



Cerrado - Appendix

MapBiomas 10m - Collection 4

Version 1

General coordinator

Ane A. Alencar

Team

Ana G. P de Souza

Bárbara C. da Silva

Dhemerson E. Conciani

Joaquim J. S. P. Pereira

Julia Z. Shimbo

Vera L. S. Arruda

Wallace V. da Silva

August, 2026

1. OVERVIEW OF THE CERRADO CLASSIFICATION METHOD

The land use and land cover (LULC) classification methodology for the Cerrado biome, developed under the MapBiomas 10m initiative, is based on high spatial resolution remote sensing data. The initiative aims to produce annual maps of dominant native vegetation (NV) types at 10m spatial resolution, primarily using Sentinel-2 imagery. For the Cerrado biome, the classification framework focuses on mapping NV types grouped into six broad categories: forest formation, savanna formation, grassland formation, wetland, herbaceous sandbank vegetation, and rocky outcrop. This document is intended as a technical appendix to the general MapBiomas 10m ATBD and provides a detailed description of the methodological procedures adopted for LULC mapping in Cerrado.

Since its inception, the MapBiomas 10m Cerrado workflow has undergone continuous methodological refinement, incorporating both conceptual advances and technical improvements while preserving key methodological developments established in previous MapBiomas 30 m collections. It is important to note that all previous MapBiomas 10 m collections were released as *beta* products, reflecting their ongoing methodological development and had not yet undergone a systematic biome-wide accuracy assessment. Collection 4.0 represents the first MapBiomas 10m product to be accompanied by an accuracy assessment report. Across all collections, the workflow follows a general structure, comprising: (i) definition of the feature space based on remote sensing metrics; (ii) generation of training samples for algorithm calibration; (iii) application of post-classification spatial and temporal filters to reduce noise and improve time-series consistency; and (iv) integration of the resulting maps with cross-cutting products. Classification results are systematically evaluated through visual inspection. A synthesis of the methodological evolution across collections is presented in Table 1.

Collection 1.0 (2016-2022) represents the first LULC dataset produced under the MapBiomas 10m initiative. Its methodological design closely followed the approaches adopted in MapBiomas 30m Collection 7.1, which are based on Landsat 30m imagery. Reference data included state-level data such as *“Inventário Florestal do Estado de São Paulo”* and *“Base Temática Digital do Estado do Tocantins”*, as well as deforestation datasets from PRODES and canopy-height information derived from GEDI. The classification was performed using the Random Forest algorithm, with hyperparameter tuning and feature-space selection conducted independently for each of the 38 Cerrado classification regions. This regionalized strategy involved preliminary experiments to identify the most suitable combination of predictor variables and algorithm parameters for each region. Spatial and temporal post-classification filters were subsequently applied following the same general logic adopted in MapBiomas 30 m Collection 7.1.

Collection 2.0 (2016-2023) introduced substantial methodological innovations, particularly regarding reference data and feature space composition. Additional reference datasets were incorporated, including “*Mapa de Uso e Cobertura da Terra do Distrito Federal*”, the “*Mapeamento dos Remanescentes de Campos de Murundus do Estado de Goiás*”, and deforestation alerts from PRODES and MapBiomias Alerta. A major conceptual shift in this collection was the transition from a regionalized classification strategy to a unified model applied across all Cerrado regions. The same Random Forest hyperparameters and feature space were used for all regions, while the classifier was enhanced by adopting a multiprobability approach. New topographic variables derived from the MERIT DEM were included, along with a suite of red-edge and chlorophyll-related spectral indices, exploiting the spectral capabilities of Sentinel-2. To further reduce classification noise, SNIC image segmentation was applied. In addition, new post-classification filters were developed, including: (i) a false regrowth filter to correct forest formation commission errors in silviculture areas; and (ii) a geomorphometric filter to reduce wetland commission errors associated with terrain-induced shadow effects. Collection 2.0 also marked the first inclusion of the rocky outcrop class in the MapBiomias 10m Cerrado products, following the conceptual definition used in Landsat-based collections, but enhanced with geomorphometric variables such as relative relief, valley depth, and the topographic position index.

Table 1. Summary of methodological evolution across MapBiomias 10m Cerrado collections. RF: Random Forest; NV: Native Vegetation.

Collection	Year Range	Method	Mapped classes	Key Improvements
1.0	2016– 2022	RF	Forest, Savanna, Wetland, Grassland, Mosaic of Uses, Agriculture and Pasture, Other Non-Vegetated Area, Water	First version of the Cerrado collection in the MapBiomias 10m initiative.
2.0	2016– 2023	RF	Forest, Savanna, Wetland, Grassland, Rocky Outcrop, Mosaic of Uses, Other Non-vegetated Area, Water	Inclusion of a new NV Class (Rocky Outcrop); New reference maps; Random Forest multiprobability approach; Improvement of post-classification filters; New false regrowth and geomorphometric filters.
3.0	2017– 2024	RF	Forest, Savanna, Wetland, Grassland, Rocky Outcrop, Herbaceous Sandbank Vegetation, Mosaic of Uses, Other Non-vegetated Area, Water	Inclusion of a new NV Class (Herbaceous Sandbank Vegetation); Introduction of Satellite Embedding dataset; New reference maps; Redefined Rocky Outcrop concept.
4.0	2017– 2025	RF	Same as Collection 3.0	New reference maps; New post-processing sequence; Redefined Rocky Outcrop concept.

Collection 3.0 (2017-2024) incorporates further refinements in both reference data and predictor variables. New reference maps were added, including the *“Mapa de Veredas e Áreas Úmidas do Sudoeste do Tocantins”*, *“Cobertura e Uso da Terra do Sudoeste do Tocantins”*, and *“Zoneamento do Parque Nacional da Chapada das Mesas”*. A key methodological advance in this collection is the incorporation of the annual Satellite Embedding product developed by Google into the classification feature space. These embeddings were used in combination with annual Sentinel-2 mosaics and the same set of topographic and geomorphometric variables adopted in Collection 2.0, ensuring methodological continuity while expanding the representational capacity of the predictors. This collection spans the period from 2017 to 2024, aligning with the temporal availability of the embedding product. Post-classification filters were comprehensively reviewed, and a new filter was developed to include the mapping of the herbaceous sandbank vegetation class. In addition, the rocky outcrop classification was refined to more strictly restrict the class to areas of exposed rock, explicitly excluding rupestrian vegetation. This conceptual adjustment aligns the 10m product with the definition adopted in MapBiomias 30m Collection 10.0.

The current Collection 4.0 introduces additional improvements to the reference database while retaining the same feature space adopted in Collection 3.0, including annual Google Satellite Embedding variables, Sentinel-2 spectral metrics, and topographic and geomorphometric predictors. Two new reference maps were included: the *“Mapa de Veredas do Mato Grosso”* and the *“Mapa de Vegetação da Fazenda Água Limpa – UnB.”* The post-classification workflow was comprehensively revised, with existing spatial and temporal filters reorganized and new filters incorporated to address specific classification inconsistencies. The order of application was also restructured according to a logical sequence similar to that adopted in MapBiomias 30 m Collection 11.0, thereby improving the interaction among filtering rules and enhancing the spatial and temporal consistency of the final maps. In addition, the conceptual definition of the rocky outcrop class was also revised. Areas associated with rupestrian vegetation were reincorporated into this class maintaining the conceptual consistency adopted in Collection 2.0. The classification and post-processing scripts employed in this collection are available at <https://github.com/mapbiomas/brazil-cerrado>.

2. METHODOLOGICAL DESCRIPTION

In MapBiomias Collection 4.0, the annual LULC classification for the Cerrado biome comprises ten classes, as defined in the official MapBiomias legend (available at: [Legend Code](#)). These classes, detailed in Table 2, include native vegetation, water bodies, and

anthropogenic land uses such as agriculture and pasture. As in previous collections, agriculture and pasture are mapped in the initial stages of the workflow but are not the focus of the Cerrado-specific methodology. These classes are produced by cross-cutting thematic teams and, in the Cerrado workflow, serve to support native vegetation mapping by improving the separation between natural and anthropogenic land cover and reducing omission and commission errors. In the post-processing stage, agriculture and pasture are merged into a single class labeled mosaic of uses. The subsequent sections describe each major component of the classification workflow, including: Annual Image Mosaic (Section 3), General Classification and Post-Classification (Sections 4 and 5), and rocky outcrop Classification (Section 6). An overview of the workflow is presented in Figure 1. The main methodological steps are summarized below:

- **Annual Image Mosaic (Sentinel-2 + Satellite Embeddings):** Annual mosaics were constructed by combining Sentinel-2 surface reflectance composites with the annual Satellite Embedding product. The Sentinel-2 mosaics provide high-resolution spectral information, while the 64-dimensional embedding vectors encode learned representations of annual surface condition trajectories, capturing spatiotemporal context beyond individual spectral observations.
- **Training Samples:** Training samples were extracted from stable pixels identified in MapBiomas 10m Collection 3.0 (2017–2024), complemented by thematic reference data, GEDI and Global Pasture Watch data. A stratified sampling design ensured proportional class representation across regions and years.
- **General Classification:** Land cover classification was performed using the Random Forest algorithm in Google Earth Engine, applying a multiprobability approach to capture class uncertainty. Post-classification filters were applied to correct temporal inconsistencies and remove spatial artifacts.
- **Rocky Outcrop Classification:** A dedicated classification workflow was developed to map rocky outcrop. Training samples were visually interpreted and validated, and classification was performed using Random Forest. Temporal and spatial filters refined the final output.
- **Internal Integration and BAP harmonization:** Two outputs were generated: (i) the general LULC map and (ii) the rocky outcrop map. The rocky outcrop classification was integrated into the general map through a final post-processing procedure. In addition, specific spatial corrections were applied within the Paraguay River Basin (BAP) to harmonize the annual maps along the Pantanal biome.
- **Integration with Cross-Cutting Themes:** Final annual maps were harmonized with cross-cutting thematic layers following predefined MapBiomas prevalence rules.

