



Cerrado - Appendix

Collection 10

Version 1

General coordinator

Ane A. Alencar

Team

Bárbara C. da Silva

Dhemerson E. Conciani

Joaquim J. S. P. Pereira

Julia Z. Shimbo

Vera L. S. Arruda

Wallace V. da Silva

August, 2025

1. OVERVIEW OF THE CERRADO CLASSIFICATION METHOD

The land cover classification methodology for the Cerrado biome, developed under the MapBiomass project, uses decision tree-based algorithms to generate annual maps of dominant native vegetation (NV) types. These types are grouped into five broad categories: Forest Formation, Savanna Formation, Wetland, Grassland Formation, and Rocky Outcrop. Since the project's inception, the methodology has undergone continuous refinement, resulting in substantial improvements from the first collection to the current version (Collection 10.0). While significant advances have been made in recent editions, efforts have also focused on preserving the methodological improvements achieved in previous collections. The Cerrado classification workflow involves several key steps. First, identify the optimal time window to generate annual Landsat mosaics. Then, remote sensing metrics were defined as potential predictors (feature space). Reference training samples were generated to calibrate the classification algorithm. Post-classification treatments were applied to reduce noise and ensure temporal consistency. Finally, the resulting maps were integrated with other cross-cutting themes. Classification results are validated through both visual inspection and sample-based accuracy assessment. A detailed summary of the methodological evolution of land use and land cover (LULC) classifications in the Cerrado biome is presented in Table 1.

Table 1. Table 1. Summary of methodological evolution across MapBiomass Cerrado collections. EDT: Empirical Decision Tree; RF: Random Forest; SR: Surface Reflectance; NV: Native Vegetation.

Collection	Year Range	Method	Mapped classes	Key Improvements
1.0	2008 – 2015	EDT	Forest	First version of the Cerrado collection
2.0	2000 – 2016	EDT	Forest, Savanna, Grassland	Inclusion of new NV classes (Savanna and Grassland formations)
2.3	2000 – 2016	RF	Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Other Non-vegetated Area, Water	Introduction of Random Forest (RF); Auxiliary classes (mosaic, non-vegetated areas, and water); Training samples from stable areas
3.0	1985 – 2017	RF	Same as Collection 2.3	Full Landsat series (since 1985); Improved training sample quality with outlier removal
3.1	1985 – 2017	RF	Same as Collection 3.0	The classification by ecoregions (38) was used to replace the regular tiles.
4.0	1985 – 2018	RF	Same as Collection 3.1	Improvement in training samples quality by using new reference maps

Collection	Year Range	Method	Mapped classes	Key Improvements
4.1	1985 – 2018	RF	Forest, Savanna, Grassland, Pasture, Agriculture, Other Non-vegetated Area, Water	Variable importance for feature selection; Enhanced post-processing temporal filter; Significant accuracy gains in NV mapping
5.0	1985 – 2019	RF	Same as Collection 4.1	New reference maps; Improvements in spatial continuity between classification regions New vegetation dynamics product (deforestation and secondary vegetation)
6.0	1985 – 2020	RF	Forest, Savanna, Wetland, Grassland, Mosaic of Agriculture and Pasture, Other Non-vegetated Area, Water	New NV class (Wetland); New reference maps; Landsat SR mosaics; Feature space refinement
7.0	1985 – 2021	RF	Forest, Savanna, Wetland, Grassland, Rocky Outcrop, Mosaic of Uses, Other Non-vegetated Area, Water	New NV class (Rocky Outcrop); New reference maps; GEDI for outlier training samples filtering; Revision of classification regions (38); Significant accuracy gains in NV mapping
7.1	1985 – 2021	RF	Same as Collection 7.0	Improvement of temporal filter rules in the last year (2021)
8.0	1985 – 2022	RF	Same as Collection 7.0	New reference maps; Regionalization of hyperparameters and classification; Revision of the temporal filtering rules; Territorial expansion of Rocky Outcrop classification
9.0	1985 – 2023	RF	Same as Collection 7.0	New reference maps; Multiprobability approach; Post-processing filters revision; New false regrowth filter; New Rocky Outcrop workflow; Significant accuracy gain in NV mapping
10.0	1985-2024	RF	Forest, Savanna, Wetland, Grassland, Rocky Outcrop, Herbaceous Sandbank Vegetation, Mosaic of Uses, Other Non-vegetated Area, Water	New reference maps; New NV Class (Herbaceous Sandbank Vegetation); Monthly Landsat SR mosaics; Herbaceous Sandbank Vegetation and Geomorphometric filters; Redefined Rocky Outcrop concept

In Collections 1.0 and 2.0, empirical decision trees were applied, with rule-based thresholds informed by expert knowledge of the spectral characteristics of each class. Collection 1.0 covered the period 2008 to 2015 and was released in 2016. Collections 2.0 and 2.3 extended the temporal range to 2000–2016, with Collection 2.3 (a revised version

of Collection 2.0) introducing the use of the Random Forest (RF) algorithm. Empirical decision trees were used to generate stable training samples for the 2000–2016 period, which in turn were used to train the RF classifier across the full time series. Collections 3.0 and 3.1 expanded the historical coverage to include the years 1985–2017, and a methodological paper was published (Alencar et al., 2020). Collections 4.0 and 4.1 (1985–2018) demonstrated a notable enhancement in mapping precision compared to previous collections. These improvements were largely driven by the refinement of the feature space, which was redesigned based on a variable importance analysis using the RF algorithm. Additionally, new post-classification filters were implemented to better account for temporal consistency and reduce noise in the time series, resulting in more accurate detection of land cover transitions. From Collection 2.3 onward, training samples were no longer generated using empirical trees but were instead based on stable samples from previous collections, contributing to improved classification accuracy.

Beginning with Collection 5.0 (1985–2019), official reference maps (PRODES Cerrado and PROBIO) were incorporated to spatially constrain training sample collection for NV classes, minimizing bias. Collection 6.0 (1985–2020) introduced several important methodological updates. The classified time series was extended to encompass the period from 1985 to 2020, a new NV class (Wetland) was introduced, the surface reflectance mosaic was implemented, and the feature space was refined. In addition to the reference maps already used in Collection 5.0, the *“Inventário Florestal do Estado de São Paulo”* was incorporated to support training sample filtering for NV classes. In Collection 7.0 (1985–2021), a new class (Rocky Outcrop) was introduced. The training dataset was refined through the application of an outlier detection filter based on GEDI Global Forest Height data (Lang et al., 2022) and the incorporation of the *“Base Temática Digital do Estado do Tocantins”* as an additional reference map. These improvements enhanced the quality and reliability of training samples. The reference maps adopted in Collection 6.0 were maintained. Unlike in Collection 6.0, where the Wetland class was treated as a pseudo cross-cutting theme, it was directly integrated into the main classification process in Collection 7.0. The Collection 7.1 was released as a reprocessed version of Collection 7.0 to correct temporal filtering issues in the final year of the series. It addressed the overestimation of deforestation peaks in 2021 caused by misclassified transitions.

Collection 8.0 (1985–2022) incorporated substantial methodological advances. The training mask was improved with the inclusion of deforestation alerts from MapBiomas Alert and SAD Cerrado (2019–2022), enhancing the filtering of anthropogenic classes. Additionally, the classification approach was strengthened by regionalizing both the hyperparameters and the classification process, allowing better adaptation to local biophysical conditions. The classification of the Rocky Outcrop class was also expanded to ensure its representation across the entire Cerrado biome, using a more robust and spatially consistent methodology. Collection 9.0 (1985–2023) continued this strategy,

updating training sample filtering with alerts from MapBiomas Alert and SAD Cerrado (2019–2023). New reference datasets were introduced, including the “*Mapa de Uso e Cobertura da Terra do Distrito Federal*” and the “*Mapeamento dos Remanescentes de Campos de Murundus do Estado de Goiás*”. In this collection, hyperparameter regionalization was discontinued, although the regional classification scheme was maintained. Moreover, the classification process adopted a multiprobability approach, which considers the probability distribution of all classes instead of relying solely on majority voting, thereby improving the reporting of classification uncertainties. Post-processing routines were revised, and a new false regrowth filter was introduced to suppress spurious transitions in recent years. The Rocky Outcrop class was also expanded using a revised classification approach, updated training samples, and new predictor variables.

In the current Collection 10.0 (1985–2024), the classification period was updated, and the strategy for training sample design was revised to improve spatial representativeness and reduce classification noise. Training samples filtering was updated to incorporate MapBiomas Alert deforestation alerts (2019–2024), with the removal of SAD Cerrado data used in previous collections. The training masks were refined by including new reference datasets, such as the “*Mapa de Veredas e Áreas Úmidas do Sudoeste do Tocantins*” and “*Cobertura e Uso da Terra do Sudoeste do Tocantins*”. A major improvement in image preprocessing was the adoption of monthly Landsat surface reflectance mosaics, which improved temporal resolution and data quality for classification. The classification continues to apply the multiprobability approach. Post-processing filters were comprehensively revised to enhance the accuracy of land cover change detection. Two new thematic filters were introduced: (i) a post-processing step to map the Herbaceous Sandbank Vegetation, and (ii) a Geomorphometric filter, designed to enhance Sandbank Vegetation mapping in coastal areas and to remove misclassified Wetland pixels in shadow or relief-prone zones. Furthermore, the methodology for mapping the Rocky Outcrop class underwent revision in this collection, subsequent to a conceptual update. The class is now strictly defined as areas with exposed rock, explicitly excluding regions with rupestrian vegetation. This approach marks a departure from the method employed in Collections 7.0 to 9.0, wherein both exposed rock and rupestrian vegetation were incorporated into the classification. The classification and post-processing scripts employed in this collection are available at <https://github.com/mapbiomas/brazil-cerrado>.

2. METHODOLOGICAL DESCRIPTION

In MapBiomas Collection 10.0, the classification of Landsat mosaics for the Cerrado biome included a comprehensive set of ten land use and land cover classes, as

defined in the official MapBiomas legend (available at: [Legend Code](#)). These classes, detailed in Table 2, include native vegetation (NV) and water bodies, as well as anthropogenic land use classes, such as Agriculture and Pasture. It is important to note that the classes Agriculture and Pasture are mapped in the early stages of the classification process. These classes are not the focus of the Cerrado-specific methodology and are derived from cross-cutting themes coordinated by teams specialized in farming land use. Within the Cerrado classification framework, these classes help improve the delineation of native vegetation by reducing the risk of omission or commission errors and minimizing confusion between anthropogenic and natural land cover types. During the post-processing stage, these two classes are merged into a single class labeled as Mosaic of Uses.

