



Cerrado - Appendix

Collection 9

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, 2024

1. OVERVIEW OF THE CERRADO CLASSIFICATION METHOD

The classification approach for the Cerrado biome in the MapBiomias project involved the use of decision trees to generate annual maps of the dominant native vegetation (NV) types, categorized into four groups: Forest Formation, Savanna Formation, Wetland, and Grassland Formation. Over time, the method for generating these maps has been refined, resulting in significant improvements from the first MapBiomias collection to the current version. The entire process of classifying the native vegetation of the Cerrado involved several steps. First, the optimal time of year to construct annual Landsat mosaics was selected. 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 remove noise and produce a consistent time series. Finally, the resulting maps were integrated with other cross-cutting themes. Classification results were evaluated through visual inspection and sample-based validation analysis. The methodological development of the Cerrado native vegetation (NV) classifications is presented in Table 1.

Table 1. This overview presents a historical account of the evolution of the Cerrado collections, starting from their initial version. In the method column, "EDT" means "Empirical Decision Tree," while "RF" means "Random Forest."

Collection	Range	Method	Mapped classes	Mainly improvements
1.0	2008 – 2015	EDT	Forest	- First collection
2.0	2000 – 2016	EDT	Forest, Savanna, Grassland	- New NV Classes (Savanna and Grassland)
2.3	2000 – 2016	RF	Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Other Non-vegetated Area, Water	- New classifier (Random Forest) - New auxiliary classes - Training samples derived from stable areas
3.0	1985 – 2017	RF	Same as Collection 2.3	- Expanded to the entire Landsat series - Improvement in training samples quality through outlier detection
3.1	1985 – 2017	RF	Same as Collection 3.0	- Ecoregions (38) substituted regular tiles as the classification unity
4.0	1985 – 2018	RF	Same as Collection 3.1	- Improvement in training samples quality by confronting with new reference maps for NV

Collection	Range	Method	Mapped classes	Mainly improvements
4.1	1985 – 2018	RF	Forest, Savanna, Grassland, Pasture, Agriculture, Other Non-vegetated Area; Water	<ul style="list-style-type: none"> - New feature space derived from variable importance analysis - Improvements in temporal with additional post-processing/ filters - Significant accuracy gain related to better mapping of NV
5.0	1985 – 2019	RF	Same as Collection 4.1	<ul style="list-style-type: none"> - Improvements in spatial continuity between classification regions - 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	<ul style="list-style-type: none"> - New NV Class (Wetland) - New classification mosaics (SR) - Upgrades in the statistical methodology to define feature space - New reference maps
7.0	1985 – 2021	RF	Forest, Savanna, Wetland, Grassland, Rocky Outcrop, Mosaic of Uses, Other Non-vegetated Area, Water	<ul style="list-style-type: none"> - New class (Rocky Outcrop) - Improvement in training samples using GEDI data to filter outliers - Accuracy gain related to better mapping of NV
7.1	1985 – 2021	RF	Same as Collection 7.0	<ul style="list-style-type: none"> - Improvement of temporal filter rules in the last year (2021)
8.0	1985 – 2022	RF	Same as Collection 7.0	<ul style="list-style-type: none"> - Regionalization of the hyperparameters, and classification - Extensive revision of the temporal filtering strategy and rules. - Expansion of the classification of the Rocky Outcrop theme
9.0	1985 – 2023	RF	Same as Collection 7.0	<ul style="list-style-type: none"> - New reference maps - Multiprobability approach, - Review of post-processing filters - False regrowth filter - New workflow for rocky outcrop classification

In the initial two collections, empirical decision trees were utilized as the classification approach, with nodes defined based on expert knowledge of the spectral features of each class. Collection 1.0 spanned the period from 2008 to 2015 and was published in 2016. Collections 2.0 and 2.3, which covered the period from 2000 to 2016, were published in 2018. The Random Forest method was employed for classification purposes in Collection 2.3. Subsequently, the empirical decision tree was employed to generate stable samples (2000–2016), which were then utilized to train the Random Forest models for classifying the entire time series. Collections 3.0 and 3.1 extended the temporal scope to encompass the period from 1985 to 2017, and a methodological paper was published (Alencar et al., 2020). Collections 4.0 and 4.1 demonstrated a notable enhancement in the precision of mapping in comparison to their predecessors. Additionally, they abandoned the use of empirical decision trees to generate training samples, instead relying on the collection of training samples based on stable samples from the previous collection (since collection 3.1).

To mitigate potential bias in the training dataset, reference maps (PRODES) of remaining native vegetation have been implemented since Collection 5.0 to delineate the area for collecting training samples for NV classes. In Collection 6.0, 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, the feature space was refined, and a greater number of reference NV maps were employed to filter the training samples, namely the "Inventário Florestal do Estado de São Paulo" and the "Base Temática Digital do Estado do Tocantins." Collection 7.0 processed the time series between 1985 and 2021, introduced a new class in the legend (Rocky Outcrop), refined the training samples by incorporating an outlier filter based on GEDI (Global Ecosystem Dynamics Investigation), and improved the hyperparameters of the RF classifier. Furthermore, the Wetland class was classified on the general map, in contrast to Collection 6.0, where it was a pseudo-cross-cutting theme.

The Collection 8.0 updated the time interval (1985-2022) and incorporated significant methodological advances, including improvements to the mask used to obtain training samples, incorporation of stable pixels from the Collection 7.1, and the use of reference maps and deforestation polygons from MapBiomias Alert and SAD Cerrado for the 2019-2022 period. The present collection (9.0), in addition to updating the period analyzed (1985-2023), maintains the strategy of filtering the training samples with deforestation polygons from SAD Cerrado (2020-2023) and MapBiomias Alert (2019-2023). Additionally, the Land Use and Land Cover Map of the Federal District (*Mapa de Uso e Cobertura da Terra do Distrito Federal*, CODEPLAN, 2019) and the Remaining Campos de Murundus Map of the State of Goiás (*Mapeamento dos Remanescentes de Campos de Murundus do Estado de Goiás* SEMAD, 2020) were incorporated as new reference maps. Moreover, the classification output was modified to consider the probability of each class,

rather than solely relying on the majority voting class, a process known as “multiprobability”, which allows us to improve the report of classification uncertainties. In the post-processing stage, all filters were revised to more accurately reflect the dynamics of biome land use and land cover. Additionally, a new false regrowth filter was introduced, based on the most recent years of the time series, to prevent spurious transitions to forest formation and wetlands at the end of the series. The rocky outcrop class was expanded to encompass the entire biome, employing a novel classification flow and revising the training samples and predictor variables. All classification and post-processing scripts utilized are available at: <https://github.com/mapbiomas-brazil/cerrado>.

2. LANDSAT IMAGE MOSAICS

The initial step in the classification of the native vegetation of the Cerrado biome entailed the generation of the mosaic of images utilized in the classification process. Prior to Collection 5.0, the classification of Cerrado NV employed Landsat 5 (TM), 7 (ETM+), and 8 (OLI) top-of-atmosphere (TOA) data. However, since Collection 6.0, the TOA data was superseded by surface reflectance (SR) data. The mosaic of images is created by composing pixels extracted from all the available images in a year. Statistical measures, including median, amplitude, standard deviation, and minimum, were computed for each pixel each year. These pixels were then aggregated annually, resulting in the production of Landsat mosaics that are subsequently used in the classification process.

A series of tests was conducted with the objective of determining the optimal period for image composition in the annual mosaics. Given the impact of seasonality on the spectral response of Cerrado vegetation, an assessment was conducted of the compositions of images captured during both the rainy and dry seasons (Figure 1). The tests included the classification of images from the end of the rainy season, when Cerrado vegetation is still vigorous and there is a higher probability of obtaining images with reduced cloud cover compared to the peak of the rainy season. Furthermore, additional tests were conducted with image compositions from the end of the dry season, covering the months between July and September. The results indicated that the use of images from the rainy season resulted in an overall greener mosaic, but also increased the commission errors in the forest class. Conversely, the use of images acquired in the last three months of the dry season resulted in a drier mosaic, leading to an underestimation of forest coverage, primarily due to the reduced potential to map deciduous forests.

