Cerrado - Appendix

MapBiomas 10m - Collection 3

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

April, 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 and implemented entirely within the Google Earth Engine (GEE) cloud computing environment. 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 five broad categories: Forest Formation, Savanna Formation, Grassland Formation, Wetland, and Rocky Outcrop.

Since its inception, the MapBiomas 10m Cerrado workflow has undergone continuous methodological refinement, incorporating both conceptual advances and technical improvements while preserving key methodological gains achieved in earlier MapBiomas 30m collections. It is important to note that all MapBiomas 10m collections are currently released as *beta* products, meaning that they are under active development and have not yet undergone a full, systematic accuracy assessment at the biome scale. Across all collections, the workflow follows a consistent structure, comprising: (i) definition of the feature space derived from remote sensing metrics; (ii) generation of reference training samples for algorithm calibration; (iii) application of post-classification spatial and temporal filters to reduce noise and ensure time-series consistency; and (iv) integration of the resulting maps with cross-cutting thematic 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 as *“Inventário Florestal do Estado de São Paulo” and “Base Temática Digital do Estado do Tocantins”*, as well as deforestation datasets from PRODES and PROBIO. The classification was performed using the Random Forest (RF) algorithm, with explicit hyperparameter tuning and feature space selection conducted independently for each of the 38 Cerrado classification regions. This regionalized strategy involved preliminary tuning experiments to identify the optimal combination of variable predictors and algorithm parameters for each region. Post-classification temporal and spatial filters were applied following the same logic used in 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 MapBiomas 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 MapBiomas 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 MapBiomas 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 Agriculture and Pasture, Other Non-vegetated Area, Water	First version of the Cerrado collection in the MapBiomas 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.

The current 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. The Satellite Embedding dataset provides a 64-dimensional embedding vector at 10m resolution for each pixel. As a result, Collection 3.0 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 MapBiomas 30m Collection 10.0. 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 3.0, the annual LULC classification for the Cerrado biome comprises ten classes, as defined in the official MapBiomas 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 primarily extracted from stable pixels identified in MapBiomas 10m Collection 2.0 (2016–2023), complemented by

thematic reference data and HAND-derived information. 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:** 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.
- **Integration with Cross-Cutting Themes:** Final annual maps were harmonized with cross-cutting thematic layers following predefined MapBiomas prevalence rules.

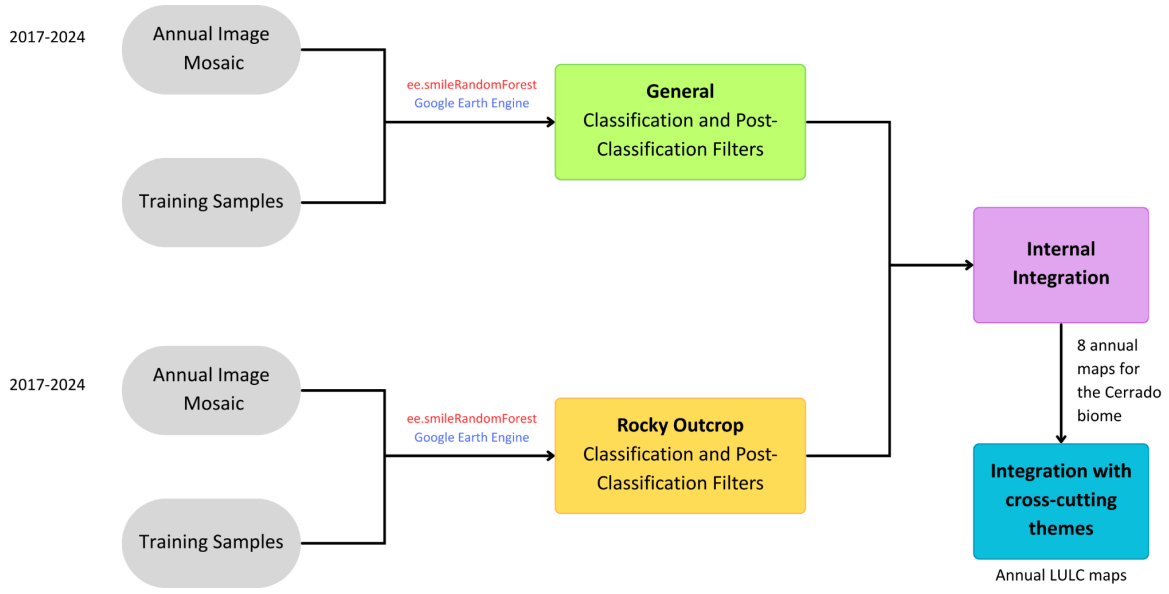
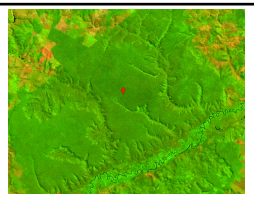
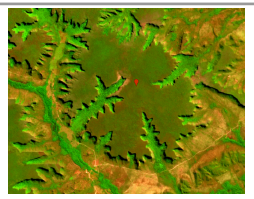
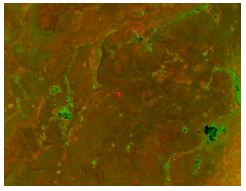
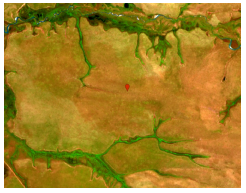
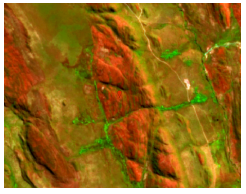
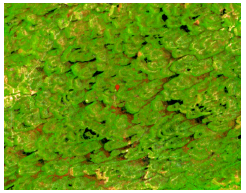
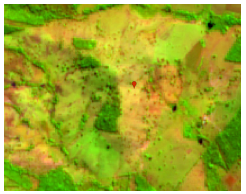
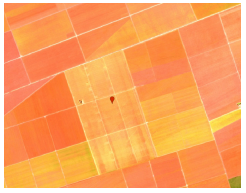
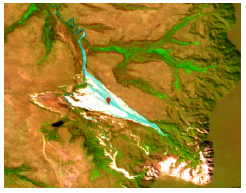
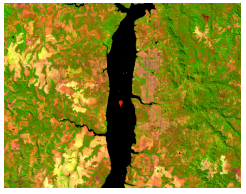


Figure 1. Workflow for the LULC classification of the Cerrado biome in Collection 3.0. Two parallel modules were applied: (i) the general classification, and (ii) a specific classification of Rocky Outcrop. Both used the Random Forest algorithm implemented in GEE. Outputs were integrated into two stages: first, the internal integration combined the Rocky Outcrop with the general map; second, the maps were integrated with cross-cutting themes.