The following sections provide a detailed description of the main components of the Collection 10.0 methodology: Landsat 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 shown in Figure 1. The core methodological steps can be summarized as follows:

- **Landsat Mosaic:** Multi-temporal surface reflectance mosaics were generated from monthly Landsat imagery to ensure optimal spectral quality, reduce atmospheric noise, and maintain temporal consistency across the time series.
- **Training Samples:** Samples were primarily extracted from stable pixels identified in Collection 9.0 (1985–2023), complemented by thematic reference data and GEDI-derived information. A stratified sampling design ensured proportional class representation across regions and years.
- **General Classification:** Land cover classification was conducted using the Random Forest algorithm in Google Earth Engine, applying a multiprobability approach to represent 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:** Two outputs were generated: (i) the general land use and land cover map and (ii) the rocky outcrop map. The latter was integrated into the general map through a final post-processing procedure.

- **Integration with Cross-Cutting Themes:** Annual biome maps were harmonized with thematic layers using predefined MapBiomass prevalence rules, ensuring consistency and avoiding class overlaps.

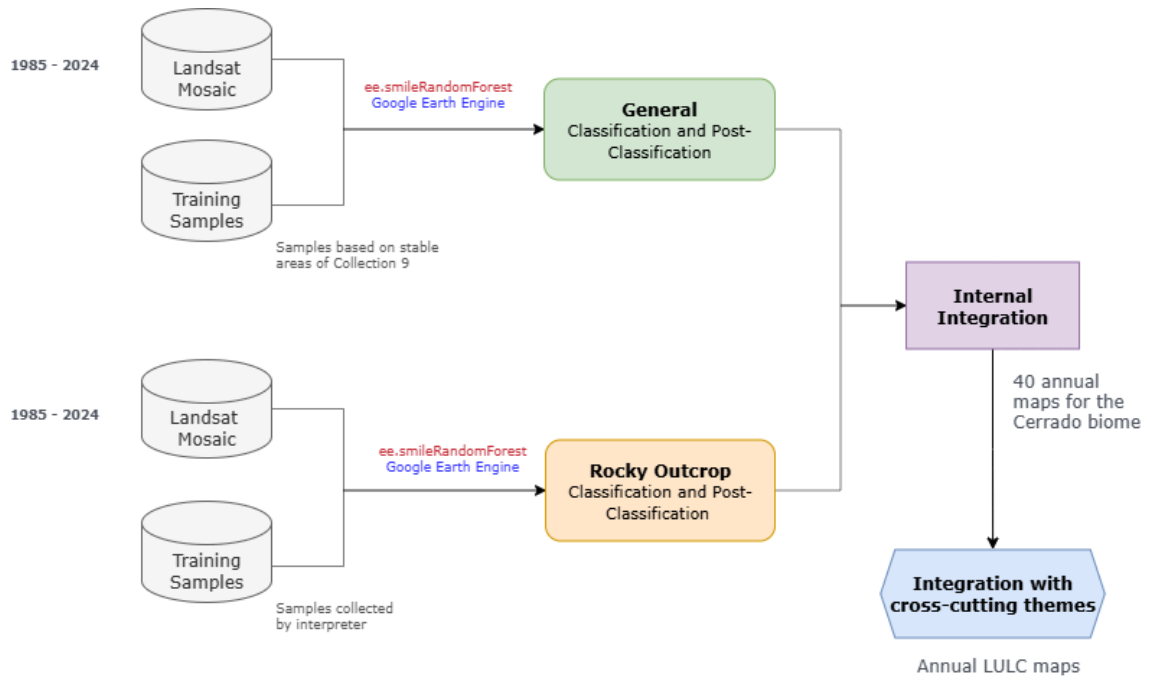

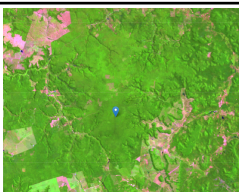

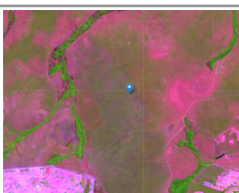

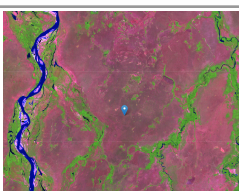









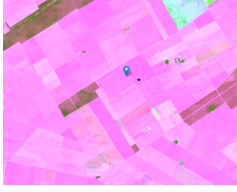



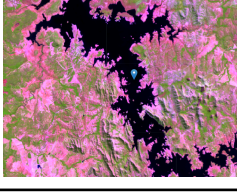


Figure 1. Workflow for the land use and land cover (LULC) classification of the Cerrado biome in Collection 10.0. Two parallel modules were applied: (i) the general classification, and (ii) a specific classification of Rocky Outcrop. Both used the Random Forest algorithm implemented in Google Earth Engine. Outputs were integrated into two stages: (i) internal integration combined the Rocky Outcrop map with the general map; and (ii) integration of the maps with cross-cutting themes to produce 40 annual LULC maps (1985 to 2024).

Table 2. Land use land cover categories used for the Landsat mosaics classification for the Cerrado biome in MapBiomass Collection 10.0.

Classes Level 1	Classes Level 2	ID	Legend Color	RGB composite (SWIR1-NIR-Red)	Description
Forest	Forest Formation	3			Vegetation types with predominance of tree species, with continuous canopy formation (Riparian Forest, Gallery Forest, Dry Forest, and Forested Savanna - <i>Mata Ciliar, Mata de Galeria, Mata Seca e Savana Florestada</i>) (Ribeiro & Walter, 2008), as well as Semi-deciduous Seasonal Forests.
	Savanna Formation	4			Savanna formations with defined tree and shrub-herbaceous stratum (Cerrado <i>Stricto Sensu</i> : Dense, Typical, Sparse and Rupestrian Savanna - <i>Cerrado Stricto Sensu: Savana Densa, Típica, Esparsa e Rupestre</i>).
Herbaceous and Shrubby Vegetation	Wetland	11			Vegetation with a predominance of herbaceous strata subject to seasonal flooding (e.g., Campo Úmido) or under fluvial/lacustrine influence (e.g., Brejo). In some regions, the herbaceous matrix is associated with arboreal species of savanna formation (e.g., Parque de Cerrado) or palm trees (Vereda, Palmeiral).
	Grassland	12			Grassland formations with a predominance of herbaceous strata (Dirty, Clean and Rupestrian Grassland - <i>Campo Sujo, Limpo e Rupestre</i>) and some areas of savanna formations such as the Rupestrian Cerrado.
	Rocky Outcrop	29			Monolithic features, bedrock, or slabs naturally exposed to the earth's surface without soil cover or rupestrian vegetation. This class typically includes stable geological formations with clear indicators of sedimentary, igneous, or metamorphic origins.

Classes Level 1	Classes Level 2	ID	Legend Color	RGB composite (SWIR1-NIR-Red)	Description
	Herbaceous Sandbank Vegetation*	50			A landscape composed of predominantly herbaceous-shrubby vegetation, with sparse shrubs, developed on sandy plains in coastal environments.
Farming	Pasture*	15			The pasture area, predominantly planted, is linked to cattle ranching activities.
	Agriculture*	18			Areas occupied with short to long-cycle crops. This encompasses both perennial and temporary crops.
Non-Vegetated Area	Other Non-Vegetated Areas	25			Areas of non-permeable surfaces (infrastructure, urban infrastructure, or mining), regions of exposed soil in natural areas (e.g., erosion and landslides), or in crop areas in the off-season.
Water	River, Lake, and Ocean	33			Rivers, lakes, dams, reservoirs, and other water bodies

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

3. LANDSAT IMAGE MOSAICS

The first step in the classification of native vegetation (NV) in the Cerrado biome is the generation of annual image mosaics, which serve as the primary input for the LULC classification process. In earlier MapBiomass collections, up to 5.0, annual mosaics were generated using top-of-atmosphere (TOA) reflectance data from Landsat 5 (TM), Landsat 7 (ETM+), and Landsat 8 (OLI). From Collection 6.0 onward, this approach was replaced using surface reflectance (SR) data from Landsat Collection 2, which provides atmospherically corrected and radiometrically consistent inputs across sensors throughout the time series. To ensure temporal continuity and consistency from 1985 onward, a sensor-prioritization strategy was adopted. Landsat 5 (TM) SR data was prioritized for the period 1985 to 2010, except in 2001 and 2002, when TM sensor issues led to the use of Landsat 7 (ETM+). Landsat 7 data remained the primary source for 2011 and 2012, followed by Landsat 8 (OLI) from 2013 onward.

Up to Collection 9.0, annual mosaics were constructed using Level-2 surface reflectance images from Landsat Collection 2. In Collection 10.0, however, a major advancement was the adoption of the composite dataset [LANDSAT/COMPOSITES/C02/T1_L2_32DAY](#), a product composed of 32-day surface reflectance composites derived from orthorectified Level-1 scenes. This product provides increased scene availability, improved geometric correction, and greater temporal continuity, especially in earlier years of the time series. For more information about this product, refer to the Earth Engine Data Catalog. The adoption of this new mosaic source in Collection 10.0 led to notable improvements in mosaic quality, particularly in regions historically affected by persistent cloud cover or limited image availability.

Since the first collection, the optimal temporal window for mosaic generation has been systematically evaluated, considering the strong seasonality of Cerrado vegetation. Given that vegetation spectral responses differ significantly between the rainy and dry seasons, comparative tests were conducted using imagery compositions from both the rainy and dry seasons (Figure 2). Imagery from the end of the rainy season, when vegetation remains vigorous and cloud cover tends to decrease, resulted in greener mosaics but also increased commission errors in forest formation class mapping. Conversely, mosaics generated from images acquired in the final months of the dry season (July to September) resulted in drier mosaics, which led to underestimation of forest coverage, mainly due to the reduced potential to detect deciduous forests.

