



Cerrado - Appendix

Collection 8

Version 1

General coordinator

Ane A. Alencar

Team

Bárbara C. da Silva

Dhemerson E. Conciani

Felipe E. B. Lenti

Joaquim J. S. P. Pereira

Julia Z. Shimbo

Luis F. Martenexen

Vera L. S. Arruda

Wallace V. da Silva

1. OVERVIEW OF THE CERRADO CLASSIFICATION

The classification approach for the Cerrado biome in the MapBiomias project involved the application of decision trees to generate yearly 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 underwent refinements, resulting in significant improvements from the first MapBiomias Collection to the current version. The overall classification process for Cerrado native vegetation encompassed several steps. Firstly, the optimal time of year for constructing annual Landsat mosaics was selected. Then, remote sensing metrics were defined as potential predictors (feature space). Reference training samples were created to calibrate the classification algorithm. Post-classification treatments were applied to eliminate noise and generate 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 evolution of Cerrado native vegetation (NV) classifications is available in Table 1.

In the initial two collections, empirical decision trees were employed as the classification approach, with nodes defined based on expert knowledge of the spectral features of each class. Collection 1.0 covered the period from 2008 to 2015 and was published in 2016, while Collections 2.0 and 2.3 covered the period from 2000 to 2016 and were published in 2018. The Random Forest (RF) method was implemented for classification in Collection 2.3. Subsequently, the empirical decision tree was used 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 expanded the covered period to 1985–2017, and a methodological paper was published (Alencar et al., 2020). Collections 4.0 and 4.1 exhibited notable enhancement in the precision of mapping compared to their predecessors and abandoned the use of empirical decision trees to generate training samples. Instead, these collections relied on collecting training samples based on stable samples from the previous 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 delimit the area for collecting training samples for NV classes. Collection 6.0 expanded the classified time series (1985–2020), included a new NV class (Wetland), implemented the surface reflectance mosaic, refined the feature space, and used more reference NV maps to filter the training samples (“Inventário Florestal do Estado de São Paulo” and “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. Additionally, the Wetland class was classified into the general map, contrary to Collection 6.0, where it was

a pseudo-cross-cutting theme, by including the Height Above the Nearest Drainage (HAND) as a predictor in the feature space. The changes made in Collection 7.1 pertain to the application of new temporal filter rules to the last year to avoid minor overestimation of NV loss, as observed in previous results.

The present Collection (8.0) updated the temporal range (1985–2022) and introduced three significant methodological advancements. Firstly, a complete regionalization of the classification workflow was achieved, encompassing hyperparameter calibration, training sample selection, and classification models. The delineation of regions was based on the ecological attributes of the Cerrado landscapes, such as seasonality (Figure 1). Secondly, an extensive revision of the functioning of temporal filters applied to the raw annual classifications was conducted. Although the rationale behind this phase remained largely unchanged since Collection 3.0, earlier collections included rules concerning new classes. Therefore, a new filtering strategy was developed to eliminate false NV loss and NV gain-type transitions throughout the time series. The third major innovation in Collection 8.0 is the expanded mapping of Rocky Outcrop areas, accompanied by a refined classification strategy specific to this theme. Additionally, an improvement was made in the classification flow by refining the spatial mask used for training sample selections. Notably, NV samples falling within MapBiomas Alert and SAD Cerrado polygons for deforestation during the 2019–2022 period were excluded. All the classification and post-process scripts used in the Cerrado biome are available at: <https://github.com/mapbiomas-brazil/cerrado>.

Table 1. Overview of Cerrado collections since their first version. In the method column, “EDT” means “Empirical Decision Tree”, and “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	<ul style="list-style-type: none"> - Improvement in training samples quality by confronting with new reference maps for NV
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 consistency through 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 the spatial contiguity among classification regions - Vegetation dynamics product (vegetation loss 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) - Improvements in the statistical methodology applied to define the 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	Forest, Savanna, Wetland, Grassland, Rocky Outcrop, Mosaic of Uses, Other Non-vegetated Area, Water	<ul style="list-style-type: none"> - Improvement of temporal filter rules in the last year (2021)
8.0	1985 – 2022	RF	Forest, Savanna, Wetland, Grassland, Rocky Outcrop, Mosaic of Uses, Other Non-vegetated Area, Water	<ul style="list-style-type: none"> - Regionalization of the workflow, including sample selection, hyperparameter calibration, and classification model - Extensive revision of the temporal filtering strategy and rules. - Expansion of the classification of the Rocky Outcrop theme

The initial four collections utilized a grid at a 1:250,000 scale as the primary classification unit, where each grid cell (n = 172 tiles) was independently analyzed by the

classification algorithm. However, this approach often resulted in inconsistent contact lines between grids, leading to undesirable classification boundaries. With the introduction of Collection 5.0, a new set of classification units was implemented based on the regional variation of biophysical and land-use attributes. To achieve this, the Cerrado 19 ecoregions proposed by Sano et al. (2019) were subdivided, taking into account Brazil's major watersheds and the regional-scale spatial pattern of land-use and land-cover classes in Collection 3.1 (2017). As a result, 38 final regions were defined, replacing the need for regular grids and better compartmentalizing the environmental heterogeneity typical of the Cerrado biome. Such heterogeneity has the potential to affect the spectral signatures of NV, even within the same NV class.

In Collection 8.0, the same classification regions as in Collection 7.0 were utilized, and the number of regions remained at 38. However, we modified their perimeters while considering the NV seasonality. To achieve this, we computed the Normalized Difference Vegetation Index (NDVI) between 2017 and 2020 for each available Sentinel 2 (SR) scene. By conducting a per-pixel subtraction of the 90th and 10th percentiles (p90-p10), we were able to discriminate regions with high NV seasonal variation. This product was then used to adjust our classification regions empirically, ensuring that areas with distinct phenology and spectral signatures did not integrate into the same classification region (Figure 1).