Table 2. Land use land cover categories used for the Sentinel-2 mosaics classification for the Cerrado biome in MapBiomas Collection 3.0.

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 & Walter, 2008), as well as Semi-deciduous Seasonal Forests.
	Savanna Formation	4			Vegetation types with a distinct stratification of tree, shrub, and herbaceous strata. This includes different physiognomies of Cerrado *sensu stricto* (Dense, Typical, Sparse, and Rupestrian Savanna) (Ribeiro & Walter, 2008).
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			Open vegetation dominated by herbaceous species, with minimal or no tree cover. Includes Dirty, Clean, and Rupestrian Grasslands, as well as some savanna formations such as Rupestrian Cerrado (Ribeiro & Walter, 2008).
	Rocky Outcrop	29			Naturally exposed rocky surfaces, including monoliths, bedrock, and slabs with little or no soil cover and minimal vegetation. These features are typically associated with stable geological formations of sedimentary, igneous, or metamorphic origin.

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 trees and 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 the post-classification method. ** Agriculture and Pasture are merged into the Mosaic of Uses during the temporal filter

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 MapBiomas 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.

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 MapBiomas 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 combines (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.

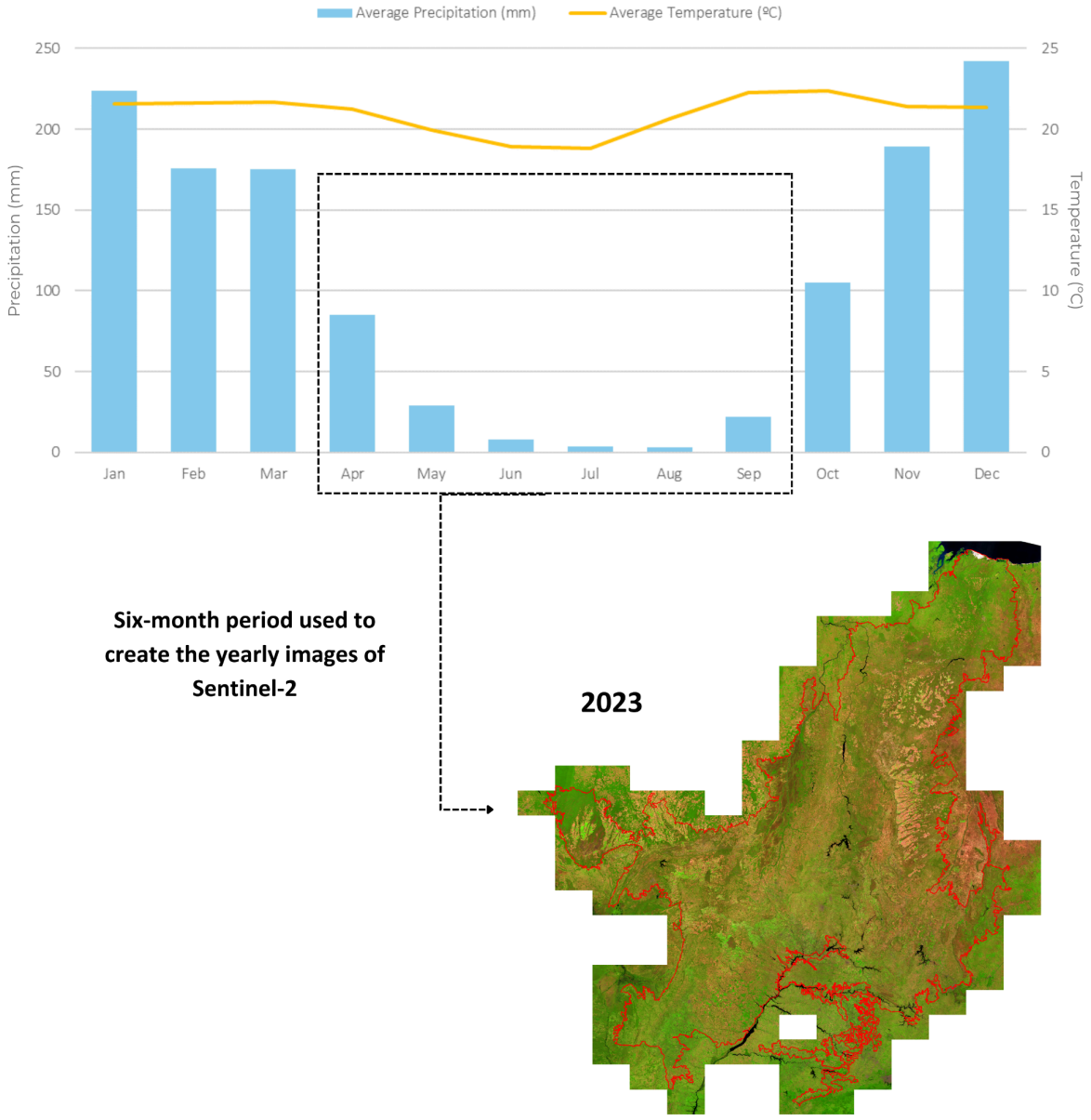


Figure 2. Time window adopted for the construction of annual classification mosaics in the Cerrado biome. Precipitation data represent monthly averages for the Cerrado region (Macena et al., 2008), and temperature data correspond to monthly averages for the Federal District (INMET).

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.