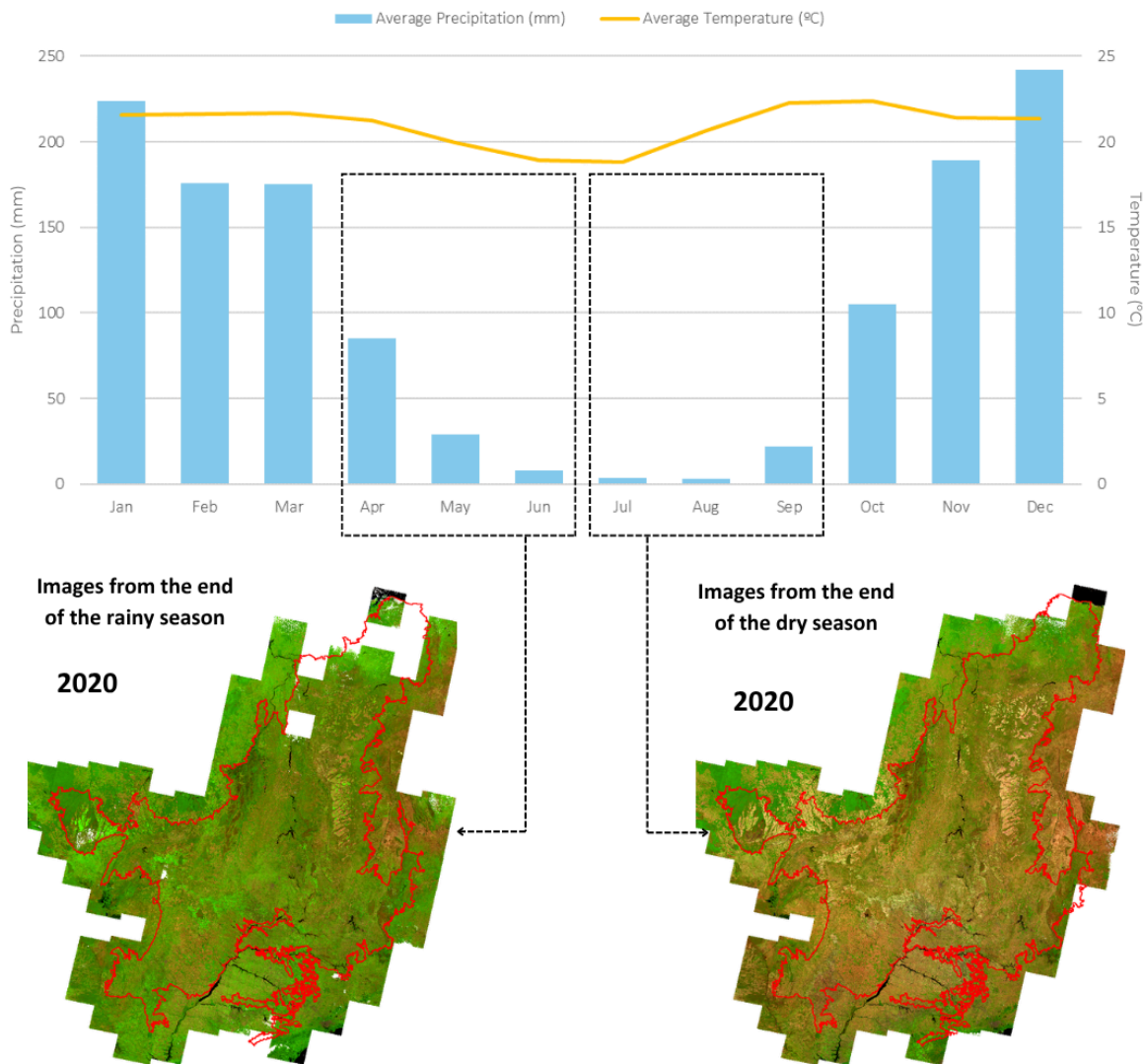


Figure 2. False-color composite of Landsat 8 (OLI) mosaics illustrating spectral differences between the end of the rainy season and the end of the dry season in the Cerrado biome. Precipitation data represent monthly averages for the Cerrado region (Macena et al., 2008), while temperature data correspond to monthly averages for the Federal District (INMET).

Based on these findings, a standardized six-month compositing window (April to September) was adopted across all 38 classification regions and all years. This broader window improved regional consistency and addressed limitations observed in tests using narrower time windows (Figure 3). The mosaics are constructed by compositing pixels from all valid Landsat scenes within this window. For each pixel, reduction metrics such as median, amplitude, standard deviation, and minimum are calculated and aggregated annually, resulting in a multi-band image that serves as input for the classification models. The final mosaics are visually inspected to ensure their quality and suitability for subsequent classification steps.

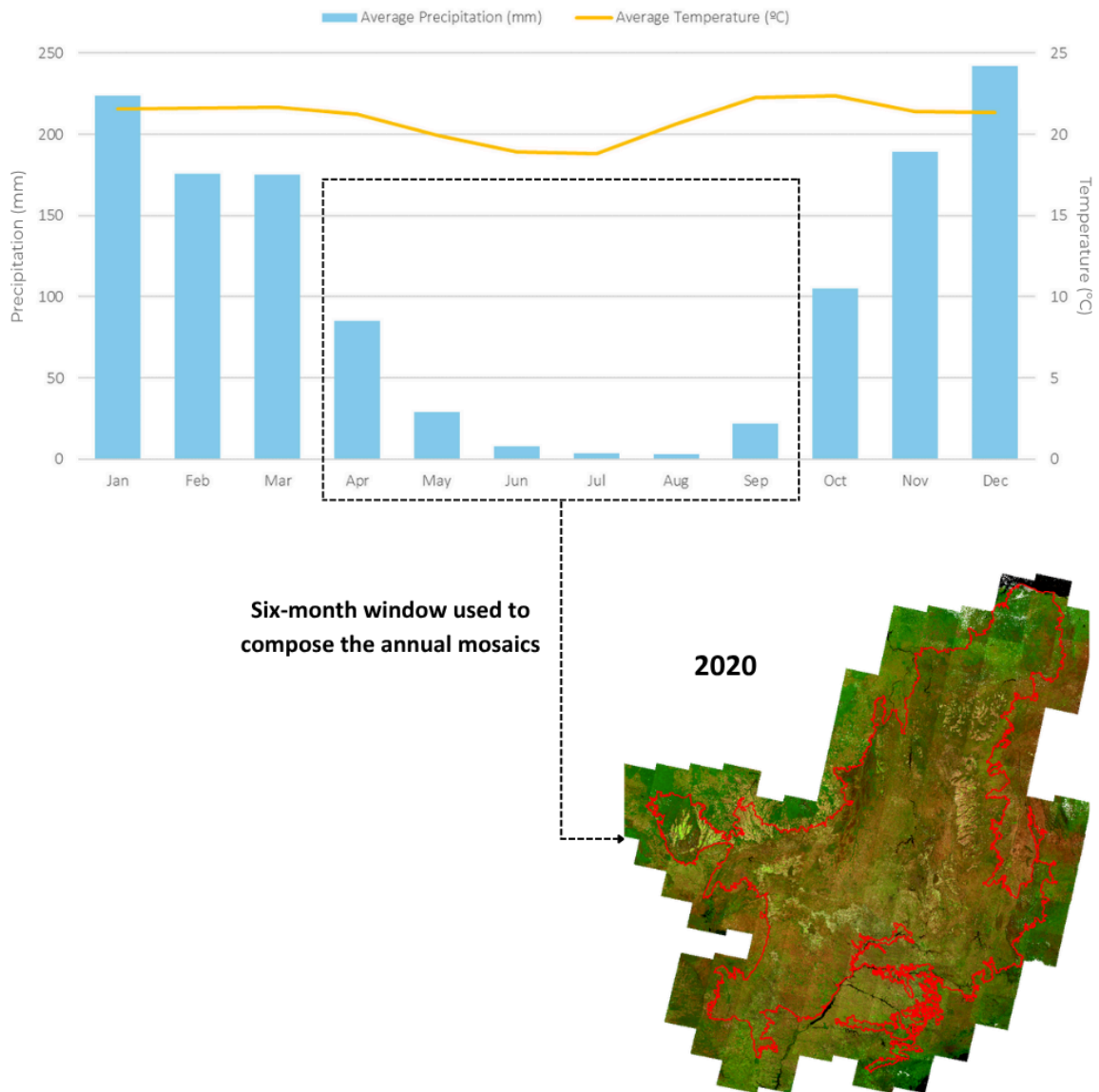


Figure 3. Time window adopted for the annual classification mosaics construction in the Cerrado biome. Precipitation data represent monthly averages for the Cerrado region (Macena et al., 2008), and temperature data correspond to monthly averages for the Federal District (INMET).

4. GENERAL MAP CLASSIFICATION

The complete workflow for the general land use and land cover classification is shown in Figure 4. This workflow integrates multiple processing steps to ensure accurate and temporally consistent mapping across the Cerrado biome. The methodological steps are detailed in the following sections, beginning with the delineation of classification regions (4.1), followed by the construction of the feature space used for model training (4.2). The sampling strategy and classification procedure are described in section 4.3,

including 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.

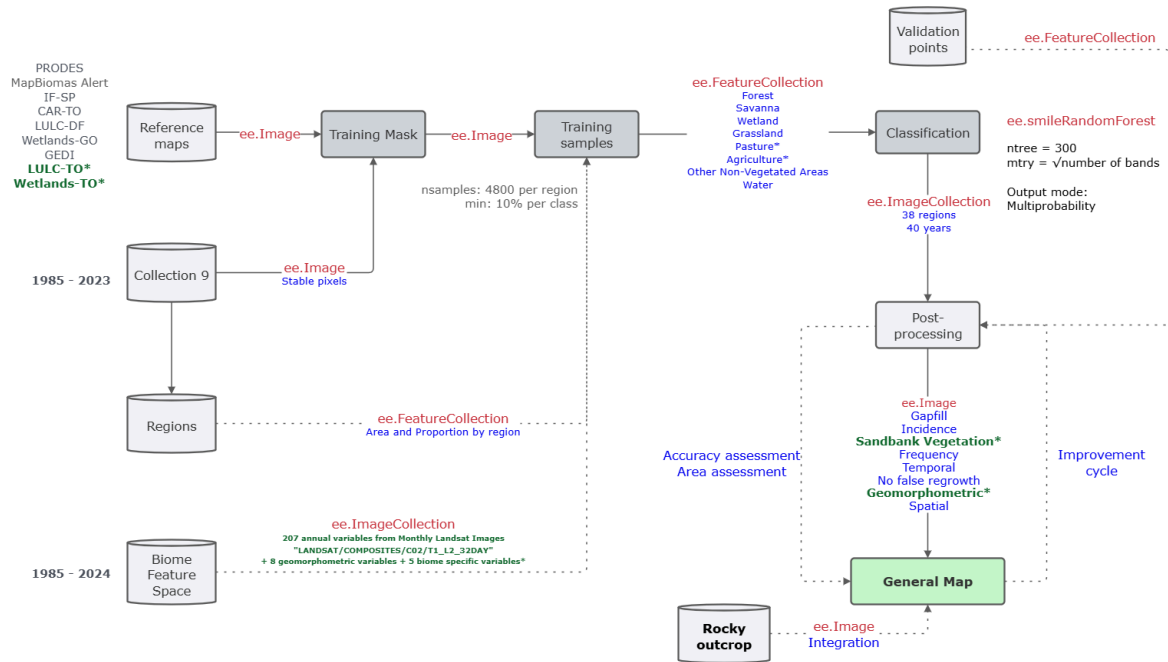


Figure 4. Each gray geometry (cylinders for databases and rectangles for processes) represents a key step in the classification schema, with the respective name inside. The gray text near databases and processes offers a description of the step, while the green text highlights the main innovations in Collection 10.0. 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.

