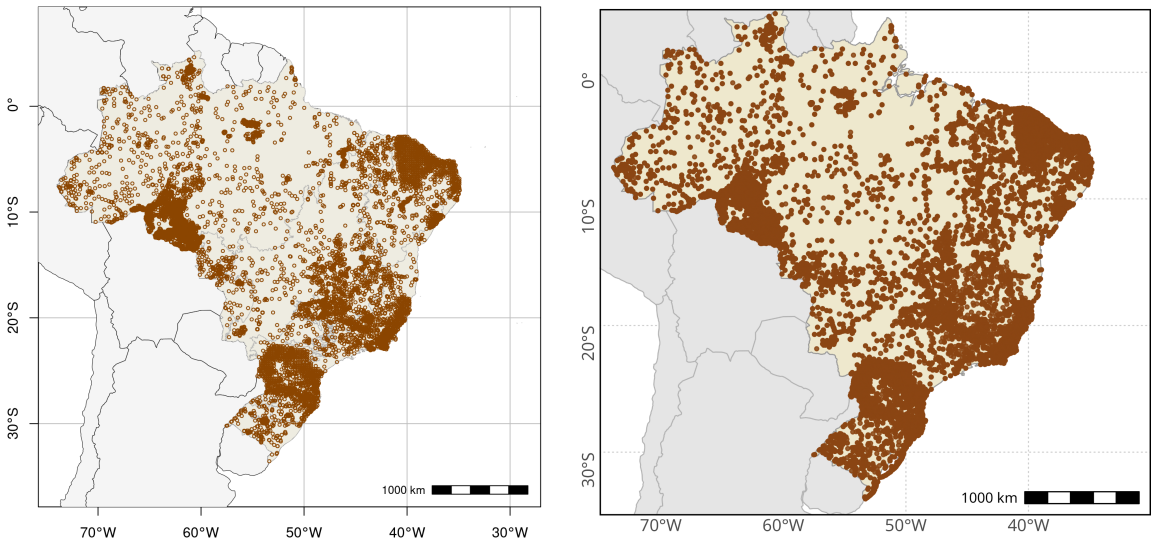

Figure 6. Spatial distribution of the soil organic carbon stock in Collection 1 (left; n = 12,666) and Collection 3 (right; n = 16,068) data.

The temporal distribution of samples in Collection 3 reveals a distinct pattern of concentration during specific timeframes. The largest volume occurred before 1985, accounting for around a quarter of the total. This was followed by a second notable peak in 1997, accounting for around 15% of samples. Other notable years include 2010, 2014 and 2023, all of which had a high number of samples, reflecting more intensive sampling campaigns during these periods. The remaining years show a sparse distribution with a low relative frequency. Collection 2 shows that 52% of SOC stock samples were within the intermediate range of 20 to 50 t/ha. Collection 3 exhibits a comparable trend, with around 48% of single points falling within this range (Figure 7).

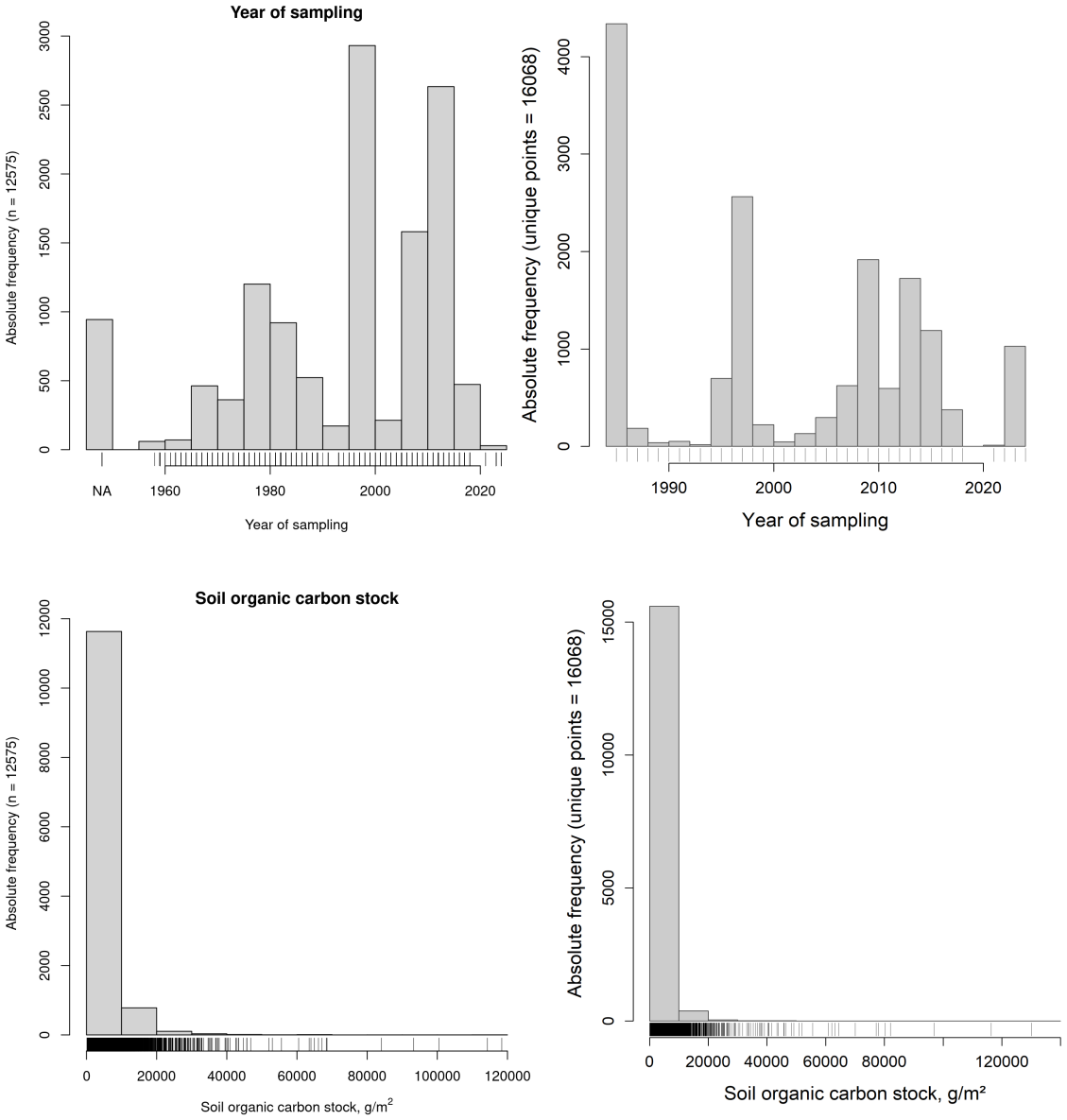

Figure 7. Temporal (top) and frequency (bottom) distribution of the SOC stocks in Collection 1 (left; 12,666) and Collection 2 (right; n = 16,068).

5.1.3 Imputation Model Performance

Collection 3 soil data required the imputation of missing coarse fragment and bulk density values, a process executed using Random Forest models. Table 8 presents the model performance, which was evaluated using bootstrap resampling (out-of-bag error). While not directly comparable due to variations in sample size between collections, the results indicate an improvement in prediction performance for this latest collection.

Table 8. Model performance of Random Forest imputation for missing coarse fragment and bulk density values (out-of-bag error) across MapBiomas Soil collections.

Collection	Content of coarse fragments			Soil bulk density		
	ME (%m)	RMSE (%m)	MEC	ME (g/cm³)	RMSE (g/cm³)	MEC
1.0	-	-	-	-	-	-
2.1	0.31	7.80	0.48	0.00	0.15	0.74
3.0	0.85	53.26	0.75	0.00	0.11	0.84

5.2 Soil Particle Size Distribution Maps

The training process of the 30 individual, 10-cm layer gradient boosted tree models was assessed using the cumulative out-of-bag (OOB) improvement, computed as the sum of per-tree OOB loss reductions. The cumulative OOB improvement after 400 trees, averaged across depth layers, was about 0.89 for log(sand/clay), 1.68 for log(coarse/clay), and 0.45 for log(silt/clay). Across depth layers, these values ranged from 0.80-1.05, 1.33-2.11, and 0.41-0.49, respectively. The top five most important variables across all models were clay_000_030cm (30x, i.e made the top five of all 30 models), sand_000_030cm (20x), depth (18x), silt_000_030cm (17x), Distance_to_rock_v33 (13x), elevation (12x), NEOSSOLO_LITOLICO (9x), Distance_to_sand_v33 (8x), koppen_I2_Am (7x), Thinsols (5x), sedimentares (3x), cerrado_sedimentos (2x), Ferralsols (2x), Sandysols (2x), Argisols (1x), and Pantanal (1x).

The LGO-CV estimates the models' ability to generalize to new, unsampled locations. Results indicate that predictive performance varies between particle size fractions and across biomes (Table 9). At the national level, Sand exhibited the highest overall predictive power with a MEC of 0.81 and a slope of 1.00, but also had the highest RMSE at 11.46%. Silt showed the lowest error magnitude, with the smallest RMSE (7.85%) and MAE (4.20%), though it had the lowest MEC (0.66). Clay demonstrated strong performance with an MEC of 0.75 and an RMSE of 10.82%.

The highest MEC for any fraction was silt in the Pantanal (0.92), followed closely by the high MEC values for sand in the Cerrado (0.85), and clay and sand in the Pampa (0.84 for both). The lowest MECs were observed for silt in the Atlantic Forest (0.41) and clay in the Pantanal (0.52). The Pantanal biome demonstrated the highest error magnitudes for both clay (RMSE = 14.51%) and sand (RMSE = 18.86%). The Atlantic Forest consistently showed the highest positive bias (overestimation) across all three fractions, with ME values of 2.64% (clay), 2.83% (silt), and 3.04% (sand).

Table 9. Prediction performance statistics of the PSD model for Brazil and per biome across MapBiomas Soil collections estimated using resampling methods.

Territory	ME (%)	MAE (%)	RMSE (%)	MEC	Slope
Clay					
Brazil	1.11	6.55	10.82	0.75	1.00
Amazon	0.07	5.67	8.94	0.81	1.00
Atlantic Forest	2.64	8.36	13.80	0.67	1.02
Caatinga	1.61	7.31	11.53	0.53	0.87
Cerrado	2.03	6.35	11.35	0.77	0.99
Coastal Zone	2.17	8.15	13.88	0.62	0.88
Pampa	1.35	5.15	8.04	0.84	1.02
Pantanal	1.69	10.26	14.51	0.52	0.89
Silt					
Brazil	0.77	4.20	7.85	0.66	0.93
Amazon	-0.09	2.98	5.36	0.82	1.02
Atlantic Forest	2.83	6.16	11.22	0.41	0.81
Caatinga	1.66	5.34	8.73	0.49	0.82
Cerrado	0.78	4.47	8.47	0.60	0.90
Coastal Zone	0.70	5.24	7.93	0.78	1.05
Pampa	0.04	4.27	6.96	0.63	1.07
Pantanal	-0.74	6.78	9.67	0.92	0.92
Sand					
Brazil	1.18	6.74	11.46	0.81	1.00
Amazon	0.55	5.90	9.48	0.84	1.00
Atlantic Forest	3.04	7.70	13.58	0.70	0.96
Caatinga	0.92	8.98	14.52	0.67	0.94
Cerrado	1.57	6.64	11.65	0.85	1.02
Coastal Zone	-1.56	8.51	14.97	0.80	0.97
Pampa	-0.01	5.71	10.18	0.84	1.01
Pantanal	-0.40	13.66	18.86	0.54	0.89

The comparison between the current soil textural group maps and existing datasets (PronaSolos, SoilGrids, and the previous collection) revealed large spatial agreement (Figure 8). The results of the cross-tabulation (Table 10) show that the Sandy group achieved the highest levels of agreement, with values of 96% compared to PronaSolos and 100% against SoilGrids.

The Very Clayey and Clay groups also demonstrated high spatial agreement with PronaSolos and SoilGrids, with values ranging from 67% to 83%. Disagreements predominantly occurred as transitions between the two textural groups (e.g., 26% of pixels classified as Very Clayey in PronaSolos were identified as Clay in this collection), which is expected given the continuous nature of clay content and the differences in spatial resolution between products. The Medium textural group showed a strong agreement of approximately 76-78% with PronaSolos and SoilGrids, and a 93% agreement with the previous collection.

The Silty group presented the lowest agreement levels, typically below 10%. This low agreement highlights the challenges in regionalizing silty textures, which often occupy restricted transitional niches in the landscape.

The visual assessment (Figure 8) confirms that while all maps capture the broad macro-regional textural trends, the Collection 3 provides a more detailed representation of textural transitions. This increased granularity is a direct result of the 30-meter resolution and the updated training matrix.