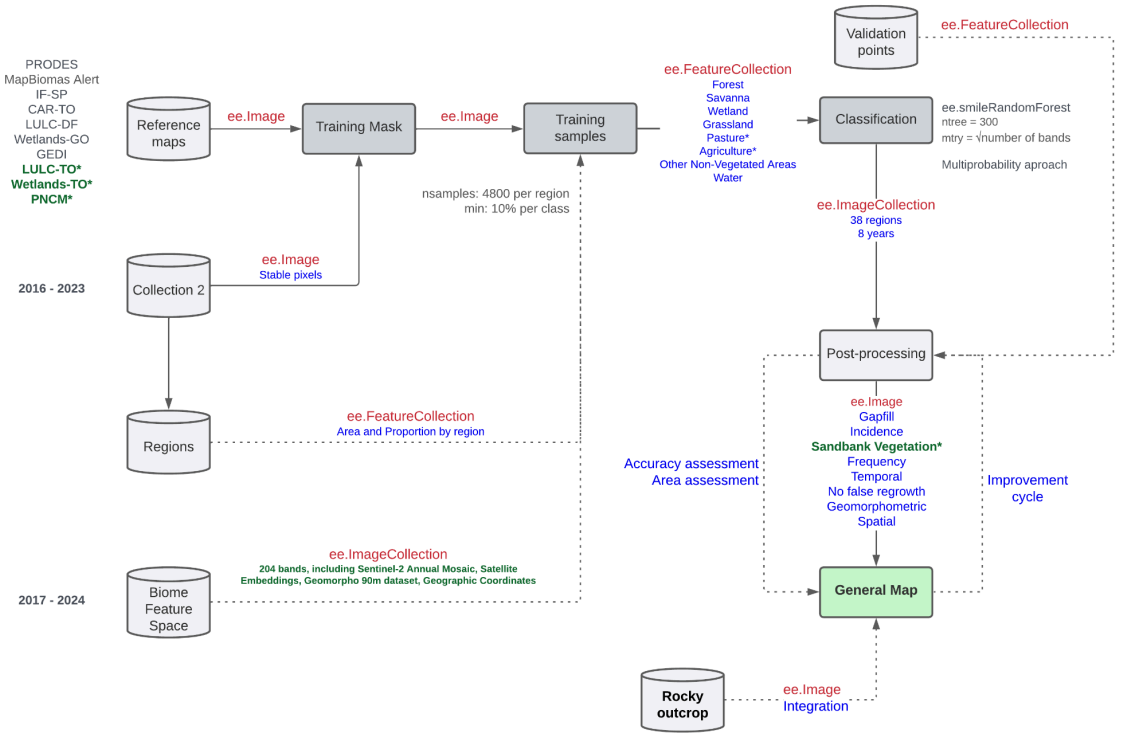


Figure 3. Each gray geometry (cylinders for databases and rectangles for processes) represents a key step in the classification schema, with the respective name inside. The gray text near databases and processes offers a description of the step, while the green text highlights the main innovations in Collection 3.0. 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 MapBiomas 10m Collection 3.0 correspond to the same regionalization scheme used since MapBiomas 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 land use and land cover 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.

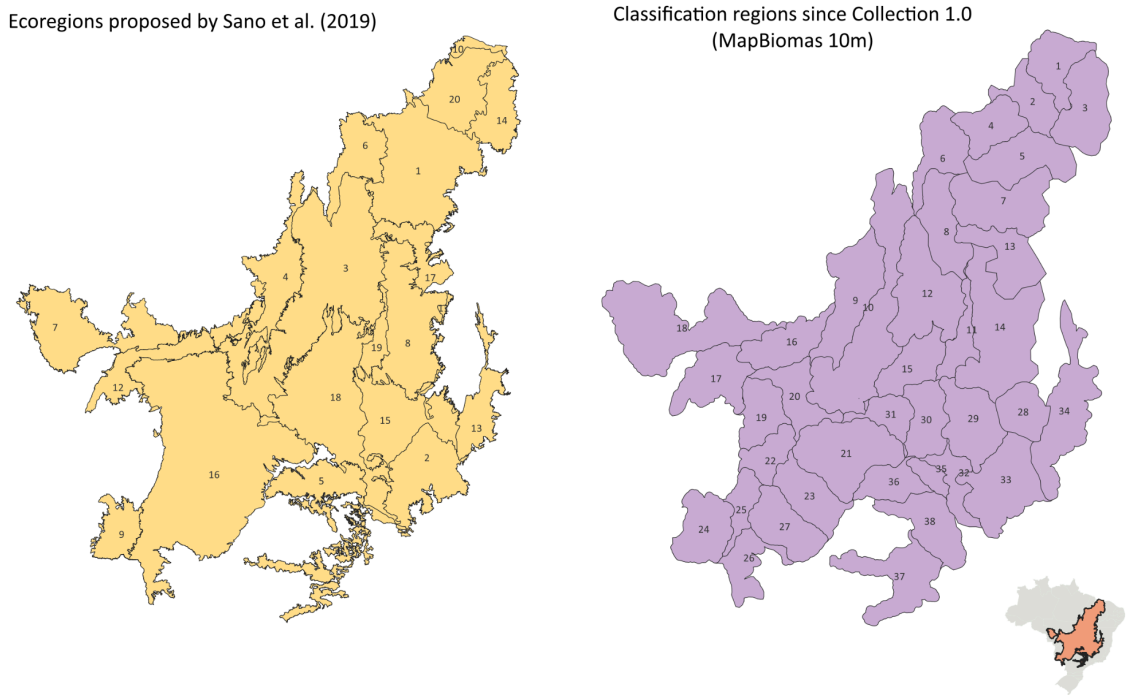


Figure 4. Classification regions, modified from Sano et al., 2019. Highlighted in orange is the location of the Cerrado biome in Brazilian territory.