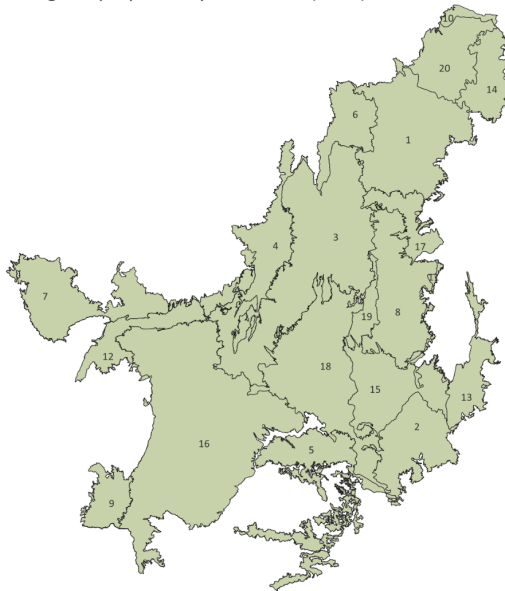
4.1. Classification regions

In the initial MapBiomass Cerrado collections (up to Collection 3.0), the classification process was organized using a regular grid based on a 1:250,000 scale, comprising 172 tiles. Each tile was independently processed by the classification algorithm. Although operationally simple, this grid-based approach often caused discontinuities and inconsistencies along tile boundaries, compromising the spatial coherence of the resulting maps. Starting from Collection 3.1, a new regionalization strategy was implemented to better capture the ecological and biophysical heterogeneity of the Cerrado biome. Classification units were defined based on the ecoregions proposed

by Sano et al. (2019) and further refined using Brazil's major hydrographic basins and the spatial distribution of land use and land cover classes from Collection 3.0 (2017). This integration resulted in the delineation of 38 classification regions, replacing the previous tile-based system and enabling a more ecologically meaningful and spectrally consistent classification.

In Collection 7.0, these classification regions were refined to account for phenological variability, particularly related to the seasonality of native vegetation (Figure 5). This refinement was based on an empirical analysis of Sentinel-2 surface reflectance (SR) data from 2017 to 2020, in which the Normalized Difference Vegetation Index (NDVI) was calculated for each pixel. The difference between the 90th and 10th percentiles (p90–p10) of NDVI values was used to identify areas with high seasonal variation. These patterns were then used to adjust the region boundaries, ensuring that areas with distinct spectral and phenological characteristics were not grouped within the same classification region. Since Collection 7.0, this revised set of 38 classification regions has been maintained and remains in use in Collection 10.0. This framework has proven effective in capturing the Cerrado's regional vegetation dynamics and enhancing the consistency and accuracy of classification results across diverse environmental contexts.

Ecoregions proposed by Sano et al. (2019)



Classification regions since Collection 7.0

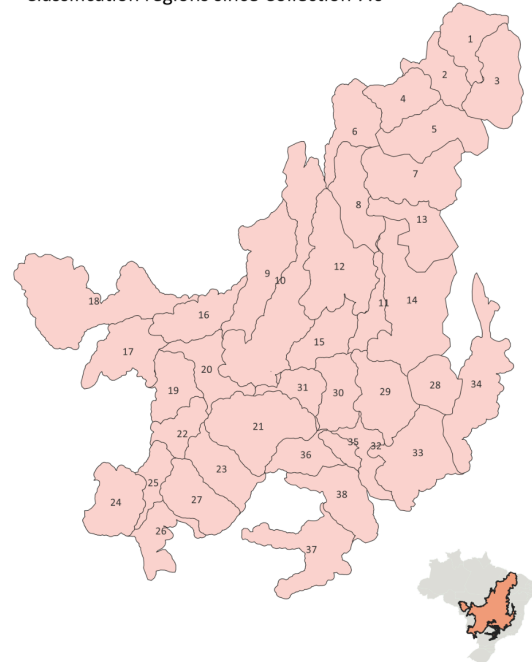


Figure 5. Classification regions, modified from Sano et al., 2019. Highlighted in orange is the location of the Cerrado biome in Brazilian territory.

4.2. Feature space

The feature space used for LULC classification in Collection 10.0 consisted of a comprehensive set of 220 predictor variables, designed to capture the spectral, temporal, and environmental complexity of the Cerrado biome. These variables included both dynamic (annual) and static (non-annual) components. The annual component was primarily derived from Landsat surface reflectance mosaics, as described in Table 3. It included the original spectral bands, a range of vegetation indices, and variables derived from spectral mixture modeling. For each of these variables, a set of statistical reduction metrics was calculated within the April–September compositing window, including minimum (5th percentile), maximum (95th percentile), median, mean, standard deviation, and amplitude. These metrics were selected to capture the seasonal variability in spectral response, providing a temporally robust input for classification.

Table 3. Feature space considered in the classification of the Cerrado biome in MapBiomas Collection 10.0. The column “Statistic” refers to the set of per-pixel statistical reducers applied to each variable within the annual temporal window (April–September): a) Amplitude – range of pixel values; b) Median – annual median; c) Median_dry – seasonal median for dates below the first quartile of NDVI values (dry period); d) Median_wet – seasonal median for dates above the first quartile of NDVI values (wet period); e) Standard deviation – annual variation; f) Minimum – 5th percentile of pixel values within the temporal window; g) Maximum – 95th percentile of pixel values within the temporal window.

Type	Name	Formula / Description	Statistics	Reference
Landsat Band	Blue, Green, Red, NIR, SWIR1, SWIR2	Original reflectance bands	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	USGS
	NDVI Normalized Difference Vegetation Index	$(\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red})$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Rouse et al., 1974
Spectral Index	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, Minimum (P5), Maximum (P95), StdDev, Amplitude	Parente et al., 2018
	GCVI Green Chlorophyll Vegetation Index	$(\text{NIR} / \text{Green} - 1)$	Median, Median_dry, Median_wet,	Burke et al., 2017

Type	Name	Formula / Description	Statistics	Reference
			Minimum (P5), Maximum (P95), StdDev, Amplitude	
	PRI Photochemical Reflectance Index	$(\text{Blue} - \text{Green}) / (\text{Blue} + \text{Green})$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Gamon et al., 1992
	MNDWI Modified Normalized Difference Water Index	$(\text{Green} - \text{SWIR1}) / (\text{Green} + \text{SWIR1})$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Xu et al., 2006
	CAI Cellulose Absorption Index	$\text{SWIR2} / \text{SWIR1}$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Nagler et al., 2003
	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, Minimum (P5), Maximum (P95), StdDev, Amplitude	Qi et al., 1994
	GCVI Green Chlorophyll Vegetation Index	$(\text{NIR} / \text{Green} - 1)$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Burke et al., 2017
	GRND Green Normalized Difference Vegetation Index	$\text{Green} / \text{Red}$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Gitelson et al., 1996
	MSI Moisture Stress Index	$\text{SWIR1} / \text{NIR}$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	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, Minimum (P5),	Gitelson et al., 2003

Type	Name	Formula / Description	Statistics	Reference
			Maximum (P95), StdDev, Amplitude	
	GNDVI Green Normalized Difference Vegetation Index	$(\text{NIR} - \text{Green}) / (\text{NIR} + \text{Green})$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Gitelson et al., 1996
	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
	GV Green Vegetation Fraction	Fractional abundance of green vegetation within the pixel	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Souza et al., 2005
	GVS Green Vegetation Shade Fraction	$\text{GV} / (\text{GV} + \text{NPV} + \text{Soil} + \text{Cloud})$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Housman et al., 2018
	NDFI Normalized Difference Fraction Index	$(\text{GVS} - (\text{NPV} + \text{Soil})) / (\text{GVS} + (\text{NPV} + \text{Soil}))$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Souza et al., 2005
Fraction	NPV Non-photosynthetic Vegetation Fraction	Fractional abundance of non-photosynthetic vegetation within the pixel	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Souza et al., 2005
	SEFI Savanna Ecosystem Fraction Index	$(\text{GV} + \text{NPV}_S - \text{Soil}) / (\text{GV} + \text{NPV}_S + \text{Soil})$	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Alencar et al., 2020
	Shade Fraction	$100 - (\text{GV} + \text{NPV} + \text{Soil} + \text{Cloud})$	Median	Housman et al., 2018
	Soil Fraction	Fractional abundance of soil within the pixel	Median, Median_dry, Median_wet, Minimum (P5),	Souza et al., 2005

Type	Name	Formula / Description	Statistics	Reference
	WEFI Wetland Ecosystem Fraction Index	$((GV + NPV) - (Soil + Shade)) / ((GV + NPV) + (Soil + Shade))$	Maximum (P95), StdDev, Amplitude Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Rosa, 2020
Other	Green Band Entropy	Entropy over the green_median band computed with a 5-pixel square kernel	Median	Adapted from GLCM texture

In addition to these annual variables, a set of static predictors has been incorporated since Collection 7.0 to provide environmental and spatial context to the classification model. This set includes HAND (Height Above Nearest Drainage) and fire history metrics derived from MapBiomass Fire Collection 3. Spatial coordinates (latitude and longitude) were also included to reduce the influence of training samples from one region on other areas. A key innovation in Collection 10.0 was the incorporation of geomorphometric variables, which aimed to support the discrimination of native vegetation types in areas characterized by more complex topography. Derived from digital elevation models, these variables include terrain morphology descriptors such as slope, aspect, curvature, and other derivatives that reflect geomorphological processes. The full list of non-annual variables is presented in Table 4.