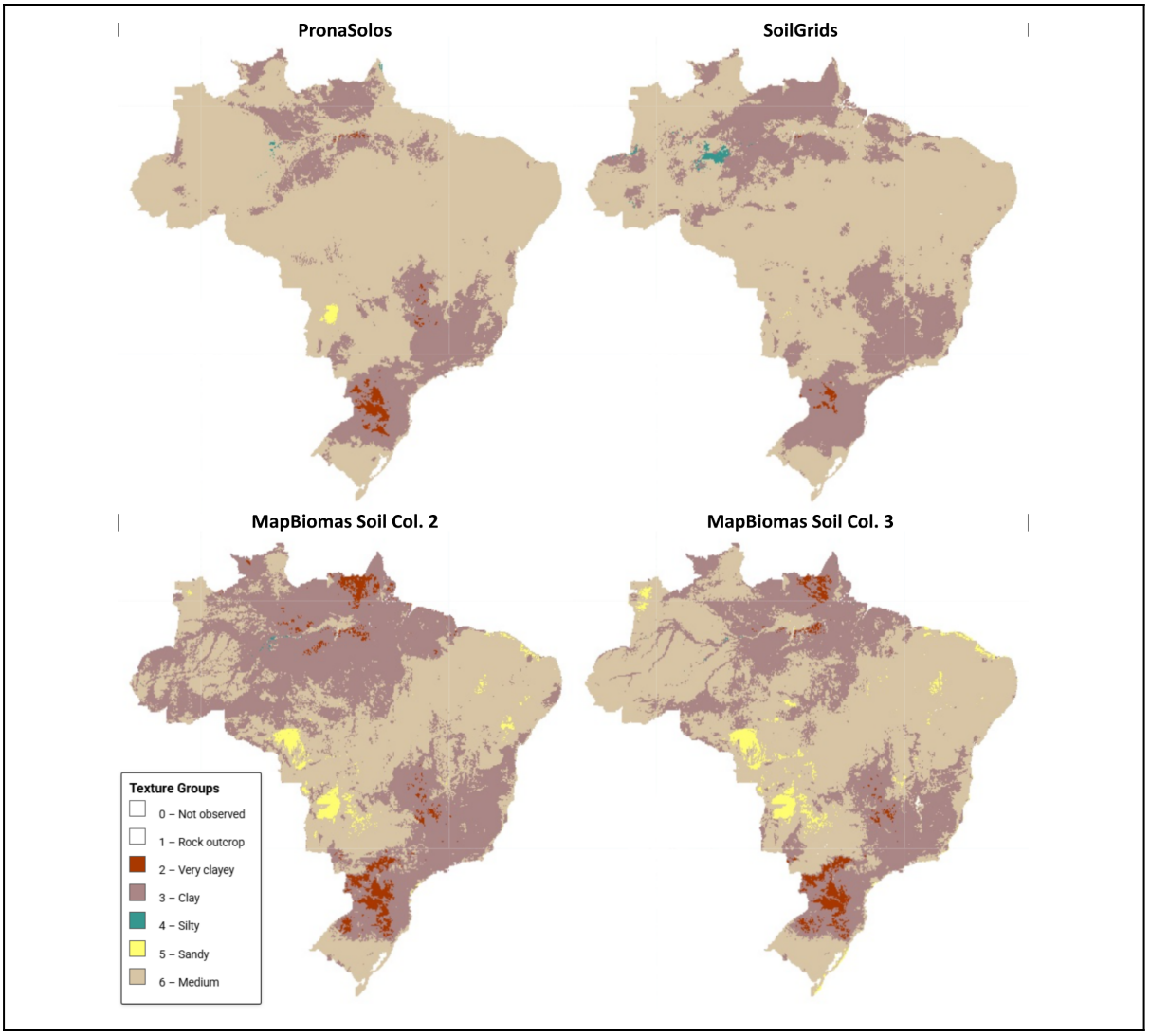

Figure 8. Soil textural group classification maps, obtained from soil particle size distributions of the 0-30 cm layer, from different sources: PronaSolos, SoilGrids, MapBiomas Soil Collection 2, and Collection 3.

Table 10. Comparison (%) for soil texture groups (0–30 cm) between MapBiomas Soil Collection 3 against PronaSolos, SoilGrids, and Collection 2.

		PronaSolos				
		Very clay	Clay	Medium	Silt	Sand
MapBiomas Soil Col. 3 vs.	Very clay	**74**	7	-	-	-
	Clay	26	**75**	20	40	-
	Medium			**76**		
	Silt	-	-	-	**6**	-
	Sand	-	-	3	-	**96**
		SoilGrids				
		Very clay	Clay	Medium	Silt	Sand
MapBiomas Soil Col. 3 vs.	Very clay	**83**	7	-	-	-
	Clay	17	**67**	17	22	-
	Medium			**78**		
	Silt	-	-	-	**2**	-
	Sand	-	-	4	-	**100**
		MapBiomas Soil Col. 2				
		Very clayey	Clay	Medium	Silt	Sand
MapBiomas Soil Col. 3 vs.	Very clayey	**60**	1	-	-	-
	Clay	39	**62**	4	18	-
	Medium			**93**		
	Silt	-	-	-	**8**	-
	Sand	-	-	3	-	**85**

The area-of-applicability map (Figure 9) reveals a highly heterogeneous sampling density across the Brazilian territory. A high concentration of training points is evident in the South, Southeast, and coastal Northeast regions, as well as along the "arc of deforestation" in the

Amazon. In these areas, the mean distances to the nearest samples are lowest, indicating that the textural estimates are derived from a dense network of local observations, which lends higher reliability to spatial predictions. In contrast, large portions of the interior Amazon basin, the Pantanal, and certain localized areas in the Central-West are characterized by a sparse sampling network. In these regions, the distances frequently exceed 100,000 meters, placing several pixels in a lower reliability status of spatial predictions.

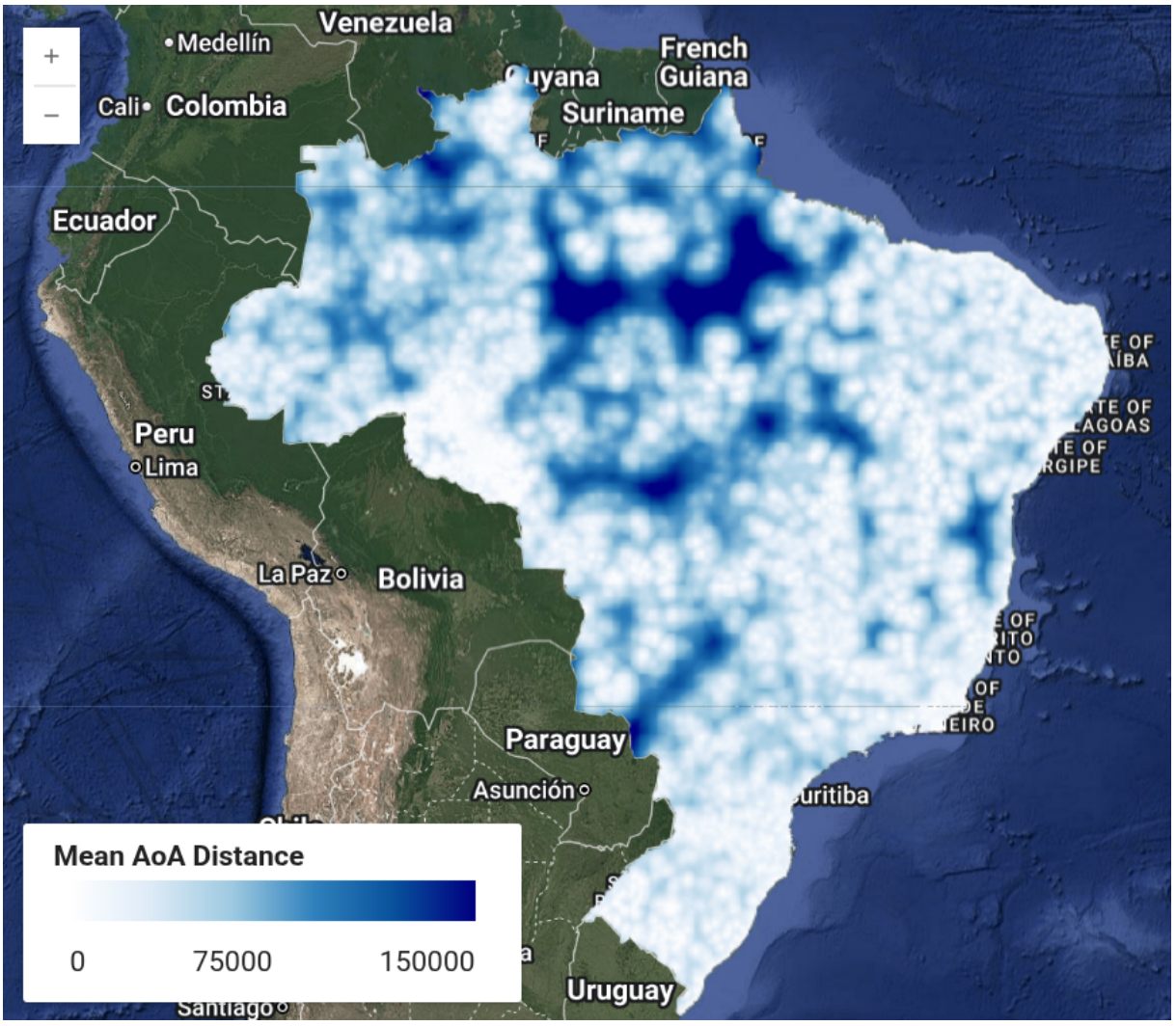

Figure 9. Mean area-of-applicability for soil particle size distribution (PSD). The map represents the geographic distance between each pixel and the nearest training sample, integrated as a mean across ten depth intervals (0–100 cm).

5.3 Soil Organic Carbon Stock Maps

The single, global SOC model showed good fit during the training phase, assessed via bootstrap resampling using the out-of-bag (OOB) training error. The trained model exhibited the following goodness-of-fit statistics: ME = 0.32 t/ha; MSE = 11.74 t/ha^2; RMSE = 26.73 t/ha; MEC = 0.73; SLOPE = 1.1. The five most important covariates were soil depth, altitude above seal level, age of wooden and herbaceous sandbank vegetation, occurrence of Espodossolos, and topsoil clay content (0-30 cm).

The prediction performance statistics of the Collection 3 SOC stock model is presented in Table 11. For context, the table also includes the statistics of Collections 1 and 2.1. Note, however, that these statistics are not directly comparable, as the collections differ in their validation samples and methodology. Despite this, the comparison suggests that the data increments and methodological changes in Collection 3 improved overall model performance in all biomes.

Table 11. Prediction performance statistics of the SOC stock model for Brazil and per biome across MapBiomas Soil collections estimated using resampling methods.

Territory	Collection	ME (t/ha)	MAE (t/ha)	RMSE (t/ha)	MEC	SLOPE
Brazil	1.0	-	-	36.70	0.37	-
	2.1	-7.17	18.96	54.34	0.25	-
	3.0	-0.83	17.09	38.44	0.42	1.11
Amazon	1.0	-	-	22.60	0.24	-
	2.1	-4.52	15.40	33.37	0.18	-
	3.0	-0.52	16.57	27.31	0.27	1.16
Atlantic Forest	1.0	-	-	46.70	0.49	-
	2.1	-8.50	23.07	62.01	0.29	-
	3.0	-0.92	19.73	40.18	0.56	1.10
Caatinga	1.0	-	-	22.60	0.46	-
	2.1	-4.96	14.11	33.96	0.34	-
	3.0	2.81	11.31	17.40	0.19	0.68
Cerrado	1.0	-	-	53.30	0.19	-
	2.1	-12.92	24.00	89.29	0.14	-
	3.0	-1.28	14.68	40.57	0.19	0.94
Coastal Zone	1.0	-	-	-	-	-
	2.1	-	-	-	-	-
	3.0	-24.13	92.77	179.89	0.22	1.45
Pampa	1.0	-	-	75.40	-0.17	-
	2.1	-5.55	14.60	51.27	0.18	-
	3.0	-0.25	11.53	24.61	0.45	1.00
Pantanal	1.0	-	-	26.90	0.23	-
	2.1	-25.18	35.79	56.33	-0.07	-
	3.0	0.25	12.49	18.16	0.59	0.97

The Pantanal biome demonstrated the largest increase in MEC, improving from -0.07 (C2.1) to 0.59. Its RMSE also dropped significantly to 18.16 t/ha. The Atlantic Forest also showed strong performance with an MEC of 0.56. The Caatinga and Cerrado biomes had the lowest MEC values, both at 0.19. The Caatinga biome is the only one with a positive ME (2.81 t/ha), suggesting a positive bias (overestimation) in the predictions for this region. The Coastal

Zone exhibits the highest error magnitudes, with an RMSE of 179.89 t/ha and an MAE of 92.77 t/ha, suggesting significant bias with an ME of -24.13 t/ha and a high SLOPE of 1.45.