4.2. Feature space

The feature space used for LULC classification in Collection 3.0 comprised a comprehensive set of 202 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 MapBiomas 10m Collection 3.0. The column "Statistic" refers to the set of per-pixel statistical reducers applied to each variable within the annual temporal window: a) 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) 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	(NIR - Red) / (NIR + Red)	Median, Median_dry, Median_wet, StdDev	Rouse et al., 1974
	EVI2 Enhanced Vegetation Index 2	2.5 × (NIR - Red) / (NIR + 2.4 × Red + 1)	Median, Median_dry, Median_wet, StdDev	Parente et al., 2018
	GCVI Green Chlorophyll Vegetation Index	(NIR / Green - 1)	Median, Median_dry, Median_wet, StdDev	Burke et al., 2017
	PRI Photochemical Reflectance Index	(Blue - Green) / (Blue + Green)	Median, Median_dry, Median_wet, StdDev	Gamon et al., 1992
	MNDWI Modified Normalized Difference Water Index	(Green − SWIR1) / (Green + SWIR1)	Median, Median_dry, Median_wet, StdDev	Xu et al., 2006
	CAI Cellulose Absorption Index	SWIR 2 / SWIR 1	Median, Median_dry, Median_wet, StdDev	Nagler et al., 2003
	MSAVI Modified Soil Adjusted Vegetation Index	$0.5 \times [2 \times NIR + 1 - \sqrt{(2 \times NIR + 1)^2 - 8 \times (NIR - Red)}]$	Median, Median_dry, Median_wet, StdDev	Qi et al., 1994

Type	Name	Formula / Description	Statistics	Reference
	GCVI Green Chlorophyll Vegetation Index	(NIR / Green - 1)	Median, Median_dry, Median_wet, StdDev	Burke et al., 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	(NIR − (Green − (Blue − Red))) / (NIR + (Green − (Blue − Red)))	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 2003
	GNDVI Green Normalized Difference Vegetation Index	(NIR − Green) / (NIR + Green)	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 1996
	NDVI Red-Edge	(RedEdge1 - Red) / (RedEdge1 + Red)	Median, Median_dry, Median_wet, StdDev	Potter et al., 2012
	VI 700 (NDCI) Vegetation Index 700	(RedEdge1 − RedEdge2) / (RedEdge1 + RedEdge2)	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 2002
	IRECI Inverted Red-Edge Chlorophyll Index	(RedEdge3 − Red) / (RedEdge1 + RedEdge2)	Median, Median_dry, Median_wet, StdDev	Guyot and Baret, 1988
	CIRE Chlorophyll Index Red-Edge	(NIR / Red Edge 1)	Median, Median_dry, Median_wet, StdDev	Gitelson et al., 2005
	TCARI Transformed Chlorophyll Absorption Reflectance Index	3 × [(RedEdge1 − Red) − 0.2 × (RedEdge1 − Green) × RedEdge1/Red]	Median, Median_dry, Median_wet, StdDev	Haboudane et al. (2002)
	SFDVI Spectral Feature Depth Vegetation Index	(Green + NIR) / 2) - (Red + RedEdge1) / 2)	Median, Median_dry, Median_wet, StdDev	Baptista, 2015
	NDRE Normalized Difference Red-Edge Index	(NIR − RedEdge) / (NIR + RedEdge)	Median, Median_dry,	Gitelson and Merzlyak (1994)

Type	Name	Formula / Description	Statistics	Reference
			Median_wet, StdDev	
	TGSI Topsoil Grain Size Index	(Red − Blue) / (Red + Blue + Green)	Median, Median_dry, Median_wet, StdDev	Xiao et al., 2006
	Hall Height	(−0.039 × Red − 0.011 × NIR − 0.026 × SWIR1 + 4.13)	Median, StdDev	Hall et al., 2006
	Hall Cover	(- Red × 0.017 - NIR × 0.007 - SWIR2 × 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 DEM (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 MapBiomas Collection 3.0 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 (ee.Image.pixelLonLat())	Identity	Geolocation function
Cosine of Longitude	cos(ee.Image.pixelLonLat() .select(['longitude']))	Identity	Geolocation function
Sine of Longitude	sen(ee.Image.pixelLonLat() .select(['longitude']))	Identity	Geolocation function
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

Name	Formula / Description	Statistics	Reference
			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 2.0, reclassified according to the Cerrado classification scheme. Pixels that maintained the same class throughout the 2016–2023 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 “Other Non-Vegetated Areas” (25) and “Mosaic of Uses” (21), were excluded. The refinement process incorporated multiple independent reference datasets, including deforestation alerts from PRODES and MapBiomas Alerta, and regional LULC maps from state agencies (e.g., São Paulo, Tocantins, Goiás, Distrito Federal) and zoning data from ICMBio (*Instituto Chico Mendes de Conservação da Biodiversidade*). To ensure spatial consistency, only homogeneous patches equal to or larger than 0.1 hectare (~11 Sentinel pixels) were retained in the final training mask. A canopy height filter based on GEDI data (Lang et al., 2022) was applied to exclude pixels with anomalous canopy heights within each class. The outlier removal procedure followed the same logic as in MapBiomas 30m Collection 7.0,

resulting in a documented accuracy improvement of +0.9%. The height-based filtering criteria were:

- 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 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 MapBiomas 10m Collection 2.0. 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 MapBiomas 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.

4.4. Variable importance and SHAP value

To support the interpretation of the classification results and better understand the contribution of different predictors, variable importance and SHAP (SHapley Additive exPlanations) analyses were performed based on the Random Forest models used in MapBiomas Collection 3.0. These analyses provide complementary perspectives on model

behavior: variable importance summarizes how frequently and effectively each predictor contributes to decision splits in the ensemble of trees, while SHAP values quantify the magnitude and direction of each variable's contribution to the final class prediction, enabling class-specific interpretation.

Variable importance was computed for each classification region and year and subsequently aggregated by calculating the mean importance of each predictor across all regions and the full temporal range (2017–2024). The results indicate that both the Sentinel-2 mosaic and the Satellite Embeddings dataset play a central role in the classification of the Cerrado biome. Among the most important predictors are elevation-related variables (MERIT DEM), spatial coordinates (latitude and longitude), spectral information from Sentinel-2 (particularly SWIR bands and water-related indices), and several embedding dimensions (variables labeled as A*) (Figure 5). Overall, the variable importance analysis highlights the strong contribution of embedding features to the global model structure, whereas class-specific influence is more clearly expressed by physically interpretable spectral variables, as revealed by the SHAP analysis