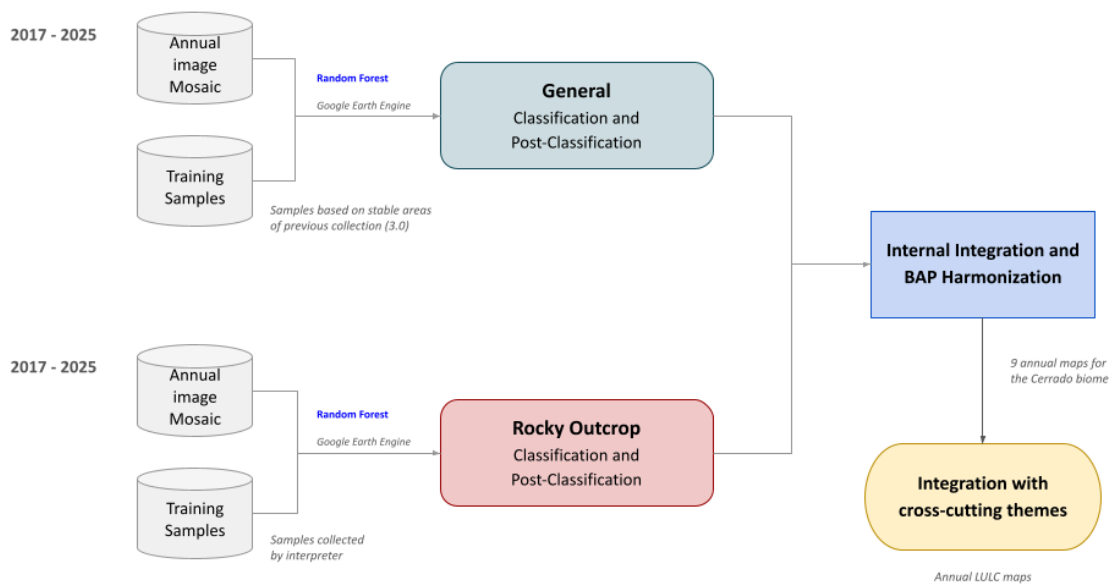
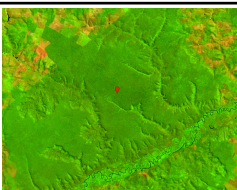
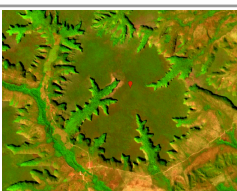
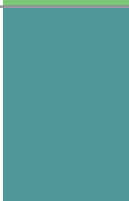
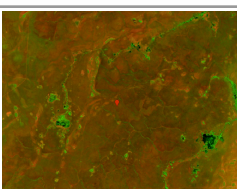
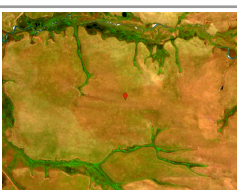
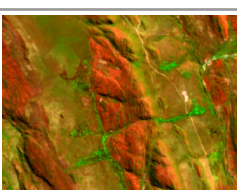


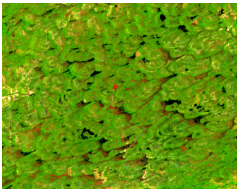
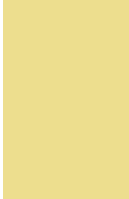
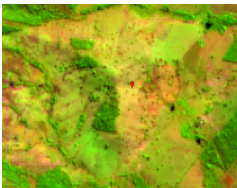
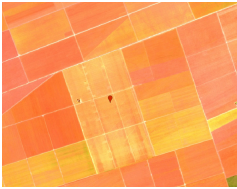
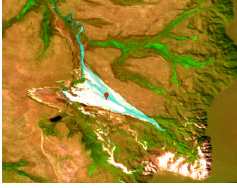
Figure 1. Workflow for the LULC classification of the Cerrado biome in Collection 4.0. Two parallel modules were applied: (i) the general classification, and (ii) a specific classification of rocky outcrop. Both used the Random Forest algorithm implemented in GEE. Outputs were integrated into two stages: first, the internal integration combined the rocky outcrop with the general map; second, the maps were integrated with cross-cutting themes.

3. ANNUAL IMAGE MOSAICS

The first step in the classification of LULC in the Cerrado biome consists of generating annual image mosaics that serve as the primary input. In the MapBiomias 10m initiative, the mosaic construction strategy builds upon methodological advances established in previous Landsat-based collections. In collections 1.0 and 2.0, annual mosaics were derived exclusively from Sentinel-2 surface reflectance imagery obtained from the **COPERNICUS/S2_SR_HARMONIZED** collection. All available spectral bands were used in the mosaic construction, including blue, green, red, red-edge 1, red-edge 2, red-edge 3, red-edge 4, near-infrared (NIR), shortwave infrared 1 (SWIR1), and shortwave infrared 2 (SWIR2). For the Cerrado biome, an additional quality-control filter was applied to remove images with cloud cover greater than 40%, reducing atmospheric contamination while maintaining sufficient image availability.

Table 2. Land use land cover categories used for the Sentinel-2 mosaics classification for the Cerrado biome in MapBiomas Collection 4.0.

Classes Level 1	Classes Level 2	ID	Legend Color	RGB composite (SWIR1-NIR-Red)	Description
Forest	Forest Formation	3			Vegetation types characterized by the predominance of tree species forming a continuous canopy. This includes Riparian Forests, Gallery Forests, Dry Forests, and Forested Savannas (Ribeiro and Walter, 2008), as well as Semi-deciduous Seasonal Forests.
	Savanna Formation	4			Vegetation types with a distinct stratification of tree, shrub, and herbaceous stratum. This includes different physiognomies of Cerrado <i>sensu stricto</i> (Dense, Typical, Sparse, and Rupestrian Savanna) (Ribeiro and Walter, 2008). The class may also include native Cerrado <i>sensu stricto</i> areas suitable for grazing, if they maintain their natural structure and composition.
Herbaceous and Shrubby Vegetation	Wetland	11			Ecosystems dominated by herbaceous vegetation subject to seasonal flooding or constant fluvial/lacustrine influence. Examples include <i>Campo Úmido</i> and <i>Brejo</i> . In some areas, the herbaceous matrix is interspersed with savanna tree species (e.g., <i>Parque de Cerrado</i>) or palm species, such as in <i>Veredas</i> and <i>Palmeirais</i> (Ribeiro and Walter, 2008).
	Grassland	12			Open vegetation dominated by herbaceous species, with minimal or no tree cover. Includes the Dirty, Clean and Rupestrian Grassland phytophysiomies (Ribeiro and Walter, 2008).
	Rocky Outcrop	29			Naturally exposed rocky surfaces, including monoliths, bedrock, and slabs with little or no soil cover, these features are typically associated with stable geological formations of sedimentary, igneous, or metamorphic origin. Includes areas of rupestrian vegetation. This class also includes areas of rupestrian vegetation associated with these environments, such as <i>Cerrado Rupestre</i> and <i>Campo Rupestre</i> (Ribeiro and Walter, 2008).

Classes Level 1	Classes Level 2	ID	Legend Color	RGB composite (SWIR1-NIR-Red)	Description
	Herbaceous Sandbank Vegetation*	50			Coastal sandy plain ecosystems characterized by predominantly herbaceous and shrubby vegetation, with sparse shrub distribution.
Farming	Pasture**	15			Pasture area, predominantly planted, linked to cattle ranching activities.
	Agriculture**	18			Areas occupied with short to long vegetative cycles of crops. This encompasses both perennial and temporary crops.
Non-Vegetated Area	Other Non-Vegetated Areas	25			Includes impermeable surfaces (e.g., roads, buildings, mining infrastructure), exposed soil in natural settings (e.g., erosion features, gullies, landslides), and croplands in the off-season.
Water	River, Lake, and Ocean	33			Rivers, lakes, dams, reservoirs, and other water bodies.

*Herbaceous Sandbank Vegetation is added via post-classification method; ** Agriculture and Pasture are merged into Mosaic of Uses during the first spatial filter

Annual mosaics were generated using a standardized compositing window from April to September. This period represents the transition between the end of the rainy season and the onset of the dry season (Figure 2). During this interval, cloud cover is substantially reduced while vegetation remains in a relatively vigorous state, resulting in improved spectral separability and reduced commission errors associated with cloud contamination and senescent vegetation. This temporal window was selected based on extensive empirical testing in earlier Landsat-based collections. Within the April–September window, multiple temporal reduction metrics were computed for each pixel, including median, median for wet and dry periods, and standard deviation.

In MapBiomass 10m Collection 3.0, the Sentinel-2 mosaic strategy remains consistent with previous collections, preserving the same compositing window, spectral bands, and temporal metrics. However, Collection 3.0 introduces a major methodological advancement through the integration of annual Satellite Embedding products developed by Google (Brown et al., 2025). These embeddings dataset [GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL](#) provide a 64-dimensional, analysis-ready representation for each 10m pixel and year, encoding learned spatiotemporal and contextual information derived from multiple Earth observation datasets. As the embedding product is intrinsically annual, no additional temporal aggregation is applied.

Consequently, the annual classification input in Collection 3.0 combined (i) Sentinel-2 mosaics summarizing seasonal spectral behavior during the April–September transition period and (ii) annual Satellite Embedding bands that capture broader contextual and structural patterns. This integrated representation enhances class discrimination in the highly heterogeneous Cerrado landscape. All annual mosaics were subjected to visual inspection to ensure data quality and suitability before classification. No modifications were introduced to the classification feature space in Collection 4.0. The annual classification inputs followed the same methodological framework and band configuration adopted in previous collections. Thus, no predictor bands were added, removed, or reparameterized in the current collection. Maintaining an unchanged feature space ensured methodological continuity and comparability between Collections 3.0 and 4.0, allowing differences in the resulting maps to be directly associated with updates to the reference maps and revisions to the training samples and post-classification workflow.