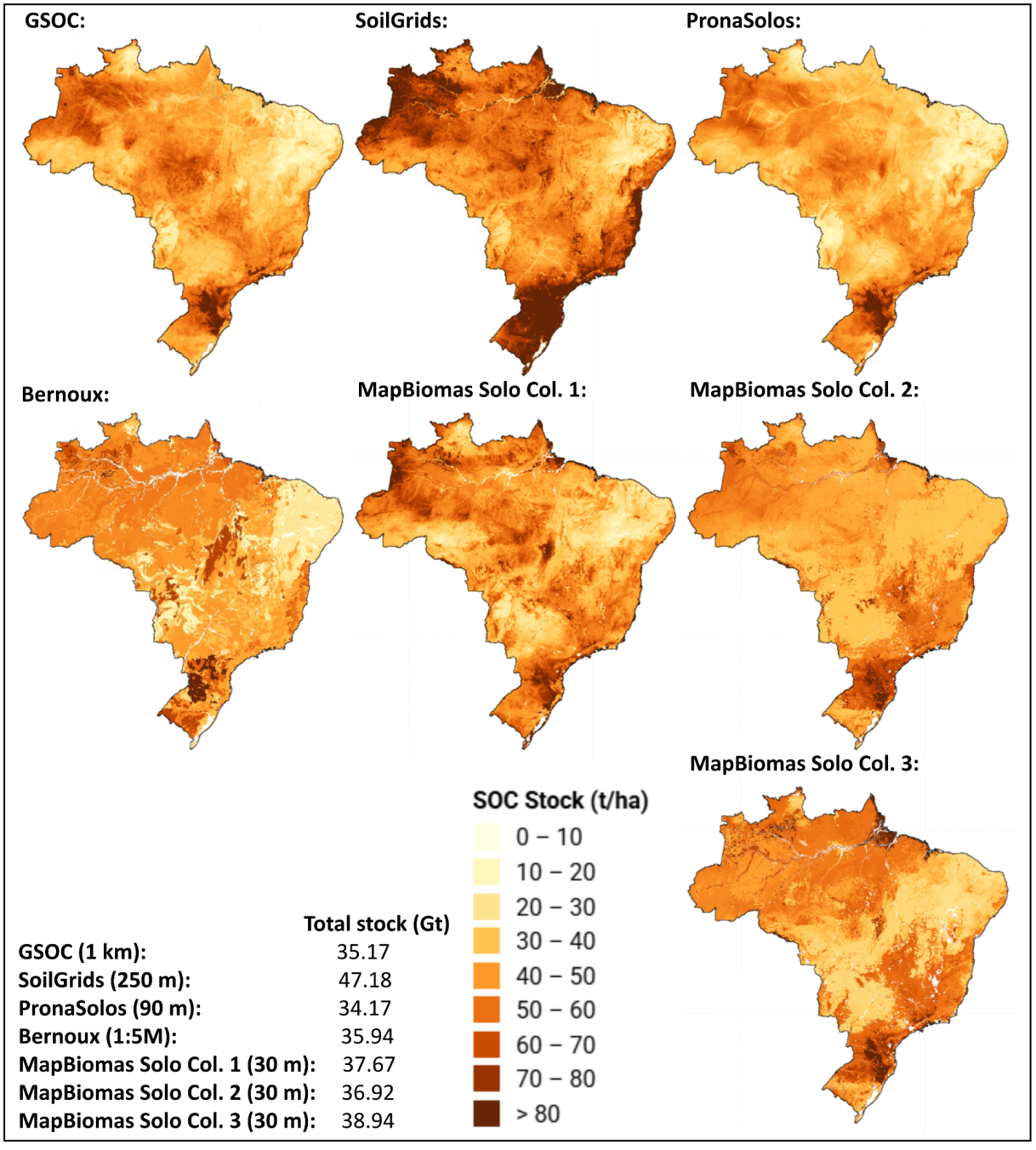

Figure 10. Spatial patterns and total estimated mass of soil organic carbon (SOC) for the 0–30 cm layer in Brazil from seven existing map products used for quali-quantitative comparison: PronaSolos map (Embrapa), FAO's GSOC, ISRIC's SoilGrids, and the Bernoux dataset, alongside the current and previous releases of the MapBiomas Soil maps (Collections 1, 2, and 3) for the year of 2000.

Broad spatial similarities are observed across all SOC stock products, with general agreement that the Amazon and Atlantic Forest biomes hold the highest SOC stocks, while the Cerrado and Caatinga biomes consistently show the lowest values (Figure 10). The semi-arid Caatinga biome in northeastern Brazil presents the lowest SOC stocks across all datasets. Another

point of convergence is the high SOC stock in the southern portion of the Atlantic Forest, with values generally exceeding 60 t/ha. The SoilGrids product stands out in this region, with estimates reaching values greater than 140 t/ha.

High SOC stocks are also evident throughout the Amazon biome across all maps, with SoilGrids providing the highest overall estimates for the region. In contrast, the Bernoux dataset produces the most pronounced estimates for floodplains within the Cerrado, specifically in the Araguaia National Park region in central Brazil. This specific geographic feature is also captured in the MapBiomas Soil maps, notably in Collections 1 and 3.

The primary differences among the products stem from the total estimated SOC mass for the 0–30 cm layer across the country. MapBiomas Soil maps show strong alignment with the PronaSolos, GSOC, and Bernoux maps, with total estimates ranging between 34 and 36 Gt. This contrasts with the global SoilGrids product, which displays a much more intense and widespread distribution of high-SOC stock areas, resulting in a higher total estimate of 47.15 Gt.

The AoA maps for SOC model (Figure 11) appear visually identical. However, it is important to note a key difference in their underlying training datasets. The SOC model incorporates approximately 1,000 additional data points sourced from the National Forest Inventory (IFN). These supplementary samples are distributed across the states of Ceará, Paraná, Rondônia, Espírito Santo, Sergipe, and parts of Rio Grande do Norte. Consequently, while the broad patterns of data scarcity remain, the SOC model possesses enhanced data support in these specific regions, suggesting a subtle but significant increase in the local reliability of its predictions compared to the PSD model.

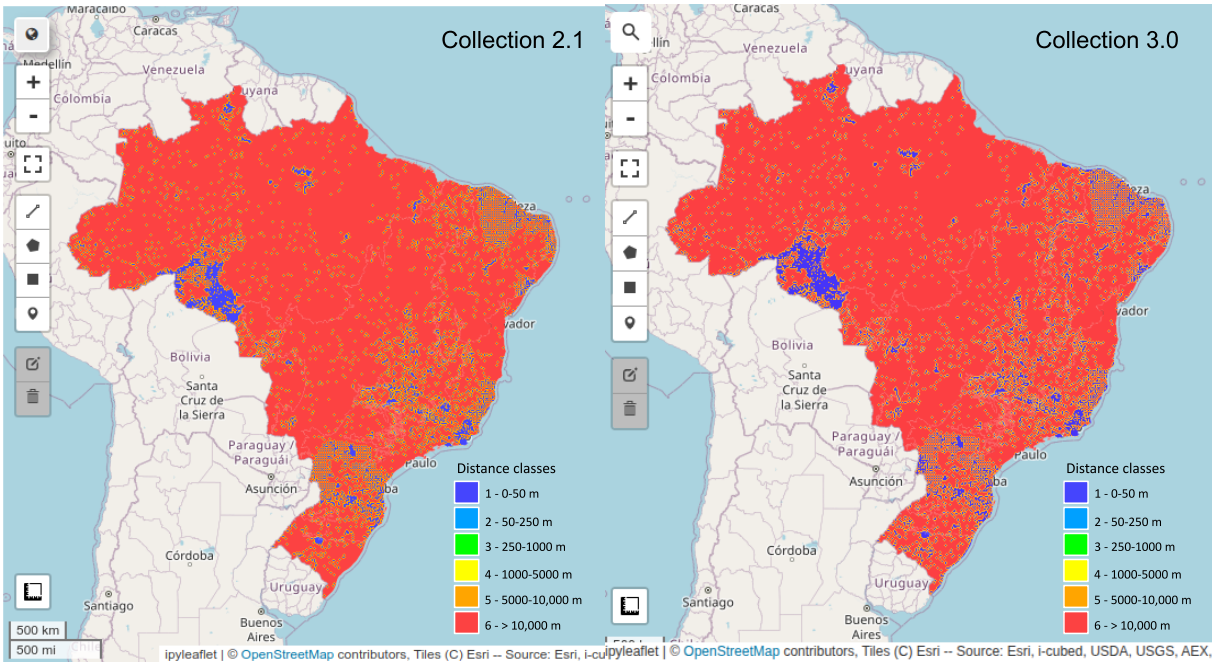

Figure 11. Distance (classes) to the nearest soil sample suggested as an indicator of the area-of-applicability of soil organic carbon stock maps of Collection 2.1 (left) and 3.0 (right).

Note about temporal inconsistencies and training artifacts: The analysis of the Collection 3 time series reveals an abrupt decrease in carbon mass between 2021 and 2022. This

discontinuity is technically classified as a processing artifact, derived from the composition of the training dataset. Specifically, the sampling of organic soils on the coast showed a spatial and temporal distribution that induced a bias in the final estimates for this interval. Such inconsistencies are under review by the MapBiomas Soils team for algorithm refinement and correction of training metrics.

6. Practical Considerations

Collection 3 provides an integrated estimate of both global and relative soil organic carbon (SOC) stocks, capturing the expected large-scale spatial patterns across the Brazilian territory. It is fundamental that these maps are interpreted as temporal and spatial **baselines**; that is, they represent the best possible approximation for each specific date, given the available field data and remote sensing conditions at the time of modeling. The plausibility of these results must be assessed by considering the present and historical environmental conditions specific to each region.

A significant advancement in Collection 3 is the expansion of the vertical profile modeling, now providing estimates for standardized depth intervals up to 100 cm. While this allows for a more comprehensive understanding of soil texture, users should note that uncertainty typically increases with depth, as the density of field observations is significantly higher in the topsoil (0–30 cm) than in deeper horizons.

The theoretical and conceptual framework used to generate the information contained in these maps is based on the digital soil mapping approach. The models take into account key processes that control soil carbon stocks in space, such as terrain morphometric variables, and over time, such as land-use changes. However, other important drivers of change, such as climate variables and land management practices, were not included in the model. This omission is due to the limited availability of field data and its heterogeneous distribution across space and time. These factors will be incorporated as new data are added to the SoilData repository and as new modeling strategies are explored.

The most critical impacts of land use in Brazil are not fully captured by this series. This is because the most significant soil degradation issues in the country predate 1985, for which satellite data are unavailable. During that period, a significant portion of agricultural land was degraded or in the process of degradation. When conservation and management techniques became widespread in the 1990s, changes in agricultural systems, such as minimum tillage and no-till farming, resulted in increased SOC stocks in some regions, particularly in areas that were initially below expected levels due to environmental conditions. In these cases, the increase in SOC stocks can be attributed to the initial soil conditions at the beginning of the series, and does not necessarily suggest that the land use is efficient in terms of SOC sequestration.

The model used for predictions is a first approximation of what we expect the distribution of SOC stocks in the top 30 cm of Brazilian soil to be. This is because models are simplified representations of reality. In these models, complex processes of addition, loss, transformation, and translocation of soil properties, and consequently SOC stocks, are represented simplistically in space and time. The model is based on a specific dataset of field observations associated with environmental covariates, whose relationships are computed

by the model. Therefore, the estimates provided represent the best approximation of the true values, containing inherent errors and uncertainties from the soil mapping process.

6.1 Limitations

The SOC stock maps from the collection 3 do present explicit spatial uncertainty estimates associated with the predicted values. However there are some limitations of the data used and the potential sources of uncertainty that are not fully addressed, including uncertainties related to the point data and covariates. These uncertainties are combined (added, canceled, or multiplied) and propagate through the final map. This process, known as uncertainty propagation, causes a distortion relative to the field truth or what we expect it to be. This variation may sometimes have the same magnitude as the natural variation of SOC stocks in the landscape and over time.

6.1.1 Inherent Limitations of Point Data for SOC Stocks

The samples available for Collection 3 exhibit a heterogeneous spatial and temporal distribution throughout the series, with minimal temporal repetition.

For each soil data point, the quantification of soil properties inevitably contains errors, as every measurement, no matter how precise, is always an approximation of the true value. In determining the carbon content of a soil sample, the analytical methods and equipment used each have distinct, limited precision. Since the data comes from different projects designed for various purposes, combining SOC data obtained using different analytical methods required a harmonization step to ensure consistency.