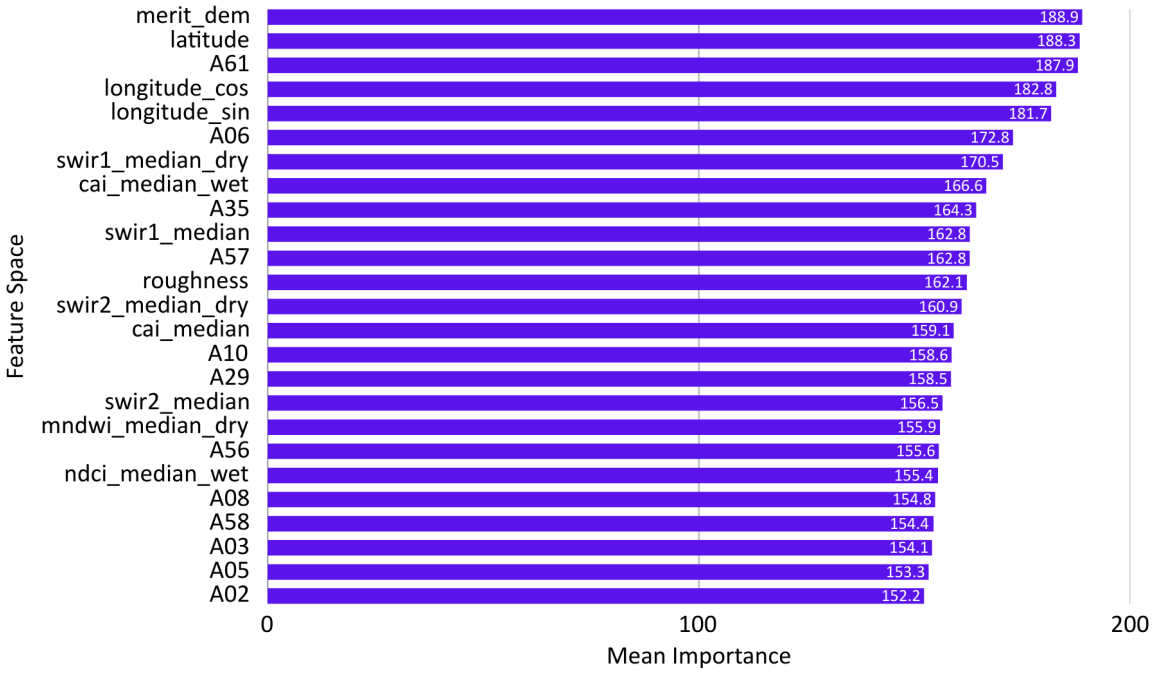


Figure 5. Mean variable importance ranking derived from the Random Forest model, aggregated across all 38 classification regions and the 2017–2024 period.

SHAP values were computed to explore how variables influence class-specific predictions (Lu et al., 2024). This analysis focuses on understanding how individual

variables affect the Random Forest output for each LULC class (Figure 6). The SHAP results reveal distinct variable importance patterns across classes, reflecting differences in biophysical controls and spectral behavior. For Forest and Savanna formations, red-edge–based indices, vegetation structural metrics, and embedding dimensions exhibit strong contributions, consistent with their sensitivity to canopy structure and chlorophyll content. Grassland and Wetland classes show greater influence from geomorphometric variables, moisture-sensitive indices, and specific embedding bands, reflecting topographic controls and hydrological conditions. SWIR bands and water indices primarily drive river, lake, and ocean class, while non-vegetated areas show a stronger contribution from embedding dimensions and spectral features associated with exposed surfaces.

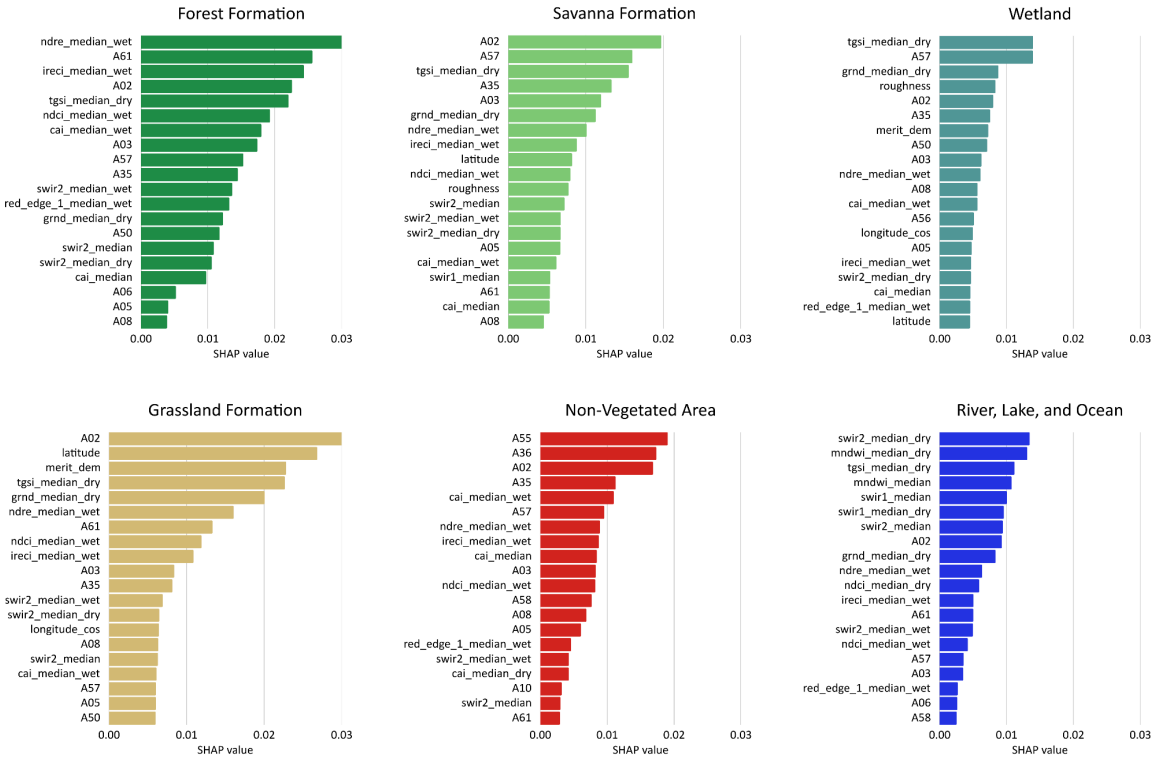


Figure 6. SHAP value distributions showing the class-specific contribution of predictor variables to the Random Forest model outputs for the Cerrado biome classification.