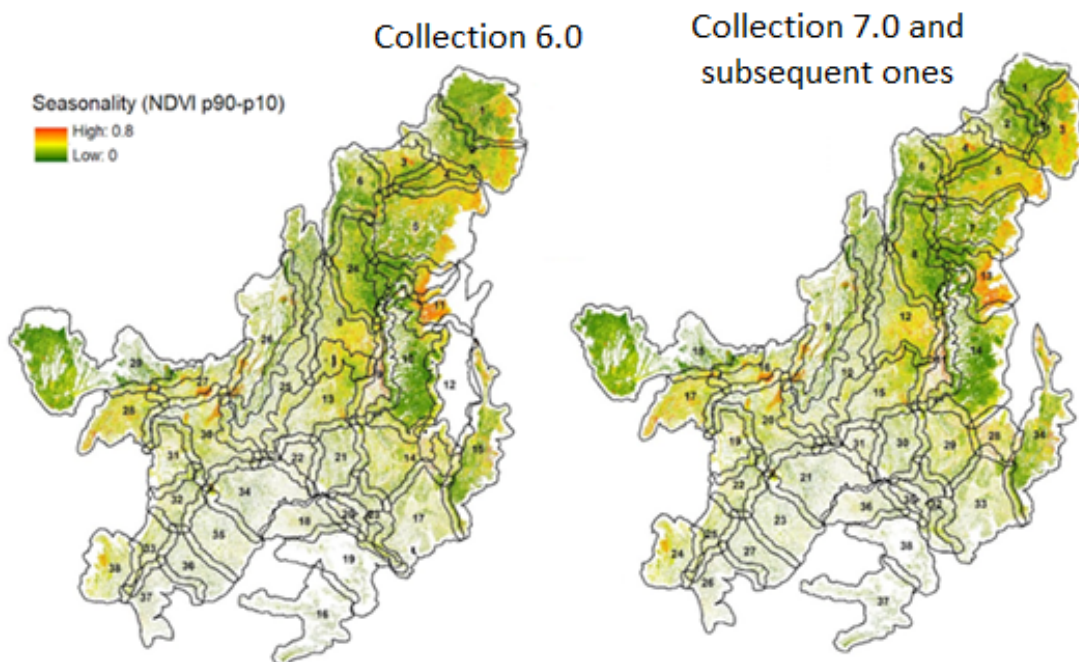


Figure 1. Classification regions used in Collection 6.0 (left) versus in Collection 7.0 and all subsequent collections (right). The regions are depicted as black polygons, with an NV seasonality map in the background. The NV seasonality is represented by different colors, with green indicating low seasonality ($p_{90}-p_{10} \geq 0$ and ≤ 0.4), and yellow and red indicating medium and high seasonality ($p_{90}-p_{10} > 0.4$), respectively. Each polygon is labeled with its corresponding ID number. Collection 8.0 retained the same classification regions as in Collection 7.0.

2. LANDSAT IMAGE MOSAICS

The initial step in classifying the native vegetation of Cerrado involved generating the mosaic of images used in the classification process. Until Collection 5.0, the classification of Cerrado NV utilized Landsat 5 (TM), 7 (ETM+), and 8 (OLI) top of atmosphere (TOA) data. However, since Collection 6.0, the TOA data was abandoned in favor of surface reflectance (SR) data. The mosaic of images is created by composing pixels extracted from all the available images during a defined period within 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.

Several tests were conducted to determine the optimum period of images to compose the annual mosaics. Due to the impact of seasonality on the spectral response of Cerrado vegetation, compositions of images from both the rainy and dry seasons were assessed. Tests included the classification of images from the end of the rainy season, when the 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. Additionally, tests were carried out with image compositions from the end of the dry season, covering the months between July and September. The results of these tests showed that using images from the rainy season would lead to a greener overall mosaic, but with an increased likelihood of commission errors in the classification of the Forest class. However, if images acquired in the last three months of the dry season were chosen, the mosaic would be drier, resulting in an underestimation of forest coverage, primarily due to the reduced potential to map deciduous forests (Figure 2).

Based on the tests described above, a large window was chosen to select the initial and final dates for generating the mosaics. These dates were standardized across all 38 classification regions and for all years under consideration. The selection criteria involved utilizing a six-month window between the months of April and September, with a maximum limit (Figure 3). The pixels identified during this defined period were found to be more effective in resolving the mapping issues that arose from the narrower window tests. To ensure the quality of yearly mosaics over the Cerrado biome, a visual inspection was conducted. In Collection 8.0, we used Landsat 5 data from 1985 to 2010, with the exception of the years 2001 and 2002 due to technical failures in the TM sensor. Instead, Landsat 7 data was used in these years. Furthermore, Landsat 7 data was also used in 2011 and 2012. Landsat 8 data was used from 2013 to 2022. As a result, we obtained 38 Landsat surface reflectance mosaics, ranging from 1985 to 2022 (Figure 4).

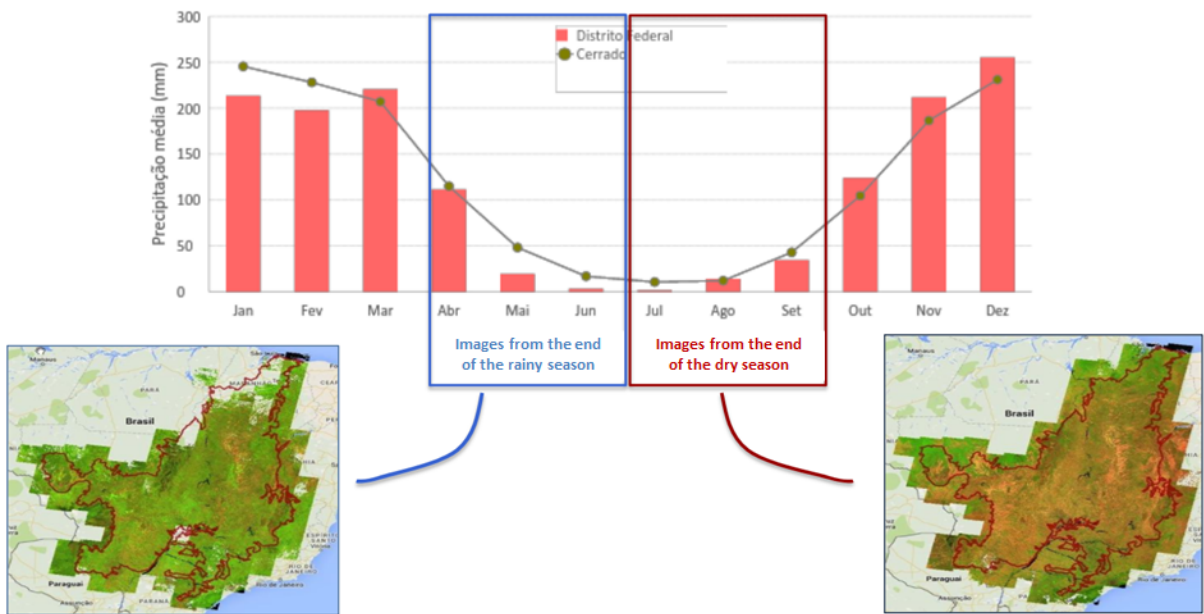


Figure 2. False-color (SWIR1-NIR-Red) composite mosaics at the end of the rainy season and the end of the dry season in the Cerrado biome.