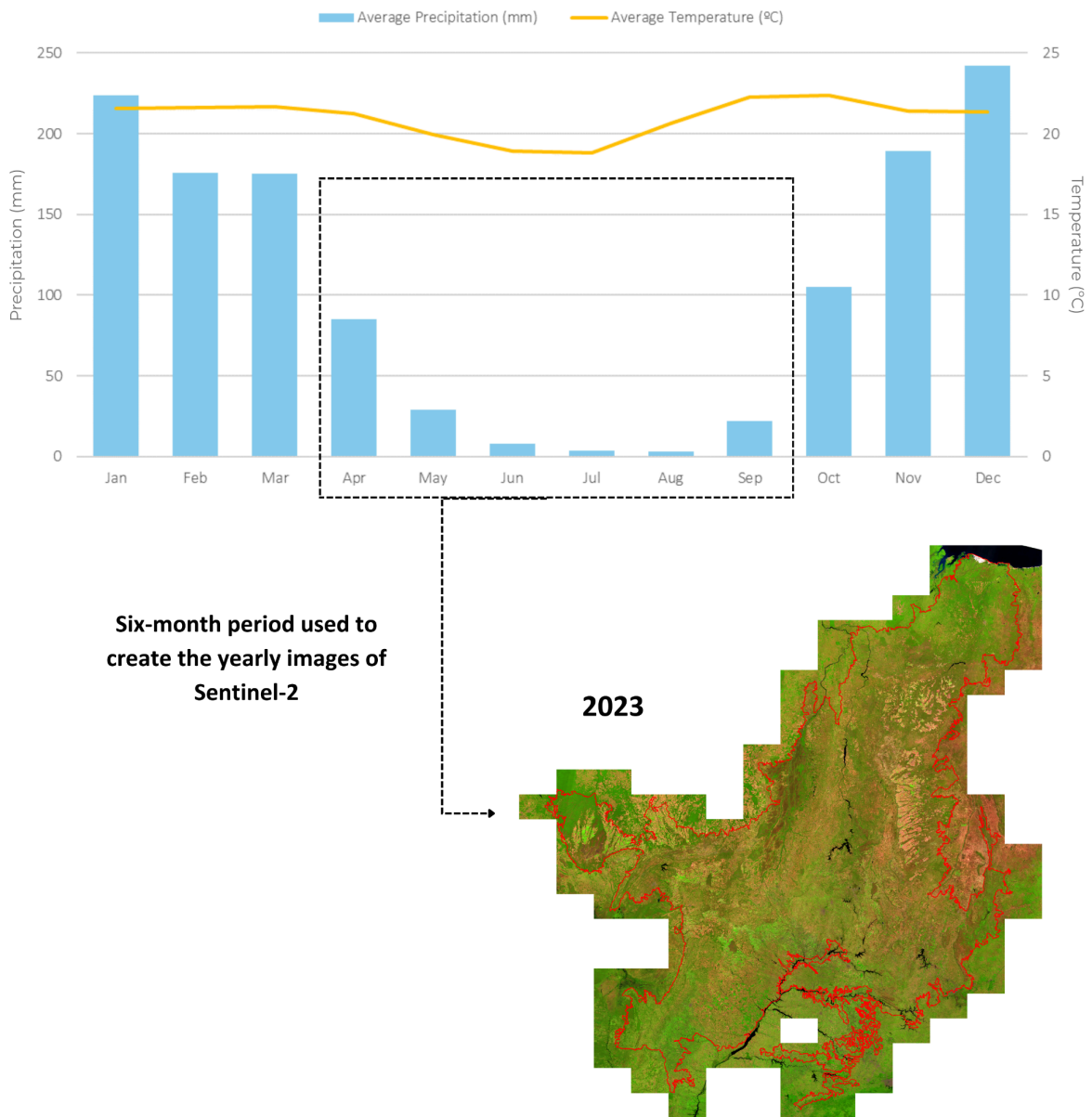


Figure 2. Time window adopted for the construction of annual classification mosaics in the Cerrado biome. Precipitation data represent monthly averages for the Cerrado region (Macena et al., 2008), and temperature data correspond to monthly averages for the Federal District (INMET).

4. GENERAL MAP CLASSIFICATION

The complete workflow for the general LULC classification is illustrated in Figure 3. This workflow integrates multiple processing stages to ensure accurate and temporally consistent mapping across the Cerrado biome. The methodological steps are described in detail in the subsequent sections, starting with the delineation of classification regions (4.1), followed by the construction of the feature space used in model training (4.2). The

sampling strategy and classification procedure are presented in section 4.3, which includes the generation of a training mask based on stable areas, the application of a stratified sampling design, and the classification using the Random Forest algorithm. Finally, the post-classification filtering steps applied to improve temporal and spatial consistency are detailed in section 5.

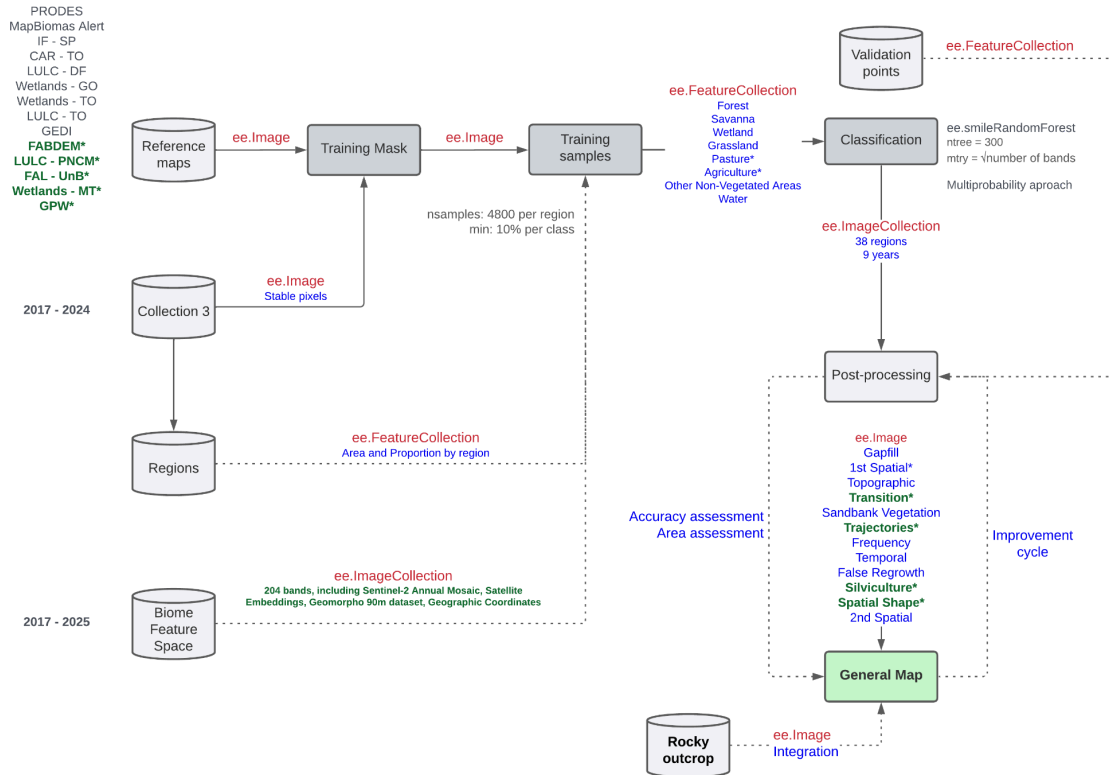


Figure 3. Each gray geometry (cylinders for databases and rectangles for processes) represents a key step in the classification schema, with the respective name inside. The gray text near databases and processes offers a description of the step, while the green text highlights the main innovations in Collection 4.0. Arrows with a continuous black line connecting the key steps represent the main direction of the processing flux, while arrows with dotted black lines represent the databases that feed the main processes. Red text inside arrows refers to the asset type in the Google Earth Engine, while blue text offers a concise description of the asset content.

4.1. Classification regions

The classification regions adopted in MapBiomias 10m Collection 4.0 correspond to the same regionalization scheme used since MapBiomias 30m Collection 7.0. In total, 38 classification regions were defined for the Cerrado biome. These regions were designed to capture spatial patterns of Cerrado's heterogeneity, particularly differences in vegetation phenology and seasonal dynamics of native vegetation. The classification units were originally defined based on the ecoregions proposed by Sano et al. (2019) and were

subsequently refined using Brazil's major hydrographic basins and the spatial distribution of LULC classes observed in MapBiomas 30m Collection 3.0 (2017) (Figure 4). This framework has proven effective in accommodating the Cerrado's regional vegetation dynamics and improving the consistency and accuracy of the classification results across diverse environmental contexts. In Collection 4.0, however, the updated Cerrado biome boundaries proposed by the *Instituto Brasileiro de Geografia e Estatística* (IBGE) required a revision of region 38 (São Paulo and Minas Gerais states). Its boundary was redesigned to incorporate the newly included portion of the biome, increasing the region's area by approximately 15,000 hectares. The remaining regions were maintained without changes.

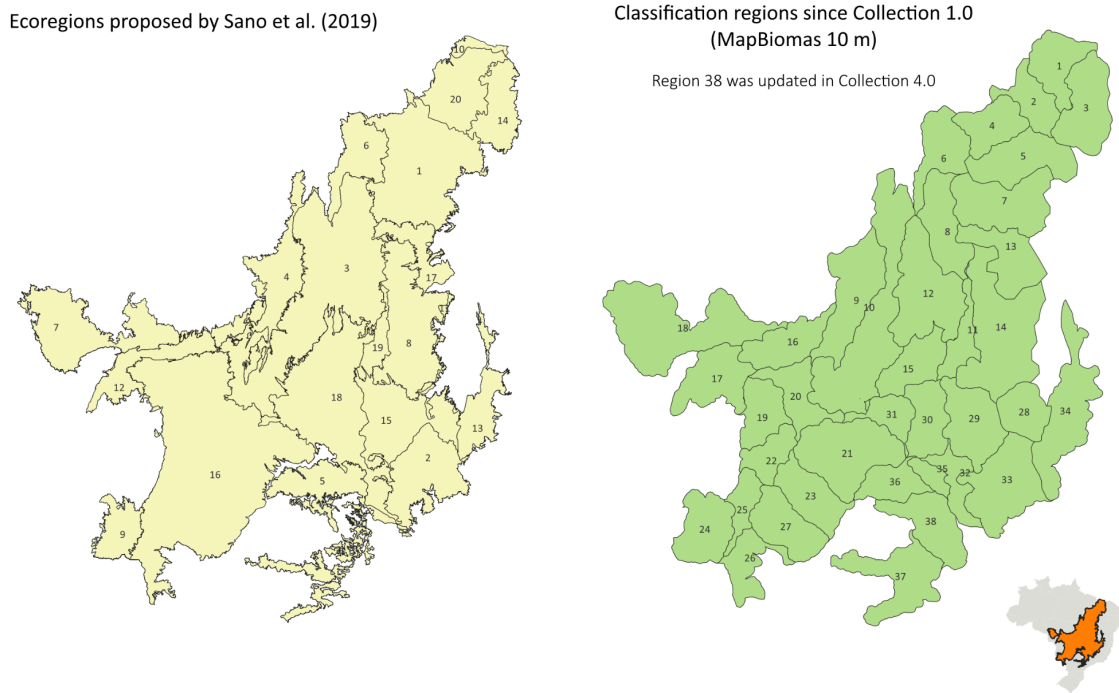


Figure 4. Classification regions, modified from Sano et al., 2019. Highlighted in orange is the location of the Cerrado biome in Brazilian territory.