For some field samples, soil density had to be estimated using a regression model, based on other soil properties and environmental covariates. Another important limitation is that the volume of soil occupied by roots is not considered in the SOC stock calculation. Although this is a critical factor, especially in forested areas, it is often omitted due to the lack of available data.

All soil point samples are linked to geographic coordinates that provide their spatial location. The accuracy of these positional data primarily depends on the precision of the equipment used to collect them. Generally, the older the field data available in SoilData, the lower the positional accuracy, with some extreme cases involving errors spanning several kilometers. This means that even if the SOC stock estimate at a given point is very close to the expected value, it could still be misaligned with environmental conditions that are substantially different. This risk is amplified as the resolution of the covariates increases.

In summary, while the data is valuable for understanding SOC stocks, these inherent limitations—such as spatial and temporal heterogeneity, lack of root volume data, and positional inaccuracies—should be carefully considered when interpreting the results and further improving the dataset.

6.1.2 Inherent Limitations of Covariates

Covariates are employed as approximations for landscape representation, and temporal predictions are particularly dependent on the quality of land use and land cover data. Errors

in data quantification and allocation inevitably propagate into the SOC stock maps. Furthermore, there is limited knowledge regarding the dynamics of land use and land cover prior to 1985.

Given that certain areas of Brazil are permanently cloud-covered (resulting in missing data), especially in the northern region during the early years of the historical series, missing data were filled using filters. However, this approach can introduce additional uncertainties that may affect the final results.

Additionally, while existing land use and land cover maps, as well as satellite imagery used to calculate NDVI, provide valuable information, they are insufficient for fully capturing the land management practices in agricultural, livestock, and forestry production areas. The time span covered by these maps is relatively short compared to the prolonged use of Brazilian soils for productive activities. The absence of data for the decades prior to the 1980s presents challenges in modeling the medium- and long-term effects of land use dynamics on soil properties. Future solutions will be developed to enhance the representation of agricultural management practices throughout the historical series.

This understanding of the limitations helps guide future improvements in data collection and model accuracy, particularly in refining the portrayal of land management practices across different time periods.

6.1.3 Inherent Limitations of the Predictive Model

Random Forest and Gradient Booster Tree are data-driven models, meaning its results are directly influenced by the quality and quantity of the available data.

Despite the high computational capacity of Google Earth Engine, which is crucial for applying large-scale prediction models like this one, limitations were identified in the implementation of the Random Forest algorithm on the platform. The main limitation is that the implemented algorithm (through the function ee.Classifier.smileRandomForest) does not allow for fine control over the model training process beyond configuring the parameters mtry, ntree, minLeafPopulation, and bagFraction. Efforts are underway to improve the flexibility of the algorithm's parameters, particularly for computing global and local uncertainty statistics of the model.

6.2 Disclaimer

The information presented in the Collection 3 is based on the best, and sometimes the only, available soil data, environmental information, and digital soil mapping techniques. Although this product has been created with the utmost care, the author(s), editor(s), and/or MapBiomas cannot be held liable for any damages resulting from the use of these data or any content contained within, in any form, whether caused by potential errors or failures, or for any consequences thereof.

The terms used and the presentation of material in this informational product do not imply the expression of any opinion by MapBiomas regarding the legal status of soil carbon stocks, particle size distribution, soil texture, and stoniness in any territory, city, area, or their authorities.

7. Final Considerations and Future Perspectives

User feedback is essential for the continuous enhancement of future versions of the MapBiomas Solo collection, enabling improvements in the accuracy and quality of data to better meet users' needs and expectations. The MapBiomas Solo development team is committed to addressing the artifacts and inconsistencies reported by users during the beta collection phase, with the goal of generating a refined and updated collection of maps for public release. This collaboration between users and developers is crucial for expanding the utility of the MapBiomas Solo map series, providing valuable tools for researchers, scientists, and policymakers across Brazil.

Several updates are planned for future collections. The first will involve the inclusion of spatially explicit uncertainty estimates for model predictions, which will be calculated at the pixel level. These uncertainty estimates, along with global quality assessments, will allow the team to make informed decisions about where to focus efforts on improving the map series. Once trained and informed, users will be able to use these uncertainty statistics to better guide their applications of the data.

The second update focuses on mapping SOC stocks in the subsoil layer, between 30 and 100 cm depth, where significant SOC stocks are expected, especially in biomes such as Pantanal and Cerrado. The third update will expand the range of explanatory dynamic variables, considering additional change drivers for SOC stocks. This will allow the temporal dynamics of SOC stocks to capture not only land cover and land use changes but also the effects of climate change, soil management practices, flooding regimes, and the frequency and intensity of wildfires.

Ultimately, the initiative aims to promote a more collaborative and open soil science community, encouraging the sharing of ideas, data, and algorithms. Collaboration is fundamental to developing a time series that accurately represents soil dynamics. Data collected by experts, who understand the relationships between soil and the landscape in Brazil’s biomes, whether retrieved from past initiatives or obtained through contemporary fieldwork, is central to the evolution of these products. Initiatives such as the Brazilian National Soil Survey and Interpretation Program (PronaSolos) will be vital partners in this effort.

8. References

Adhikari, K., Kheir, R. B., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A. B., & Greve, M. H. (2013). High-Resolution 3-D Mapping of Soil Texture in Denmark. *Soil Science Society of America Journal*, *77*(3), 860–876. https://doi.org/10.2136/sssaj2012.0275

Alencar, A. A., Azevedo, T., Conciani, D. E., Damasceno, C., Souza Jr, C., Cardoso, D., Soares Filho, J., Oliveira Jr, L., Miranda, D. dos R., Souza, D., Costa, D. P., Silva, E. M. da, Santos, F. N. dos, Brito, L. P. de, Galano, S., Franca-Rocha, W. de J. da, Costa, B., Zimbres, B., Martenexen, F., ... Hasenack, H. (2024). *MapBiomas Degradation Module BETA - Algorithm Theoretical Basis Document (ATBD)* [Dataset]. MapBiomas Data. https://doi.org/10.58053/MapBiomas/XPA9ZB

Alencar, A. A., Conciani, D. E., Rosa, E. R., Martin, E. V., Andrade, G., Hasenack, H.,

Martenexen, L. F. M., Ribeiro, J. P. F. M., Shimbo, J., Rosa, M., Dias, M., Crusco, N., Santos, N., Monteiro, N. C., Duverger, S. G., Azevedo, T., Piontekowski, V. J., Arruda, V. L. da S., Silva, W. V. da, & Rocha, W. da F. (2024, July 12). *MapBiomas Fire Brazil Collection 3: Annual Burned Area Maps of Brazil (1985-2023). Algorithm Theoretical Basis Document (ATBD)*. MapBiomas Data, V2. https://doi.org/10.58053/MapBiomas/0KJBRA

Alvares, C. A., Stape, J. L., Sentelhas, P. C., de Moraes Gon, J. L., & Sparovek, G. (2013). Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, *22*(6), 711–728. https://doi.org/10.1127/0941-2948/2013/0507

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, *5*(1), 180040. https://doi.org/10.1038/sdata.2018.40

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., & Domisch, S. (2020). Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. *Scientific Data*, *7*(1), 162. https://doi.org/10.1038/s41597-020-0479-6

Batjes, N. H., Ribeiro, E., & van Oostrum, A. (2020). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, *12*(1), 299–320. https://doi.org/10.5194/essd-12-299-2020

Bernoux, M., Carvalho, M. C. S., Volkoff, B., & Cerri, C. C. (2002). Brazil's soil carbon stocks. *Soil Science Society of America Journal*, *66*, 888–896. https://doi.org/10.2136/sssaj2002.8880

Brasil. (2015). *Intended Nationally Determined Contribution Towards Achieving the Objective of the United Nations Framework Convention on Climate Change*. República Federativa do Brasil. https://antigo.mma.gov.br/images/arquivo/80108/BRAZIL%20iNDC%20english%20FINAL.pdf

Bronick, C. J., & Lal, R. (2005). Soil structure and management: A review. *Geoderma*, *124*(1–2), 3–22. https://doi.org/10.1016/j.geoderma.2004.03.005

Camargo, F. A. O., Alvarez, V. H., & Baveye, P. C. (2010). Brazilian soil science: From its inception to the future, and beyond. *Revista Brasileira de Ciência Do Solo*, *34*, 589–599. https://doi.org/10.1590/S0100-06832010000300001

Chagas, C. S., Carvalho Junior, W., Bhering, S. B., Tanaka, A. K., & Baca, J. F. M. (2004). Organization and structure of the Brazilian soil information system (SigSolos – version 1.0). *Revista Brasileira de Ciência Do Solo*, *28*(5), 865–876. https://doi.org/10.1590/S0100-06832004000500009

Chagas, C. S., De Carvalho Junior, W., Bhering, S. B., & Calderano Filho, B. (2016). Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *CATENA*, *139*, 232–240. https://doi.org/10.1016/j.catena.2016.01.001

Cooper, M., Mendes, L. M. S., Silva, W. L. C., & Sparovek, G. (2005). A national soil profile database for Brazil available to international scientists. *Soil Science Society of America Journal*, *69*(3), 649–652. https://doi.org/10.2136/sssaj2004.0140

Cousin, I., Nicoullaud, B., & Coutadeur, C. (2003). Influence of rock fragments on the water retention and water percolation in a calcareous soil. *CATENA*, *53*(2), 97–114. https://doi.org/10.1016/S0341-8162(03)00037-7

FAO. (2020). *Technical specifications and country guidelines for Global Soil Organic Carbon Sequestration Potential Map (GSOCseq)*. Food and Agriculture Organization of the United Nations. http://www.fao.org/3/cb0353en/cb0353en.pdf

FAO. (2022a). *Global Soil Organic Carbon Map – GSOCmap v.1.6—Technical Report* (p. 238). Food and Agriculture Organization of the United Nations. https://doi.org/10.4060/cb9015en

FAO. (2022b). *Global status of black soils*. FAO. https://doi.org/10.4060/cc3124en

GlobalSoilMap. (2015). *Specifications Tiered GlobalSoilMap products*. https://www.isric.org/sites/default/files/GlobalSoilMap_specifications_december_2015_2.pdf

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G. R., & Filho, E. I. F. (2019). Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma*, *340*, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

Gudmundsson, L. (2025). *qmap: Statistical Transformations for Post-Processing Climate Model Output* (Version 1.0-6) [Computer software]. https://cran.r-project.org/web/packages/qmap/index.html

Hanson, B., Sugden, A., & Alberts, B. (2011). Making Data Maximally Available. *Science*, *331*(6018), 649–649. https://doi.org/10.1126/science.1203354

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed., p. 746). Springer.

Hengl, T., Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS One*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Heuvelink, G. B. M., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G. F., & Sanderman, J. (2020). Machine learning in space and time for modelling soil organic carbon change. *European Journal of Soil Science*. https://doi.org/10.1111/ejss.12998

IBGE. (2012). *Manual Técnico da Vegetação Brasileira* (2a edição revista e ampliada). Instituto Brasileiro de Geografia e Estatística. https://biblioteca.ibge.gov.br/visualizacao/livros/liv63011.pdf

IBGE. (2015). *Manual Técnico de Pedologia* (3rd ed., p. 430). Instituto Brasileiro de Geografia e Estatística, Coordenação de Recursos Naturais e Estudos Ambientais. https://biblioteca.ibge.gov.br/visualizacao/livros/liv95017.pdf

IBGE. (2019a). *Biomas e sistema costeiro-marinho do Brasil: Compatível com a escala 1:250 000* (Vol. 45). Instituto Brasileiro de Geografia e Estatística, Coordenação de Recursos Naturais e Estudos Ambientais. https://biblioteca.ibge.gov.br/visualizacao/livros/liv101676.pdf

IBGE. (2019b). *Províncias estruturais, compartimentos de relevo, tipos de solos e regiões fitoecológicas*. Instituto Brasileiro de Geografia e Estatística. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101648

IPCC (with Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Jan C. Minx, Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., Eickemeier, P., Savolainen, J., Schlömer, S., Stechow, C. von, & Zwickel, T.). (2014). *Climate change 2014: Mitigation of climate*

change Working Group III contribution to the fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press. https://www.ipcc.ch/report/ar5/wg3/

Kämpf, N. (1971). *Mineralogia e Gênese de alguns Solos da Região Nordeste do Planalto Riograndense* (p. 105) [Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia]. http://www.ufrgs.br/agronomia/materiais/19714dt.pdf

Lal, R. (2013). Soil carbon management and climate change. *Carbon Management*, *4*(4), 439–462. https://doi.org/10.4155/cmt.13.31

Lal, R., & Shukla, M. K. (2004). *Principles of Soil Physics*. Marcel Dekker, Inc.