Table 4. Static (non-annual) variables used in the classification process of MapBiomass Collection 10.0 for the Cerrado biome. "Identity" in the statistics column indicates that the variable is used directly, without temporal reduction. All geomorphometric variables are from the Geomorpho 90m dataset (Amattulli et al., 2020).

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
Time Since the Last Fire	Current year - Year of the last fire	Identity	Alencar et al., 2022

Name	Formula / Description	Statistics	Reference
Height Above the Nearest Drainage	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 data to represent east-facing slopes	Identity	Geomorpho 90m Amatulli et al., 2020
Northness	Derived from the aspect data to represent north-facing slopes	Identity	Geomorpho 90m Amatulli et al., 2020
Longitudinal Curvature	Second derivative of elevation in the x-direction	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 primarily derived from stable pixels identified in MapBiomas Collection 9.0 and reclassified according to the Cerrado classification scheme. Pixels that maintained the same class throughout the 1985–2023 time series were selected and further refined using additional validation sources. To improve reliability, classes with high uncertainty or transitional characteristics, such as “Other Non-Vegetated Areas” (25) and “Mosaic of Uses” (21), were excluded. These included deforestation alerts from PRODES and MapBiomas Alerta, as well as regional reference maps from state agencies (e.g., São Paulo, Tocantins, Goiás, Distrito Federal). To ensure spatial consistency, only homogeneous patches equal to or larger than 1 hectare (~11 Landsat pixels) were retained in the final training mask. A canopy height filter based on GEDI data was applied to exclude class-specific outliers. The outlier removal procedure followed the same logic adopted in Collection 7.0, which resulted in a documented accuracy improvement of +0.9%. The height-based filtering criteria were defined as follows:

- Forest Formation (3, including Mangrove and Flooded Forest): height < 4 m
- Savanna Formation (4): height < 2 m or > 8 m

- Wetland (11): height > 15 m
- Grassland Formation (12): height > 6 m
- Pasture (15): height > 8 m
- Agriculture (18, including temporary and perennial crops): height > 7 m
- Non-Vegetated Areas (25, including urban areas, mining, and beach): height > 0 m
- Water (33): height > 0 m

A stratified random training sampling strategy was employed to ensure proportional representation of all LULC classes across different regions. Sampling was based on the 2005 map from Collection 9.0 to estimate class areas per classification region. A total of 4,800 samples were allocated per region, with at least 480 samples per class to ensure representation of less frequent classes. For the water class, a minimum threshold of 240 samples was applied to reduce overestimation. This approach aimed to improve classification accuracy and ensure that underrepresented classes were adequately accounted for during the classification procedure.

Classification was performed regionally for each year using the Random Forest algorithm implemented in Google Earth Engine (GEE) via the `ee.Classifier.smileRandomForest` function. Based on the results of previous collections, the number of decision trees (`ntrees`) was set to 300 for all regions, and the number of variables per split was set to $\sqrt{\text{number of bands}}$ (`mtry`). The model operated in multiprobability mode, generating a probability distribution for each class. The final land use land cover class for each pixel was assigned based on the class with the highest probability, increasing the model's robustness to spectral confusion and uncertainty.

5. GENERAL MAP POST-CLASSIFICATION

Given the pixel-based classification methodology and the annual processing of the time series (1985–2024), a structured post-classification filtering framework was applied to improve 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 typically associated with per-pixel classifiers applied to long temporal series. The post-processing framework included a sequence of filters: temporal gap-filling, transition incidence, temporal consistency, and spatial coherence. Each filter was designed to address specific classification limitations and, together, they contributed to the overall quality and reliability of Collection 10.0.

5.1. Temporal Gap-Fill Filter

The gap-filling filter was designed to address missing data, commonly caused by cloud and shadow contamination or image availability gaps, by propagating valid classifications across the temporal dimension. The method implemented a bidirectional temporal search. First, the algorithm conducted a forward fill, replacing each no-data pixel with the nearest valid classification from subsequent years. In a second pass, a backward fill was applied to capture remaining gaps, 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 1985–2024 period.

5.2. Incidence filter

To address excessive class transitions over time, particularly oscillations between natural and anthropogenic cover types, an incidence filter was applied. All classes were grouped into three broad categories: (i) Natural (Forest Formation, Savanna Formation, Grassland Formation, Wetland), (ii) Anthropogenic (Pasture, Agriculture, Non-Vegetated Areas), and (iii) Other (Water and Not Observed). For each pixel, the number of transitions between the Natural and Anthropogenic groups was computed. Pixels with more than ten such transitions and fewer than seven connected similar pixels were classified as unstable and likely influenced by edge effects or spectral mixing. In these cases, the entire pixel time series was replaced by its most frequent class to simplify the trajectory. A similar correction was applied to core pixels (defined as those with more than seven similar neighbors) when transitions exceeded fourteen (approximately one-third of the time series). This filtering strategy helped distinguish edge-related spectral noise from legitimate land cover dynamics within core patches of vegetation.

5.3. Herbaceous Sandbank Vegetation

The Herbaceous Sandbank Vegetation (Restinga Herbácea, code: 50) was not included in the initial Random Forest classification due to the lack of training samples and its geographically restricted occurrence along the Brazilian coast. This class was mapped in a post-classification step using a targeted rule-based approach based on spectral, topographic, and edaphic criteria. The primary spectral criterion was the Soil Adjusted Vegetation Index (SAVI), derived from annual Landsat mosaics (1985–2024). SAVI values between 13,000 and 14,500 were empirically identified as optimal for distinguishing this vegetation type from other native formations, based on comparisons between the SAVI

distributions for Herbaceous Sandbank Vegetation and other NV classes (forest, savanna, wetland, and grassland) (Figure 6). Although some overlap with the savanna was observed, the Herbaceous Sandbank class exhibited a more distinct SAVI range overall. To reduce potential confusion, the filter was restricted to areas mapped as Quartzarenic Neosols (Entisols) by IBGE (Brazilian Institute of Geography and Statistics), and only areas with HAND values below 3.5 meters were used to represent the low-elevation, poorly drained environments characteristic of this vegetation type. The filter was applied only to pixels originally classified as Savanna, Grassland, or Wetland. It is important to note that this is a BETA class, introduced experimentally in Collection 10.0 to assess its feasibility and potential for future integration into the main classification workflow. The methodology and spatial delineation will be refined in future MapBiomass Cerrado collections, based on users' feedback and improved training data.

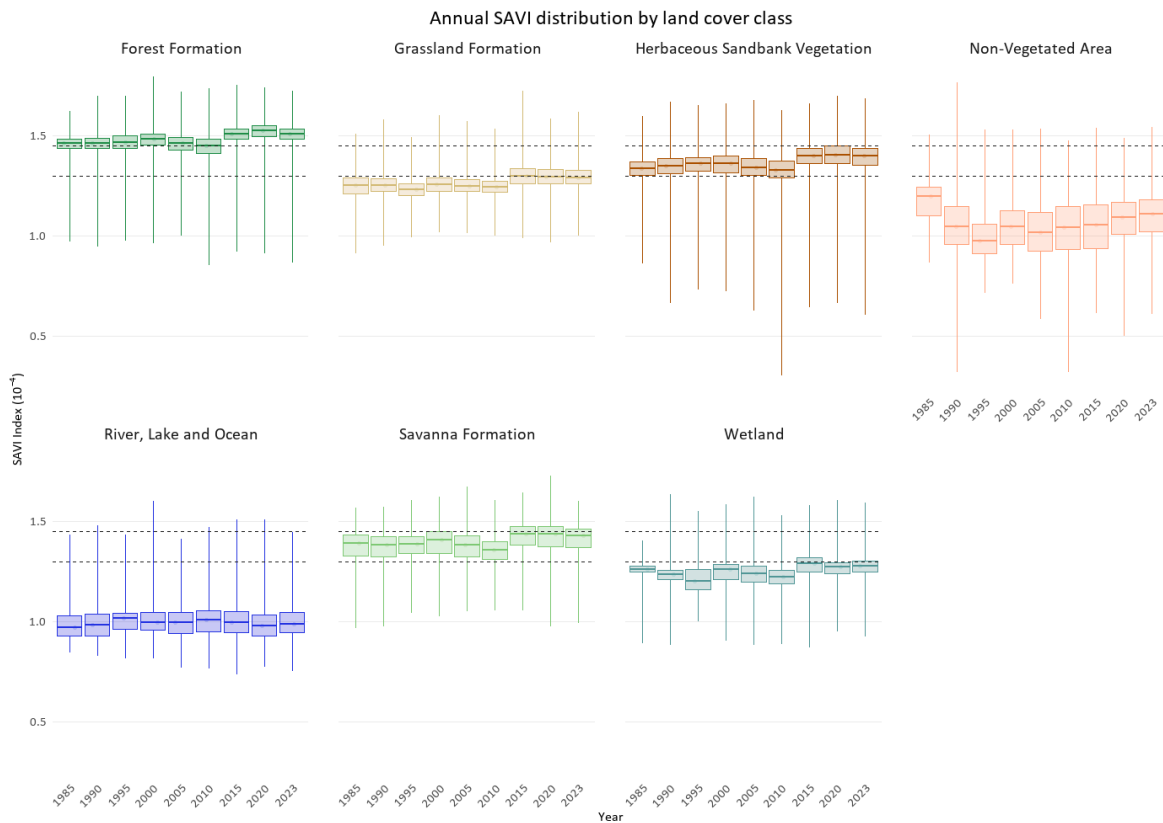


Figure 6. Annual distribution of the Soil Adjusted Vegetation Index (SAVI $\times 10^{-4}$) for different land cover classes in the Cerrado biome between 1985 and 2023. Each boxplot represents the temporal variability of SAVI values for a given year and native vegetation class. Dashed lines indicate thresholds used to guide class discrimination in post-classification filters.

5.4. Frequency

The frequency filter was applied to pixels classified as native vegetation in at least 90% of the 1985–2024 time series. Its objective was to improve temporal consistency and reduce uncertainties caused by intermittent misclassifications. For each pixel, the frequency of assignment to native vegetation classes was calculated, and specific thresholds were applied to confirm class stability. Pixels that met the 90% native vegetation criterion were further evaluated using class-specific frequency thresholds. Forest Formation was confirmed when present in $\geq 70\%$ of the years; Wetland and Herbaceous Sandbank when $\geq 60\%$; Grassland Formation when $> 50\%$; and Savanna Formation when $> 40\%$. It is important to note that the frequency filter was particularly effective in mitigating edge noise and inconsistencies during the initial and final years of the time series, which are more susceptible to image availability and cloud effects.