4.2. Feature space

The feature space used for LULC classification in Collection 4.0 comprised a set of 204 predictor variables, designed to capture the spectral and temporal complexity of the Cerrado biome using high spatial resolution data (Table 3). It comprises a combination of dynamic (annual) and static (non-annual) predictor variables derived from Sentinel-2 imagery, Satellite Embeddings, and ancillary environmental datasets. The dynamic

component of the feature space includes two complementary data sources with distinct temporal structures. For the Sentinel-2–based variables, a set of statistical reduction metrics was calculated, including median and standard deviation. These metrics were selected to summarize intra-annual spectral variability while reducing noise related to cloud contamination and short-term phenological fluctuations. The Satellite Embedding product, in contrast, is provided exclusively as annual data, with one embedding vector per year for each 10 m pixel. Given this intrinsic temporal structure, the embedding dimensions were incorporated directly into the feature space as annual predictors, without the application of additional temporal reduction metrics.

Table 3. Feature space used in the Cerrado biome classification for MapBiomass 10m Collection 4.0. The column “Statistics” refers to the set of per-pixel statistical reducers applied to each variable within the annual temporal window: a) Median – annual median; b) Median_dry – seasonal median for dates below the first quartile of NDVI values (dry period); c) Median_wet – seasonal median for dates above the first quartile of NDVI values (wet period); d) Standard deviation – annual variation; e) Identity - the variable is used directly, without temporal reduction.

Type	Name	Formula / Description	Statistics	Reference
Embeddings Bands	Bands A00 to A63	Satellite Embedding V1	Identity	Brown et al., 2025
Sentinel Bands	Bands Blue, Green, Red, NIR, SWIR1, SWIR2	Original reflectance bands	Median, Median_dry, Median_wet, StdDev	USGS
Spectral Index	NDVI Normalized Difference Vegetation Index	$(\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red})$	Median, Median_dry, Median_wet, StdDev	Rouse et al., 1974
	EVI2 Enhanced Vegetation Index 2	$2.5 \times (\text{NIR} - \text{Red}) / (\text{NIR} + 2.4 \times \text{Red} + 1)$	Median, Median_dry, Median_wet, StdDev	Parente and Ferreira, 2018
	PRI Photochemical Reflectance Index	$(\text{Blue} - \text{Green}) / (\text{Blue} + \text{Green})$	Median, Median_dry, Median_wet, StdDev	Gamon et al., 1992
	MNDWI Modified Normalized Difference Water Index	$(\text{Green} - \text{SWIR1}) / (\text{Green} + \text{SWIR1})$	Median, Median_dry, Median_wet, StdDev	Xu, 2006
	CAI Cellulose Absorption Index	SWIR 2 / SWIR 1	Median, Median_dry,	Nagler et al., 2003

Type	Name	Formula / Description	Statistics	Reference
			Median_wet, StdDev	
	MSAVI Modified Soil Adjusted Vegetation Index	$0.5 \times [2 \times \text{NIR} + 1 - \sqrt{(2 \times \text{NIR} + 1)^2 - 8 \times (\text{NIR} - \text{Red})}]$	Median, Median_dry, Median_wet, StdDev	Qi et al., 1994
	GCVI Green Chlorophyll Vegetation Index	$(\text{NIR} / \text{Green} - 1)$	Median, Median_dry, Median_wet, StdDev	Burke and Lobell, 2017
	GRND Green Normalized Difference Vegetation Index	Green / Red	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 1996
	MSI Moisture Stress Index	SWIR1 / NIR	Median, Median_dry, Median_wet, StdDev	Zarco-Tejada et al., 2003
	GARI Green Atmospherically Resistant Vegetation Index	$(\text{NIR} - (\text{Green} - (\text{Blue} - \text{Red}))) / (\text{NIR} + (\text{Green} - (\text{Blue} - \text{Red})))$	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 2003
	GNDVI Green Normalized Difference Vegetation Index	$(\text{NIR} - \text{Green}) / (\text{NIR} + \text{Green})$	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 1996
	NDVI Red-Edge	$(\text{RedEdge1} - \text{Red}) / (\text{RedEdge1} + \text{Red})$	Median, Median_dry, Median_wet, StdDev	Potter et al., 2012
	VI 700 (NDCI) Vegetation Index 700	$(\text{RedEdge1} - \text{RedEdge2}) / (\text{RedEdge1} + \text{RedEdge2})$	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 2002
	IRECI Inverted Red-Edge Chlorophyll Index	$(\text{RedEdge3} - \text{Red}) / (\text{RedEdge1} + \text{RedEdge2})$	Median, Median_dry, Median_wet, StdDev	Guyot and Baret, 1988
	CIRE Chlorophyll Index Red-Edge	$(\text{NIR} / \text{Red Edge 1})$	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 2005
	TCARI Transformed Chlorophyll Absorption Reflectance Index	$3 \times [(\text{RedEdge1} - \text{Red}) - 0.2 \times (\text{RedEdge1} - \text{Green}) \times \text{RedEdge1} / \text{Red}]$	Median, Median_dry, Median_wet, StdDev	Haboudane et al. (2002)

Type	Name	Formula / Description	Statistics	Reference
	SFDVI Spectral Feature Depth Vegetation Index	$(\text{Green} + \text{NIR}) / 2 - (\text{Red} + \text{RedEdge1}) / 2$	Median, Median_dry, Median_wet, StdDev	Baptista, 2015
	NDRE Normalized Difference Red-Edge Index	$(\text{NIR} - \text{RedEdge}) / (\text{NIR} + \text{RedEdge})$	Median, Median_dry, Median_wet, StdDev	Gitelson and Merzlyak (1994)
	TGSI Topsoil Grain Size Index	$(\text{Red} - \text{Blue}) / (\text{Red} + \text{Blue} + \text{Green})$	Median, Median_dry, Median_wet, StdDev	Xiao et al., 2006
Vegetation Structure	Hall Height	$(-0.039 \times \text{Red} - 0.011 \times \text{NIR} - 0.026 \times \text{SWIR1} + 4.13)$	Median, StdDev	Hall et al., 2006
	Hall Cover	$(-\text{Red} \times 0.017 - \text{NIR} \times 0.007 - \text{SWIR2} \times 0.079 + 5.22)$	Median, StdDev	Hall et al., 2006

In addition to the annual variables, a set of static predictors was incorporated to provide environmental and spatial context to the classification model. This set includes the Height Above Nearest Drainage (HAND) index, as well as spatial coordinates (latitude and longitude), which were used to reduce the influence of training samples across distant regions and support regional consistency in the classification. Furthermore, the inclusion of geomorphometric variables enhances the discrimination of native vegetation types in areas characterized by complex terrain. These geomorphometric predictors were derived from the MERIT digital elevation model (Yamazaki et al., 2017) and describe key aspects of terrain morphology, including slope, aspect, curvature, and other derivatives associated with geomorphological processes. The complete list of non-annual variables used in the classification is presented in Table 4.

Table 4. Static (non-annual) variables used in the classification process of MapBiomass Collection 4.0 for the Cerrado biome. "Identity" in the statistics column indicates that the variable is used directly, without temporal reduction.

Name	Formula / Description	Statistics	Reference
Latitude	Extracted from pixel latitude (<code>ee.Image.pixelLonLat()</code>)	Identity	Geolocation function
Cosine of Longitude	<code>cos(ee.Image.pixelLonLat().select(['longitude']))</code>	Identity	Geolocation function
Sine of Longitude	<code>sen(ee.Image.pixelLonLat().select(['longitude']))</code>	Identity	Geolocation function

Name	Formula / Description	Statistics	Reference
Height Above the Nearest Drainage (HAND)	HAND Global 30m	Identity	Donchyts et al., 2016
Elevation (DEM)	MERIT DEM elevation (in meters)	Identity	Yamazaki et al., 2017
Aspect	Terrain aspect	Identity	Geomorpho 90m Amatulli et al., 2020
Convergence Index	Terrain convergence	Identity	Geomorpho 90m Amatulli et al., 2020
Roughness	Surface roughness index	Identity	Geomorpho 90m Amatulli et al., 2020
Eastness	Derived from the aspect to represent east-facing slopes	Identity	Geomorpho 90m Amatulli et al., 2020
Northness	Derived from the aspect to represent north-facing slopes	Identity	Geomorpho 90m Amatulli et al., 2020
Longitudinal Curvature	Second derivative of elevation calculated along the direction of maximum slope (downslope direction).	Identity	Geomorpho 90m Amatulli et al., 2020
Profile Curvature	Second derivative of elevation measured in the vertical plane aligned with the direction of steepest slope.	Identity	Geomorpho 90m Amatulli et al., 2020
DXX Second-order partial derivative in the x-direction	Second-order partial derivative of elevation with respect to the x-direction (east–west axis).	Identity	Geomorpho 90m Amatulli et al., 2020
CTI Compound Topographic Index	Wetness index combining slope and upstream area	Identity	Geomorpho 90m Amatulli et al., 2020

4.3. Training mask, stratified sampling, and classification approach

The training mask was derived primarily from stable pixels identified in MapBiomas 10m Collection 3.0, reclassified according to the Cerrado classification scheme. Pixels that maintained the same class throughout the 2017–2024 time series were selected and then refined using additional validation sources. To improve its reliability, classes subject to high uncertainty or transitional characteristics, such as “mosaic of uses” (21) and “other non-vegetated areas” (25), were excluded.

The stable areas identified were subsequently refined using a set of ancillary reference datasets to improve the quality and reliability of the stable training masks. The refinement incorporated deforestation alerts from PRODES (2000–2025) and MapBiomas Alerta (2019–2025), FABDEM terrain information, and regional LULC reference maps

produced by state institutions for São Paulo, Tocantins, Goiás, Mato Grosso and Distrito Federal, as well as zoning data from ICMBio (*Instituto Chico Mendes de Conservação da Biodiversidade*). Ancillary reference datasets have been incorporated into the Cerrado training-sample workflow across multiple MapBiomias collections, as summarized in Table 5. This reference database is updated for each collection, with new datasets incorporated when they become available. In some cases, others may be discontinued or replaced according to Cerrado LULC collections updates.

Table 5. Ancillary reference datasets used to refine stable training masks for the Cerrado biome across MapBiomias collections.