Mitran, T., Suresh, J., Sujatha, G., Sreenivas, K., Karak, S., Kumar, R., Chauhan, P., & Meena, R. S. (2024). Digital Soil Mapping: A Tool for Sustainable Soil Management. In Md. M. Rahman, J. C. Biswas, & R. S. Meena (Eds.), *Climate Change and Soil-Water-Plant Nexus: Agriculture and Environment* (pp. 51–95). Springer Nature. https://doi.org/10.1007/978-981-97-6635-2_3

Ottoni, M. V., Filho, T. B. O., Schaap, M. G., Lopes-Assad, M. L. R. C., & Filho, O. C. R. (2018). Hydrophysical Database for Brazilian Soils (HYBRAS) and pedotransfer functions for water retention. *Vadose Zone Journal*, *17*(1), 1–17. https://doi.org/10.2136/vzj2017.05.0095

Ottoni, M. V., Lopes-Assad, M. L. R. C., Pachepsky, Y., & Filho, O. C. R. (2014). A hydrophysical database to develop pedotransfer functions for Brazilian soils: Challenges and perspectives. In W. G. Teixeira, M. B. Ceddia, M. V. Ottoni, & G. K. Donnagema (Eds.), *Application of soil physics in environmental analyses* (pp. 467–494). Springer International Publishing. https://doi.org/10.1007/978-3-319-06013-2_20

Peralta, G., Di Paolo, L., Luotto, I., Omuto, C., Mainka, M., Viatkin, K., & Yigini, Y. (2022). *Global soil organic carbon sequestration potential map (GSOCseq v1.1)—Technical manual*. FAO. https://doi.org/10.4060/cb2642en

Poesen, J., & Lavee, H. (1994). Rock fragments in top soils: Significance and processes. *CATENA*, *23*(1–2), 1–28. https://doi.org/10.1016/0341-8162(94)90050-7

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, *7*(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

Samuel-Rosa, A., Dalmolin, R. S. D., & Miguel, P. (2013). Building predictive models of soil particle-size distribution. *Revista Brasileira de Ciência Do Solo*, *37*(2), 422–430. https://doi.org/10.1590/S0100-06832013000200013

Samuel-Rosa, A., Dalmolin, R. S. D., Moura-Bueno, J. M., Teixeira, W. G., & Alba, J. M. F. (2020). Open legacy soil survey data in Brazil: Geospatial data quality and how to improve it. *Scientia Agricola*, *77*(1), e20170430. https://doi.org/10.1590/1678-992x-2017-0430

Samuel-Rosa, A., & Vasques, G. M. (2017). Dados para aplicações pedométricas em larga escala no Brasil. *Boletim Informativo da SBCS*, *43*(3), 22–24. https://www.sbcs.org.br/wp-content/uploads/2018/01/boletimsbcs32017ebook_03_01_2018_10_45_30_id_36404.pdf

Santos, H. G. dos, Jacomine, P. K. T., Anjos, L. H. C. dos, Oliveira, V. Á. de, Lumbreras, J. F., Coelho, M. R., Almeida, J. A. de, Araújo Filho, J. C. de, Oliveira, J. B. de, & Cunha, T. J. F. (2018). *Sistema brasileiro de classificação de solos* (5th ed., p. 531). Embrapa. https://www.embrapa.br/solos/sibcs

Santos, R. D. dos, Santos, H. G. dos, Ker, J. C., Anjos, L. H. C. dos, & Shimizu, S. H. (2015). *Manual de Descrição e Coleta de Solo no Campo* (7th ed., p. 102). Sociedade Brasileira de Ciência do Solo. https://www.sbcs.org.br/

Sato, M. V. (2015). *Primeira aproximação da biblioteca espectral de solos do Brasil: Caracterização de espectros de solos e quantificação de atributos* (p. 102) [Master's thesis, Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz]. https://doi.org/10.11606/d.11.2015.tde-15102015-152045

SEEG. (2022). *Sistema de Estimativas de Emissões e Remoções de Gases de Efeito Estufa: Setor Mudança de Uso da Terra e Florestas* (4th ed.). Instituto de Pesquisa Ambiental da Amazônia. https://seeg-br.s3.amazonaws.com/Notas%20Metodologicas/SEEG_9%20%282022%29%20com%20Municipios/Nota_Metodologica_MUT_SEEG9_2022.05.23.pdf

Sharma, P. V. (1997). *Environmental and Engineering Geophysics* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139171168

Souza Jr, C., Schirmbeck, J., Ferreira Gama, B., Medeiros Brandão, I., Ribeiro Ferreira, J. G., Costa, D., Vasconcelos, R., Galano, S., Franca-Rocha, W., Vélez Martin, E., Wolfarth Schirmbeck, L., Pereira, J. J., E. Conciani, D., Reis Rosa, E., Dias, M., S. A.Neto, H., M. C.Silva, I., dos Reis Azevedo, R., Shimbo, J., ... Azevedo, T. (2025). *MapBiomas Water Brazil General"Handbook"—Algorithm Theoretical Basis Document (ATBD)- Collection 3* [Dataset]. MapBiomas Data. https://doi.org/10.58053/MapBiomas/NH82QZ

Tetegan, M., Richer De Forges, A. C., Verbeque, B., Nicoullaud, B., Desbourdes, C., Bouthier, A., Arrouays, D., & Cousin, I. (2015). The effect of soil stoniness on the estimation of water retention properties of soils: A case study from central France. *CATENA*, *129*, 95–102. https://doi.org/10.1016/j.catena.2015.03.008

Twala, B. E. T. H., Jones, M. C., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950–956. https://doi.org/10.1016/j.patrec.2008.01.010

Vasques, G. M., Coelho, M. R., Dart, R. de O., Cintra, L. C., & Baca, J. F. M. (2021). *Soil Clay, Silt and Sand Content Maps for Brazil at 0-5, 5-15, 15-30, 30-60, 60-100 and 100-200 cm Depth Intervals with 90 m Spatial Resolution. Version 2021 – Technical Report*. Embrapa Solos. https://geoinfo.dados.embrapa.br/catalogue/#/document/2640

Vasques, G. M., Coelho, M. R., Dart, R. O., Baca, J. F. M., & Mendonça-Santos, M. de L. (2021). *Mapa de estoque de carbono orgânico do solo a 0-30 cm do Brasil na resolução espacial de 1 km—Versão 2021* [Map]. Embrapa Solos. https://geoinfo.dados.embrapa.br/catalogue/#/document/4752

Vasques, G. M., Dart, R. de O., Baca, J. F. M., Ceddia, M. B., & Mendonça-Santos, M. de L. (2017). *Mapa de estoque de carbono orgânico do solo (COS) a 0-30 cm do Brasil* [Map]. Embrapa Solos. https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1085197/mapa-de-estoque-de-carbono-organico-do-solo-cos-a-0-30-cm-do-brasil

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. *Water Resources Research*, *55*(6), 5053–5073. https://doi.org/10.1029/2019WR024873

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., & Bates, P. D. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, *44*(11), 5844–5853. https://doi.org/10.1002/2017GL072874

Appendix 1

Static and dynamic environmental covariates used to train random regression forest models for soil organic carbon (SOC) stock and particle size distribution (PSD)

The environmental covariates used in this collection are documented in this appendix, categorized by their primary source and thematic nature. Each covariate group includes a detailed description, the number of individual predictors (p), and the temporal dimension (t), along with technical specifications such as original resolution and data source.

Each variable presented in this appendix includes a reference to its specific processing directory, distinguishing between scripts used for generating pre-exported assets and those used for general data preparation during the modeling phase. The processing workflow and the location of the source code for each variable are explicitly indicated within the official MapBiomas GitHub repository (https://github.com/mapbiomas/brazil-soil/tree/main/soil_30m_landsat).

Soil Classes (Hengl et al., 2017)

Soil class probabilities (WRB, single and multiple classes, %) from SoilGrids (**p = 10, t = 1**). Data were originally at 250-m spatial resolution and processed through a focal mean interpolation with a 1000-m radius applied three times, followed by a blend with the original data. The resulting image was then bilinearly resampled to 30-m spatial resolution. These are continuous, static (single-band) variables.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/ISRIC_2017_WRB_250M

	Covariate ID	Definition	PSD	SOC
1	Argisols	Probability of occurrence of Alisols + Luvisols + Acrisols + Lixisols + Planosols (%), soil classes distinguished by an increase in clay content with depth.	Y	N
2	Ferralsols	Probability of occurrence of Ferralsols (%), which are characterized as highly weathered soils exhibiting diffuse horizon boundaries, a clay assemblage predominantly composed of low activity clays (primarily kaolinite), and an elevated content of sesquioxides	Y	Y
3	Histosols	Probability of occurrence of Histosols (%), defined as soils primarily composed of organic soil material	Y	Y
4	Humisols	Probability of occurrence of Chernozems + Kastanozems + Phaeozems + Umbrisols (%), soil classes distinguished by elevated organic matter levels and dark-colored surface horizons, frequently correlated with fertile, carbon-rich soils	Y	Y
5	Nitisols	Probability of occurrence of Nitisols (%), a soil class with well-drained, deeply weathered profiles, and gradual transitions between horizons, with a subsurface horizon containing more than 30% clay	Y	Y

6	Plinthosols	Probability of occurrence of Plinthosols (%), representing soils with a high content of coarse fragments or indurated materials, such as plinthite, petroplinthite, or ferruginous concretions. These soils typically develop under alternating wet and dry conditions, where iron accumulation and hardening processes lead to restricted rooting depth and reduced soil permeability.	Y	Y
7	Sandysols	Probability of occurrence of Arenosols + Podzols (%), soil classes distinguished by sandy-textured topsoil horizons	Y	Y
8	Thinsols	Probability of occurrence of Leptosols + Regosols (%), soil classes identified by their shallow depth and restricted development, and typically located in regions characterized by steep gradients, erosional processes, or geologically recent formations	Y	Y
9	Vertisols	Probability of occurrence of Vertisols (%), characterized as heavy clay soils with a significant proportion of swelling clays	Y	Y
10	Wetsols	Probability of occurrence of Gleysols + Planosols + Stagnosols (%), soil classes characterized by impeded drainage, water saturation, and hydromorphic attributes, commonly identified in areas of low elevation or those prone to inundation	Y	Y

Pedology (IBGE, 2015)

Pedological regions (**p = 18, t = 1**) represent the official soil classification units of Brazil, according to the Brazilian Soil Classification System (SiBCS). IBGE pedological map was processed to generate a set of 18 binary indicator variables at a 30 m spatial resolution. The classification follows the first categorical level (Soil Order) for most classes, with the exception of Neossolos, which were disaggregated into their second categorical level (Suborder) to identify Flúvicos, Quartzarênicos, Regolíticos, and Litólicos units. Each band represents the presence (1) or absence (0) of a specific soil unit, while non-pedological classes, such as water bodies, were excluded. This dataset is considered static and categorical.

Processing code and/or asset:

- soil_30m_landsat/collection_03beta/covariate_export/IBGE_2023_PEDOLOGIA_250 MIL_2025.js

	Covariate ID	Definition	PSD	SOC
1	ARGISSOLO	Soils with a diagnostic textural B horizon, showing a significant increase in clay content from the surface to deeper layers.	Y	Y
2	CAMBISSOLO	Incipient soils with a diagnostic incipit B horizon, characterized by intermediate weathering and the presence of primary minerals.	Y	Y
3	CHERNOSSOLO	Highly fertile soils with a dark, organic-rich surface layer (chernozêmico A horizon) and high base saturation, common in temperate or specific lithological contexts.	Y	Y

4	ESPODOSSOLO	Soils characterized by the subsurface accumulation of organic matter and aluminum (espódico B horizon), typically sandy and associated with hydromorphic coastal environments.	Y	Y
5	GLEISSOLO	Hydromorphic soils formed under prolonged water saturation, resulting in gleying processes (reduction of iron) and typical grayish colors.	Y	Y
6	LATOSSOLO	Highly weathered, deep, and homogeneous soils with a diagnostic latossólico B horizon, characterized by low nutrient reserve and high oxide content (Fe and Al).	Y	Y
7	LUVISSOLO	Soils with a textural B horizon and high base saturation, often shallow and common in semi-arid regions with limited weathering.	Y	Y
8	NEOSSOLO_FLUVICO	Soils formed from recent alluvial deposits, showing stratified layers (fluvic character) without a diagnostic B horizon.	Y	Y
9	NEOSSOLO_LITOLICO	Very shallow soils where the consolidated rock or a lithic contact occurs within the first 50 cm of the surface.	Y	Y
10	NEOSSOLO_QUARTZARENICO	Deep, excessively drained soils with a sandy texture throughout the profile.	Y	Y
11	NEOSSOLO_REGOLITICO	Poorly developed soils formed from weathered rock material (saprolite), lacking a diagnostic B horizon and deeper than 50 cm.	Y	N
12	NITOSSOLO	Deep, clayey soils with a nítico B horizon, showing a shiny or waxy appearance (ped face luster) and high structural stability.	Y	Y
13	ORGANOSSOLO	Organic soils formed by the accumulation of plant debris under waterlogged or high-altitude cold conditions, with very high organic carbon content.	Y	Y
14	PLANOSSOLO	Soils with a sudden textural change and a restrictive subsurface layer (plânico B horizon), leading to temporary waterlogging and perched water tables.	Y	Y
15	PLINTOSSOLO	Soils characterized by the presence of plinthite, a clay-rich, iron-mottled material that can harden irreversibly into petroplinthite upon drying.	Y	Y
16	VERTISSOLO	Soils with high expansive clay content that shrink and swell with moisture changes, forming deep cracks during dry periods and typically high fertility.	Y	N

Black Soils (FAO, 2022b)

Black soil probability (single class, %) from FAO GBSmap (**p = 1, t = 1**). The original 1-km spatial resolution data were smoothed using a three-iteration focal mean with a 1000-m

radius, then blended with the original data. The resulting image was bilinearly resampled to 30-m spatial resolution. This is a continuous, static (single-band) variable.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/FAO_2022_BLACKSOIL_1KM

	Covariate ID	Definition	PSD	SOC
1	black_soil_prob	Probability of occurrence of black soils (%), defined as mineral soils exhibiting moderate to high soil organic carbon content	Y	Y

Particle Size Distribution (MapBiomas Soil)

Percent content of clay, silt, and sand from MapBiomas Soil Collection 2 (**p = 3, t = 1**) and Collection 3 (**p= 3, t = 1**), provided at a 30-meter spatial resolution. These are continuous, static variables.