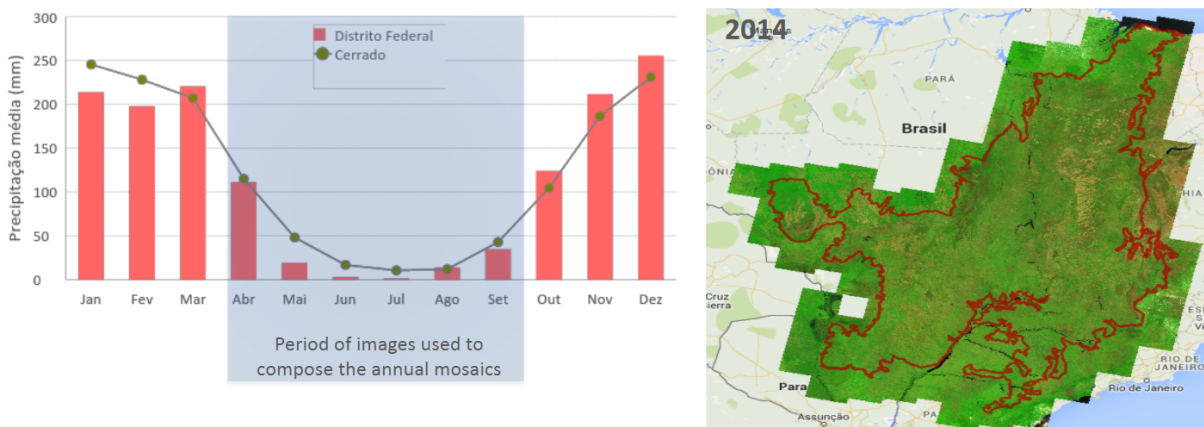


Figure 3. Time-window used to build the yearly classification mosaics used in the MapBiomias Collection 8.0 in the Cerrado biome.

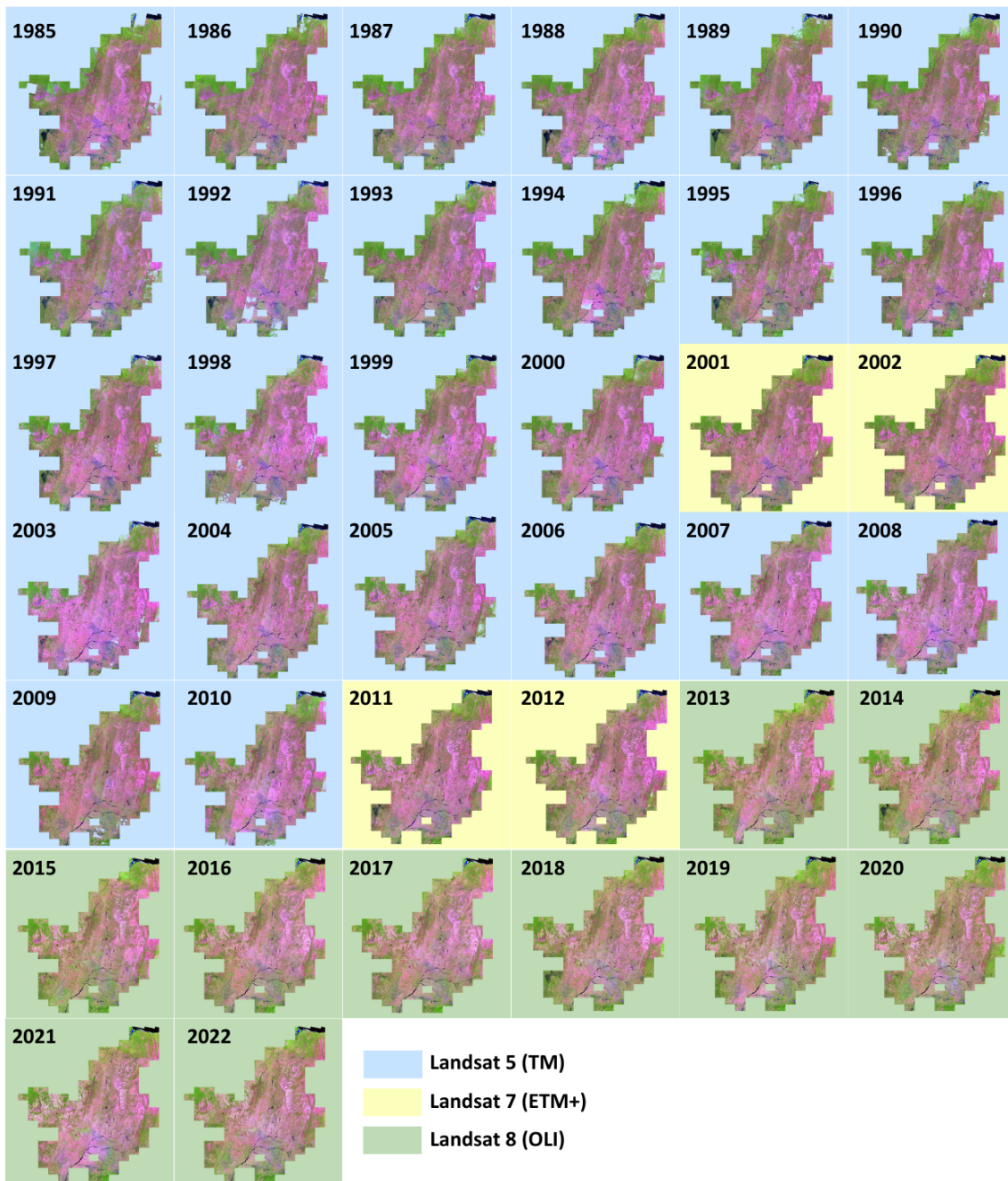


Figure 4. Annual Landsat mosaics for the Cerrado biome from 1985 to 2022. These mosaics are derived from the medians of SWIR1-NIR-Red bands. The blue box represents mosaics from Landsat 5 (TM), the yellow box represents mosaics from Landsat 7 (ETM), and the green box represents mosaics from Landsat 8 (OLI).