5.5. Temporal

The temporal filter implemented in Collection 10.0 is a key post-classification step designed to reduce spurious transitions and reinforce the temporal logic of land cover dynamics. It addresses short-term inconsistencies and stabilizes both the beginning and the end of the 1985–2024 classification series. In this step, Agriculture (18) and Pasture (15) were reclassified to “Mosaic of Uses” (21), consolidating anthropic land use as a unique class. The process then follows a series of sequential operations:

- 5-year and 4-year window filtering: The filter inspects pixels from 1986 to 2021 (5-year window) and 1986 to 2022 (4-year window) to correct any values that have a particular class in the previous year (year -1), change in the current year, and return to the original class in the most recent year (year +2 or +3). The correction follows the priority order: Savanna Formation (4), Wetland (11), Forest Formation (3), Grassland Formation (12), Herbaceous Sandbank Vegetation (50), Mosaic of Uses (21), Other Non-Vegetated Areas (25), and River, Lake, and Ocean (33).
- 3-year window filtering: This complementary step applies a 3-year moving window (1986–2023) to correct isolated inconsistencies, where the class in a year differs from both year-1 and year+1. The same class priority order is applied to ensure temporal coherence across the time series.
- Correction of the last year (2024): Two specific rules were applied to ensure the reliability of the most recent year: a) Pixels not classified as Mosaic of Uses (21) in 2024 but consistently classified as such in both 2022 and 2023 were revised to reflect the previous consistent classification, avoiding unverified regeneration; b) Pixels classified as Other Non-Vegetated (25) in 2024, but not in the two preceding years, were corrected by retaining their 2023 classification.

- Stabilization of the first year (1985): If a pixel was classified as native vegetation (Forest, Savanna, Wetland, Grassland, or Sandbank) in both 1986 and 1987, but not in 1985, the 1985 classification was adjusted to match the subsequent consistent class, ensuring temporal consistency from the start of the series.
- Removal of small patches of recent vegetation regrowth (2024): To avoid overestimating regeneration, only areas of native vegetation regrowth between 2023 and 2024 larger than 1 hectare (≥ 11 connected pixels) were retained. Smaller patches were considered noise and replaced by their 2023 classification.

5.6. No false regrowth filter

The false regrowth filter was originally developed in Collection 9.0 to correct specific cases of spurious vegetation regeneration that persisted despite the application of the temporal filter. In Collection 10.0, this procedure was refined and expanded to include additional rules and classes, further improving the temporal consistency of LULC classification. The filter consists of a set of targeted post-classification rules designed to address known sources of misclassification, including spectral confusion and model instability, affecting specific native vegetation and land use types.

- False forest regrowth in long-term silviculture areas: This rule addresses the spectral confusion between Forest Formation (3) and long-term silvicultural areas, which are mapped as Mosaic of Uses (21). Pixels continuously classified as class 21 for more than 15 consecutive years are interpreted as areas of stable anthropogenic use. If these pixels are later classified as native forest formation, this transition is assumed to be erroneous, likely caused by spectral resemblance between mature silviculture and forest formation. The classification is therefore reverted to class 21, preventing the artificial inflation of forest regeneration statistics in the last years.
- Correction of false wetland classifications in early years: In the early years of the time series (1985 and 1986), surface reflectance mosaics exhibit higher spectral saturation, even when using vegetation indices. This often results in spurious peaks for Wetlands (11), generating spurious classifications. To address this effect, 1987 was adopted as a temporal reference year, assuming improved data quality and stability from that point onward. Pixels classified as wetlands in either 1985 or 1986, but not in 1987 (or vice versa), were corrected to match the 1987 classification.
- Stabilization of early-term non-vegetated areas: Similar to the overestimation of wetlands observed in the early years, the classification of Non-Vegetated Areas (25) also presented inconsistencies at the beginning of the Landsat time series, particularly in pasture-dominated landscapes. These artifacts are attributed to the

spectral behavior of dry, exposed soils and sparse vegetation. In cases where class 25 appears in 1985 or 1986 but not in 1987 (or vice versa), the earlier classification was adjusted to match 1987, considered more stable..

- Correction of isolated occurrences of Herbaceous Sandbank Vegetation: Herbaceous Sandbank Vegetation (50) is a native formation often located within protected or low-access regions. Due to the limited anthropogenic pressure in these zones, this class is expected to present a relatively stable temporal behavior, with only subtle changes over the 1985–2024 period. However, because of its spectral similarity to other classes (e.g., savanna and grassland), it is susceptible to misclassification, particularly when it appears abruptly and is not persistent. To correct such cases, a rule was applied: if class 50 appears in a single year with no presence in previous years, the pixel is reverted to its prior classification.

5.7. Geomorphometric

This filter was developed to correct false Wetland (11) classifications in areas with steep slopes, where the presence of this class is geomorphologically unlikely. These misclassifications are common in regions with complex topography or shadow effects, which can cause spectral confusion in optical satellite imagery. To address this issue, a slope map derived from the MERIT Digital Elevation Model (DEM) was employed. Slope values were initially calculated in degrees and converted to percent. A slope threshold was defined to represent the upper limit for terrain expected to support wetlands in the Cerrado biome. For each year from 1985 to 2024, pixels classified as Wetland and located on slopes equal to or greater than 9% were flagged for correction. These pixels were replaced by the most frequent land cover class within an 8-pixel radius, using a neighborhood filter.

5.8. Spatial filter

The spatial filter implemented in Collection 10.0 aims to improve classification accuracy by minimizing isolated errors, particularly along the edges of homogeneous pixel groups. This filter identifies spatially contiguous groups of pixels that share the same class. A minimum mapping unit of eight connected pixels (~0.72 hectares) was established. In practice, this means that a pixel must be connected to at least seven neighboring pixels of the same class to retain its classification. Pixels not meet this criterion are considered isolated and are reclassified using neighborhood-based rules, typically by assigning the most frequent surrounding class. This procedure is essential for suppressing classification noise and removing small, fragmented areas, thereby improving the spatial coherence of the final maps.

6. ROCKY OUTCROP CLASSIFICATION

The classification of rocky outcrops in the Cerrado biome has undergone successive improvements over MapBiomas Collections, both in methodological refinement and in the conceptual definition of the class. Collection 7.0 introduced a beta version of the rocky outcrop class using a stepwise classification approach. This initial procedure involved an interpreter-based sample collection, followed by a two-stage classification process. A preliminary map was first generated and then used as a reference to train a second classification model. In Collection 8.0, the same stepwise logic was maintained; however, the spatial scope of the rocky outcrop class was significantly expanded, correcting the overly restricted extent observed in Collection 7.0.

Collection 9.0 introduced a major methodological advancement by redesigning the entire classification workflow. A direct classification strategy using Random Forest was implemented, enabling the mapping of rocky outcrops across the entire Cerrado biome with improved visual accuracy and consistency. The most substantial changes occurred in Collection 10.0, which introduced both conceptual and methodological refinements. The class definition was revised: whereas in earlier collections included rupestrian vegetation within the rocky outcrop class, Collection 10.0 redefined it to include only exposed rock. This revised definition allows for a more objective and geologically consistent representation of rocky outcrops. Methodologically, the classification was further refined by incorporating additional geomorphometric variables and a new spectral index, enhancing the discrimination of exposed rocky surfaces..

To avoid overestimation and ensure the class's specificity, the rocky outcrop classification is performed separately from the main land use and land cover map. This independent approach enables the use of tailored mapping criteria, better suited for identifying the distinct physical and spectral features characteristics of rocky outcrops in the Cerrado. This class typically includes stable geological formations with clear indicators of sedimentary, igneous, or metamorphic origins (Figure 7). The detailed classification workflow is presented in Figure 8. The subsequent sections describe the methodological steps in detail, including the feature space (6.1), training samples, classification algorithm, and parameters (6.2), and the post-processing filters (6.3).

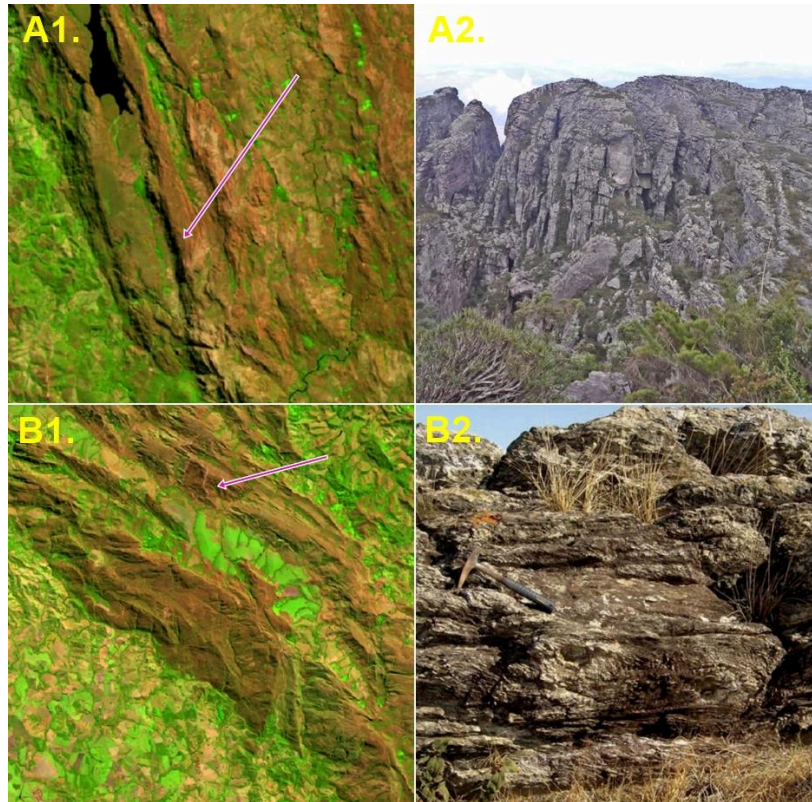


Figure 7. Example of landscapes mapped as Rocky Outcrop in the Collection 9.0. A) “Serra do Espinhaço”: A1) Landsat false-color composition (SWIR1-NIR-Red) for the year 2021. The pink arrow indicates the approximate location of the field photograph; A2) Field photograph (credits to TMbux). B) “Serra da Canastra”: B1) Landsat false-color composition (SWIR1-NIR-Red) for the year 2021. The pink arrow indicates the approximate location of the field photograph; B2) Field photograph (credits to Mario L.S.C. Chaves).