Processing code:

- soil_30m_landsat/collection_03beta/carbon/0_covariate_source
- soil_30m_landsat/collection_03beta/texture/0_covariate_source

	Covariate ID	Definition	PSD	SOC
1	clay_000_030cm	Clay content (0–0.002 mm) in the fine earth fraction (0-30 cm) (%) from MapBiomas Soil Collection 2	Y	N
2	silt_000_030cm	Silt content (0.002–0.05 mm) in the fine earth fraction (0-30 cm) (%) from MapBiomas Soil Collection 2	Y	N
3	sand_000_030cm	Sand content (0.05–2.0 mm) in the fine earth fraction (0-30 cm) (%) from MapBiomas Soil Collection 2	Y	N
4	argila_000_030cm	Clay content (0–0.002 mm) in the fine earth fraction (0-30 cm) (%) from MapBiomas Soil Collection 3	N	Y
5	silte_000_030cm	Silt content (0.002–0.05 mm) in the fine earth fraction (0-30 cm) (%) from MapBiomas Soil Collection 3	N	Y
6	areia_000_030cm	Sand content (0.05–2.0 mm) in the fine earth fraction (0-30 cm) (%) from MapBiomas Soil Collection 3	N	Y

Land Surface Variables (Amatulli et al., 2018, 2020; Yamazaki et al., 2017, 2019)

Land surface variables (**p = 13, t = 1**) derived from MERIT DEM and Geomorpho90m datasets with an original spatial resolution of 3 arc seconds (approximately 90 meters at the equator). The data underwent a process of gap-filling and subsequent resampling to a 30-meter spatial resolution using nearest neighbor interpolation. This variable is static and continuous.

Processing code:

- soil_30m_landsat/collection_03beta/carbon/0_covariate_source
- soil_30m_landsat/collection_03beta/texture/0_covariate_source

	Covariate ID	Definition	PSD	SOC
1	convergence	Convergence index, highlighting the convergent areas as channels and divergent areas as ridges. (3 x 3 window)	Y	Y
2	cross_sectional	Cross-sectional (transversal) curvature represents the intersection between the slope normal and the line perpendicular to the slope direction (aspect). This metric is derived from the MERIT Digital Elevation Model using a 15 × 15-pixel convolution window.	Y	Y
3	cti	Compound topographic index, also known as the topographic wetness index, calculated as the logarithm of the upstream cumulative catchment area divided by the local slope tangent, multiplied by 10. (3 x 3 window)	Y	Y
4	dev_magnitude	Maximum multiscale deviation, representing ruggedness. (3 x 3 window)	Y	Y
5	dev_scale	Scale of the maximum multiscale deviation, representing ruggedness. (3 x 3 window)	Y	Y
6	eastness	Eastness index, representing eastward exposure, calculated as the sine of the slope multiplied by the sine of the aspect, scaled by 100 (%). (3 x 3 window)	Y	Y
7	elevation	Elevation above sea level (m).	Y	Y
8	elev_stdev	Elevation standard deviation, representing ruggedness. (3 x 3 window)	Y	N
9	longitudinal_curvature	Longitudinal curvature represents the intersection between the slope normal and the slope direction (aspect). This metric is derived from the MERIT Digital Elevation Model using a 15 × 15-pixel convolution window.	Y	Y
10	northness	Northness index, representing northward exposure, calculated as the sine of the slope multiplied by the cosine of the aspect, scaled by 100 (%). (3 x 3 window)	Y	Y
11	roughness	Roughness, representing the maximum absolute difference in elevation between a focal cell and its eight surrounding cells. (3 x 3 window)	Y	N
12	slope	Slope (%), the rate of elevation change in the direction of water flow, calculated using the equation tan((3.141593 / 180) * degrees) * 100. (3 x 3 window)	Y	Y
13	spi	Stream power index, representing the potential erosive power of flowing water, calculated as the product of the upstream catchment area and the tangent of the local slope angle, transformed using the equation log10(spi + 1) * 100. (3 x 3 window)	Y	Y

Köppen climate classification (Alvares et al., 2013)

The Köppen climate classification (**p = 19, t = 1**), originally at a 100-meter spatial resolution, was resampled to a 30-meter spatial resolution using nearest neighbor interpolation. Each climate class at the first, second, and third categorical level was then converted to a binary (0 or 1) variable. This resulted in a static, categorical dataset.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/IPEF_2013_KOPPEN_100M

	Covariate ID	Definition	PSD	SOC
1	koppen_l1_A	A - Tropical	Y	Y
2	koppen_l2_Af	Af - Tropical without dry season	Y	Y
3	koppen_l2_Am	Am - Tropical monsoon	Y	Y
4	koppen_l2_As	As - Tropical with dry summer	Y	Y
5	koppen_l2_Aw	Aw - Tropical with dry winter	Y	Y
6	koppen_l1_B	B - Dry	Y	N
7	koppen_l2_Bs	BS - Dry semi-arid steppe: annual precipitation is below the threshold but above 50%	Y	N
8	koppen_l3_BSh	BSh - Dry semi-arid low latitude and altitude	Y	Y
9	koppen_l1_C	C - Humid subtropical	Y	Y
10	koppen_l2_Cf	Cf - Humid subtropical oceanic climate, without dry season	Y	Y
11	koppen_l3_Cfa	Cfa - Humid subtropical oceanic climate, without dry season with hot summer	Y	Y
12	koppen_l3_Cfb	Cfb - Humid subtropical oceanic climate, without dry season with temperate summer	Y	Y
13	koppen_l2_Cw	Cw - Humid subtropical with dry winter	Y	Y
14	koppen_l3_Cwa	Cwa - Humid subtropical with dry winter and hot summer	Y	Y
15	koppen_l3_Cwb	Cwb - Humid subtropical with dry winter and temperate summer	Y	Y
16	koppen_l3_Cwc	Cwc - Humid subtropical with dry winter and short and cool summer	Y	N
17	koppen_l2_Cs	Cs - Humid subtropical with dry summer	Y	N
18	koppen_l3_Csa	Csa - Humid subtropical with dry summer and hot	Y	N
19	koppen_l3_Csb	Csb - Humid subtropical with dry summer and temperate	Y	N

Biome (IBGE, 2019a)

Biomes (**p = 7, t = 1**) are defined as geographically distinct areas with similar climates, soils, and vegetation. The data, originally at a scale of 1:250,000, was rasterized to a 30 m spatial resolution. Subsequently, the biome classes were converted into binary variables (0 or 1) to enable their use as static, categorical variables within the model.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/IBGE_2019_BIOMAS_ZN_30m_DUMMY

	Covariate ID	Definition	PSD	SOC
1	Amazonia	Amazon biome, characterized by a predominantly hot and humid climate, a prevalence of forest physiognomy, geographic continuity, a peri-equatorial location, and the context of the Amazon Basin itself	Y	Y
2	Caatinga	Caatinga biome, a semi-arid space characterized by open and sparse woodland (Savanna-Steppe), with a predominance of humid and sub-humid tropical climates	Y	Y
3	Cerrado	Cerrado biome, characterized by a predominance of formations with savanna physiognomies, a tropical hot sub-humid climate with two distinct seasons – one dry and one rainy – and fire	Y	Y
4	Mata_Atlantica	Mata Atlântica biome, an environmental complex that incorporates mountain ranges, plateaus, valleys, and plains along the entire eastern Brazilian Atlantic continental strip	Y	Y
5	Pampa	Pampa biome, also known as Estepe, characterized by a rainy climate without a dry season, but with cold winters and negative temperatures, causing plant physiological seasonality typical of a cold, dry climate	Y	Y
6	Pantanal	Pantanal biome, characterized by the geomorphological unit known as the Pantanal Plain, constituting the world's largest inland floodplain, distinguished by soils of predominantly low permeability and an area with an effectively negligible gradient	Y	Y
7	Zona_Costeira	The Coastal Zone (Zona Costeira), is a transitional strip between land and sea that extends across 17 Brazilian states. It is characterized by the interaction of geographic factors such as topography, climate, and ocean dynamics. This area holds great ecological importance, as it encompasses ecosystems such as mangroves, coral reefs, cliffs, and coastal vegetation.	Y	Y

Phyto Ecological Regions (IBGE, 2012, 2019b)

Phyto ecological regions (**p = 11, t = 1**) are areas where specific groups of plant genera and life forms recur in similar climatic conditions. These regions may have varying soil types but share a well-defined relief. The data, originally at a scale of 1:250,000, was rasterized to a 30 m spatial resolution. The phytoecological regions (or vegetation types) were then converted into binary variables (0 and 1) for use in the model. The transition classes "Contact" (Ecotone and Encrave) features were decomposed and assigned to their respective pure

phytoecological based on the metadata field `nm_contat`. This approach allows transitional zones to be represented in multiple binary bands, capturing the complex soil-vegetation relationships inherent in ecological tension areas. This dataset is considered static and categorical.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/IBGE_2023_FITOFISIONOMIAS_250MIL

	Covariate ID	Definition	PSD	SOC
1	Campinarana	Campinarana, a type of Amazonian vegetation, can be dense or open arboreal, shrubby, or grass-woody. It is typically found in areas with leached accumulations and plains containing Espodossolos and Neossolos Quartzarênicos, featuring biological forms adapted to these nearly always waterlogged soils and a super-humid hot climate.	Y	Y
2	Estepe	The Estepe phytoecological region is predominantly a grassland environment subject to or experiencing a cold season, designating typically temperate fields with rainfall throughout the year, encompassing most of the fields in the southern region.	Y	Y
3	Floresta_Estacional_Decidual	Floresta Estacional Decidual (Seasonal Deciduous Forest), refers to a type of deciduous forest where more than 50% of the trees lose their leaves during an unfavorable period. This occurs in two distinct situations: in the tropical zone, it follows a rainy season with a dry period, and in the subtropical zone, without a dry period, but with a cold winter.	Y	Y
4	Floresta_Estacional_Semidecidual	Floresta Estacional Semidecidual (Seasonal Semideciduous Forest), a semi-deciduous forest established by a seasonal climate causing partial leaf fall. In the tropics, this is due to a dry winter and intense summer rains. In subtropical zones, it's due to a very cold winter causing physiological rest and leaf fall. The percentage of deciduous trees is usually between 20% and 50%.	Y	Y
5	Floresta_Estacional_Sempre_Verde	Floresta Estacional Sempre Verde (Seasonal Evergreen Forest), a vegetation type that maintains high greenness throughout the dry season. This forest is mainly composed of Amazonian species that exhibit minimal to no leaf shedding during dry periods. This is due to higher soil moisture content or the trees' presumed capacity to access water from deeper soil layers during dry seasons.	Y	Y
6	Floresta_Ombrofila_Aberta	Floresta Ombrófila Aberta (Open Ombrophilous Forest), located in a hot and humid climate with abundant rainfall and a short dry season (2-3 months). This forest is dominated by rosette phanerophytes and woody lianas	Y	Y

		and includes forest communities with palms, bamboo, sororoca, and lianas.		
7	Floresta_Ombrofila_Densa	Floresta Ombrófila Densa (Dense Ombrophilous Forest), characterized by a predominance of trees, woody lianas, and epiphytes. It occurs in a hot and humid climate with rainfall exceeding 2,300 mm and average annual temperatures between 22°C and 25°C in the northern region of the country.	Y	Y
8	Floresta_Ombrofila_Mista	Floresta Ombrófila Mista, also known as "mata-de-araucária" or "pinheiral," is a forest with a mix of tropical and temperate tree species. It grows at altitudes above 500-600 meters in a rainy climate without a dry season. Average temperatures are 18°C with cold winters lasting 3-6 months where average temperatures are below 15°C.	Y	Y
9	Formacao_Pioneira	Formações Pioneiras, refers to pioneer vegetation, which is the first vegetation to occupy rejuvenated land. This land includes areas with repeated marine sand deposits on beaches and coastal dunes, fluviomarine alluvial deposits at river mouths, and alluvial and lacustrine riparian soils.	Y	Y
10	Savana	Savanna, predominantly covering the Brazilian Central Plateau, is characterized by a seasonal climate with a dry period of 3 to 5 months. Its vegetation, consisting of trees, shrubs, and grasses adapted to dry conditions, is typically found in aluminized soils.	Y	Y
11	Savana_Estepica	Savana-Estépica, a type of vegetation defined by grasslands with deciduous and thorny trees. It has varying climates: in the Northeast Sertão, it has two annual dry periods; in the Chaco Mato-Grossense-do-Sul Disjunction, it has double seasonality with both a physiological drought and a rainy period; and in the southern Savana-Estepe, there is no dry period but it is affected by cold fronts.	Y	Y