3. CLASSIFICATION

The classification process employed in Collection 8.0 was carried out through the application of region-specific Random Forest models, each individually calibrated for

different regions and years, utilizing training samples derived from consistent classifications in Collection 7.1. To ensure accuracy, a GEDI-based methodology was employed to remove outliers from stable pixels. The resulting classification scheme comprised ten land use and land cover (LULC) classes, which included the integration of the Rocky Outcrop theme. In addition to classifying the native vegetation (NV), anthropogenic classes like Pasture and Agriculture were also charted to complete the Cerrado landscape mosaic. In the subsequent subsections, the procedures adopted in the Collection 8.0 classification are presented: training samples and parameters (3.1), hyperparameter calibration and feature space selection (3.2), rock outcrop classification (3.3) and a comprehensive description of the overall classification scheme (3.4).

3.1. Training samples and parameters

During the construction of Collection 8.0, the classification process involved the utilization of Random Forest models, with each region and year having its own model. These models were calibrated using training samples extracted from stable areas in the Collection 7.0 classification over the 37-year period. To ensure accuracy and reliability, the training samples were derived from reference spatial datasets for Non-Vegetated areas and NV loss, in addition to employing a GEDI-based methodology that effectively eliminated outliers from stable pixels. The procedure used for eliminating outliers remained consistent with that of Collection 7.1. Moreover, to enhance precision, the canopy height model proposed by Lang et al. (2022) played a crucial role in excluding stable pixels with erroneous canopy height values for each NV class. The remarkable accuracy gain achieved with the introduction of this procedure in Collection 7.0, which showed an accuracy improvement of +0.9%. To identify and remove spurious pixels, the following criteria were employed:

- A. Forest Formation with canopy height lower than 6 meters
- B. Savanna Formation with canopy height higher than 12 meters
- C. Wetland with canopy height higher than 15 meters
- D. Grassland Formation with canopy height higher than 6 meters

After applying these adjustments, each classification unit (region) was allocated a sample size of 7,000 training samples, proportionally distributed according to the stable area of each classification in Collection 7.1, using the year 2000 as the reference. This allocation ensured that each class was adequately represented in the training dataset. For classes that comprised less than 10% of a region, a minimum of 700 samples was assigned to ensure sufficient representation. The class “River, Lake, and Ocean” had a specific

minimum number of samples ($n = 250$) to minimize class-specific commission errors, just as in Collection 7.1. This approach aimed to improve the accuracy of the classification results and ensure that classes with low representation were adequately accounted for during the classification procedure.

3.2. Hyperparameters calibration and Feature Space selection

In Collection 8.0, the first iteration of regionalized feature space selection for Cerrado classification was carried out. The feature space consisted of a universe of 119 variables that are common to all biomes, including annual mosaic bands listed in Table 2, as well as six variables specific to the Cerrado, previously used in Collection 7.1, as listed in Table 3. From this set of variables, 80 were selected to define the feature space for each region. The process began by randomly selecting eight annual mosaics, comprising 20% of the time series length, equivalent to 38 years for Collection 8.0. Within these mosaics, training samples were organized consistently with the main classification workflow, as illustrated in Figure 5. This regionalized feature space selection allowed for the adaptation of the classification models to the specific characteristics and variations of each Cerrado region, leading to improved classification accuracy.

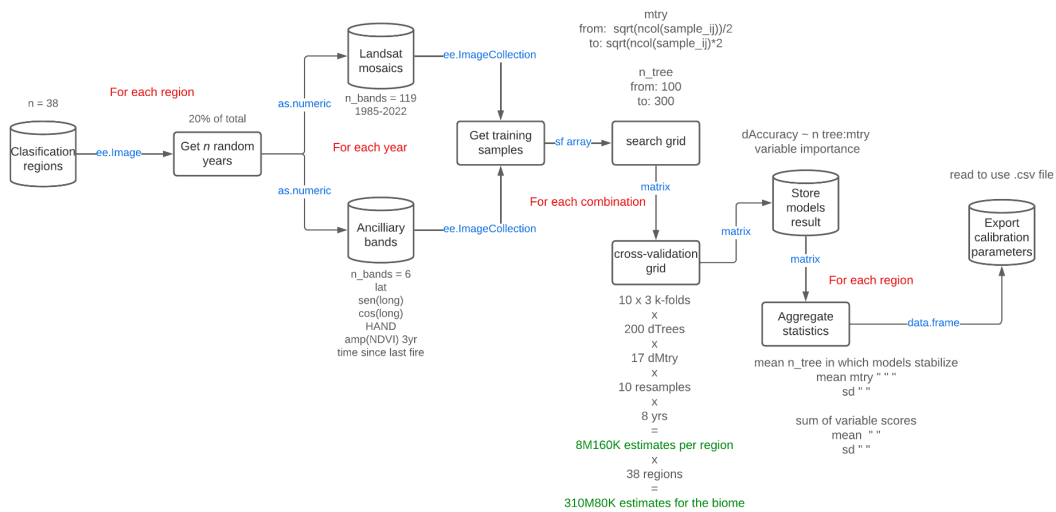


Figure 5. Hyperparameter tuning and feature-space regionalization workflow. Cylinder boxes represent input/output datasets; rectangle boxes point to key processing steps. Arrows indicate the flow of processing. Each arrow contains blue text labeling the data type/structure used in the step. Gray texts near boxes offer details about the data (cylinder) or parameters (rectangle). The red text indicates the level of the loop in which each step runs.

During the hyperparameter calibration process for Collection 8.0, Random Forest classifiers were tested with various configurations of hyperparameters, including the number of classification trees ($n_{tree} = 100\text{--}300$) and the number of randomly sampled variables for each tree ($m_{try} = 5\text{--}22$). A cross-validation analysis was executed for every hyperparameter combination into a 10×3 schema and repeated 10 times for each combination with subsampling and permutation. Subsequently, the resulting model performance statistics were compiled and retained. To determine the optimal hyperparameters for the final models, we identified the mean number of trees at which the model's overall accuracy stabilized and selected the mean value for m_{try} that yielded the highest overall accuracy. We ranked the variables based on their importance score, and the top-performing 80 variables were selected as the feature space for each region.