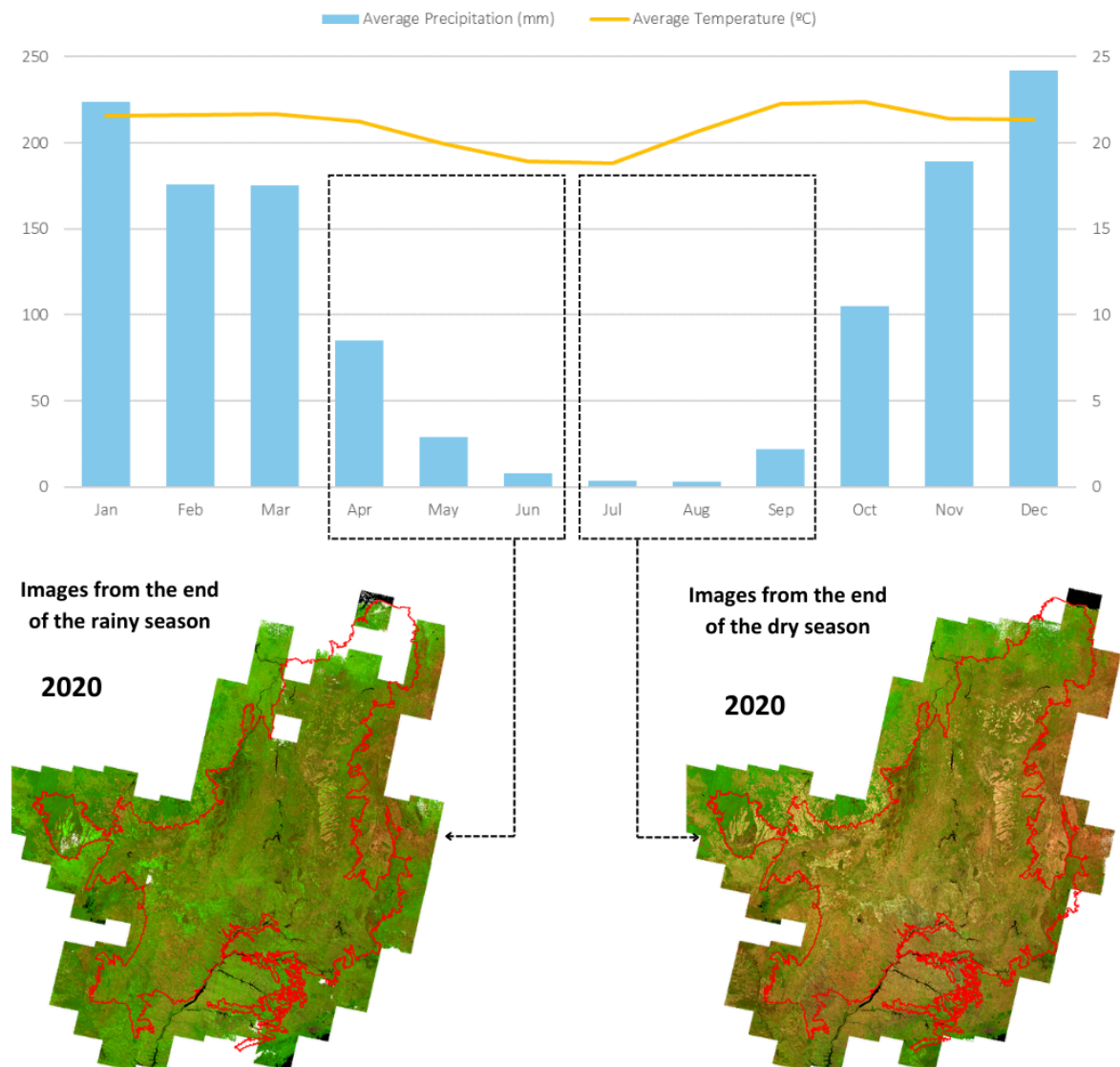


Figure 1. Precipitation data is the monthly averages for the Cerrado (Macena et al., 2008). Temperature data is the monthly averages for the Federal District (INMET). False-color composite Landsat 8 mosaics at the end of the rainy season and at the end of the dry season in the Cerrado.

In consideration of the aforementioned tests, a broader window was selected for the purpose of defining the initial and final dates for the generation of the mosaics. These dates were standardized across all 38 classification regions and for all years under consideration. The selection criteria entailed the utilization of a six-month window between April and September, with a maximum limit (Figure 2). The application of this methodology during the designated period yielded more effective results in addressing the mapping issues observed in the narrower window tests. To guarantee the quality of the yearly mosaics over the Cerrado biome, a visual inspection was conducted. In Collection 9.0, Landsat 5 data from 1985 to 2010 was utilized, with the exception of 2001

and 2002, during which Landsat 7 data was employed due to technical failures in the TM sensor. Moreover, Landsat 7 data was employed for 2011 and 2012, while Landsat 8 data was utilized from 2013 to 2023. As a result, 39 Landsat surface reflectance mosaics, spanning from 1985 to 2023, were obtained (Figure 3).

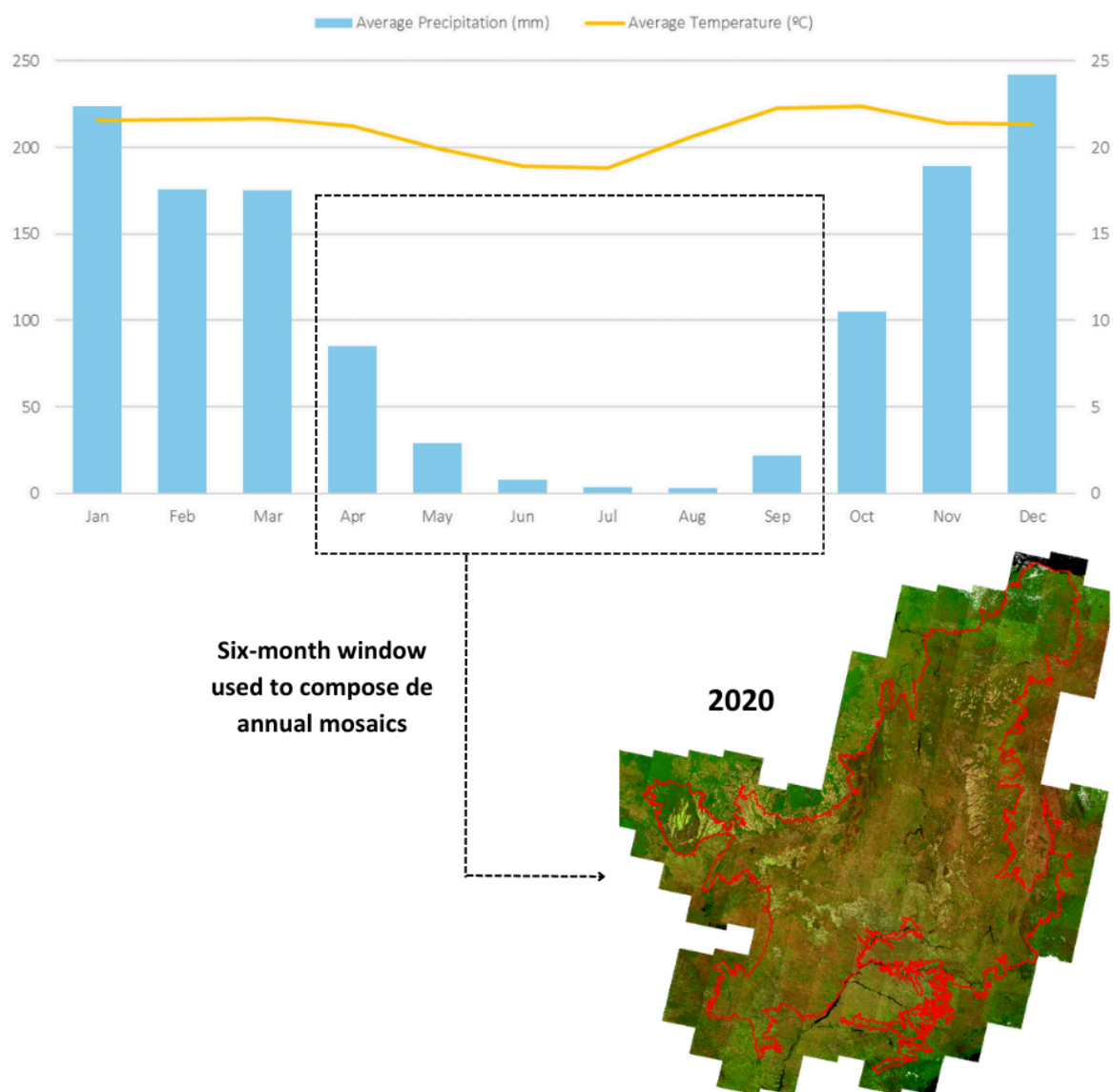


Figure 2. Precipitation data is the monthly averages for the Cerrado (Macena et al., 2008). Temperature data is the monthly averages for the Federal District (INMET). Time-window used to build the yearly classification mosaics used in the MapBiomias Collection 9.0 in the Cerrado biome.