5. GENERAL MAP POST-CLASSIFICATION

Given the pixel-based classification methodology and the annual processing of the time series (2017–2024), 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: temporal gap-filling, temporal consistency, and spatial coherence. Each filter was designed to address specific classification limitations and collectively contributed to the overall quality and reliability of Collection 3.0.

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 2017–2024 period.

5.2. Herbaceous Sandbank Vegetation

The class Herbaceous Sandbank Vegetation (*Restinga Herbácea*, 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. The mapping of this class was conducted in a post-classification step, using a targeted rule-based approach based on ecological, spectral, and geological criteria. In the first stage, reference data for herbaceous sandbank vegetation and non-sandbank vegetation were manually collected through visual interpretation, resulting in two polygon datasets representing areas of confirmed occurrence and absence of sandbank vegetation. These polygons were used to generate random training points, with class labels assigned to sandbank vegetation (class 50) and non-sandbank vegetation. The sample points were associated with multiple years (2017–2024), allowing the construction of a multi-year training dataset. This dataset was subsequently used to extract spectral information from annual satellite products.

In the second stage, the discrimination between sandbank vegetation and other LULC classes was evaluated using annual Satellite Embedding data. Training samples were spatially restricted to areas of coastal geological deposits (SGB/CPRM), to ensure consistency with the known geological setting of sandbank vegetation. Feature separability was assessed by computing F-scores for each embedding dimension, quantifying the relative discriminatory power of individual embedding bands between sandbank and non-sandbank classes (Figure 7). The embedding band A18 has the highest

separability score and was selected as the primary spectral discriminator. An optimal decision threshold was then derived using Youden's J statistic, maximizing the trade-off between sensitivity and specificity. The final sandbank vegetation mask was generated by combining three complementary criteria: (i) a spectral condition based on the selected embedding band and its optimal threshold; (ii) an ecological constraint based on low HAND values (< 3), representing floodplain and low-relief; and (iii) a geological constraint restricting occurrences to coastal deposits. The resulting mask delineates areas with a high likelihood of herbaceous sandbank vegetation. This mask was subsequently applied to the annual land cover maps (2017–2024), converting pixels originally classified as Savanna, Grassland, or Wetland into Herbaceous Sandbank Vegetation where all criteria were met. This targeted remapping strategy ensures ecological plausibility and spatial coherence while minimizing commission errors in adjacent vegetation classes.

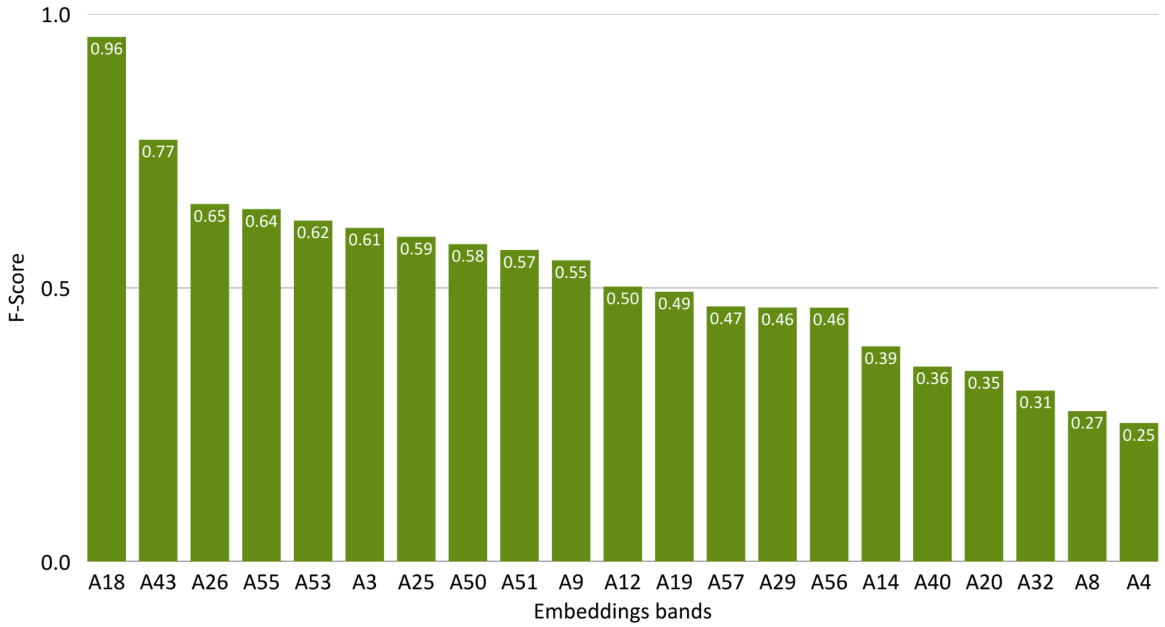


Figure 7. F-score–based separability analysis of Satellite Embedding dimensions used to discriminate Herbaceous Sandbank Vegetation from other LULC classes. Higher F-score values indicate greater discriminatory power of individual embedding bands.

5.3. Frequency

The frequency filter was applied to pixels classified as native vegetation in at least 90% of the 2017–2024 time series. Its objective was to enhance temporal consistency and reduce uncertainties caused by intermittent misclassifications. For each pixel, the frequency of assignment to native vegetation classes was calculated, followed by the

application of specific thresholds to confirm class stability. Pixels that met the 90% native vegetation criterion were further evaluated according to per-class frequency thresholds. Forest Formation was confirmed when present in ≥70% of the years, Wetland when ≥95%, Savanna Formation when >60%, and Herbaceous Sandbank and Grassland Formation when >40%. It is important to note that the frequency filter was particularly effective in mitigating edge noise and inconsistencies in the initial and final years of the time series, which are more susceptible to image availability and cloud effects.

5.4. Temporal

The temporal filter implemented in Collection 3.0 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 (2017–2024). 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 (2018-2023). 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 (2024): The filter searches for pixel values that were not classified as Mosaic of Uses (21) in 2024, but were classified as such in 2023 and 2022. The 2024 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) 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 2023 and 2024 larger than 1 hectare are retained. Smaller patches are assumed to be noise and are replaced by the 2023 class.