Table 2. Feature space subset of the 119 variables considered in the classification of the Cerrado biome in the MapBiomas Collection 8.0. 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
	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, stdDev	Burke et al., 2017
	Hall Cover	$Hall\ Cover = (- Red \times 0.017 - NIR \times 0.007 - SWIR2 \times 0.079 + 5.22)$	median, stdDev	Hall et al., 2006
Spectral Index	Normalized Difference Vegetation Index	$NDVI = (NIR - Red) / (NIR + Red)$	amplitude, median, median_dry, median_wet, stdDev	Rouse et al., 1974
	Normalized Difference Water Index	$NDWI = (NIR - SWIR1) / (NIR + SWIR1)$	amplitude, median, median_dry, median_wet, stdDev	Gao et a., 1996
	Photochemical Reflectance Index	$PRI = (Blue - Green) / (Blue + Green)$	median, median_dry, median_wet	Gamon et al., 1992
	Soil-Adjusted Vegetation Index	$SAVI = 1.5 \times (NIR - Red) / (NIR + Red + 0.5)$	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	$GVS = GV / (GV + NPV + Soil + Cloud)$	amplitude, maximum, median, median_dry, median_wet, minimum, stdDev	Housman et al., 2018
Fraction	Normalized Difference Fraction Index	$NDFI = (GVS - (NPV + Soil)) / (GVS + (NPV + Soil))$	amplitude, maximum, median, median_dry, median_wet, minimum, stdDev	Souza et al., 2005

Non-photosynthetic Vegetation Fraction	$NPV = \text{Fractional abundance of non-photosynthetic vegetation within the pixel}$	amplitude, maximum, median, median_dry, median_wet, minimum, stdDev	Souza et al., 2005	
Savanna Ecosystem Fraction Index	$SEFI = (GV + NPV_S - Soil) / (GV + NPV_S + Soil)$	median, median_dry, stdDev	Alencar et al., 2020	
Shade Fraction	$Shade = 100 - (GV + NPV + Soil + Cloud)$	median	Housman et al., 2018	
Soil Fraction	$Soil = \text{Fractional abundance of soil within the pixel}$	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

Variable importance for each regional preliminary model was assessed by measuring the mean decrease in accuracy when a particular variable was excluded from the model. We utilized the Mean Decrease in Gini coefficient, which quantifies the contribution of each variable to the uniformity of nodes and leaves in the Random Forest models. A higher Mean Decrease Gini value for a specific variable indicates its greater impact in determining the accuracy of the model. This process allowed us to refine the feature space and create region-specific Random Forest models, enhancing the precision of the classification results for the Cerrado biome in Collection 8.0. Moreover, as a result of the distinct attributes of the Cerrado biome, we have incorporated supplementary bands into our feature set, as demonstrated in Table 3.

Table 3. Complementary bands added to the Cerrado feature space in Collection 8.0.

Name	Formula	Statistics	Reference
Latitude	$ee.Image.pixelLonLat().select(['latitude'])$	identity	Geolocation
cos(Longitude)	$cos(ee.Image.pixelLonLat().select(['longitude']))$	identity	Geolocation
sen(Longitude)	$sen(ee.Image.pixelLonLat().select(['longitude']))$	identity	Geolocation
Time Since the Last Fire	$TSLF = \text{Current year} - \text{Year of the last fire}$	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

3.3. Rocky Outcrop classification

In Collection 7.0, we have incorporated the BETA version of the Rocky Outcrop classification. This classification comprises a set of rock outcrops, which are remarkably stable, and encompass sedimentary, igneous, or metamorphic features. It is worth noting that certain parts of "campos rupestres" may also be included in this classification. The Rocky Outcrop class is characterized by monolithic or clustered elevations that stand out as isolated features in the surrounding landscape. It is frequently observed in regions with arid and semi-arid climates, where vegetation is sparse or even absent, as depicted in Figure 6. Due to its distinctive features and geological significance, rocky outcrops often attract anthropogenic activities, particularly mineral extraction.

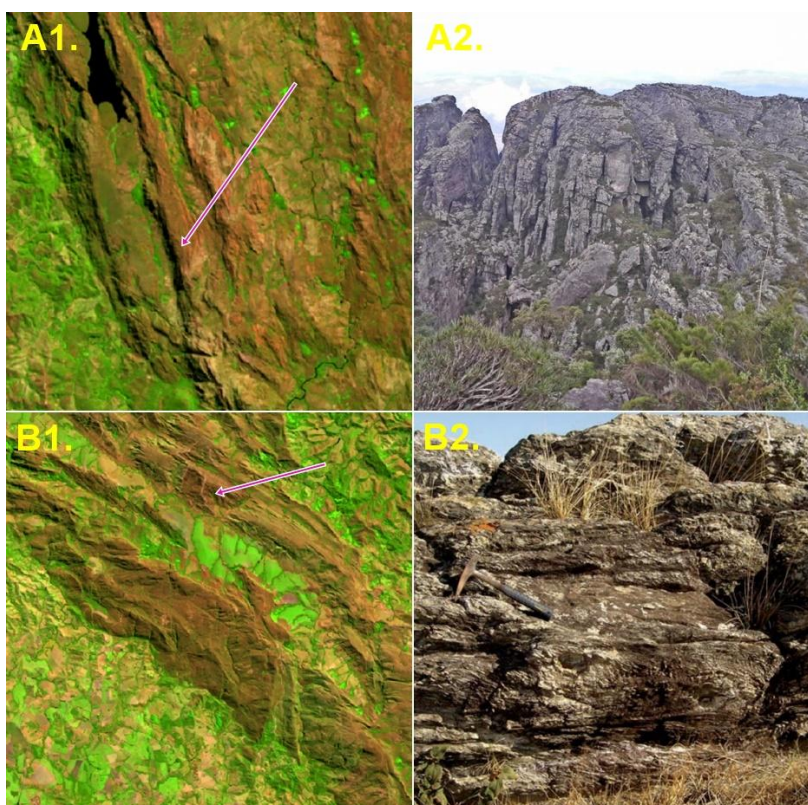


Figure 6. Example of landscapes mapped as Rocky Outcrop in the Collection 8.0. A) “Serra do Espinhaço”; A1) Landsat false-color composition (SWIR1-NIR-Red) for the year 2021—Pink arrow point the approximated localization of the field photo; A2) Field photo (credits to TMbux); B) “Serra da Canastra”; B1) false-color composition (SWIR1-NIR-Red) for the year of 2021; B2) Field photo (credits to Mario L.S.C Chaves).