6.1. Feature space

The feature space used for rocky outcrop classification in Collection 10.0 builds upon the 230 predictor variables previously applied in the general land use and land cover mapping (Section 4.2 and Table 3). To address the specific demands of rocky outcrop mapping, additional variables were incorporated to improve discrimination and reduce misclassification with spectrally or structurally similar classes. These additions include an extended set of geomorphometric variables derived from the MERIT Digital Elevation Model (DEM) and the Geomorpho90m dataset. These variables help characterize terrain form and landscape position, which are particularly relevant for identifying rocky outcrops, as these formations typically occur in steep, rugged, and elevated environments.

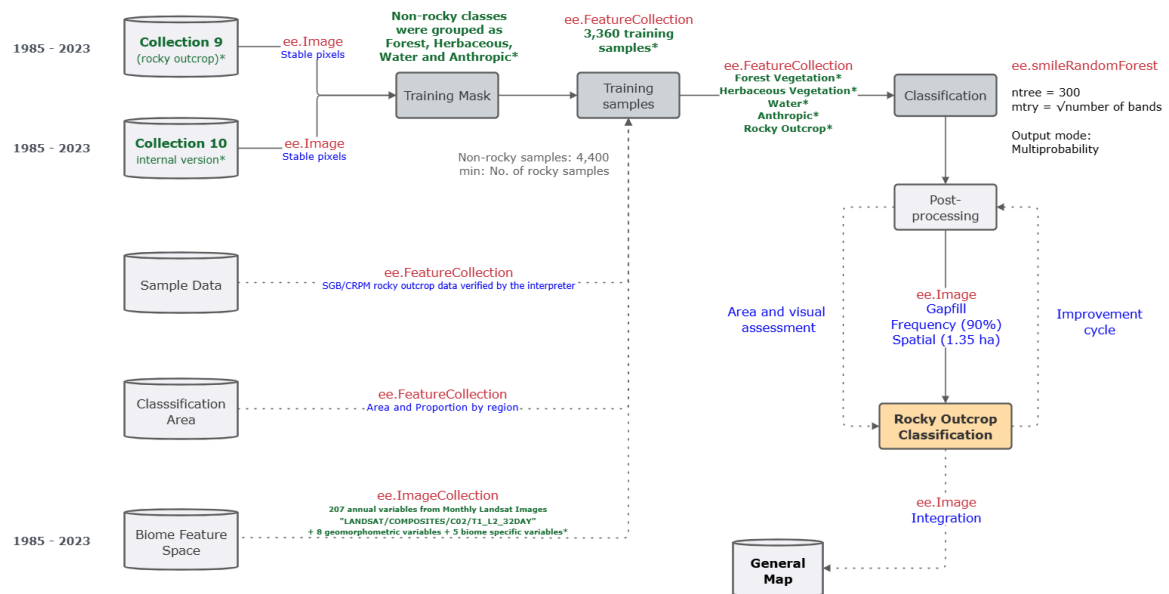


Figure 8. Each gray geometry (cylinders for databases and rectangles for processes) represents a key step in the classification schema, with the respective name inside. The gray text near databases and processes offers a description of the step, while the green text highlights the main innovations in Collection 9.0. 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.

Additionally, a set of temporal metrics was calculated using a three-year moving window to assess vegetation variability over time. These metrics include the NDVI amplitude between wet and dry seasons, the variance of the 25th percentile of NDVI, and the variance of the median NBR. These indicators capture patterns of spectral stability and disturbance, supporting the identification of exposed rock surfaces, which typically exhibit no vegetative cover and minimal seasonal variation. An additional spectral index, the Topsoil Grain Size Index (TGSi), was included to distinguish rocky substrates from bare soils or degraded vegetation. A full description of variables used exclusively for rocky outcrop mapping is provided in Table 5.

Table 5. Complementary bands added to the Cerrado rocky outcrop classification feature space. The column “Statistic” refers to the set of per-pixel statistical reducers applied to each variable within the annual temporal window (April–September): a) Amplitude – range of pixel values; b) Median – annual median; c) Median_dry – seasonal median for dates below the first quartile of NDVI values (dry period); d) Median_wet – seasonal median for dates above the first quartile of NDVI values (wet period); e) Standard deviation – annual variation; f) Minimum – 5th percentile of

pixel values within the temporal window; g) Maximum – 95th percentile of pixel values within the temporal window. h) Identity – the variable is used directly without temporal reduction.

Type	Name	Statistics	Reference
Terrain	Elevation	Identity	MERIT DEM Yamazaki et al., 2017
	Aspect	Identity	Geomorpho 90m Amatulli et al., 2020
	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
	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
	TRI Topographic Ruggedness Index	Identity	Geomorpho 90m Amatulli et al., 2020
	CTI Compound Topographic Index	Identity	Geomorpho 90m Amatulli et al., 2020
Spectral Index	TGSI Topsoil Grain Size Index (Red – Blue) / (Red + Blue + Green)	Median, Median_dry, Median_wet, Minimum (P5), Maximum (P95), StdDev, Amplitude	Xiao et al., 2006
Temporal data	NDVI amplitude (3-year)	Range	Celebrezze et al., 2025
	NDVI P25 variance (3-year)	Variance	Celebrezze et al., 2025
	NBR median variance (3-year)	Variance	Celebrezze et al., 2025

6.2. Training samples, classification algorithm, and parameters

A total of 3,360 training samples were used across the entire classification area (Figure 9), combining visually interpreted samples by specialists and additional samples provided by the Brazilian Geological Service (SGB/CPRM). All samples were thoroughly reviewed to ensure alignment with the updated conceptual definition adopted in Collection 10.0. The training mask for general land cover categories was derived from the stable pixel map generated in the current collection (10.0), based on temporal consistency from 1985 to 2023. This mask included only pixels with stable classifications throughout the time series and was grouped into broader thematic classes: Forest (Forest and Savanna formations), Herbaceous vegetation (Grassland), Water (Wetland and Water bodies), and Anthropogenic (Mosaic of Uses). In contrast, reference pixels for the rocky outcrop class were derived from stable areas mapped in Collection 9.0, as this class was not included in the main classification workflow.

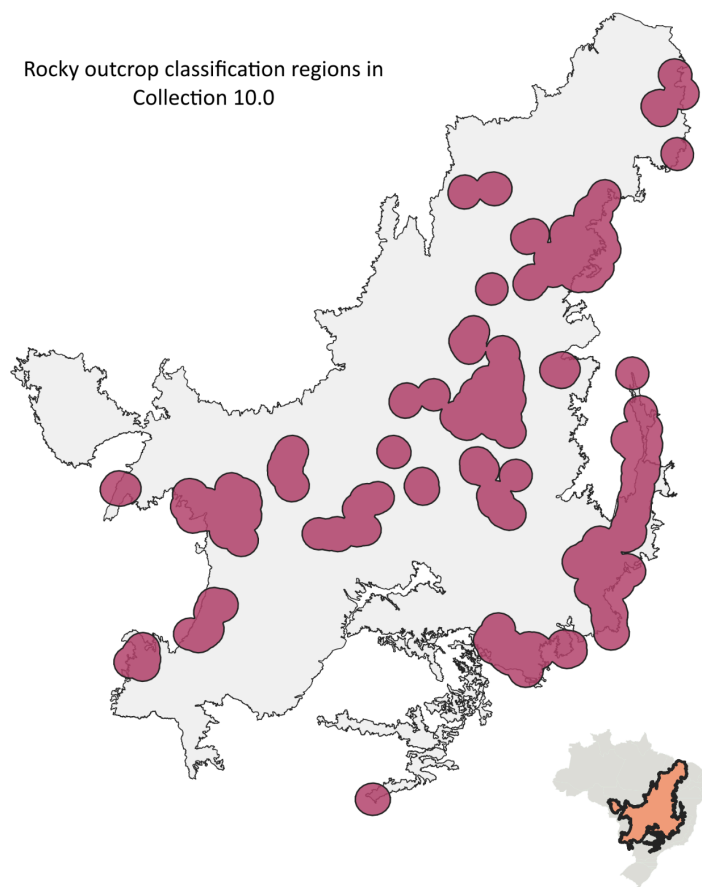


Figure 9. The rocky outcrop classification area used in collection 9.0. Highlighted in orange is the location of the Cerrado biome in Brazilian territory.

Training points distribution followed a stratified sampling approach, with the number of samples per class proportional to its area within the classification extent. This ensured balanced class representation, with up to 4,400 samples for major classes and a minimum equal to the number of rocky outcrop samples. Classification was conducted for each year using the Random Forest algorithm implemented in GEE via the `ee.Classifier.smileRandomForest` function. Based on previous collections, the number of decision trees (ntrees) was set to 300 for all regions, and the number of variables per split was set to $\sqrt{\text{number of bands}}$ (mtry). The model operated in multiprobability mode, generating a probability distribution across all classes. The final class for each pixel was assigned based on the highest probability, enhancing model robustness to spectral confusion and classification uncertainty.

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 filters to improve classification consistency. Three main filters were applied: gap-filling, frequency, and spatial smoothing.




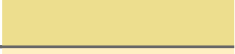


- The gap-fill filter addressed inconsistencies and missing classifications across the time series. Unlike conventional unidirectional approaches, it operated bidirectionally (both forward and backward in time), filling undefined pixels using the classification of subsequent and preceding years. This ensured temporal continuity in cases where rocky outcrop presence was stable but temporarily unclassified due to spectral noise or image limitations.
- The frequency filter reinforced 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 90% of the years within the observation period. This threshold effectively filtered out areas associated with rupestrian vegetation, which exhibits greater spectral variability and temporal dynamics.
- Finally, a spatial filter was used to remove isolated pixels or small misclassified patches inconsistent with the typical spatial pattern of rocky outcrops. The filter removed connected components smaller than 15 pixels (1.35 hectares), using an 8-neighbor connectivity criterion.

7. INTEGRATION

The integration process is a critical step to ensure consistency and completeness in the annual land use and land cover maps. It is carried in two sequential stages. The first stage involves an internal integration within the Cerrado biome, in which the rocky outcrop classification is overlaid onto the native vegetation map generated in the main classification workflow. The second stage integrates cross-cutting themes developed by the MapBiomias initiative. These themes include additional layers such as urban infrastructure, agriculture, mining, and others. To harmonize the thematic layers with the biome-level classifications, a set of predefined prevalence rules is applied. These rules establish which classes take precedence in case of overlap, ensuring a standardized decision logic across biomes and years. The specific prevalence rules used in this integration process are detailed in Table 6.