Reference dataset	Source	Main information	Use history
PROBIO	INPE	Cerrado native vegetation mapping	Collection 1.0; discontinued from Collection 2.0
PRODES Cerrado	INPE	Annual Cerrado native vegetation loss	Since collection 1.0; updated for each collection
SAD Cerrado	IPAM	Cerrado deforestation alerts	Collection 2.0; discontinued from Collection 3.0
MapBiomias Alerta	MapBiomias	Validated deforestation alerts	Since collection 2.0; updated for each collection
Global Canopy Height	Lang et al. (2023)	Canopy-height information	Since collection 1.0
São Paulo State Forest Inventory Mapping	SEMIL/SP	High-resolution vegetation and land-cover mapping for São Paulo	Since collection 1.0
Tocantins CAR Thematic Digital Vector Database	SEMARH/TO	Detailed thematic reference data for land-cover interpretation in Tocantins	Since collection 1.0
Distrito Federal Land Use Land Cover Map	SEMA/DF	Regional vegetation and anthropogenic land-cover	Since collection 2.0
Remnants of Murundus in Goiás	SEMAD/GO	Reference mapping of <i>campos de murundus</i> in Goiás	Since collection 2.0
Southeastern Tocantins Cartographic Database – Wetlands	SEPLAN/TO	Regional wetlands mapping in southeastern Tocantins	Since collection 3.0
Southeastern Tocantins Cartographic Database – Land Use and Land Cover	SEPLAN/TO	Regional land use and land cover mapping for southeastern Tocantins	Since collection 3.0
Tocantins Land Use and Land Cover Map	SEPLAN/TO	Regional land use and land cover mapping for Tocantins	Since collection 3.0

Reference dataset	Source	Main information	Use history
Chapada das Mesas National Park Zoning	MMA / ICMBio	Vegetation mapping for Chapada das Mesas National Park	Since collection 3.0
Mapping of Veredas in Mato Grosso	SEMA/MT	Regional mapping of <i>veredas</i>	Introduced in Collection 4.0
FAL–UnB Land Use and Land Cover Map	Leal and Cartagenes (2021)	Local land use and land cover mapping for <i>Fazenda Água Limpa</i> , Federal District	Introduced in Collection 4.0
GPW Annual Short Vegetation Height	Global Pasture Watch	Annual short-vegetation height	Introduced in Collection 4.0

Following this reference-based refinement, additional spatial and structural criteria were applied to further improve the consistency of the training masks. Only homogeneous patches equal to or larger than 0.25 hectare (25 Sentinel pixels) were retained in the final training mask. A canopy height filter based on GEDI (Lang et al., 2022) and Global Pasture Watch (GPW) data was applied to exclude anomalous pixels. The outlier removal procedure followed the same logic adopted in MapBiomass 30m Collection 11.0. The height-based filtering criteria were:

- Forest Formation (3): tree canopy height < 4 meters
- Savanna Formation (4): tree canopy height < 2 meters or > 8 meters
- Wetland (11): tree canopy height > 15 meters
- Grassland Formation (12): short-vegetation height > 3 meters
- Pasture (15): short-vegetation height > 4 meters
- Agriculture (18): short-vegetation height > 3 meters
- Non-Vegetated Area (25): tree canopy height > 0 meters
- Water (33): canopy height > 0 meters

A stratified random training sampling strategy was employed to ensure representative coverage of all LULC classes across the 38 classification regions. The spatial allocation of samples was constrained by the reference training mask, ensuring that samples were drawn exclusively from areas with high classification confidence. Within each classification region, the number of samples allocated to each class was computed proportionally to its mapped area, estimated from the 2020 map of MapBiomass 10m Collection 3.0. Particularly, the class-specific sample size was calculated using a proportional allocation equation in which the relative area of each class was multiplied by a reference value of 4,800 samples per region. A minimum threshold of 480 samples per class was enforced to prevent underrepresentation of less extensive classes. This

approach aimed to improve classification accuracy and ensure that underrepresented classes were adequately accounted for during the classification procedure.

The classification was performed independently for each classification region and year using the Random Forest algorithm implemented in Google Earth Engine (GEE) through the `ee.Classifier.smileRandomForest` function. Based on empirical testing observed in previous MapBiomass collections, the number of decision trees was fixed at 300 for all regions, while the number of variables considered at each split was set to the square root of the total number of predictor bands. The classifier was operated in multiprobability mode, producing a probability distribution across all target classes for each pixel. The final class assignment was determined by selecting the class with the highest predicted probability, an approach that increases robustness to spectral confusion and better represents classification uncertainty, particularly in heterogeneous landscapes.

5. GENERAL MAP POST-CLASSIFICATION

Given the pixel-based classification methodology and the annual processing of the time series (2017–2025), a structured post-classification filtering framework was applied to enhance the spatial and temporal consistency of the final land cover maps. The main objective of this stage was to correct classification artifacts and reduce spurious transitions commonly associated with per-pixel classifiers operating over long temporal windows. The post-processing framework included a sequence of filters designed to address specific classification limitations and collectively contributed to the overall quality and reliability of Collection 4.0.

5.1. Temporal Gap-Fill Filter

The gap-filling filter was designed to address missing data, commonly caused by cloud and shadow contamination or gaps in image availability, by propagating valid classifications across the temporal dimension. The implemented method applied a bidirectional temporal search. Initially, the algorithm conducted a forward fill: each no-data pixel was replaced with the nearest valid classification from subsequent years. In a second pass, a backward fill was performed to capture any remaining gaps by assigning the most recent valid classification from previous years. This two-step process ensured that the final classified time series had minimal missing values, resulting in more complete and temporally coherent land cover maps. Persistent gaps remained only in exceptional cases where a pixel was consistently unclassified throughout the entire 2017–2025 period.

5.2. First spatial filter

The first spatial filter was applied to reduce isolated pixels and small classification patches before the thematic and temporal filters. At this step, pasture (15) and agriculture (18) were reclassified as mosaic of uses (21), consequently, these anthropogenic land-use types were treated as a single class throughout the subsequent post-classification procedures. For each annual map, connected components were calculated, and patches containing approximately 0.5 hectare or less (equivalent to 50 Sentinel pixels) were replaced by the modal class within a 9×9 -pixel neighborhood. Eight-neighbor connectivity was used for most classes, whereas four-neighbor connectivity was adopted for mosaic of uses to avoid connecting anthropogenic patches only through diagonal contact. Forest formation (3), wetland (11), and water (33) were protected from this correction to preserve small but ecologically relevant features.

5.3. Topographic

The topographic filter was designed to correct classifications that were inconsistent with terrain conditions, particularly errors associated with relief shadows and steep slopes. Slope was calculated from the MERIT Digital Elevation Model and expressed as a percentage. Wetland (11) pixels occurring on slopes equal to or greater than 12% were reclassified as forest formation (3). Water (33) pixels on slopes equal to or greater than 15% were also reclassified as forest formation (3) when the pixel had been mapped as native vegetation in at least two years of the historical series, indicating that the apparent water occurrence was probably associated with terrain shadow. Water (33) occurrences on slopes equal to or greater than 50% were corrected independently of their historical vegetation frequency. Mosaic of uses pixels located on slopes equal to or greater than 50% were replaced by the modal class of the surrounding neighborhood.

5.4. Spatial–temporal transition

The spatial–temporal transition filter was applied to remove short and spatially restricted oscillations among broad land-cover groups. The classes were grouped into Native Vegetation (forest formation, savanna formation, wetland, and grassland formation), Anthropogenic Cover (mosaic of uses and other non-vegetated areas) and Other Classes (water and no data). For each intermediate year, the filter evaluated a three-year window and identified A – B – A trajectories in which a pixel temporarily changed thematic group and returned to its previous group in the following year. When the temporary transition formed a connected patch of 25 pixels or fewer, approximately 0.25 hectare at 10-m resolution, the intermediate classification was replaced by the

detailed class mapped in the preceding year, converting into an A – A – A trajectories. This procedure reduces edge effects and small transition artifacts without suppressing spatially consolidated land-cover changes. Up to Collection 3.0, this post-classification procedure used a different approach and was referred to as the “Incidence Filter.”

5.5. Herbaceous Sandbank Vegetation

Herbaceous sandbank vegetation (50) is a beta class introduced for the Cerrado biome in Collection 3.0 and maintained in current collection through a dedicated post-classification procedure. Reference polygons were manually collected to represent areas of herbaceous sandbank vegetation and non-sandbank vegetation. Random points extracted from these polygons were used to train a binary Gradient Tree Booster (GTB) classifier configured with 50 trees. Predictor variables were selected based on expert knowledge of the spectral and structural characteristics associated with herbaceous sandbank vegetation. The selected variables included median Red-Edge 1 and SWIR1 reflectance; median and wet-season GCVI index; wet- and dry-season EVI2; NDWI amplitude; wet-season WEFI; and dry-season SEFI. In addition, HAND, canopy height, and FABDEM were included to represent hydrological, vegetation structure, and topographic conditions, respectively. The combination of the selected predictor variables and the GTB classifier resulted in a spatial mask representing areas with potential occurrence of herbaceous sandbank vegetation. The purpose was not to generate annual maps of herbaceous sandbank, but to produce a static spatial mask representing areas with potential occurrence of this vegetation type. The GTB prediction was restricted to the coastal sandy-deposit areas mapped by the *Serviço Geológico Brasileiro* (CPRM/SGB), ensuring that the potential occurrence mask was limited to environments considered compatible with this class. This restricted mask was then used to remap the annual general LULC classification from 2017 to 2025. For each year, pixels located within the potential occurrence mask and originally classified as savanna formation (4), wetland (11), or grassland formation (12) were reassigned to herbaceous sandbank vegetation (50). Pixels outside the mask, as well as pixels belonging to other LULC classes, retained their original classification. Therefore, although the potential occurrence mask remained spatially constant throughout the time series, the annual extent of herbaceous sandbank varied according to the distribution of eligible classes in each annual general map.

5.6. Trajectory

The trajectory filter corrected specific temporal patterns associated primarily with grassland formation (12), which may occur as an unstable intermediate class between

native vegetation and anthropogenic land use. One-year grassland occurrences in A) native vegetation → grassland formation – mosaic of uses, B) mosaic of uses → grassland formation → native vegetation, and C) mosaic of uses → grassland formation → mosaic of uses trajectories were replaced by the class mapped in the subsequent year. One- or two-year of grassland immediately preceding a consolidated mosaic of uses period were also reclassified as mosaic of uses. The same rule was applied to isolated grassland occurrences appearing at the end of time series otherwise dominated by mosaic of uses. In addition, temporary occurrences of other non-vegetated areas (25) enclosed by native classes were replaced by the subsequent valid native classification, reducing artificial gaps within otherwise continuous native-vegetation trajectories.