5.5. No false regrowth filter

The false regrowth filter was originally developed in the MapBiomas 30m Collection 9.0 to correct spurious signals of native vegetation regeneration that persisted even after the temporal filter. In MapBiomas Collection 3.0, this approach was adapted and expanded through a set of rule-based temporal post-classification procedures designed to reduce ecologically implausible or inconsistent LULC class transitions. The filter operates on the full annual time series (2017–2024) and applies a sequence of class-specific temporal rules that target abrupt appearances, short-lived interruptions, and unrealistic regeneration patterns:

- False Forest Formation Regrowth: This rule corrects spurious forest formation regeneration in silviculture areas by enforcing long-term persistence patterns at the beginning or end of the time series. Corrections are constrained using a stable reference classification to avoid introducing unrealistic transitions.
- False Wetland Regeneration (Temporal Interruption): This rule removes short-term wetland interruptions characterized by the pattern wetland → mosaic → wetland (11 → 21 → 11), which are interpreted as classification artifacts rather than true land-cover change.
- False Wetland Regeneration (Abrupt Appearance): To enforce temporal consistency in wetland dynamics, a second wetland-specific rule prevents wetlands from appearing abruptly in a given year if they were absent in the immediately preceding year. In such cases, the wetland class is replaced by the class observed in the previous year, thereby avoiding isolated or unsupported wetland emergence.
- False Savanna, Grassland, and Herbaceous Sandbank Vegetation Regeneration (Abrupt Appearance): This rule ensures these classes from appearing without temporal continuity. Pixels classified as Savanna, Grassland, or Sandbank Vegetation in a given year but not in the previous year are reclassified to the preceding land cover class.

5.6. Geomorphometric

This filter was developed to correct false classifications of Wetland (class 11) and Water (class 33) in areas with steep terrain, where the occurrence of these classes is geomorphologically inconsistent. Such misclassifications are common in regions with pronounced relief and shadow effects, which can lead to spectral confusion in optical satellite imagery. To address this issue, a slope map derived from the MERIT DEM was used. Slope values were calculated in degrees and converted to percent. A slope threshold was defined to represent the maximum limit for areas expected to contain wetlands or

water in the Cerrado biome. For each year from 2017 to 2024, pixels classified as Wetland and located on slopes equal to or greater than 12%, as well as pixels classified as Water on slopes greater than 20%, were identified as inconsistent. These pixels were reclassified using a spatial neighborhood approach, in which the original class label was replaced by the most frequent land use and land cover class within a 35-pixel radius.

5.7. Spatial filter

The spatial filter applied in MapBiomas Collection 3.0 aims to reduce salt-and-pepper noise and eliminate small, isolated patches resulting from classification artifacts, particularly at the edges of homogeneous pixel groups. This filter identifies spatially contiguous pixels that share the same class. For each annual classification map (2017–2024), the number of connected pixels belonging to the same class was computed using a four-neighbor connectivity criterion. A minimum mapping unit of 60 connected pixels (~0.6 hectares) was defined as the threshold for spatial consistency. Pixels belonging to connected components with a size less than or equal to this threshold were considered spatially inconsistent. These pixels were reclassified using a neighborhood-based approach, in which the original class label was replaced by the local focal mode calculated within a 3 × 3 pixel window. To further stabilize the spatial structure of the maps, the filtering procedure was applied sequentially in two iterations. In each iteration, connected components were recalculated based on the updated classification. This procedure is essential to suppress classification noise and eliminate small, fragmented areas, thus increasing 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 different MapBiomas Collections, both in terms of methodological refinement and conceptual definition of the class. In Collection 1.0, rocky outcrops were not mapped as an independent land cover class; instead, areas with exposed rock were implicitly included within the Grassland Formation class. This approach limited the explicit representation of rocky environments. The Collection 2.0 marked the first explicit mapping of the Rocky Outcrop class in the 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 Brazilian Geological Service (SGB/CPRM), and subsequently validated by expert interpreters.

In MapBiomas Collection 3.0, the Rocky Outcrop classification was further refined both conceptually and methodologically. Training samples were visually reviewed to ensure consistency with Sentinel-2 imagery. The feature space was revised to rely exclusively on annual Satellite Embedding data combined with geomorphometric variables. In addition, the conceptual definition of the class was updated to explicitly represent areas of exposed rock only, excluding rupestrian vegetation. This refinement aligns the 10 m product with the conceptual definition adopted in the MapBiomas 30m Collection 10.0. As in Collection 2.0, the Rocky Outcrop class in Collection 3.0 is mapped through an independent classification workflow, separate from the general LULC mapping. This strategy enables the application of tailored criteria that better capture the distinct physical, spectral, and geomorphological characteristics of rocky outcrops in the Cerrado. The class typically represents stable geological formations with clear sedimentary, igneous, or metamorphic signatures. An overview of the classification workflow is presented in Figure 8, and the following sections detail the methodological components, including the feature space (Section 6.1), training samples and classification parameters (Section 6.2), and post-classification filters (Section 6.3).

6.1. Feature space

The feature space adopted for the Rocky Outcrop classification in Collection 3.0 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. 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.