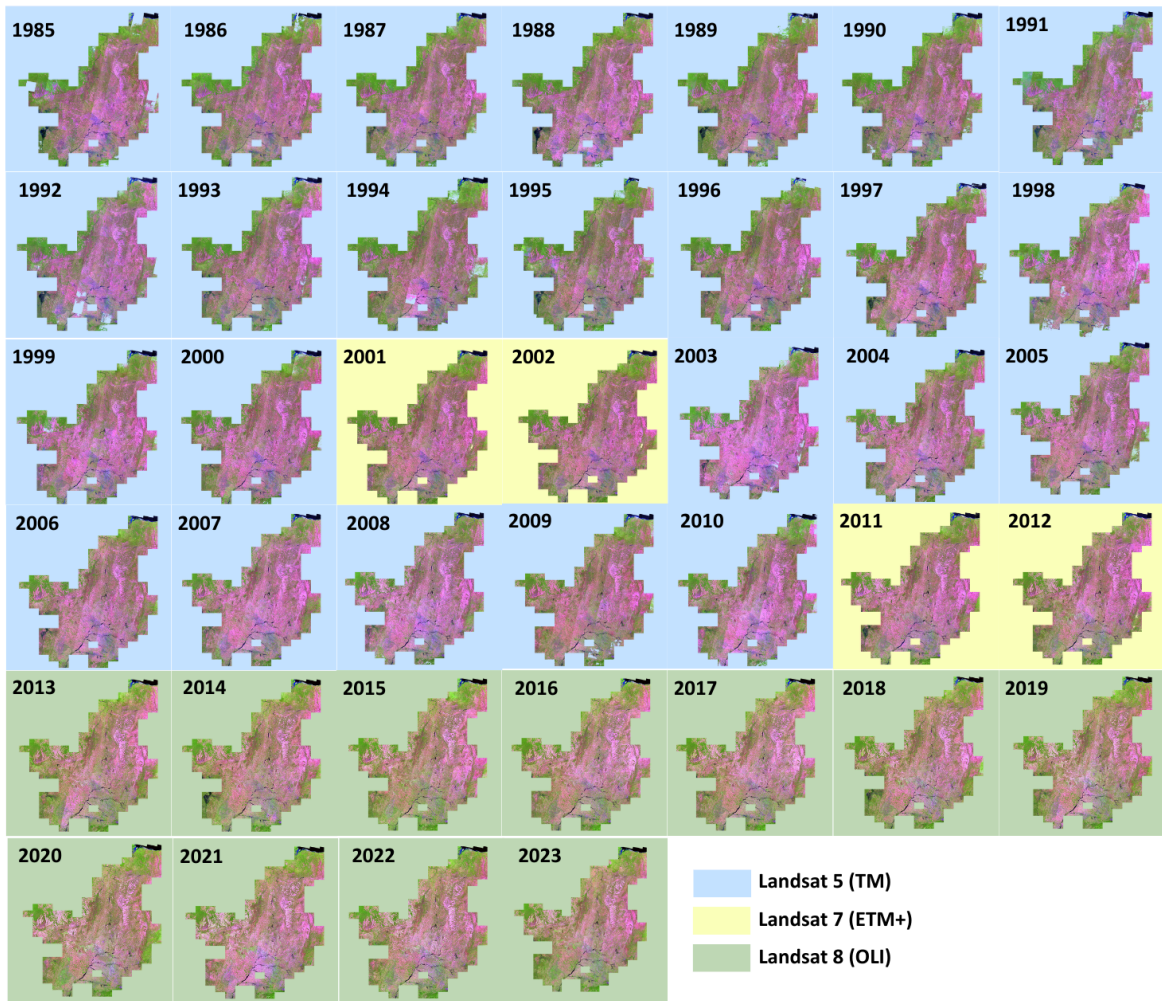

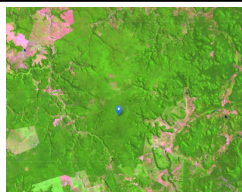

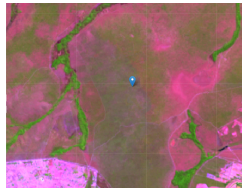



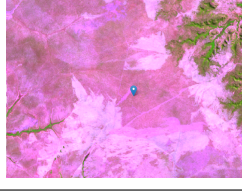





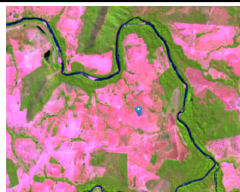

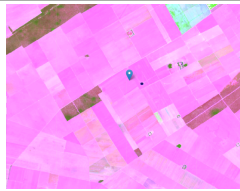



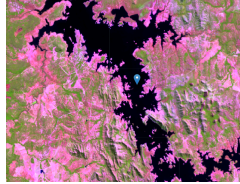
Figure 3. Annual Landsat mosaics for the Cerrado biome from 1985 to 2023. These mosaics are derived from the medians of SWIR1-NIR-Red bands.

3. CLASSIFICATION SCHEME

In the context of MapBiomias Collection 9.0, the classification of Landsat mosaics for the Cerrado biome encompassed a comprehensive set of nine land use and land cover (LULC) classes, as specified in the MapBiomias legend, which is available at <https://brasil.mapbiomas.org/en/codigos-de-legenda/>. All these classes are described in Table 2. This classification scheme was not limited to identifying only native vegetation (NV) and water bodies but also included anthropogenic classes, such as Agriculture and Pasture, to provide a comprehensive representation of the Cerrado landscape. It is important to note that the “Agriculture” and “Pasture” classes are mapped together with the native vegetation and water body classes. However, these classes are transformed into a “Mosaic of Uses” in the post-processing stages. This approach was essential in reducing the potential for omission or commission errors within the NV classes, thereby ensuring a more accurate and holistic classification.

Table 2. Land use land cover categories used for the Landsat mosaics classification for the Cerrado biome in MapBiomas Collection 9.0.

Classes Level 1	Classes Level 2	ID	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) (Ribeiro & Walter, 2008), as well as Semi-deciduous Seasonal Forests.
	Savanna Formation	4			Savanna formations with defined tree and shrub-herbaceous stratum (Cerrado Stricto Sensu: Dense, Typical, Sparse and Rupestrian Savanna).
Non-Forest Natural Formation	Wetland	11			Vegetation with a predominance of herbaceous strata subject to seasonal flooding (e.g. Campo Umido) 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 fields) and some areas of savanna formations such as the rupestrian cerrado.
	Rocky Outcrop	29			Monolithic features, bedrock or slabs naturally exposed on the earth's surface without soil cover, often with partial presence of rupestrian vegetation and high slope.

Classes Level 1	Classes Level 2	ID	Color	RGB composite (SWIR1-NIR-Red)	Description
Farming	Pasture*	15			Pasture area, predominantly planted, linked to farming production activities.
	Agriculture*	18			Areas occupied with short to long vegetative cycle agricultural 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, reservoir and other water bodies

* The "Agriculture" and "Pasture" classes are transformed into the "Mosaic of Uses" class during the post-processing stage.