5.7. Frequency

The frequency filter was designed to reduce short-term temporal oscillations among native vegetation classes while preserving persistent changes in vegetation type. First, pixels classified as native vegetation in at least 95% of the temporal series (2017 - 2025) were identified as temporally stable native areas. Within these pixels, the historical frequency of each native class was calculated, and a reference class was assigned using class-specific thresholds: savanna formation (>60%), herbaceous sandbank vegetation (≥50%), grassland formation (>40%), wetland (≥95%), and forest formation (≥70%). When more than one threshold was satisfied, the classes were evaluated hierarchically, with forest formation having the highest priority, followed by wetland, grassland formation, herbaceous sandbank vegetation, and savanna formation. Before applying any replacement, the annual classification was evaluated for temporal persistence. A native class was considered persistent when it occurred for at least three consecutive years, considering backward, centered, or forward three-year windows. Persistent occurrences were retained, whereas non-persistent native classifications were replaced by the reference class assigned from the long-term frequency analysis. The filter therefore targets short-lived transitions among native vegetation classes without modifying persistent temporal patterns or pixels classified as non-native land cover.

5.8. Temporal

The temporal filter implemented in Collection 4.0 is a critical post-classification step designed to reduce spurious transitions and reinforce the temporal logic of LULC dynamics. This filter applies a set of temporal consistency rules to correct short-term spurious transitions and ensure the stability over time (2017–2025). It operates by comparing each pixel's class over multi-year windows and applying logic to eliminate

implausible transitions, enforce class persistence, and refine the first and last years of the time series. The filter follows these four main steps:

- 3-year window filtering: This rule identifies and corrects brief one-year transitions surrounded by the same class before and after (2018-2024). The objective is to correct pixel values that present a specific class in the previous year (year -1), change in the current year, and return to the initial class in the last year of the window (year +1). It is applied to each land use and cover class in the following order: savanna formation (4), grassland formation (12), forest formation (3), wetland (11), herbaceous sandbank vegetation (50), mosaic of uses (21), river, lake, and ocean (33), and other non-vegetated areas (25).
- Correction of the last year (2025): The filter searches for pixel values that were not classified as mosaic of uses (21) in 2025, but were classified as such in 2024 and 2023. The 2025 class is corrected to match the previous year, avoiding any regeneration that cannot be confirmed in the last year.
- Stabilization of the first year (2017): If a pixel was classified as native vegetation (forest, Savanna, wetland, grassland, or sandbank vegetation) in both 2018 and 2019 but not in 2017, the classification is corrected to reflect native vegetation also in 2017. This ensures temporal consistency from the beginning of the series.
- Removal of small patches of recent vegetation regrowth (2024): To avoid overestimating regeneration, only areas of native vegetation regrowth between 2024 and 2025 larger than 0.1 hectare are retained. Smaller patches are assumed to be noise and are replaced by the 2024 class.

5.9. False-regrowth and long-term trajectory

The false-regrowth filter complements the general temporal filter by correcting class-specific and long-term trajectory inconsistencies that are not adequately addressed by three-year moving windows. Its main objective is to suppress implausible vegetation-regrowth events, remove short interruptions within consolidated land-cover trajectories, and improve the temporal consistency of native vegetation and anthropogenic classes throughout the 2017–2025 series. The procedure comprises the following sequential operations:

- Harmonization of Wetland–Grassland trajectories: A stable classification as wetland throughout the entire 2017–2024 time series and ending as grassland formation (12) in 2025, all grassland occurrences within the series were reclassified as wetland. This rule reduced inconsistent transitions between these spectrally related classes and produced a more coherent long-term trajectory.
- Stabilization of long Grassland trajectories: Pixels classified as mosaic of uses (21) or Other Non-Vegetated Area (25) in 2017 were reclassified as grassland formation

throughout the complete series when they presented at least 5 annual occurrences of grassland formation and were classified as grassland formation in the 2024 reference year. The occurrences were counted across the time series and were not required to be consecutive.

- Removal of temporary anthropogenic occurrences within native vegetation: Blocks of one to three consecutive years classified as mosaic of uses (21) and enclosed by native vegetation were interpreted as temporary commission errors. This rule considered savanna formation (4), wetland (11), and grassland formation (12) as the surrounding native classes. The mosaic of uses block was replaced by the native class observed immediately after the temporary anthropogenic interval.
- Suppression of short false-regrowth intervals: Native-vegetation intervals lasting from one to three consecutive years were reclassified as mosaic of uses (21) when they occurred between two consolidated anthropogenic periods, each consisting of at least four consecutive years of mosaic of uses (21). Forest formation (3), savanna formation (4), wetland (11), and grassland formation (12) were considered native-vegetation classes in this rule.

5.10. Spatial-shape

The spatial-shape filter was introduced in Collection 4.0 to remove small, fragmented, or geometrically irregular patches classified as mosaic of uses (21). The procedure was applied independently to each annual map and used an object-based approach in which spatially connected class-21 pixels were grouped into individual patches using eight-neighbor connectivity. For each patch, the filter calculated its area, the proportion of its rectangular bounding box effectively occupied by classified pixels, and the presence of an internal core composed of a complete 3 × 3-pixel neighborhood. Patches with an area equal to or smaller than 1 hectare were considered for correction when they met at least one of the following conditions: a bounding-box fill ratio below 0.65, indicating an elongated or irregular shape; absence of a 3 × 3-pixel core, indicating a thin or fragmented structure; or a total size of three pixels or fewer, representing isolated speckle noise. Pixels belonging to the selected patches were replaced by the modal LULC class within a circular neighborhood of 150m. This procedure reduces small anthropogenic artifacts and improves the spatial coherence of the annual maps without modifying larger or more spatially consolidated patches of mosaic of uses.

5.11. Second spatial filter

The second spatial filter was applied as the final post-classification step to remove residual spatial artifacts remaining after the temporal, thematic, and trajectory-based

corrections. The procedure followed the same general logic adopted in the first spatial filter (section 5.2) and was applied independently to each annual map. Spatially connected pixels belonging to the same class were grouped into patches, and patches containing approximately 0.25 hectare or less, corresponding to 25 Sentinel pixels, were identified as potential classification noise. Pixels within these small patches were replaced by the modal land-cover class calculated within a 9×9 -pixel neighborhood. Forest formation (3), wetland (11), and river, lake, and ocean (33) were excluded from this correction to preserve small but ecologically relevant features.

6. ROCKY OUTCROP CLASSIFICATION

The classification of rocky outcrops in the Cerrado biome has undergone successive improvements over different MapBiomias Collections, both in terms of methodological refinement and conceptual definition of the class. In Collection 1.0, rocky outcrops were not mapped as a land cover class; instead, areas with exposed rock were implicitly included within the grassland formation class. Collection 2.0 marked the first explicit mapping of the rocky outcrop class in the MapBiomias 10m collections for Cerrado biome. To avoid overestimation, the classification followed a dedicated workflow, independent from the general LULC mapping. The feature space used in Collection 2.0 included the same spectral bands and vegetation indices adopted in the general classification, complemented by terrain-related predictors. Training samples were obtained through visual interpretation and from datasets provided by the *Serviço Geológico Brasileiro* (SGB/CPRM), and subsequently validated by expert interpreters.

In Collection 3.0, the rocky outcrop classification was further refined from both methodological and conceptual perspectives. The feature space was revised to rely exclusively on annual Google Satellite Embedding bands combined with geomorphometric variables. Training samples were also visually reviewed to improve their consistency with Sentinel-2 imagery and the revised classification criteria. In addition, the conceptual definition of the class was narrowed to represent predominantly exposed rock surfaces, explicitly excluding areas dominated by rupestrian vegetation. This adjustment aligned the MapBiomias 10 m product with the definition adopted in MapBiomias 30 m Collection 10.0. As in Collection 2.0, the class was mapped through an independent classification workflow, allowing the use of criteria specifically designed to represent the spectral, structural, and geomorphological characteristics of rocky environments in the Cerrado.

Collection 4.0 reinstates the broader conceptual definition adopted in Collection 2.0, reincorporating areas associated with rupestrian vegetation into the rocky outcrop class. Despite this conceptual revision, the methodological framework remains consistent with that established in Collection 3.0. The classification continues to be performed

independently from the general LULC mapping, using annual Google Satellite Embedding bands and geomorphometric variables as predictors. To ensure consistency with the revised class definition, the training dataset was comprehensively reviewed and updated. New samples were collected through visual interpretation. The independent classification strategy enables the application of class-specific criteria that better capture the distinct spectral, structural, and geomorphological characteristics of rocky outcrops and rupestrian vegetation in the Cerrado. These environments are generally associated with stable geological formations and may occur over sedimentary, igneous, or metamorphic substrates. An overview of the classification workflow is presented in Figure 5, and the following sections detail the methodological components.

6.1. Feature space

The feature space adopted for the rocky outcrop classification in Collection 4.0 differs fundamentally from that used in the general LULC classification. Given the distinct physical nature, spatial configuration, and relative temporal stability of rocky outcrops, the classification strategy prioritizes contextual and geomorphological information over traditional spectral metrics. In this workflow, only the annual Satellite Embedding bands were used as spectral predictors, same as Collection 3.0. These embeddings provide context-aware representations of surface conditions, capturing spatial patterns, texture, and neighborhood information that are particularly relevant for identifying rocky outcrops, which typically exhibit well-defined shapes, sharp boundaries, and stable spectral–structural signatures. The use of embeddings avoids reliance on vegetation-sensitive indices and reduces confusion with rupestrian vegetation.

The feature space includes the full set of 64 embedding bands available in the annual Google Satellite Embedding product. To further enhance class separability and constrain the classification to geomorphologically plausible settings, the embedding bands were combined with a comprehensive set of topographic and geomorphometric variables derived from MERIT DEM and include elevation, aspect and its sine and cosine components, profile and tangential curvature, convergence, roughness, eastness, northness, terrain ruggedness index (TRI), topographic position index (TPI), second-order terrain derivatives (DXX), and compound topographic index (CTI). In addition, spatial coordinates (latitude and longitude, transformed into sine and cosine components) were included to provide spatial context. The complete list of predictor variables used in the rocky outcrop classification is presented in Table 5.