Table 6. General prevalence rules - MapBiomias Collection 10.0

Class	Pixel value	Prevalence order	Color
Photovoltaic Power Plant	75	1	
Mining	30	2	
Beach, Dune, and Sand Spot	23	3	
Mangrove	5	4	
Aquaculture	31	5	
Hypersaline Tidal Flat	32	6	
Urban Infrastructure	24	7	
Forest Plantation	9	8	
Rocky Outcrop	29	9	
Sugar Cane	20	10	
Soybean	39	11	
Rice	40	12	
Cotton	62	13	
Other Temporary Crops	41	14	
Coffee	46	15	
Citrus	47	16	
Other Perennial Crops	48	17	
Herbaceous Sandbank Vegetation	50	18	
River, Lake, and Ocean	33	19	
Forest Formation	3	20	

Class	Pixel value	Prevalence order	Color
Savanna Formation	4	21	
Wetland	11	22	
Grassland Formation	12	23	
Pasture	15	24	
Mosaic of Uses	21	25	
Other Non-Vegetated Areas	25	26	

Notably, exceptions to this general prevalence rule apply within areas:

- Within protected areas, native vegetation classes (Forest Formation, Savanna Formation, Wetland, and Grassland — classes 3, 4, 11, and 12) take precedence over specific crop classes, including Cotton (62), Citrus (47), and Coffee (46).
- Similarly, in the case of pasture (15) within protected areas, native vegetation classes (3, 4, 11, and 12) are preserved as the prevailing classification.
- Outside protected areas, however, pasture (15) takes precedence over Savanna (4), Wetland (11), and Grassland (12) classes.

8. ACCURACY METRICS

The accuracy analysis of Collection 10.0 was conducted using a dataset provided by LAPIG/UFG, consisting of 20,851 validation samples for the Cerrado, covering the period from 1985 to 2024. The samples were classified by experts with knowledge of Cerrado vegetation, ensuring reliable validation. The assessment included calculations of overall and per-class accuracy, as well as omission and commission errors, and quantity and allocation disagreements. These metrics were derived from confusion matrices comparing the reference dataset with sample pixels from the integrated Collection 10.0 map. Figures 10 to 12 show the overall accuracy, allocation error, and quantity error for Cerrado land use and land cover maps. At Level 1, Collection 10.0 achieved an overall accuracy of 87.6%, nearly the same as Collection 9.0 (87.6%). However, Collection 10.0 showed a slight reduction in allocation error and a minor increase in quantity error compared to Collection 9.0. This suggests that Collection 10.0 improved spatial agreement even as class proportion deviations slightly rose. At Level 3, Collection 10.0 reached an overall accuracy of 80.4%, slightly higher than Collection 9.0 (80.2%). In this case, quantity disagreement decreased to 7.0%, while allocation disagreement increased slightly to 12.6%, indicating a small trade-off between spatial allocation and class proportion accuracy. These results position Collection 10.0 as the most refined MapBiomass product for the Cerrado, demonstrating significant improvements in spatial allocation consistency without sacrificing overall accuracy. All classification accuracy metrics are available at: <https://brasil.mapbiomas.org/analise-de-acuracia/>.

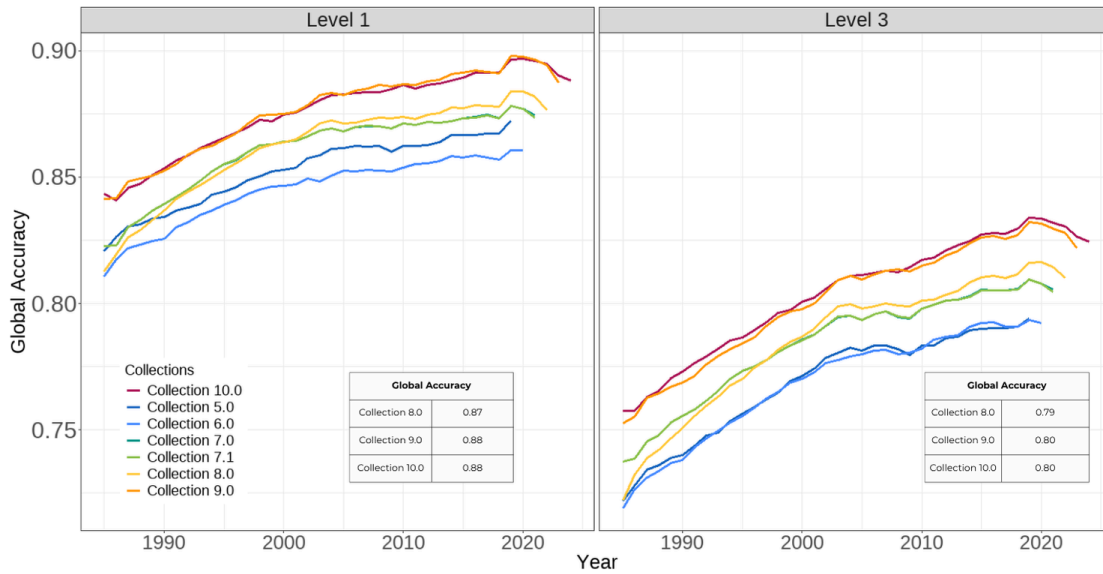


Figure 10. Global accuracy for the Cerrado biome at legend level 1 and level 3. The x-axis represents the years (from 1985 to 2024), while the y-axis represents the global accuracy value (1 = high accuracy). The colored lines indicate the accuracy per year of the current collection (10.0 - red line) and the previous collections (9.0, 8.0, 7.1, 6, and 5 - orange to blue lines). The overall average accuracies for the last three collections are indicated in the table.

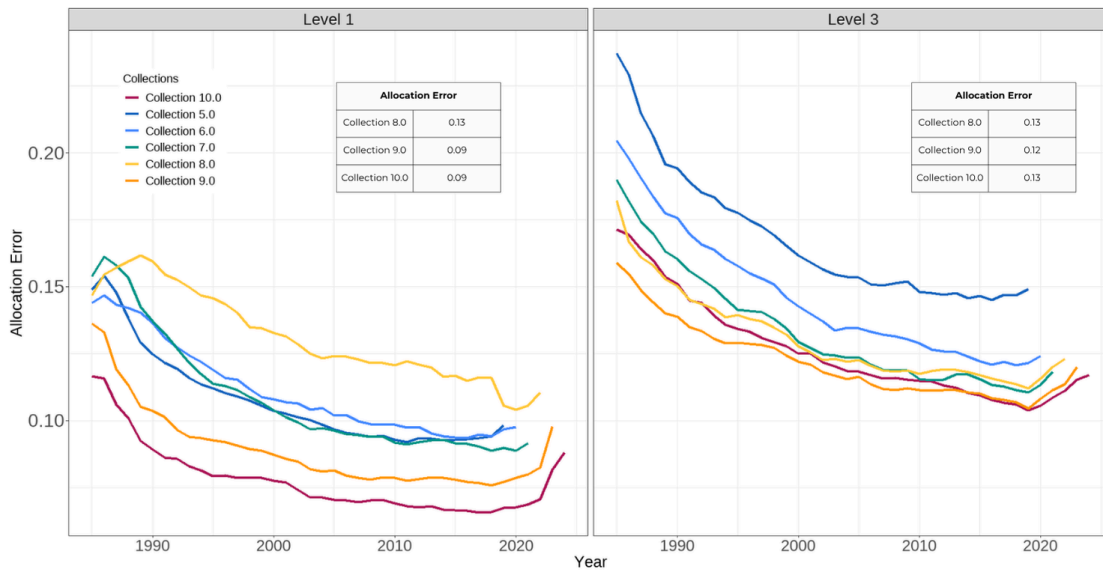


Figure 11. Allocation error for the Cerrado biome at legend level 1 and level 3. The x-axis represents the years (from 1985 to 2024), while the y-axis represents the allocation error value. The colored lines indicate the accuracy per year of the current collection (10.0 - red line) and the previous collections (9.0, 8.0, 7.1, 6, and 5 - orange to blue lines). The average allocation error for the last three collections is indicated in the table.



Figure 12. Quantity error for the Cerrado biome at legend level 1 and level 3. The x-axis represents the years (from 1985 to 2024), while the y-axis represents the quantity error value. The colored lines indicate the accuracy per year of the current collection (10.0 - red line) and the previous collections (9.0, 8.0, 7.1, 6, and 5 - orange to blue lines). The average quantity error for the last three collections is indicated in the table.

9. REFERENCES

- Alencar, A., Z. Shimbo, J., Lenti, F., Balzani Marques, C., Zimbres, B., Rosa, M., Arruda, V., Castro, I., Fernandes, Márcio Ribeiro, J. P., Varela, V., Alencar, I., Piontekowski, V., Ribeiro, V., M. C. Bustamante, M., Eyji Sano, E., & Barroso, M. (2020). Mapping Three Decades of Changes in the Brazilian Savanna Native Vegetation Using Landsat Data Processed in the Google Earth Engine Platform. *Remote Sensing*, 12(6), 924.
- Alencar, A.A.C.; Arruda, V.L.S.; Silva, W.V.d.; Conciani, D.E.; Costa, D.P.; Crusco, N.; Duverger, S.G.; Ferreira, N.C.; Franca-Rocha, W.; Hasenack, H.; Martenexen, L.F.M.; Piontekowski, V.J.; Ribeiro, N.V.; Rosa, E.R.; Rosa, M.R.; dos Santos, S.M.B.; Shimbo, J.Z.; Vélez-Martin, E. (2022) Long-Term Landsat-Based Monthly Burned Area Dataset for the Brazilian Biomes Using Deep Learning. *Remote Sens.* 14, 2510. <https://doi.org/10.3390/rs14112510>.
- 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.
- 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.

- Celebrezze, J. V., Alegbeleye, O. M., Glavich, D. A., Shipley, L. A., & Meddens, A. J. (2025). Classifying rocky land cover using random forest modeling: Lessons learned and potential applications in Washington, USA. *Remote Sensing*, 17(5), 915.
- 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., 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., 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.
- 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.
- Lang, N., Jetz, W., Schindler, K., & Wegner, J. D. (2022). A high-resolution canopy height model of the Earth. *arXiv preprint arXiv:2204.08322*.
- 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.
- 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>.