The subsequent subsections provide detailed information on the procedures adopted in the Collection 9.0 classification: Regions for classification (4.1), Feature space selection (4.2), Training samples, classification algorithm, and parameters (4.3). Figure 4 provides an overview of the methodology for Cerrado native vegetation classification in Collection 9.0. The general framework can be summarized as follows:

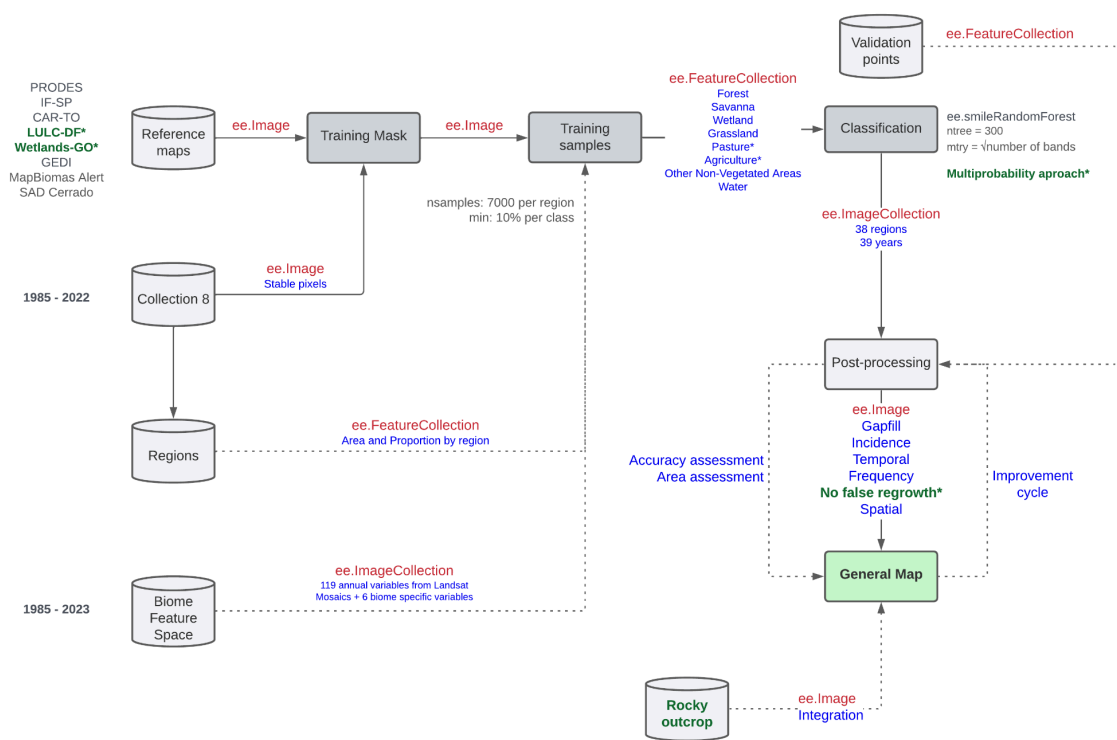


Figure 4. Each gray geometry (cylinders for databases and rectangles for processes) represents a key step in the classification schema, with the respective name inside. The gray text near databases and processes offers a short description of the step, while the green text highlights the main innovations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

- Training Samples:** Training samples were based on stable areas in Collection 8.0 (1985-2022), reference maps, and filtering techniques including vegetation structure (GEDI) and terrain data. The proportion of area covered by each class was evaluated in order to ensure that the training dataset was balanced for each iteration of the classification model, on a regional and yearly basis.

- **Classification Process:** Classification was performed using the Random Forest algorithm, implemented in the 'ee.Classifier.smileRandomForest' function with the output mode set to "Multiprobability" on the Google Earth Engine (GEE) platform.
- **Classification of Rocky Outcrop:** A visual inspection-based training dataset was created for the specific purpose of training the classifier on the Rocky Outcrop class. This entailed the introduction of new predictor variables and a novel framework designed to enhance the accuracy of this class.
- **Map Creation:** Two distinct maps were produced: A general map encompassing all LULC classes mapped across the Cerrado biome and a specialized map focused on the mapping of rocky outcrop areas.
- **Integration:** The classification results of the rocky outcrop theme were integrated into the general map.

4. GENERAL MAP CLASSIFICATION

4.1. Classification regions

In the initial four collections, a grid at a 1:250,000 scale served as the primary unit for classification. Each grid cell ($n = 172$ tiles) underwent independent analysis by the classification algorithm. However, this approach frequently resulted in inconsistent boundaries between grids, which in turn produced undesirable classification outcomes. In Collection 5.0, a revised set of classification units was introduced, based on regional variations in biophysical and land-use attributes. This restructuring involved the subdivision of the Cerrado in 19 ecoregions proposed by Sano et al. (2019), Brazil's major watersheds and the land use and land cover classes observed in Collection 3.1 (2017). As a result, 38 classification regions were delineated, superseding the preceding grid-based methodology and effectively compartmentalizing the environmental heterogeneity of the Cerrado biome. Such heterogeneity can exert a considerable influence on the spectral signatures of native vegetation (NV), even within the same NV class.

In Collection 7.0, adjustments were made to classification regions, taking into account NV seasonality. This involved calculating the Normalized Difference Vegetation Index (NDVI) between 2017 and 2020 for each available Sentinel 2 (SR) scene. By performing a per-pixel subtraction of the 90th and 10th percentiles (p90-p10), regions exhibiting substantial natural vegetation seasonal variation were identified. This information was then employed to empirically refine the classification regions, ensuring that areas with distinct phenological and spectral characteristics were not grouped together within the same classification region. Subsequent collections, including the

current one, have maintained the classification regions established in Collection 7.0, with the number remaining consistent at 38 (see Figure 5).

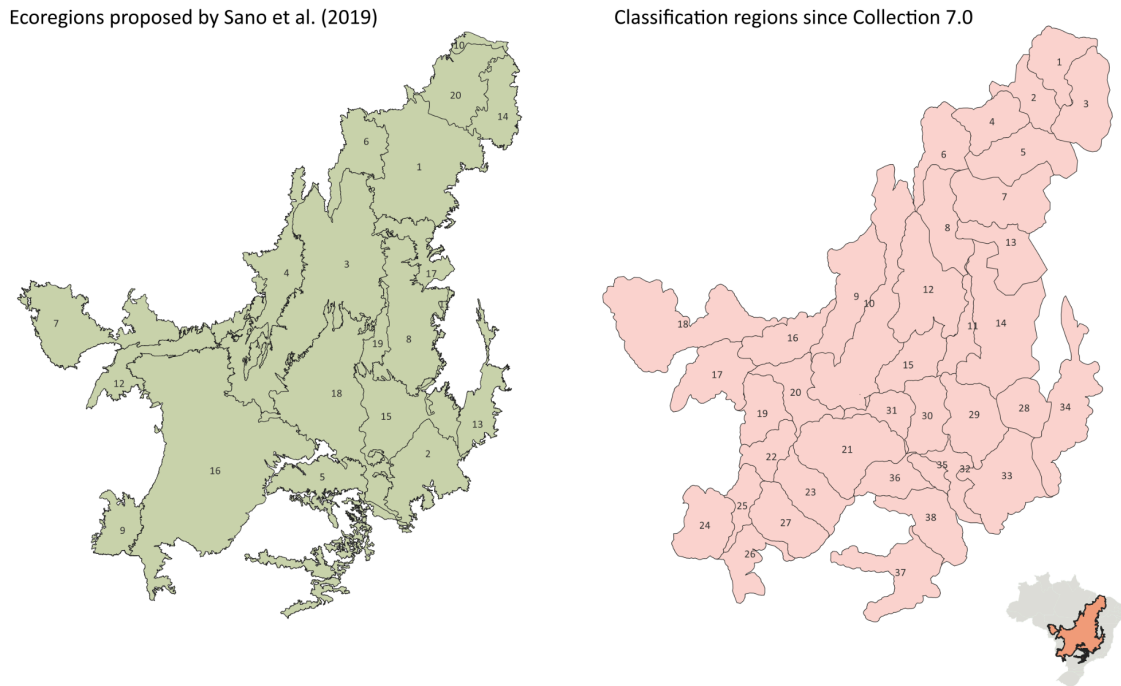


Figure 5. Classification regions, modified from Sano et al., 2019. Highlighted in orange is the location of the Cerrado biome in Brazilian territory.

4.2. Feature space

The feature space for the classification process included a comprehensive set of 119 variables common to all biomes. These variables included the annual mosaic bands listed in Table 3, as well as six variables specific to the Cerrado biome that had been used in previous collections (Collections 7.0, 7.1, and 8.0), as listed in Table 4. The variables included the original Landsat reflectance bands, various vegetation indices, and variables derived from spectral mixture modeling. For each of these variables, a number of statistical measures were calculated, including median, dry period median, wet period median, minimum, amplitude, and standard deviation. The inclusion of such a diverse set of variables was designed to capture the complex spectral and temporal characteristics of the Cerrado biome. All of these data were considered as predictor variables for land use and land cover classification per region in the Cerrado.

Table 3. Feature space considered in the classification of the Cerrado biome in the MapBiomias Collection 9.0. Column "statistic" refers to the set of per pixel statistical reducers used for each

variable: a) amplitude: variation of the index considering the pixel values within the temporal mapping window; b) median: per year median considering the temporal window; c) median_dry: seasonal median below NDVI first quartile; d) median_wet: seasonal median above NDVI first quartile; e) standard deviation: pixel standard deviation considering values within the temporal window; f) lower annual pixel value within the temporal window.