Despite the independent implementation of the Rocky Outcrop's classification, the methodological approach used was similar to that applied for the overall map. The classification schema for Rocky Outcrops is presented in Figure 7, depicting the specific criteria and features used to distinguish this class from other land cover types in the Cerrado biome. The approach aimed to accurately identify and delineate the distinct rocky outcrop areas in the region, accounting for their geological characteristics and ecological importance. The classification provides a representation of the distribution and extent of rocky outcrops within the Cerrado biome.

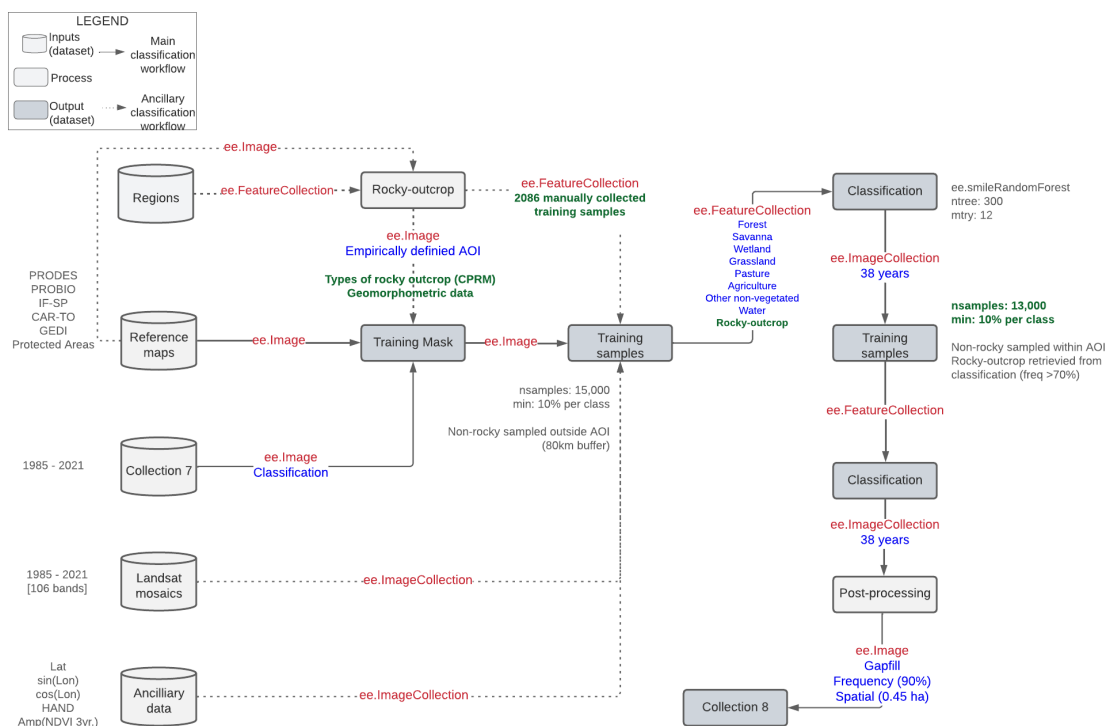


Figure 7. Overview of the methodology for the Rocky Outcrop classification. Each gray geometry (cylinders for databases and rectangles for processes) represent a key step in the classification schema—with the respective name inside. The gray text near databases and processes offers a short description of the step, while the green text highlights the main differences among stepwise classification. Arrows with a continuous black line connecting the key steps represent the main direction of the processing flux. In contrast, 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.

In Collection 8.0, the Rocky Outcrop classification underwent refinements and improvements to enhance its representation in the Cerrado biome. The BETA version of this classification was initially introduced in Collection 7.0, and since then, efforts were made to further optimize and refine the process. The overall processing flow for the Rocky

Outcrop classification in Collection 8.0 remained similar to that used in Collection 7.0. However, to increase the accuracy and coverage of rocky outcrop mapping, additional classification areas were incorporated, allowing a more comprehensive identification of these geological features in the Cerrado biome. Moreover, adjustments were made to the selection of samples per year, fine-tuning the classification process for improved results. Through these refinements and improvements, Collection 8.0 provides an updated and enhanced representation of rocky outcrops in the Cerrado biome.

A series of tests were performed to determine the most suitable strategy for the inclusion of the Rock Outcrop in the native vegetation (NV) map of the Cerrado without compromising the overall quality of the map. The classification of rock outcrops was addressed in an independent workflow, delimiting an empirical area of interest (AOI). The processing was divided into two steps, initially by visual inspection and later using the stable pixels from the classification obtained in the first round independently.

During the first stage, an interpreter examined Landsat images covering all classification regions of the Cerrado to identify the main rocky outcrops in the biome. Three types of outcrops were established based on this analysis: Bedrock, Slope/Erosion Front, and Lajedos, considering the different geological formations present in the biome. The visual inspection was supported by reference data from CPRM (Companhia de Pesquisa de Recursos Minerais) and Geomorphometric data (Geomorpho 90). Using this information, a dataset comprising 2,086 stable rock outcrop points was created, covering the period from 1985 to 2021. These samples were used in the first round of classification, defining an AOI with an 80 km buffer around each sample point.

To avoid misclassification of native vegetation, especially in pasture areas, a multi-class approach was adopted. This involved considering stable samples of the same classes present in the general map and additional stable samples of rock outcrops limited to the AOI. To balance the samples, the entire AOI was treated as a single classification unit, distributing 13,000 samples per year. The proportion of bedrock samples was fixed at 10%, while the remaining 90% was proportionally distributed according to other land covers and land uses. Post-processing filters, such as gap-fill, frequency, and spatial, were implemented to improve the classification results from this first step. For more detailed information, please refer to Section 4 of this ATBD.











The second processing step was conducted using the results of the first classification round. Instead of using the rock outcrop samples obtained by visual inspection, the stable pixels from the rock outcrop classification of the first round were chosen and treated as new samples, as were the other classes. The balance of 10% for the rock outcrop samples was maintained, considering their relative proportional area within the AOI. For the second round of classification, use “`ee.classifier.SmileRandomForest`” was used, and the post-processing filters were applied again. From the results obtained, only the rock outcrop class (29) was kept, discarding all other classes classified in that

particular workflow. The rock outcrop class was then integrated into our overall map independently of the other classes present in the native vegetation map.