Spatial coordinates

Spatial coordinates (**p = 1, t = 1**), specifically latitude and longitude in decimal degrees, for each pixel representing a location within the Brazilian territory. These coordinates have been transformed to a strictly positive range. The variables in this dataset are static and continuous. Processing code:

- soil_30m_landsat/collection_03beta/carbon/0_covariate_source
- soil_30m_landsat/collection_03beta/texture/0_covariate_source

Covariate ID	Definition	PSD	SOC
longitude	Longitude in decimal degrees transformed to the positive range and multiplied by 1000	N	N

Structural Provinces and subprovinces (IBGE, 2019b)

Structural provinces (**p = 14, t = 1**) are large natural geological areas that present their own stratigraphic, magmatic, tectonic and metamorphic evolution and are different from those of neighboring provinces, portraying the corresponding geological substrate. The mapping of provinces was originally at a scale of 1:250,000 and rasterized to a spatial resolution of 30 m; the classes (structural provinces and subprovinces) were converted into binary variables (0 and 1). A 3 km morphological dilation was applied to each binary band to ensure spatial continuity and mitigate rasterization artifacts. These are static, categorical variables.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/IBGE_PROVINCIAIS_ESTRUTURAIS_250mil

	Covariate ID	Definition	PSD	SOC
1	Amazonia_Provincia	Structural Province of the Amazon, encompassing the Northern Region and part of the Central-West Region of Brazil, presents radiometric ages that decrease from east to west and a diversified geology, which resulted in lands and environments rich in minerals, such as the Carajás Mineral Province.	Y	Y
2	Amazonas_Solimoes_Provincia	Amazonas-Solimões Structural Province, a vast sedimentary province with diverse sedimentary geological formations, distributed throughout Acre, Amapá, Amazonas, Pará, Roraima and Rondônia. This diversity influences the hydrography and the formation of low-lying areas, such as the Amazon Basin.	Y	Y
3	Borborema_Provincia	Structural Province Borborema presents a complex geological substrate, with ancient terrains and mobile belts, as well as rifts and alkaline complexes. This diversity, influenced by neighboring geological units, shapes the variety of soils and relief in northeastern Brazil.	Y	Y
4	Cobertura_Cenozoica_Provincia	Cenozoic Cover Structural Province, located on the South American Platform, presents post-Gondwana continental sedimentary sequences, including detrital and lateritic deposits. Influenced by global climatic variations of the Cenozoic, the area underwent processes of erosion and formation of extensive detrito-lateritic covers.	Y	Y
5	Costeira_Margem_Continental_Provincia	Coastal and Continental Margin Structural Province, contains Cenozoic deposits and coastal marginal basins, formed due to the opening of the Atlantic Ocean. The passive margin contains segments with different tectonic natures, which influenced the depositional sequences. The Coastal Plain resulted from Quaternary marine transgressions.	Y	Y
6	Gurupi_Provincia	Gurupi Structural Province, located mainly in Pará, part	Y	N

		of the Gurupi Belt, whose nature and age are still debated. It may be either a remnant of Paleoproterozoic structures or a Brasiliano belt delimiting a Neoproterozoic craton, its evolution remains uncertain.		
7	Mantiqueira_Provincia	Mantiqueira Structural Province, formed during the Brasiliano Orogenic Cycle, characterized by granulitized and deformed terrains, and intrusions of basic dikes from the Mesozoic. Its limits encompass a vast area, influencing the relief and diversity of landscapes.	Y	Y
8	Parana_Provincia	Paraná Structural Province, marked by the Paleozoic Paraná Basin, the Mesozoic Bauru-Caiuá Basin and volcanic and plutonic activities, presents six depositional supersequences influenced by tectonic and marine events. This geology shapes the relief, including the Planalto dos Campos Gerais and the Arco de Ponta Grossa.	N	N
9	Parecis_Provincia	Parecis Structural Province, located in Mato Grosso and Rondônia, a Gondwana sedimentary basin with sedimentation from the Ordovician to the Permian and diverse depositional environments. It is marked by eolian sediments and basaltic volcanism in the Mesozoic, fluvial and eolian sediments and kimberlitic intrusions in the Cretaceous, and covered by sandstones, pelites, and lateritic crust in the Cenozoic.	Y	Y
10	Parnaiba_Provincia	Structural Province Parnaíba, located in six states, characterized by Paleozoic and Mesozoic basins, with exposed sedimentary rocks and vulcanoplutonism, influencing the landscape and soils of the region.	Y	Y
11	Reconcavo_Tucano_Jatoba_ Provincia	Recôncavo-Tucano-Jatobá Structural Province, with an extensive sedimentary cover, formed during the fragmentation of Gondwana in the Lower Cretaceous. It is composed of basins with sediments from the Paleozoic, which influence the distribution of aquifers and the formation of the regional relief.	Y	Y
12	Sao_Francisco_Provincia	São Francisco Structural Province, formed by the fragmentation of the supercontinent Gondwana, constituted by an Archean basement, intracratonic covers and foreland basins.	Y	Y
13	Sao_Luis_Provincia	São Luís Structural Province, located mainly in Maranhão, is a fragment of the West African Craton, separated after the fragmentation of Gondwana. Its age and evolution are still debated, but its geotectonic importance justifies its definition as a distinct province, separated from the Parnaíba Province.	Y	N
14	Tocantis_Provincia	Structural Province of Tocantins, located in the central region of Brazil, is formed by a system of Neoproterozoic Brazilian orogens. It is covered by Phanerozoic basins and Cenozoic covers, and its complex geological history	Y	Y

		influences the characteristics of the relief and soils in the region.		

Additionally, subprovinces (**p = 5, t = 1**) were derived from the province map to represent tectonic subdivisions of the provinces, resulting from orogenic events of continental collisions or their fragmentation that occurred in the evolution of each province to its current constitution. They are also known as tectonic domains of the structural provinces. subprovinces, only pure lithological units were considered, excluding mixed or transitional units. A 3 km morphological dilation was applied to each binary band to ensure spatial continuity and mitigate rasterization artifacts. These are static, categorical variables.

	Covariate ID	Definition	PSD	SOC
1	metamorficas	The Metamorphic Subprovinces encompass a set of metamorphic rocks with varying degrees of metamorphism. These subprovinces comprise dominant rocks with processes of uplift temperature and pressure that resulted in metamorphism. They mainly cover the regions of the massive Brazilian shields.	Y	Y
2	plutonicas	These Plutonicas Subprovincia group together domains of igneous rocks that form when magma cools and solidifies slowly within the earth crust. The rocks present in these subprovinces comprise intrusive rocks and visible on the surface only after erosion removes the rocks that converge them.	Y	N
3	sedimentares	Sedimentary Subprovinces comprise a large group of sedimentary rocks formed by the accumulation and consolidation of sediments, such as fragments of other rocks, organic matter, or precipitated minerals. These sub-provinces are mainly composed of clastic rocks, but chemical and organic rock environments also occur. They occur mainly along the sedimentary basins of the Amazon, Paraná, and Parnaíba, as well as in smaller inland basins.	Y	Y
4	sedimentos	Subprovinces composed of sediments encompass sectors with a predominance of surfaces formed by unconsolidated sediments, defined as solid materials (particles of rocks, minerals, or remains of organisms) that are broken off from pre-existing rocks, transported by natural agents such as wind, water, and ice, and subsequently deposited in new locations. The sediments constitute recent formations and occur mainly along watercourses and coastal environments.	Y	Y
5	vulcanicas	Volcanic subprovinces encompass areas composed of igneous rocks formed by the rapid cooling of magma (or lava) when it reaches the Earth's surface during volcanic flows and eruptions. These subprovinces mainly include sectors composed of extrusive rocks generated by large lava flows (predominantly basaltic formations).	Y	Y

Land cover and land use (MapBiomas Coverage C10) (MapBiomas Project, 2025)

Land cover and land use data were obtained from MapBiomas Collection 9 (**p = 16, t = 40**). Each land cover/land use class was converted to a numerical value representing the duration

(in years) of that class on a particular pixel. For the year 1985, a 20-year stable period was assumed for the corresponding land cover/land use class. These are continuous, dynamic variables. Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/MB_2024_AGELULC_C10_v2

	Covariate ID	Definition	PSD	SOC
1	formacaoFlorestal	MapBiomas Class 3. Forest Formations includes various ombrophilous forests, seasonal forests, savannas, and areas with fluvial and/or lacustrine influence. It also covers areas affected by fires or logging, secondary forests, and bamboo forests.	N	Y
2	outrasFormacoesFlorestais	MapBiomas Classes 5 and 6. Mangroves (Class 5) are coastal ecosystems characterized by dense, perennial forest and/or shrub formations, frequently flooded by tides. Floodable Forest (Class 6) corresponds to the Alluvial Open Ombrophilous Forest, which develops along watercourses in plains and terraces subject to periodic or permanent flooding.	N	Y
3	formacaoCampestre	MapBiomas Class 12. Grassland Formations are predominantly composed of herbaceous and shrubby vegetation, including areas of Caatinga, Cerrado, Atlantic Forest, Pampa, and Pantanal. They occur in shallow and deep, well-drained or humid soils, including rocky and sandy terrains. They may present patches of invasive exotic vegetation or areas used for forage.	N	Y
4	formacaoSavanica	MapBiomas Class 4, Savanna Formation, is characterized by herbaceous vegetation and small trees, scattered shrubs, and low-growth vegetation.	N	Y
5	campoAlagadoAreaPantanosa	MapBiomas classes 11 and 32. Wetland (11) includes herbaceous vegetation areas subject to flooding or river/lake influence, potentially including marshy areas with hygrophilous vegetation and aquatic plants, and areas with water surfaces covered by macrophytes or sediments. Hypersaline Tidal Flat (32) includes hypersaline areas without tree or shrub vegetation, located in the transition between mangrove and mainland or dune fields.	N	Y
6	restingas	MapBiomas classes 49 (Wooded Sandbank Vegetation) and 50 (Herbaceous Sandbank Vegetation) represent forest and herbaceous formations found on sandy soils or coastal dunes.	N	Y
7	afloramento	MapBiomas Class 29, Rocky Outcrop, identifies areas where the geological substrate is exposed at the surface with little to no vegetation cover.	N	Y
8	vegNatural	Grouping of natural vegetation classes (Herbaceous and Shrubby Vegetation and Forest) from MapBiomas.	N	Y

9	lavouraTemporaria	MapBiomas agriculture classes, encompassing monoculture areas of Temporary Crops: Soybean (first harvest) (39), Sugarcane (20), Irrigated Rice (40), Cotton (62), and Other Temporary Crops (41).	N	N
10	lavouraPerene	MapBiomas agriculture classes, encompassing monoculture areas of Perennial Crops: Coffee (46), Citrus (47), and Oil Palm (35), along with Other Perennial Crops such as cashew (48).	N	N
11	lavouras	MapBiomas Agriculture classes, encompassing monoculture areas of both temporary and perennial crops.	N	Y
12	pastagem	MapBiomas Class 15. Cultivated pastures are linked to agriculture and livestock, while natural pastures include fields that may or may not be used for grazing. However, in the Amazon, recent deforestation may initially be classified as pasture.	N	Y
13	silvicultura	MapBiomas Class 9, Forest Plantation, encompasses the cultivation of tree species for commercial purposes, such as pine, eucalyptus, and araucaria.	N	Y
14	mosaicoDeUsos	MapBiomas Class 21, Mosaic of Uses, encompasses a diverse range of land cover types. This category includes undifferentiated agricultural areas that fluctuate between pasture and agriculture. It may also feature peri-urban zones, recovering abandoned pastures, anthropized areas (excluding environmental protection areas and indigenous territories), fallow land, and horticulture.	N	Y
15	agropecuaria	Grouping of agricultural land use classes: Agriculture (18), Pasture (15), Forest Plantation (9), and Mosaic of Uses (21) from MapBiomas.	N	Y
16	areia	MapBiomas Class 23, Beach, Dune and Sand, represents non-vegetated areas dominated by sandy substrates, typically found in coastal or inland dune or sandy systems.	N	Y