Type	Name	Formula	Statistics	Reference
Landsat band	Blue	Band 1 (L5 and L7) Band 2 (L8)	median, median_dry, median_texture, median_wet, minimum, stdDev	USGS
	Green	Band 2 (L5 and L7) Band 3 (L8)	median, median_dry, median_texture, median_wet, minimum, stdDev	USGS
	Red	Band 3 (L5 and L7) Band 4 (L8)	median, median_dry, median_wet, minimum, stdDev	USGS
	NIR	Band 4 (L5 and L7) Band 5 (L8)	median, median_dry, median_wet, minimum, stdDev	USGS
	SWIR 1	Band 5 (L5 and L7) Band 6 (L8)	median, median_dry, median_wet, minimum, stdDev	USGS
	SWIR 2	Band 7 (L5 and L7) Band 8 (L8)	median, median_dry, median_wet, minimum, stdDev	USGS
Spectral Index	Cellulose Absorption Index	$CAI = SWIR2 / SWIR1$	median, median_dry, stdDev	Nagler et al., 2003
	Enhanced Vegetation Index 2	$EVI 2 = 2.5 \times (NIR - Red) / (NIR + 2.4 \times Red + 1)$	amplitude, median, median_dry, median_wet, stdDev	Parente et al., 2018
	Green Chlorophyll Vegetation Index	$GCVI = (NIR / Green - 1)$	median, median_dry, median_wet,	Burke et al., 2017

Type	Name	Formula	Statistics	Reference
	Hall Cover	$\text{Hall Cover} = (-\text{Red} \times 0.017 - \text{NIR} \times 0.007 - \text{SWIR2} \times 0.079 + 5.22)$	stdDev median, stdDev	Hall et al., 2006
	Normalized Difference Vegetation Index	$\text{NDVI} = (\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red})$	amplitude, median, median_dry, median_wet, stdDev	Rouse et al., 1974
	Normalized Difference Water Index	$\text{NDWI} = (\text{NIR} - \text{SWIR1}) / (\text{NIR} + \text{SWIR1})$	amplitude, median, median_dry, median_wet, stdDev	Gao et a., 1996
	Photochemical Reflectance Index	$\text{PRI} = (\text{Blue} - \text{Green}) / (\text{Blue} + \text{Green})$	median, median_dry, median_wet	Gamon et al., 1992
	Soil-Adjusted Vegetation Index	$\text{SAVI} = 1.5 \times (\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red} + 0.5)$	median, median_dry, median_wet, stdDev	Huete, 1988
	Green Vegetation Fraction	GV = Fractional abundance of green vegetation within the pixel	amplitude maximum, median, median_dry, median_wet, minimum, stdDev	Souza et al., 2005
	Green Vegetation Shade Fraction	$\text{GVS} = \text{GV} / (\text{GV} + \text{NPV} + \text{Soil} + \text{Cloud})$	amplitude, maximum, median, median_dry, median_wet, minimum, stdDev	Housman et al., 2018
Fraction	Normalized Difference Fraction Index	$\text{NDFI} = (\text{GVS} - (\text{NPV} + \text{Soil})) / (\text{GVS} + (\text{NPV} + \text{Soil}))$	amplitude, maximum, median, median_dry, median_wet, minimum, stdDev	Souza et al., 2005
	Non-photosynthetic Vegetation Fraction	NPV = Fractional abundance of non-photosynthetic vegetation within the pixel	amplitude, maximum, median, median_dry, median_wet, minimum, stdDev	Souza et al., 2005
	Savanna Ecosystem Fraction Index	$\text{SEFI} = (\text{GV} + \text{NPV}_S - \text{Soil}) / (\text{GV} + \text{NPV}_S + \text{Soil})$	median, median_dry, stdDev	Alencar et al., 2020

Type	Name	Formula	Statistics	Reference
	Shade Fraction	Shade = 100 - (GV + NPV + Soil + Cloud)	median	Housman et al., 2018
	Soil Fraction	Soil = Fractional abundance of soil within the pixel	amplitude, maximum, median, median_dry, median_wet, minimum, stdDev	Souza et al., 2005
	Wetland Ecosystem Fraction Index	WEFI = ((GV + NPV) - (Soil + Shade)) / ((GV + NPV) + (Soil + Shade))	amplitude, median, median_wet, stdDev	Rosa, 2020
Terrain	Slope	ALOS DSM: Global 30 m	identity	Tadono et al., 2014

Table 4. Complementary bands added to the Cerrado feature space since Collection 7.0.

Name	Formula	Statistics	Reference
Latitude	ee.Image.pixelLonLat().select(['latitude'])	identity	Geolocation
cos(Longitude)	cos(ee.Image.pixelLonLat().select(['longitude']))	identity	Geolocation
sin(Longitude)	sin(ee.Image.pixelLonLat().select(['longitude']))	identity	Geolocation
Time Since the Last Fire	TSLF = Current year - Year of the last fire	identity	Alencar et al., 2022
Height Above the Nearest Drainage	HAND Global 30m	identity	Donchyts et al., 2016
3yr NDVI Amplitude	NDVI from current year to -2 years: min(median_dry) - max(median_wet)	identity	Alencar et al., 2020

4.3. Training samples, classification algorithm, and parameters

Land use and land cover classification of the Cerrado biome was performed regionally on an annual basis using the GEE platform. The training samples for each region were derived from stable areas identified in the Collection 8.0 classification over a 38-year period. To ensure accuracy and reliability, these training samples were supplemented with reference maps of native vegetation and deforestation, in addition to a GEDI-based methodology that effectively removed outliers from stable pixels. In addition, the canopy height model proposed by Lang et al. (2022) was instrumental in excluding stable pixels with incorrect canopy height values for each NV class, significantly improving classification

accuracy. All of these procedures have been used since Collection 7.0. To identify and remove erroneous pixels, the following criteria were applied to the canopy height data:

- Forest Formation with canopy height lower than 4 meters
- Savanna Formation with canopy height lower than 2 meters and higher than 8 meters
- Wetland with canopy height higher than 15 meters
- Grassland Formation with canopy height higher than 6 meters

Following the implementation of these adjustments, each classification unit (region) was assigned a sample size of 7,000 training samples, distributed proportionately according to the area of each class in Collection 8.0, with the year 2005 (mid of the time series) used as the reference data. A minimum of 700 samples was assigned to ensure sufficient representation. The class "Water" was assigned a specific minimum number of samples ($n = 175$) to minimize class-specific commission errors and overestimation. This approach aimed to improve classification accuracy and ensure that underrepresented classes were adequately accounted for during the classification procedure.

Two parameters were identified as requiring adjustment in the training of a Random Forest Model. 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 was trained in Google Earth Engine using the function `ee.Classifier.smileRandomForest`, and applied to the Landsat mosaics by using the "Multiprobability" approach. The multiprobability output generated an array of probabilities for each class, which were ordered by the likelihood of the pixel belonging to each observed class. Following the classification, the class with the highest probability for each pixel was selected in order to determine its land use and cover class for a given year.

5. GENERAL MAP POST-CLASSIFICATION

The pixel-based classification method, employed with individual runs for each year in a long time series, showed the need to implement spatial and temporal post-classification filters to ensure consistency and eliminate classification errors. These processes encompassed several filters, including the gap-fill, incidence, temporal, frequency, and spatial filters. Each of these filters was designed to eliminate spurious transitions in the classification and improve the accuracy of the final map. These filters, run in the order described below, played a critical role in improving Collection 9.0.

5.1. Gap-Fill filter

The Temporal Gap-Fill Filter had a significant impact in addressing the issue of missing data or gaps resulting from the presence of cloud-covered or cloud-shadowed pixels in the images. The objective of the filter was to fill the no-data values with the temporally nearest future valid classification available for each pixel. In cases where no future valid classification was available, the no-data value was replaced with the previous year's valid classification. As a result of the temporal Gap-Fill filter, the final classified map should contain very few gaps, with the exception of instances where a specific pixel remained consistently classified as no-data throughout the entire temporal series.

5.2. Incidence filter

The incidence filter was developed to deal with excessive changes between classes observed over a 39-year time series, with a particular focus on transitions between native vegetation and anthropogenic areas, and vice versa. The first step was to group the land use and land cover classes into three main categories: Natural (Forest Formation, Savanna Formation, Grassland, Wetland), Anthropogenic (Pasture, Agriculture, and Other Non-vegetated Areas), and Other (River, Lake and Ocean, and Not Observed). Based on these three groups, transitions between natural and anthropogenic were counted, taking into account the number of changes per pixel over the time series.