3.4. Classification scheme

In the context of MapBiomias Collection 8.0, the classification of Landsat mosaics for the Cerrado biome encompassed a total of ten land use and land cover (LULC) classes, as detailed in the MapBiomias legend (Table 5). Beyond classifying NV, we also included the anthropogenic classes of Pasture and Agriculture to complete the Cerrado landscape mosaic and ensure there were no omission or commission errors over NV classes. However, as these two classes were also charted by their corresponding cross-cutting themes, we reclassified all pixels consisting of Pasture or Agriculture that we had previously charted into the class of "Mosaic of Uses" (21). Finally, we integrated the resultant map with the cross-cutting themes. The general methodological scheme applied to the Cerrado NV classification in Collection 8.0 is presented in Figure 8.

Table 5. Land cover and land use categories used for the Landsat mosaics classification for the Cerrado biome in MapBiomias Collection 8.0. Classes with " * " are those classified individually but converted into the Use Mosaic class before integration.

Legend class of Collection 8.0	ID	Color
Forest Formation	3	
Savanna Formation	4	
Wetland	11	
Grassland	12	
Pasture*	*	
Agriculture*	*	
Mosaic of Uses	21	
Other Non-Vegetated Areas	25	
Rocky Outcrop	29	
River, Lake and Ocean	33	

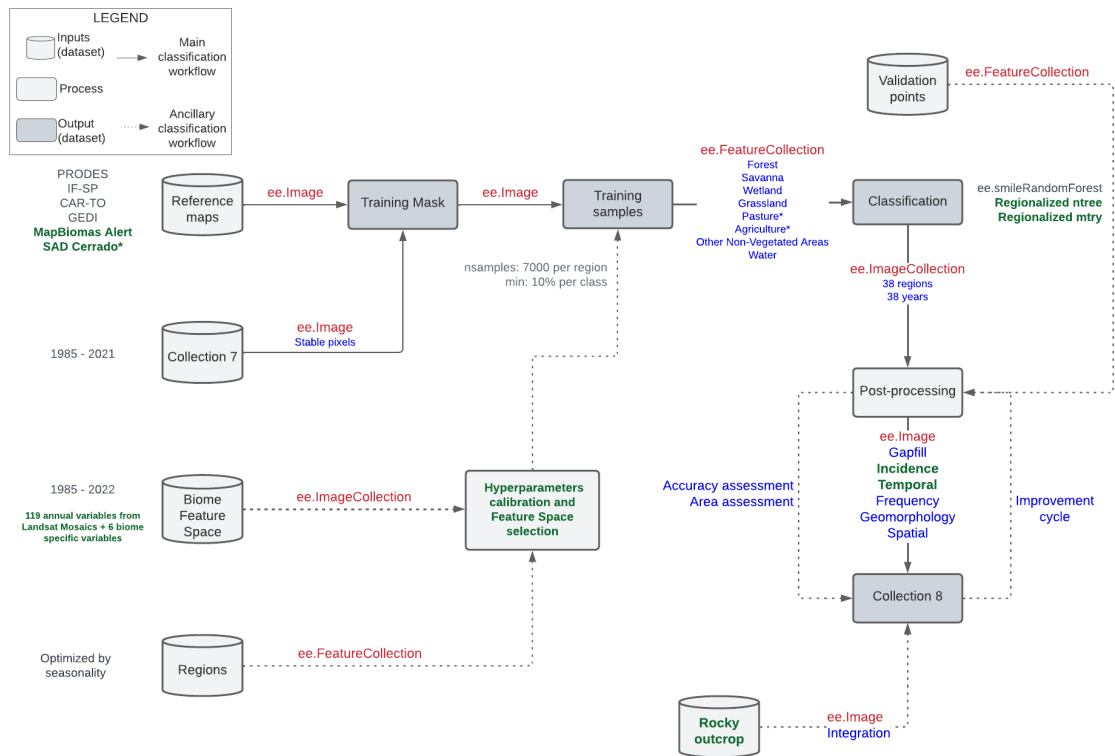


Figure 8. Overview of the methodology for Cerrado native vegetation classification in Collection 8.0. Each gray geometry (cylinders for databases and rectangles for processes) represent a key step in the classification schema—with the respective name inside. The gray text near databases and processes offers a short description of the step, while the green text highlights the main innovations in Collection 8.0. Arrows with a continuous black line connecting the key steps represent the main direction of the processing flux. In contrast, 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.

The development of the Collection 8.0 annual maps for the Cerrado biome from 1985 to 2022 was executed through a series of well-defined stages.

1. Stable areas were defined by taking into account Collection 7.1 (1985-2021), reference spatial datasets, and GEDI-based filtering for native vegetation areas. Urban area pixels were employed as a proxy for collecting samples of the Non-Vegetated Area class.
2. A visual-inspection-based training dataset was created specifically for the Rocky Outcrop class.
3. The area proportion of all classes was assessed to balance the sample set for each run of the classification model per region and per year.

4. A minimum sample size per class was set to 700 (250 for the "River, Lake and Ocean" class, covering less than 10% of the region), and a maximum sample size per class was set to 7,000.
5. The best hyperparameters for each classification region were obtained using a heuristic tune-grid, and each classifier (per region and per year) was trained using balanced samples and the novel regionalized feature space.
6. Classification was performed using Random Forest, as implemented in the 'ee.classifier.smileRandomForest' function into the Google Earth Engine platform.
7. Two distinct maps were created, one encompassing all classes in Collection 7.0 (known as the general map), and the second focused solely on mapping Rocky Outcrop within the newly expanded area of interest.
8. The classification of the Rocky Outcrop theme was integrated into the Native Vegetation map.

4. POST-CLASSIFICATION

The pixel-based classification method, employed with individual runs for each year in a long temporal series, necessitated the implementation of post-classification spatial and temporal filters to ensure consistency and eliminate classification errors. The post-classification process encompassed several filters, including the gap-fill procedure and incidence, temporal, frequency, and spatial filters, each designed to refine the classification results and enhance the accuracy of the final map. These filters, detailed below, played a crucial role in improving the overall reliability of Collection 8.0.

4.1. Temporal Gap-Fill filter

The Temporal Gap-Fill Filter played a crucial role in addressing missing data or gaps resulting from cloud-covered or cloud-shadowed pixels in the images. The filter aimed to fill these 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 instead replaced with the previous valid classification. As a result of the Temporal Gap-Fill Filter, the final classified map should generally contain very few gaps, only persisting in cases where a specific pixel remained consistently classified as no-data throughout the entire temporal series.