Additionally, areas with stable land cover/land use throughout the entire period (**p = 3**), both natural and anthropogenic, were mapped as categorical (binary), static variables. Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/MB_2024_STABLEAREAS_C10
- soil_30m_landsat/collection_03beta/covariate_export/MB_2024_DISTANCE_TO

	Covariate ID	Definition	PSD	SOC
1	Area_Estavel	Static binary map illustrating natural and anthropic areas where land use and land cover remained consistent throughout the timeframe represented by the first level of the MapBiomas Collection 10 coverage dataset legend.	N	Y

2	Distance_to_sand	Distance to rock represents the euclidean distance (in meters) to the nearest rocky outcrop. The target locations were defined based on the historical presence of the "Rocky Outcrop" class (class 29), considering pixels that presented this class in at least one year between the time series. The distance was calculated using a Euclidean kernel with a maximum saturation threshold of 7,000 m. This variable acts as a proxy for shallow soils and lithic components in the landscape. This dataset is static and continuous within its range.	Y	Y
3	Distance_to_rock	Distance to sand represents the Euclidean distance (in meters) to the nearest sandy formation, specifically beaches, dunes, and sandbanks. The source areas were identified using the historical occurrence of class 23 at least once in the series. This variable is essential for predicting sand-dominated soil classes, which are characterized by low physical protection of organic matter and, consequently, fragile SOC stocks. This dataset is static and continuous.	Y	Y

Water surface (MapBiomas Water C4) (Souza Jr et al., 2025)

Water surface data from MapBiomas Water Collection 4 (used in MapBiomas Collection 10) were utilized **(p = 1, t = 40)**. For each year from 1985-2024, pixels classified as 'Water' were identified, creating yearly binary masks where 'Water' pixels received a value of 1 and all others received 0. For each year of the series, the number of preceding years up to that point in which a pixel was classified as water was computed. This is a dynamic, continuous variable.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/MB_2024_WATER_C10

	Covariate ID	Definition	PSD	SOC
1	mb_waterRecurrence	Yearly accumulated water surface recurrence. This raster layer indicates the total number of years a pixel was classified as 'water' from 1985 to the evaluation year.	N	N

A static version of the water surface recurrence was also generated from the MapBiomas Water Collection 3 data **(p = 1, t = 1)**. Using the same yearly binary masks from 1985-2024, a final static image was created by summing the values of all 40 annual layers. The resulting continuous, static variable represents the total number of years each pixel was classified as water throughout the entire period, with values ranging from 0 (never water) to 40 (always water).

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/MB_2024_WATER_C10

	Covariate ID	Definition	PSD	SOC
1	mb_water_39y_recurrence	Water surface recurrence from 1985 to 2023, indicating the total number of years a pixel was identified as water throughout the entire 39-year period.	Y	N

Fire Scars (MapBiomas Fire C3) (Alencar, Conciani, et al., 2024)

Data from MapBiomas Fire Collection 3 were used to generate two dynamic fire-related covariates **(p = 2, t = 40)**. First, a fire recurrence variable was created from the annual fire scar maps (1985-2024). For each year in the time series, a cumulative count of how many times a pixel had been mapped as burned from 1985 to the evaluation year was calculated. Second, a 'time after fire' variable was derived from the 'Year of Last Fire' dataset. For each evaluation year, this covariate calculates the time elapsed by taking the difference between the evaluation year and the year of the last recorded fire at that pixel. Both are dynamic, continuous variables.

Processing code:

- soil_30m_landsat/collection_03beta/carbon/0_covariate_source
- soil_30m_landsat/collection_03beta/texture/0_covariate_source

	Covariate ID	Definition	PSD	SOC
1	mb_fire_recurrence_dynamic	Yearly accumulated fire frequency	N	N
2	mb_fire_time_after_fire	Time since the last fire event	N	N

Degradation Vectors (MapBiomas Degradation Beta) (Alencar, Azevedo, et al., 2024)

Data on native vegetation degradation were sourced from the MapBiomas Degradation Beta collection **(p = 1, t = 40)**. The covariate focuses on the 'edgeness' of natural vegetation fragments relative to areas of anthropic use. It measures the potential degradation risk based on six distance classes (30, 60, 90, 120, 150, and 300 meters), assigning a risk value from 6 (highest risk, ≤ 30 m) to 0 (no risk, > 300 m). This results in a dynamic, continuous variable that quantifies the proximity-based degradation pressure on native vegetation for each year.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/MB_2023_DEGRADATION_EDGES_SUM

	Covariate ID	Definition	PSD	SOC
1	mb_summed_edges	Edgeness of natural formations with respect to anthropic land uses, measured across six distance classes to represent potential degradation risk.	N	N

Vegetation Indices (Landsat 5, 7, and 8)

Three dynamic vegetation indices (NDVI, SAVI, and EVI) were derived as dynamic covariates **(p = 3, t = 40)** from annual surface reflectance mosaics of the Landsat 5, 7, 8, and 9 archives (Collection 2, Tier 1) for the period 1985-2024. The composition of these mosaics involved using the quality assessment band to identify and filter contaminated pixels and applying a median reducer function within Google Earth Engine to mitigate cloud and shadow artifacts. Gaps in the annual time series, resulting from persistent cloud cover, were filled via temporal interpolation using values from the immediately preceding and succeeding years. Subsequently, a temporal weighting was applied to the annual index values. For each target year, a weighted average of the index values from the six preceding years was calculated using an exponential decay function (alpha = 0.7). This procedure models the antecedent influence of vegetation conditions on soil organic carbon dynamics (Heuvelink et al., 2020). The resulting products are dynamic, continuous covariates.

Processing code:

- soil_30m_landsat/collection_03beta/covariate_export/LANDSAT_2024_INDICES_DECAY

	Covariate ID	Definition	PSD	SOC
1	mb_ndvi_median_decay	Temporally-weighted Normalized Difference Vegetation Index (NDVI), calculated as an exponential decay-weighted average of the previous six years of annual NDVI values.	N	Y
2	mb_savi_median_decay	Temporally-weighted Soil Adjusted Vegetation Index (SAVI), calculated as an exponential decay-weighted average of the previous six years of annual SAVI values.	N	N
3	mb_evi2_median_decay	Temporally-weighted Enhanced Vegetation Index (EVI), calculated as an exponential decay-weighted average of the previous six years of annual EVI values.	N	Y

Pedogenetic Environments

Pedogenetic environments are co-occurrence bands derived from the spatial intersection of key categorical layers, including pedology (IBGE, 2015), structural subprovinces (lithology) (IBGE, 2019b), and biomes (IBGE, 2019a). These combinations **(p = 27, t = 1)** generate unique binary indicator variables (0 and 1) that represent specific ecological–geological contexts. The rationale behind this approach is that soil-landscape relationships are often non-linear and region-specific; for instance, the carbon stabilization potential of a Latossolo in the Amazon biome differs significantly from a Latossolo in the Cerrado biome due to distinct climatic and biological drivers. By explicitly defining these intersections, the model can better account for spatial non-stationarity and regional pedogenetic processes. This dataset is considered static and categorical.

Processing code:

- soil_30m_landsat/collection_03beta/carbon/0_covariate_source
- soil_30m_landsat/collection_03beta/texture/0_covariate_source

	Covariate ID	Definition	PSD	SOC
1	amazonia_sedimentos	Areas of co-occurrence between the Amazon biome and the Sediments class.	Y	Y
2	caatinga_sedimentos	Areas of co-occurrence between the Caatinga biome and the Sediments class.	Y	Y
3	cerrado_sedimentos	Areas of co-occurrence between the Cerrado biome and the Sediments class.	Y	Y
4	mata_atlantica_sedimentos	Areas of co-occurrence between the Atlantic Forest biome and the Sediments class.	Y	Y
5	pampa_sedimentos	Areas of co-occurrence between the Pampa biome and the Sediments class.	Y	N
6	pantanal_sedimentos	Areas of co-occurrence between the Pantanal biome and the Sediments class.	Y	Y
7	latossolo_metamorficas	Areas of co-occurrence between the Latossolo class and Metamorphic rocks.	Y	N
8	latossolo_plutonicas	Areas of co-occurrence between the Latossolo class and Plutonic rocks.	Y	N
9	latossolo_sedimentares	Areas of co-occurrence between the Latossolo class and Sedimentary rocks.	Y	Y
10	latossolo_sedimentos	Areas of co-occurrence between the Latossolo class and the Sediments class.	Y	Y
11	latossolo_vulcanicas	Areas of co-occurrence between the Latossolo class and Volcanic rocks.	Y	Y
12	argissolo_metamorficas	Areas of co-occurrence between the Argissolo class and Metamorphic rocks.	Y	Y
13	argissolo_sedimentares	Areas of co-occurrence between the Argissolo class and Sedimentary rocks.	Y	Y
14	argissolo_sedimentos	Areas of co-occurrence between the Argissolo class and the Sediments class.	Y	Y
15	pantanal_gleissolo	Areas of co-occurrence between the Pantanal biome and the Gleissolo class.	Y	Y
16	pantanal_neossolo_quartzarenico	Areas of co-occurrence between the Pantanal biome and the Neossolo Quartzarênico class.	Y	Y
17	pantanal_planossolo	Areas of co-occurrence between the Pantanal biome and the Planossolo class.	Y	Y
18	pantanal_plintossolo	Areas of co-occurrence between the Pantanal biome	Y	Y

		and the Plintossolo class.		
19	caatinga_latossolo	Areas of co-occurrence between the Caatinga biome and the Latossolo class.	Y	Y
20	raso_plutonica	Areas of co-occurrence between shallow soils and Plutonic rocks.	Y	N
21	raso_sedimentares	Areas of co-occurrence between shallow soils and Sedimentary rocks.	Y	Y
22	raso_vulcanica	Areas of co-occurrence between shallow soils and Volcanic rocks.	Y	Y
23	sibcs_argiloso	Areas of co-occurrence between soil classes with predominantly clayey texture.	Y	Y
24	sibcs_btextural	Areas of co-occurrence between soil classes with a textural B horizon.	Y	Y
25	sibcs_esqueleto	Areas of co-occurrence between soils with significant presence of gravel or stoniness.	Y	Y
26	sibcs_homogeneo	Areas of co-occurrence between deep and uniform soil classes.	Y	Y
27	sibcs_rasos	Areas of co-occurrence between different shallow soil classes.	Y	Y

Sampling and Model Control Variables

This group of variables is designed to harmonize the diverse sources of soil profile data and mitigate potential biases within the training dataset. Given that the soil samples were collected over several decades and by different institutions (e.g., legacy surveys and recent forest inventories), these predictors allow the machine learning algorithm to distinguish between varying data qualities, temporal mismatches, and specific sampling designs. Furthermore, they incorporate synthetic observations (pseudo-samples) to ensure the model accurately represents extreme landscape conditions, such as rocky outcrops and sandy formations, where traditional soil sampling is often absent. This dataset is considered static and categorical.Processing code:

- soil_30m_landsat/collection_03beta/carbon/0_covariate_source
- soil_30m_landsat/collection_03beta/texture/0_covariate_source

	Covariate ID	Definition	PSD	SOC
1	profundidade	This variable identifies the vertical position of the soil sample. During model training, it uses the specific depth of the observed profile; for spatial prediction, it is set to the target the mid-layer depth (in centimeters) to estimate soil properties at standardized intervals.	Y	Y
2	IFN_index	A binary indicator used to control for sampling bias from the	N	Y

		National Forest Inventory (IFN). It allows the model to differentiate these high-density samples from traditional soil surveys, preventing overestimation.		
3	PSEUDOROCK_index	A control variable that identifies synthetic samples (pseudo-observations) replicated in areas of rocky outcrops.	N	Y
4	PSEUDOSAND_index	A control variable identifying synthetic samples replicated in sandy formations (beaches and dunes).	N	Y
5	YEAR_index	A temporal flag used to synchronize legacy soil samples with the MapBiomas historical series. For samples collected prior to the availability of annual satellite imagery the model assigns the land use and environmental covariates from the year 1985. This index allows the machine learning algorithm to identify these specific observations, accounting for potential temporal mismatches between the original sampling date and the earliest available remote sensing data.	N	Y