Pixels with less than seven connected pixels and with ten or more changes were identified as edge pixels with noise. For these edge pixels, the classification was redefined to the most frequent class (mode) in its original trajectory. For pixels connected to more than seven pixels and having more than 13 transitions ($\frac{1}{3}$ of the time series), the correction was also applied to the most frequent class. This approach was chosen because edge pixels exhibit excessive class changes due to spectral mixing in Landsat pixels containing more than one thematic target.

5.3. Temporal filter

The temporal filter implemented in Collection 9.0 plays a critical role in addressing temporal inconsistencies. This process ensures consistency and accuracy in the analysis of land use and land cover change over time and minimizes classification errors due to invalid temporal transitions. In this step, Agriculture and Pasture were reclassified to "Mosaic of Uses" (21) to filter farming as a unique class. Then, it follows a series of sequential steps:

- First, the filter evaluates all pixels in a 5-year (1986-2020) and 4-year (1986-2021) moving 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). Each transition is evaluated according to a

specific order of priority: Savanna Formation (4), Forest Formation (3), Grassland Formation (12), Wetland (11), Mosaic of Uses (21), River, Lake, and Ocean (33), and Other Non-vegetated Areas (25).

- The second stage is similar to the first, but consists of a 3-year moving window (1986-2022) that corrects for all intermediate years, taking into account previous and subsequent years (-1 and +1 years). This stage complements the first by adjusting the classifications in years where subtle changes may occur that are not captured by the wider moving window. This correction is applied in the same order of classes used in the first step.
- The third step involves checking initial classifications of native vegetation (Forest, Savanna, Wetland, and Grassland) that may not have been correctly identified in the base year of 1985, but were correctly classified in the subsequent years of 1986 and 1987. The 1985 value is then corrected to reflect this classification.
- The fourth step involves checking the values of pixels that were not classified as Mosaic of Uses (21) in 2023, but were classified as such in 2022 and 2021. The value in 2023 is corrected to be consistent with previous years to avoid uncorrected regeneration in the most recent year.
- Finally, the filter allows for regeneration of native vegetation (NV) in the last year in areas of at least 1 hectare. Pixels indicating regeneration between 2022 and 2023 are evaluated, and areas smaller than 1 hectare are discarded to ensure classification consistency.

5.4. Frequency filter

The frequency filtering process was applied exclusively to pixels that were classified as native vegetation in at least 90% of the time series. Subsequently, criteria are applied to the native vegetation classes in order to achieve stability. In cases where the Forest Formation class is present for more than 75% of the time series, this class is confirmed for the pixel in question. For Wetland, a minimum frequency of 60% is required, while for Savanna and Grassland, the minimum frequency is 50%. These values were selected to achieve a more stable classification of the native vegetation, refining and minimizing the uncertainties associated with occasional temporal fluctuations in the pixel classification over time. It is important to note that the frequency filter also helped to remove noise present in the first and last years of the classification, which cannot be adequately addressed by the temporal filter alone.

5.5. No false regrowth filter

The false regrowth filter is applied exclusively to the Forest Formation (3) and Wetland (11) classes. The principal objective is to preclude the artificial expansion of

these areas as a consequence of classification errors in silvicultural areas (for Forest Formation) or areas of Mosaic of Uses/Pasture (for Wetland).

For the Forest Formation class, the filter is designed to identify and correct classification errors in silviculture areas that may be mistakenly categorized as forest regeneration. The method considers pixels that have been classified as "Mosaic of Uses" (class 21) for a continuous period of more than 15 years and which, in the subsequent year, have been reclassified as "Forest Formation." Such abrupt alterations are frequently indicative of a classification error, and pixels exhibiting this behavior are reclassified as anthropogenic areas, retaining the "Mosaic of Uses" classification (21). This procedure guarantees that the artificial expansion of forest formation areas is kept to a minimum, thus providing a more precise representation of land use land cover dynamics.

With regard to the Wetland class, the filter is applied with a view to avoiding any artificial increase in this class at the end of the time series, particularly in areas that show greater dynamism in land use or seasonality. The filter examines pixels that were designated as Wetland in the current year, but not mapped as Wetland in the previous year. The analysis concentrates on the final five years of the time series, aiming to identify and rectify these pixels to prevent an overestimation of the Wetland area. Consequently, pixels exhibiting these attributes are reclassified in accordance with the predominant class in preceding years, thus avoiding the erroneous detection of wetlands in the Cerrado.

5.6. Spatial filter

The spatial filter implemented in Collection 9.0 plays a significant role in enhancing the precision of the classification process by addressing misclassifications at the boundaries of pixel groups. The "connectedPixelCount" function, inherent to the Google Earth Engine platform, is employed to identify connected components (neighbors) sharing the same pixel value. This approach entails the consideration of isolated pixels that lack the minimum requisite number of connected identical neighbors for further assessment. The spatial filter establishes a minimum connection value of six adjacent pixels, which corresponds to an area of approximately 0.54 hectares. This indicates that for a pixel to retain its classification, it must possess a minimum of six adjacent pixels that share an identical value. By establishing a minimum mapping unit, the spatial filter assists in the elimination of spurious noise and artifacts caused by isolated pixels that do not align with the prevailing land cover patterns within the Cerrado biome.

6. ROCKY OUTCROP CLASSIFICATION

In Collection 7.0, the beta version of the rocky outcrop classification was incorporated. In Collection 8.0, the classification underwent substantial improvements to

enhance its representation in the Cerrado biome. This land cover class includes a set of rocky outcrops that are notable for their stability and which exhibit features indicative of sedimentary, igneous, or metamorphic processes (Figure 6). It is important to note that some areas of “campos rupestres” can also be included within this classification.

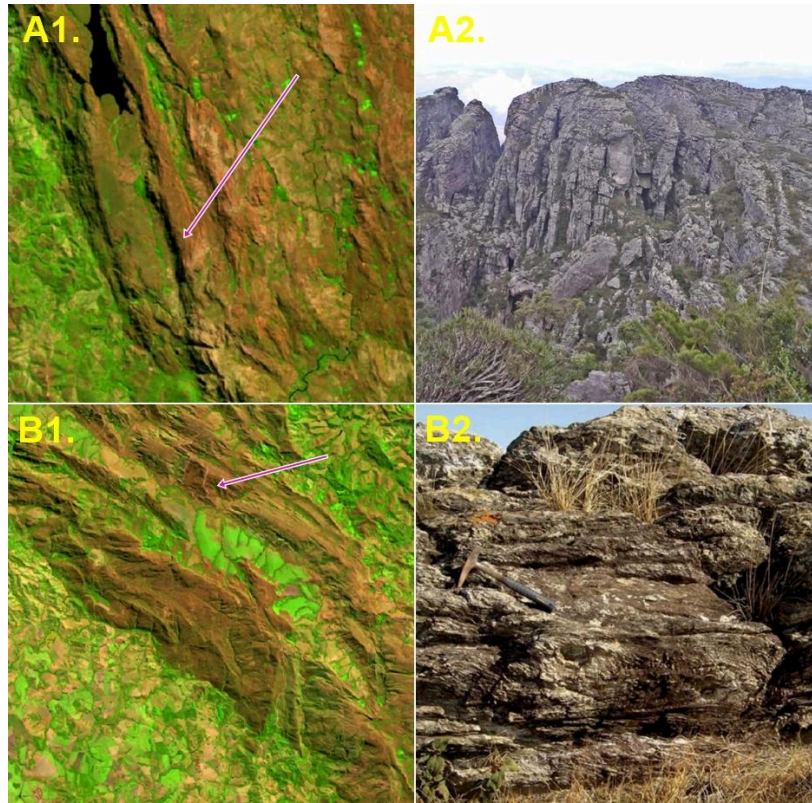


Figure 6. Example of landscapes mapped as Rocky Outcrop in the Collection 9.0. A) “Serra do Espinhaço”: A1) Landsat false-color composition (SWIR1-NIR-Red) for the year 2021. The pink arrow indicates the approximate location of the field photograph; A2) Field photograph (credits to TMbux). B) “Serra 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).