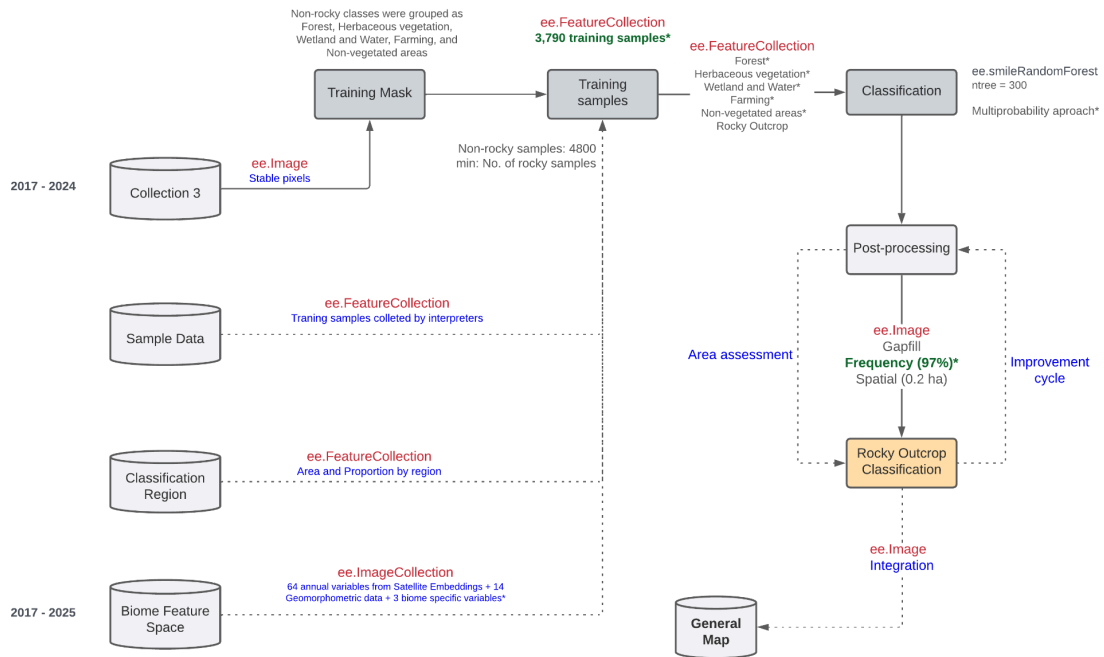


Figure 5. Each gray geometry (cylinders for databases and rectangles for processes) represents a key step in the classification schema, with the respective name inside. The gray text near databases and processes offers a description of the step, while the green text highlights the main innovations in Collection 4.0. Arrows with a continuous black line connecting the key steps represent the main direction of the processing flux, while arrows with dotted black lines represent the databases that feed the main processes. Red text inside arrows refers to the asset type in the Google Earth Engine, while blue text offers a concise description of the asset content.

Table 5. Complementary bands added to the Cerrado rocky outcrop classification feature space. Identity – the variable is used directly without temporal reduction.

Type	Name	Statistics	Reference
Embeddings Bands	Bands A00 to A63	None	Brown et al., 2025
	Elevation	Identity	Geomorpho 90m Amatulli et al., 2020
	Aspect	Identity	Geomorpho 90m Amatulli et al., 2020
Terrain	Aspect (cosine)	Identity	Geomorpho 90m Amatulli et al., 2020
	Aspect (sine)	Identity	Geomorpho 90m Amatulli et al., 2020
	Profile curvature	Identity	Geomorpho 90m Amatulli et al., 2020

Type	Name	Statistics	Reference
	Tangential curvature	Identity	Geomorpho 90m Amatulli et al., 2020
	Convergence index	Identity	Geomorpho 90m Amatulli et al., 2020
	Roughness	Identity	Geomorpho 90m Amatulli et al., 2020
	Eastness	Identity	Geomorpho 90m Amatulli et al., 2020
	Northness	Identity	Geomorpho 90m Amatulli et al., 2020
	DXX Second-order Terrain Derivatives	Identity	Geomorpho 90m Amatulli et al., 2020
	TRI Topographic Ruggedness Index	Identity	Geomorpho 90m Amatulli et al., 2020
	TPI Topographic Position Index	Identity	Geomorpho 90m Amatulli et al., 2020
	CTI Compound Topographic Index	Identity	Geomorpho 90m Amatulli et al., 2020
Geographic Coordinates	Latitude	Identity	Geolocation function
	Cosine of Longitude	Identity	Geolocation function
	Sine of Longitude	Identity	Geolocation function

6.2. Training samples, classification algorithm, and parameters

A total of 3,790 training samples were used across the entire classification area, combining samples visually interpreted by specialists. All samples were carefully reviewed to ensure consistency with the updated conceptual definition of the class adopted in Collection 4.0. For the remaining LULC classes, training samples were generated using a reference training mask derived from stable pixels of the MapBiomass 10m Collection 3.0. This mask was constructed by identifying pixels that exhibited no class changes throughout the 2017–2024 time series. Stable pixels were then grouped into broader thematic categories representing non-rocky surfaces: forest (1), herbaceous vegetation (2), wetland and water (3), farming (4), and non-vegetated areas (5). Training samples for non-rocky classes were subsequently generated using a stratified random sampling strategy, with the number of samples per class allocated proportionally to their mapped area within the area of interest (Figure 6). These samples were then combined with the visually interpreted rocky outcrop samples to form the final training dataset.

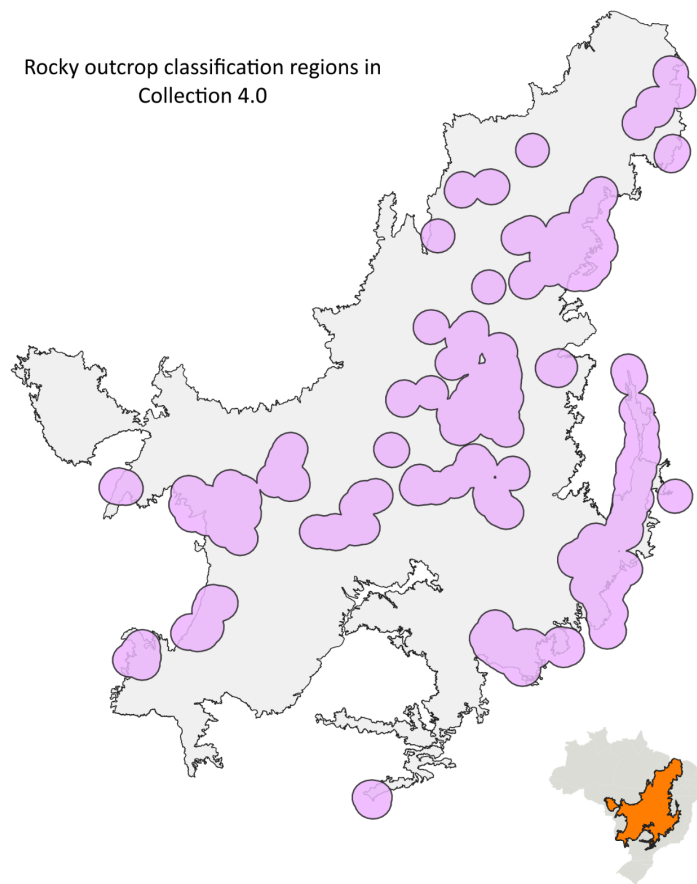


Figure 6. Area of interest used for rocky outcrop classification in MapBiomias Collection 4.0. The classification region corresponds to a 55-km buffer around the specialist-interpreted samples. The location of the Cerrado biome within Brazil is highlighted in orange.

Unlike the general land use and land cover classification, which is performed independently for each classification region, the rocky outcrop classification follows a non-regionalized approach. The classification is applied within a single area of interest defined by a spatial buffer of 55,000 meters around all training samples collected across the biome (Figure 6). This buffered area delineates the spatial extent over which the classifier is applied and reflects the known distribution and geological continuity of rocky outcrops, while avoiding unrealistic extrapolation to areas with no supporting evidence. Classification was performed for each year using the Random Forest algorithm implemented in GEE via the `ee.Classifier.smileRandomForest` function. Based on the results of previous collections, the number of decision trees was set to 300 for all regions, while the number of variables considered at each split was set to the square root of the total number of predictor bands. The model operated in multiprobability mode,

generating a probability distribution for each class. The final class for each pixel was assigned based on the highest probability.

6.3. Post-classification filters

The post-classification refinement of the rocky outcrop map followed the same methodological framework established in Section 5, incorporating temporal and spatial filtering procedures to improve classification consistency. Three main filters were applied: gap-fill, frequency, and spatial smoothing.

- The gap-fill filter addressed inconsistencies or missing classifications across the time series. Unlike conventional unidirectional approaches, this filter operated bidirectionally (both forward and backward in time), filling undefined pixels by referencing the classification of subsequent and preceding years. This ensured temporal continuity in cases where rocky outcrop presence was stable but momentarily unclassified due to spectral noise or image limitations.
- The frequency filter was designed to reinforce the temporal stability of rocky outcrops, which are geologically persistent features unlikely to change. A pixel was retained as a rocky outcrop only if it was classified as such in at least 97% of the years within the observation period. This threshold effectively filtered out areas associated with rupestrian vegetation, which exhibits more spectral variability and temporal dynamics.
- Lastly, a spatial filter was used to eliminate isolated pixels or small misclassified patches inconsistent with the typical spatial pattern of rocky outcrops. The filter removed connected components smaller than 50 pixels, equivalent to approximately 0.5 hectares, based on an 8-neighbor connectivity criterion.

7. INTEGRATION

The integration stage combines the thematic outputs generated within the Cerrado workflow and harmonizes the resulting biome-level maps with the cross-cutting land use and land cover themes produced by the MapBiomas network. The process is implemented in three sequential stages described as follows:

7.1. Internal integration

The annual rocky outcrop classification was integrated into the corresponding general Cerrado map. rocky outcrop (29) replaced the underlying general-map class within

the validated rocky outcrop classification area, including pixels previously classified as native vegetation or anthropogenic land cover. Water-related restrictions were subsequently applied to prevent the insertion of rocky outcrop in persistent water bodies. After integration, isolated overlap artifacts containing two pixels or fewer were corrected using four-neighbor connectivity and the modal class within a two-pixel neighborhood. The resulting annual maps were then used in the Paraguay Basin harmonization and cross-cutting-theme integration stages.