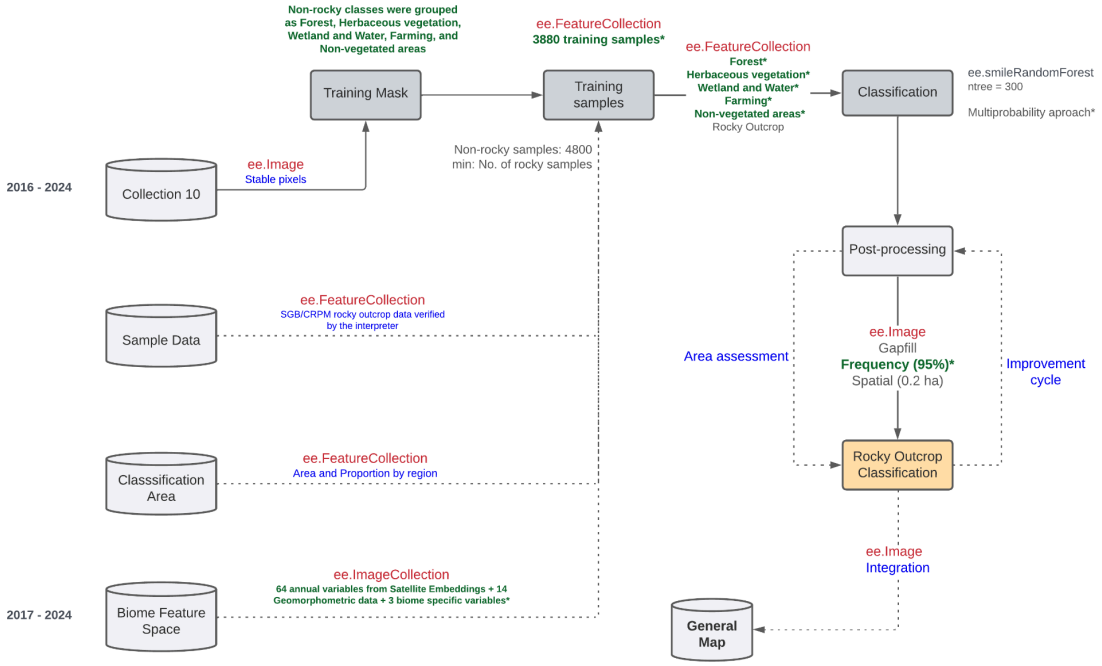


Figure 8. Each gray geometry (cylinders for databases and rectangles for processes) represents a key step in the classification schema, with the respective name inside. The gray text near databases and processes offers a description of the step, while the green text highlights the main innovations in Collection 3.0. 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
Terrain	Elevation	Identity	Geomorpho 90m Amatulli et al., 2020
	Aspect	Identity	Geomorpho 90m Amatulli et al., 2020

Type	Name	Statistics	Reference
	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
	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 3,880 training samples were used across the entire classification area, combining samples visually interpreted by specialists and additional samples provided by the Brazilian Geological Service (SGB/CPRM). All samples were carefully reviewed to ensure consistency with the updated conceptual definition of the class adopted in Collection 3.0. For the remaining LULC classes, training samples were generated using a reference training mask derived from stable pixels of the MapBiomas 30m Collection 10.0.

This mask was constructed by identifying pixels that exhibited no class changes throughout the 2016–2024 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. These samples were then combined with the visually interpreted rocky outcrop samples to form the final training dataset.

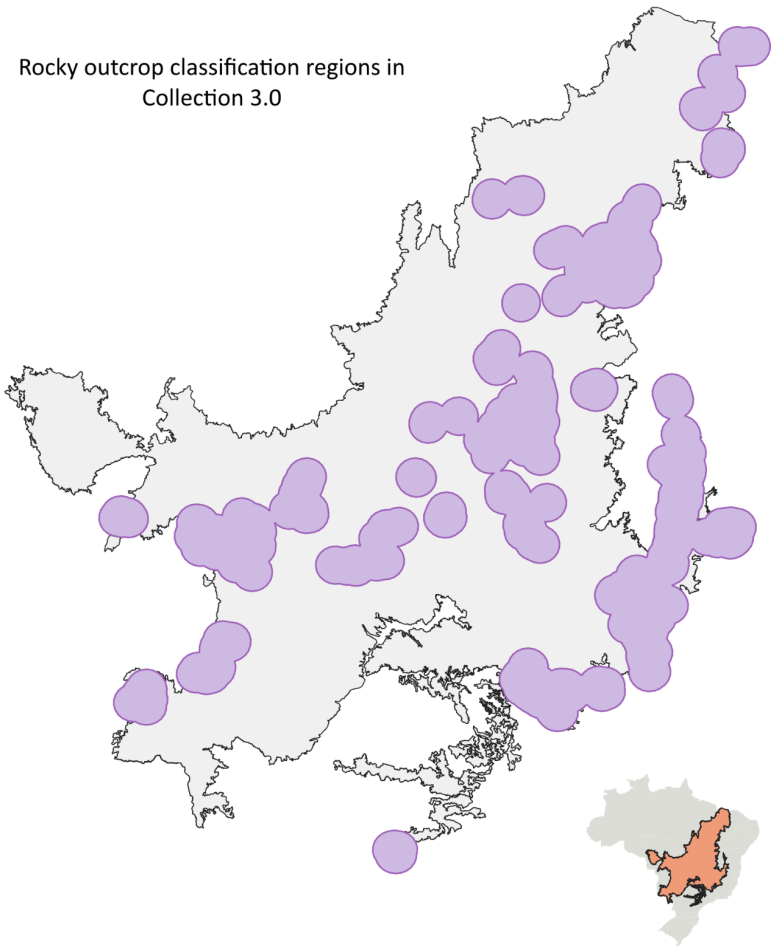


Figure 9. The rocky outcrop classification area used in collection 3.0. Highlighted in orange is the location of the Cerrado biome in Brazilian territory.

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 9). 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 95% 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 20 pixels, equivalent to approximately 0.2 hectares, based on an 8-neighbor connectivity criterion.

7. INTEGRATION

The integration process is a crucial step to ensure consistency and completeness in the annual LULC maps. It is conducted in two sequential stages. The first stage consists of

an internal integration within the Cerrado biome. In this step, the rocky outcrop classification is overlaid onto the native vegetation map produced in the main classification workflow. The second stage involves the integration with cross-cutting themes developed by the MapBiomas initiative. These themes include other layers such as urban infrastructure, agriculture, mining, and others. 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 specific prevalence rules adopted in this integration process are outlined in Table 6.

Table 6. General prevalence rules for Cerrado biome - MapBiomas 10m Collection 3.0

Class	Pixel value	Prevalence order	Color
Photovoltaic Power Plant (beta)	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	
Savanna Formation	4	21	
Wetland	11	22	
Grassland Formation	12	23	

Class	Pixel value	Prevalence order	Color
Pasture	15	24	
Mosaic of Uses	21	25	
Other Non-Vegetated Areas	25	26	

It is important to note that exceptions to this general rule apply in the context of protected areas for the Cerrado biome 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).

8. 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

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.

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.

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.

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.

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.

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., Schüler, 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.