The classification process for rocky outcrops is distinct from that employed for the general map. This approach is adopted to prevent overestimation of the rocky outcrop class and to guarantee that the mapping criteria are tailored to this specific class, which exhibits distinctive characteristics in comparison to the other land use and land cover classes in the Cerrado. The objective is to accurately identify and delineate the various rocky outcrop areas in the Cerrado region, considering their geological characteristics and ecological importance. The classification flowchart is presented in Figure 7. The subsequent sections describe the methodological steps in detail, including the selection of the feature space (6.1), the training samples, classification algorithm and parameters (6.2), and the post-processing filters (6.3).

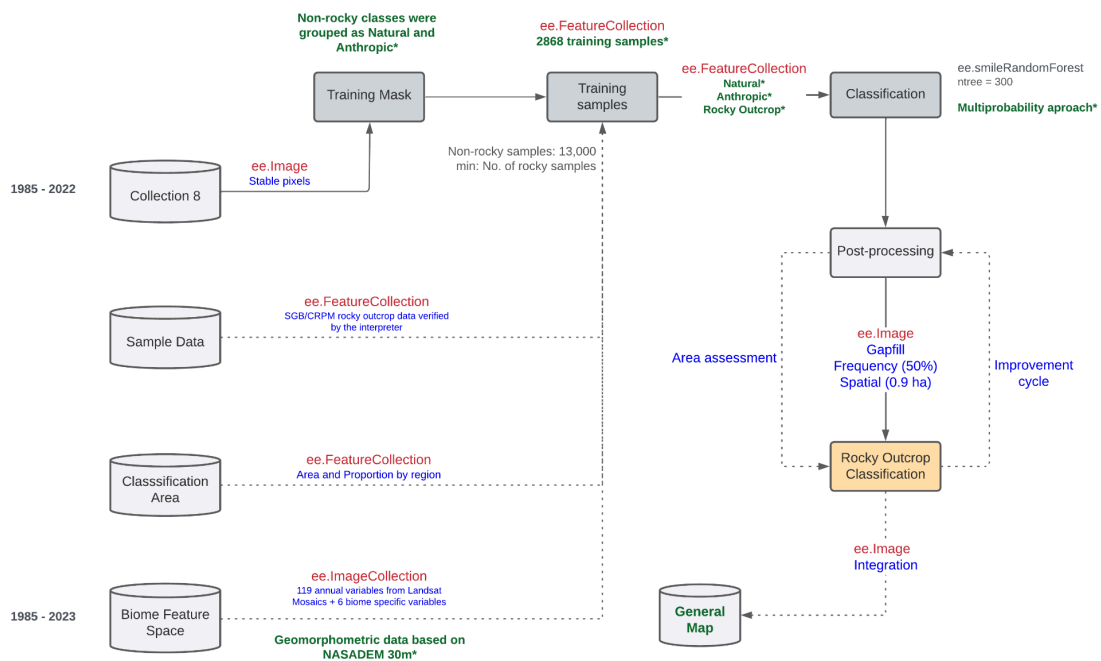


Figure 7. Each gray geometry (cylinders for databases and rectangles for processes) represents a key step in the classification schema, with the respective name inside. The gray text near databases and processes offers a short description of the step, while the green text highlights the main innovations 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 short description of the asset content.

6.1. Feature space

The Landsat images utilized for classification encompass the identical 119 bands delineated in section 4.2 and Table 3, in addition to the variables exclusive to the Cerrado biome, as detailed in Table 4. However, for the classification of rocky outcrops, four additional predictor variables representing terrain attributes were included: relative relief (representing the difference between the highest contour value and the lowest contour value of any given place or region), valley depth (difference in elevation between the valley and the upstream ridge), topographic position index (TPI; measures topographic slope positions), and elevation (in meters). All data were processed using 30-meter NADASEM images, which are accessible via the Google Earth Engine platform. These variables are described in Table 5 and represent characteristic topographical attributes that assist in identifying areas of rocky outcrops, taking into account their prevalence in more rugged terrain, including areas of higher elevation and steep slopes.

Table 5. Complementary bands added to the Cerrado rocky outcrop classification feature space

Type	Name	Dataset	Statistics	Reference
Terrain	Relative relief	NASADEM 30 m	identity	Ganerød et al., 2023
	Valley depth	NASADEM 30 m	identity	Ganerød et al., 2023
	Topographic Position Index (TPI)	NASADEM 30 m	identity	Ganerød et al., 2023
	Elevation	NASADEM 30 m	identity	Ganerød et al., 2023

6.2. Training samples, Classification algorithm, and parameters

Due to the specific characteristics of rocky outcrops and to avoid commission errors, the classification area was manually refined following a comprehensive visual evaluation of the entire biome. This approach ensured the inclusion of all rocky outcrops within the Cerrado biome in this classification area. It is notable that the current Collection 9.0 encompasses all rocky outcrops within the Cerrado, representing a significant expansion in comparison to the more limited territorial scope of Collections 7.0 and 8.0. The mapped areas in this collection are illustrated in Figure 8.

The training samples included those visually collected by an interpreter, as well as samples provided by the [SGB/CPRM](#) (Brazilian Geological Service) and subsequently verified by the interpreter. In total, 2,868 samples were collected, encompassing the entire classification area (Figure 8). To guarantee accuracy, these training samples were augmented with samples derived from the stable pixels in Collection 8.0. These stable pixels were grouped into Natural (Forest, Savanna, Wetland and Grassland formations) and Anthropogenic (Pasture, Agriculture and Mosaic of Uses) classes, with a maximum of 13,000 samples per class and a minimum of the total number of rocky outcrop samples. The model was trained in Google Earth Engine using the function `ee.Classifier.smileRandomForest`, with `ntree = 300`, and applied to the Landsat mosaics by using the “Multiprobability” approach, similar to “General Map” classification workflow.

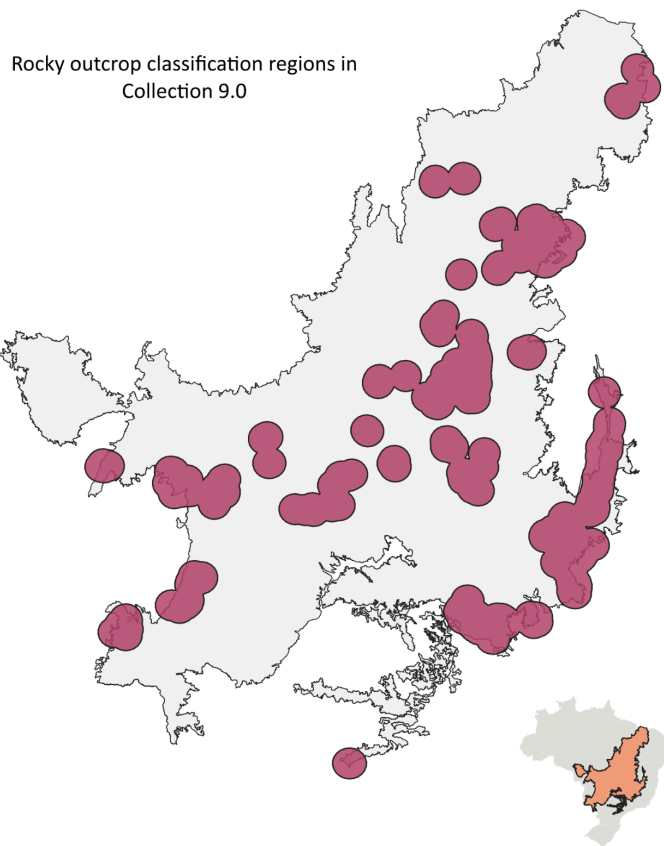


Figure 8. The rocky outcrop classification area used in collection 9.0. Highlighted in orange is the location of the Cerrado biome in Brazilian territory.

6.3. Post-classification filters

























The post-processing filters that have been implemented are consistent with the logic that was discussed earlier in section 5, namely Gap-Fill, Frequency, and Spatial filters.

- Gap-Fill: filter is based on a temporal approach, whereby classifications from subsequent years are employed to fill in pixels with no data. The objective is to guarantee the continuity and temporal consistency of rocky outcrop classification.
- Frequency filter: to regulate the rocky outcrop class over time, given that this class does not exhibit significant dynamics of change in terms of cover or use. Consequently, a pixel is classified as a rocky outcrop if it is present in at least 50% of the time series observations.
- Spatial filter: employed to eliminate spurious pixels that may appear in the classification result. This filter removes isolated pixels using a connectivity criterion of 10 pixels, equivalent to an area of 0.9 ha

7. INTEGRATION

The integration stage entails the overlaying of data generated on an annual basis. In the initial phase, the rocky outcrop classification is superimposed on the land use and land cover maps. An exception rule is applied to avoid the rocky outcrop class overlapping the grassland class in the Chapada dos Veadeiros region. Subsequently, the cross-cutting themes are integrated with the biome maps for each year, spanning the period from 1985 to 2023. This procedure is conducted in accordance with a set of clearly defined MapBiomias prevalence rules (Table 6).