7.2. Paraguay Basin harmonization

A specific boundary-harmonization procedure has been applied in the *Bacia do Alto Paraguai (BAP)* since Collection 2.0. To reduce abrupt thematic discontinuities, the annual Cerrado classification is compared with the corresponding Pantanal map within the BAP boundary. The correction is restricted to pixels classified as mosaic of uses (21) in the Cerrado map. Where the Pantanal classification provides evidence of native vegetation, the Cerrado class is replaced according to the following correspondence:

- Cerrado mosaic of uses and Pantanal forest formation → savanna formation;
- Cerrado mosaic of uses and Pantanal savanna formation → savanna formation;
- Cerrado mosaic of uses and Pantanal wetland → wetland;
- Cerrado mosaic of uses and Pantanal grassland formation → grassland formation.

7.3. Integration with cross-cutting themes

Each annual Cerrado map is combined with the cross-cutting thematic products generated by MapBiomas, including agriculture, pasture, mining, urban, and other land use and land cover themes. To harmonize the thematic data with the biome-level classifications, a set of predefined prevalence rules is applied. These rules define which classes take precedence when overlaps occur, ensuring a standardized decision logic across biomes and years. The prevalence rules adopted are described in Table 6. It is important to note that exceptions to this general rule apply in the context of protected areas regarding class prevalence:

- In protected areas, the native vegetation (3, 4, 11, and 12) is prevalent within the Cotton (62), Citrus (47), and Coffee (46) classes.
- In the case of pasture (15) within protected areas, native vegetation (3, 4, 11, and 12) is also preserved.
- Outside protected areas, pasture (15) is the prevailing land use class, superseding savanna, wetland, and grassland classes (4, 11, and 12).

In the most recent MapBiomass 30 m collections, which include annual maps through 2025, pasture no longer takes precedence over savanna formation (4) outside protected areas. Therefore, the precedence of pasture over savanna formation described here should not be interpreted as representing the current integration logic adopted in the MapBiomass 30 m products.

Table 6. General prevalence rules - MapBiomass Collection 4.0

Class	Pixel value	Prevalence order	Color
Photovoltaic Power Plant	75	1	
Wind farm (beta)	91	2	
Mining	30	3	
Beach, Dune, and Sand Spot	23	4	
Mangrove	5	5	
Aquaculture	31	6	
Hypersaline Tidal Flat	32	7	
Urban Infrastructure	24	8	
Forest Plantation	9	9	
Rocky Outcrop	29	10	
Temporary Crops	19	11	
Perennial Crops	36	12	
Herbaceous Sandbank Vegetation	50	13	
River, Lake, and Ocean	33	14	
Forest Formation	3	15	
Savanna Formation	4	16	
Wetland	11	17	
Grassland Formation	12	18	
Pasture	15	19	
Mosaic of Uses	21	20	
Other Non-Vegetated Areas	25	21	

8. ACCURACY METRICS

MapBiomass 10m Collections 1.0 to 3.0 did not include formal accuracy estimates, and the corresponding products were therefore considered *beta* versions. Beginning with Collection 4.0, a dedicated effort was undertaken to report classification accuracy at the biome level. This assessment was initially conducted using Level 1 of the MapBiomass

legend, which comprises five broad land use and land cover classes: Forest, Herbaceous and Shrubby Vegetation, Farming, Non-Vegetated Area, and Water. The implementation of biome-specific accuracy assessment represented an important methodological advance by providing a quantitative evaluation of map reliability and enabling comparisons among subsequent collections. A complete description of the accuracy assessment methodology, including the sampling design, validation procedures, and reported results, is provided in the general MapBiomias ATBD.

9. REFERENCES

- 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.
- Baptista, G. M. M. (2015). Aplicação do Índice de Vegetação por Profundidade de Feição Espectral (SFDVI – Spectral Feature Depth Vegetation Index) em dados RapidEye. In *Anais do XVII Simpósio Brasileiro de Sensoriamento Remoto (SBSR)* (pp. 1–8). INPE, João Pessoa, PB, Brasil.
- Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., & Kohli, P. (2025). AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv*. <https://arxiv.org/abs/2507.22291>.
- Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proc. Natl. Acad. Sci. USA*, 114, 2189–2194.
- Donchyts, G., Winsemius, H., Schellekens, J., Erickson, T., Gao, H., Savenije, H., & van de Giesen, N. (2016). Global 30m Height Above the Nearest Drainage (HAND). *Geophysical Research Abstracts*, Vol. 18, EGU 2016, 17445-3.
- Gamon, J. A., Peñuelas, J., & Field, C. B. (1992). A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. *Remote Sensing of Environment*, 41(1), 35-44.
- Gitelson, A. A., & Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology*, 22, 247–252. [https://doi.org/10.1016/1011-1344\(93\)06963-4](https://doi.org/10.1016/1011-1344(93)06963-4)
- Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, 58, 289–298. [https://doi.org/10.1016/S0034-4257\(96\)00072-7](https://doi.org/10.1016/S0034-4257(96)00072-7).
- Gitelson, A. A., Kaufman, Y. J., Stark, R., & Rundquist, D. (2002). Novel algorithms for remote estimation of vegetation fraction. *Remote Sensing of Environment*, 80(1), 76–87. [https://doi.org/10.1016/S0034-4257\(01\)00289-9](https://doi.org/10.1016/S0034-4257(01)00289-9)

Gitelson, A. A., Viña, A., Arkebauer, T. J., Rundquist, D. C., Keydan, G., & Leavitt, B. (2003). Remote estimation of leaf area index and green leaf biomass in maize canopies. *Geophysical Research Letters*, 30(5), 1248.

Gitelson, A. A., Viña, A., Ciganda, V., Rundquist, D. C., & Arkebauer, T. J. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters*, 32, L08403. <https://doi.org/10.1029/2005GL022688>

Guyot, G., & Baret, F. (1988). Utilisation de la haute résolution spectrale pour suivre l'état des couverts végétaux. In *Proceedings of the 4th International Colloquium on Spectral Signatures of Objects in Remote Sensing* (pp. 279–286). ESA.

Haboudane, D., Miller, J. R., Tremblay, N., Zarco-Tejada, P. J., & Dextraze, L. (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sensing of Environment*, 81(2–3), 416–426. [https://doi.org/10.1016/S0034-4257\(02\)00018-4](https://doi.org/10.1016/S0034-4257(02)00018-4)

Hall, R. J., Skakun, R. S., Arsenault, E. J., & Case, B. S. (2006). Modeling forest stand structure attributes using Landsat ETM+ data: Application to mapping of aboveground biomass and stand volume. *Forest ecology and management*, 225(1-3), 378-390.

Housman, I., Chastain, R., & Finco, M. (2018). An Evaluation of Forest Health Insect and Disease Survey Data and Satellite-Based Remote Sensing Forest Change Detection Methods: Case Studies in the United States. *Remote Sensing*, 10, 1184.

INMET (Instituto Nacional de Meteorologia). (2024). Portal INMET. Brasília, DF. Accessed on: <https://portal.inmet.gov.br/>

Lang, N., Jetz, W., Schindler, K., & Wegner, J. D. (2022). A high-resolution canopy height model of the Earth. *arXiv preprint arXiv:2204.08322*.

Lu, R., Liu, S., Duan, H., Kang, W., & Zhi, Y. (2024). Combining the SHAP Method and Machine Learning Algorithm for Desert Type Extraction and Change Analysis on the Qinghai–Tibetan Plateau. *Remote Sensing*, 16(23), 4414. <https://doi.org/10.3390/rs16234414>

Macena, F., Assad, E., Steinke, E., Müller, A. (2008). Clima do Bioma Cerrado. *In: Albuquerque, A., Silva, A. G. (2008). Agricultura tropical: quatro décadas de inovações tecnológicas, institucionais e políticas*. Brasília, DF: Embrapa Informação Tecnológica, 2008, 1137 p.

Nagler, P. L., Inoue, Y., Glenn, E. P., Russ, A. L., & Daughtry, C. S. T. (2003). Cellulose absorption index (CAI) to quantify mixed soil–plant litter scenes. *Remote Sensing of Environment*, 87(2-3), 310-325.

Parente, L., & Ferreira, L. (2018). Assessing the Spatial and Occupation Dynamics of the Brazilian Pasturelands Based on the Automated Classification of MODIS Images from 2000 to 2016. *Remote Sensing*, 10, 606.

Potter, C., Li, S., Huang, S., & Crabtree, R. L. (2012). Analysis of sapling density regeneration in Yellowstone National Park with hyperspectral remote sensing data. *Remote Sensing of Environment*, 121, 61–68. <https://doi.org/10.1016/j.rse.2012.01.009>

Qi, J., Chehbouni, A., Huete, A. R., Kerr, Y. H., & Sorooshian, S. (1994). A modified soil-adjusted vegetation index. *Remote Sensing of Environment*, 48(2), 119–126. [https://doi.org/10.1016/0034-4257\(94\)90134-1](https://doi.org/10.1016/0034-4257(94)90134-1)

Rosa, M. R. (2020). Metodologia de classificação de uso e cobertura da terra para análise de três décadas de ganho e perda anual da cobertura florestal nativa na Mata Atlântica (Doctoral Dissertation, Universidade de São Paulo).

Rouse, R. W. H., Haas, J. A. W., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. Third Earth Resources Technology Satellite (ERTS) Symposium, 309–317.

Sano, E. E., Rodrigues, A. A., Martins, E. S., Bettiol, G. M., Bustamante, M. M. C., Bezerra, A. S., Couto, A. F., Vasconcelos, V., Schüller, J., & Bolfe, E. L. (2019). Cerrado Ecoregions: A spatial framework to assess and prioritize Brazilian savanna environmental diversity for conservation. *Journal of Environmental Management*, 232, 818–828.

Souza, C. M., Roberts, D. A., & Cochrane, M. A. (2005). Combining spectral and spatial information to map canopy damage from selective logging and forest fires. *Remote Sensing of Environment*, 98, 329–343.

Xiao, J., Shen, Y., Tateishi, R., & Bayaer, W. (2006). Development of topsoil grain size index for monitoring desertification in arid land using remote sensing. *International Journal of Remote Sensing*, 27(11), 2411–2422.

Xu, H. (2006). Modification of normalized difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033. <https://doi.org/10.1080/01431160600589179>

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>