Table 6. General prevalence rules - MapBiomias Collection 9.0

Class	Pixel value	Prevalence order	Color
Mining	30	1	
Beach, Dune and Sand Spot	23	2	
Mangrove	5	3	
Aquaculture	31	4	
Hypersaline Tidal Flat	32	5	
Urban Infrastructure	24	6	
Rocky Outcrop	29	7	
Sugar Cane	20	8	
Soybean	39	9	
Rice	40	10	
Cotton	62	11	
Other Temporary Crops	41	12	
Forest Plantation	9	13	
Coffee	46	14	
Citrus	47	15	
Other Perennial Crops	48	16	
River, Lake and Ocean	33	17	
Other Non Vegetated Areas	25	18	
Forest Formation	3	19	
Savanna Formation	4	20	
Wetland	11	21	
Grassland Formation	12	22	
Pasture	15	23	
Mosaic of Uses	21	24	

It should be noted that exceptions apply in protected areas:

- 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. ACCURACY METRICS

The accuracy analysis of Collection 9.0 was conducted using a dataset provided by LAPIG/UFG, comprising approximately 20,000 reference samples for the Cerrado biome, ranging the period from 1985 to 2022 and representing various classes from the MapBiomias legend. The samples were classified by interpreters with expertise in Cerrado vegetation, thereby ensuring a high level of knowledge in the classification process. Figure 9 depicts the distribution of the number of samples per classification region.

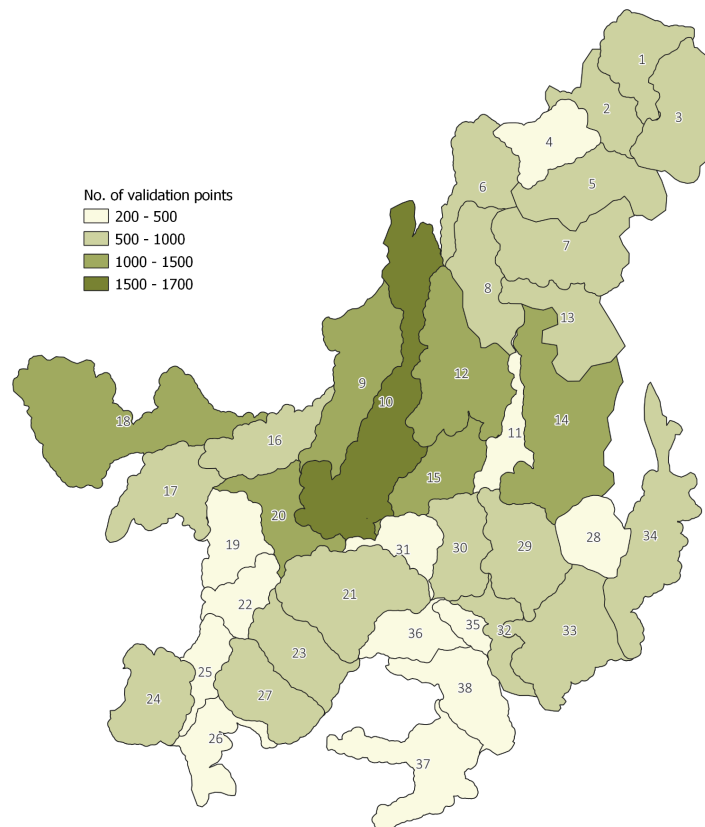


Figure 9. Number of validation samples per Cerrado biome classification region (year: 2005)

The accuracy analysis included the calculation of global and per-class accuracy, as well as the identification of errors of omission and commission, and errors of quantity and allocation. These calculations were based on the confusion matrix, which compared the reference dataset with the sample pixels from the integrated (public) version of Collection 9.0. The results demonstrated that the mean overall accuracy of Collection 9.0 was 87% at Level 1 and 80% at Level 3, representing an increase of 1.2% and 2.5%, respectively, in comparison to Collection 8.0. These findings situate Collection 9.0 as the most accurate of the previous collections (Figure 10). The observed accuracy metrics underscore the complexity of the Cerrado biome classification, but also show that the enhancements incorporated into Collection 9.0 had a notable advancement in the overall quality of the mapping. All accuracy metrics are accessible at <https://mapbiomas.org/accuracy-statistics>.

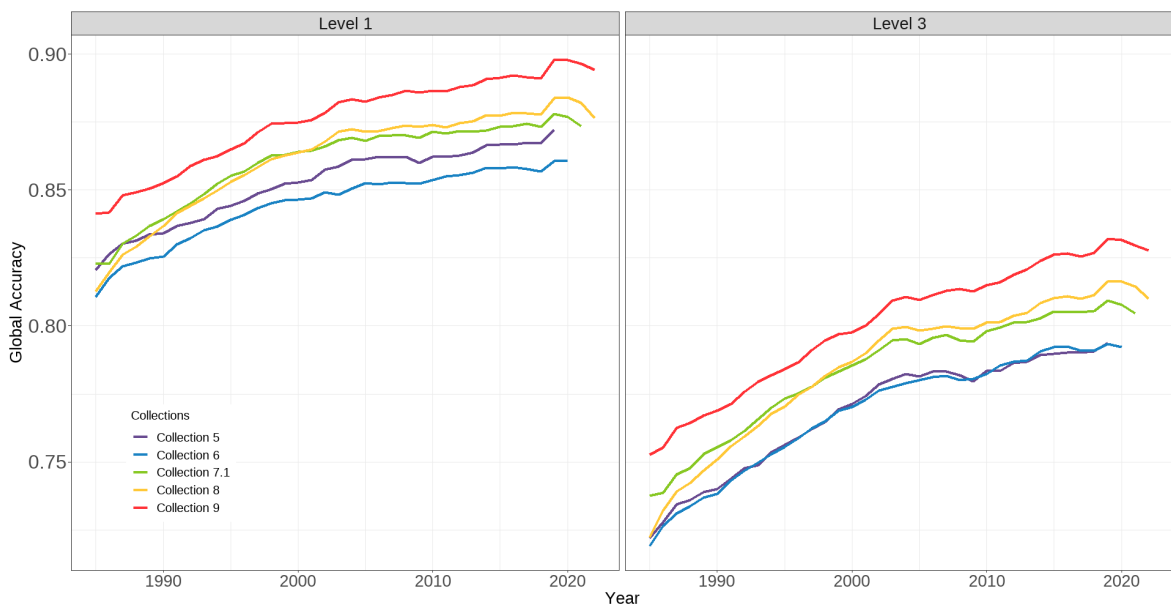


Figure 10. Global accuracy for the Cerrado biome at legend level 1 and level 3. The x-axis represents the years (from 1985 to 2022), while the y-axis represents the global accuracy value (from 0 = low accuracy to 1 = high accuracy). The colored lines indicate the accuracy per year of the current collection (9.0 - red line) and the previous collections (8.0, 7.1, 6, 5, 4.1 and 3.1 - yellow to purple lines). The overall average accuracies over the whole period for the last three collections are indicated next to the respective lines.

9. REFERENCES

Alencar, A., Z. Shimbo, J., Lenti, F., Balzani Marques, C., Zimbres, B., Rosa, M., Arruda, V., Castro, I., Fernandes Márcico 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., Arruda, V.L.S., Silva, W., Conciani, D., Costa, D., Crusco, N., Duverger, S., Ferreira, N., Franca-Rocha, W., Hasenack, H., Martenexen, L.F., Piontekowski, V.J., Ribeiro, N., Rosa, E.R., Rosa, M., 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>

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.

Gao, B. C. (1996). NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3), 257–266.

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.

Ganerød, A. J., Bakkestuen, V., Calovi, M., Fredin, O., Rød, J. K. (2023). Where are the outcrops? Automatic delineation of bedrock from sediments using Deep-Learning techniques. *Applied Computing and Geosciences*, 18, 100119.

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.

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote sensing of environment*, 25(3), 295-309.

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.

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.

Tadono, H., Ishida, F., Oda, S., Naito, S., Minakawa, K., Iwamoto, H. (2014). Precise Global DEM Generation by ALOS PRISM. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-4, 71-76.