4.2. Incidence filter

In order to improve the accuracy and reliability of the classification results, we implemented an Incident Filter to address excessive changes between classes observed over the 38-year temporal series. Unlike previous versions, this filter specifically targeted excessive changes related to natural-to-anthropogenic and anthropogenic-to-natural transitions. Noise associated with excessive changes involving only NV classes was addressed separately using the Frequency Filter (see Section 4.4).

First, we aggregated the annual maps into three classes: Natural (Forest Formation, Savanna Formation, Grassland, Wetland), Anthropogenic (Pasture, Agriculture, Mosaic of Use), and Other (River, Lake, Ocean, and Non-Observed). We then masked the annual classifications for Natural and Anthropogenic use only and derived a map indicating the number of changes each pixel underwent during the time series, excluding transitions involving the “Other” class. In this layer, pixel labeling was executed by taking into account the quantity of neighboring pixels that underwent similar changes in number (i.e., patch size). Pixels that changed more than ten times and were connected to fewer than seven pixels with the same number of changes were identified as border pixels with noise. For these pixels, their classification was reset to the most frequent class in their original trajectory (i.e., before aggregation). Pixels that changed more than ten times and were connected to more than six pixels with the same number of changes were considered patches of spurious transition. For these pixels, their classification was Anthropogenic Use.

This approach was chosen because the underlying errors causing noise in border pixels are different from those causing larger patches with excessive class changes. Border pixels exhibit excessive class changes due to spectral mixture in Landsat pixels that contain more than one thematic target. On the other hand, larger patches with excessive class changes are likely due to NV commission errors, as such regimes of loss and regrowth are ecologically unrealistic. The threshold of ten changes was determined to achieve an average persistence time of seven years for Natural or Anthropogenic use, considering the 38-year time series. This is aligned with the growth cycle of planted forests and perennial crops in the Cerrado, which can be confused with natural vegetation by the classifier due to their spectral similarities. Figure 9 provides an illustrative example of the rationale behind this filter and highlights the improvements achieved in Collection 8.0

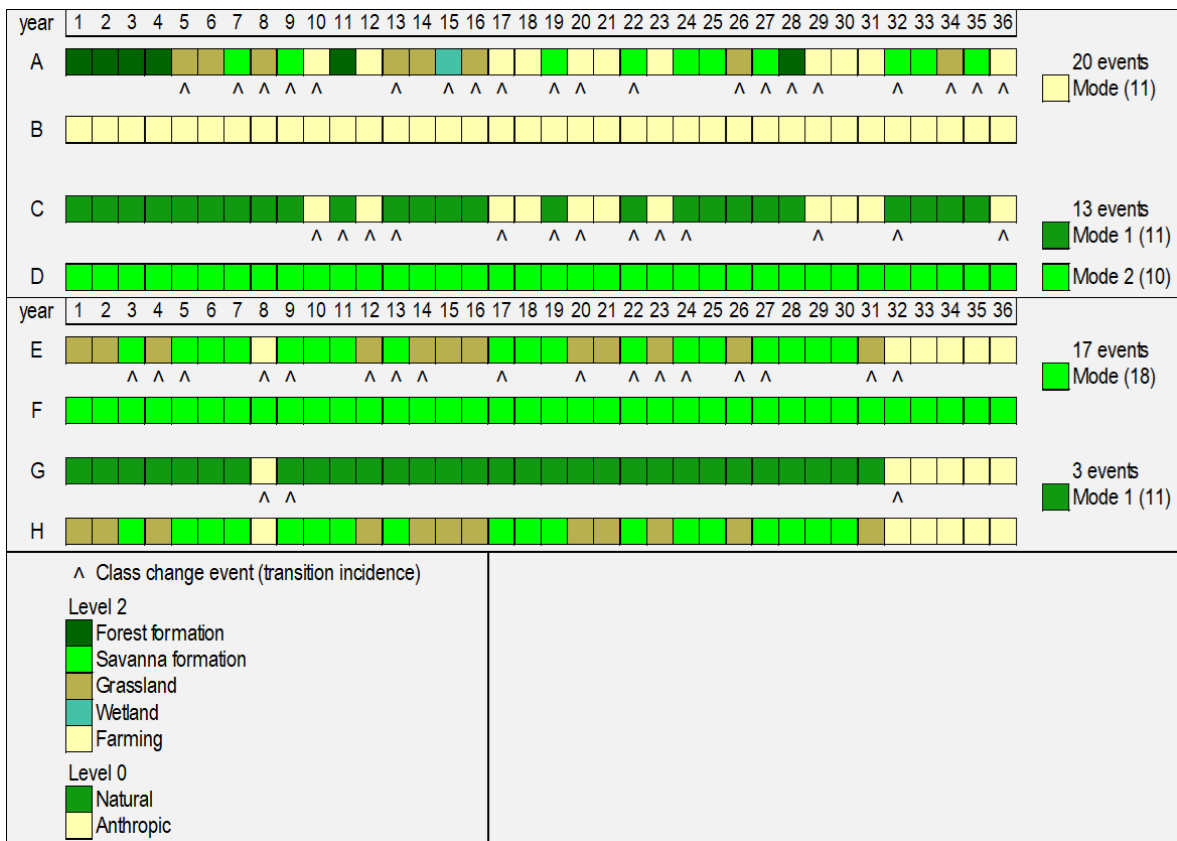


Figure 9. Two examples of the Incidence Filter used to eliminate excessive changes in classification trajectories. Sequences A and E depict the original classification trajectories as observed post the Gap Fill filtering procedure. Sequences B and F showcase the resultant trajectories after applying the Collection 7.1 Incidence Filter. Sequences C and G represent the aggregated versions of sequences A and E, serving as inputs for the Incidence Filter integrated into Collection 8. Sequences D and H demonstrate the post-filtered classifications yielded by the implementation of the Incidence Filter in Collection 8.

4.3. Temporal filter

The temporal filter implemented in Collection 8.0 plays a crucial role in addressing temporal inconsistencies (Figure 10). The filter was conducted in two distinct phases to ensure a comprehensive evaluation of temporal spurious transitions. In the first phase, the temporal filter focuses on identifying and correcting spurious transitions between the NV and Anthropic use classes. In the second phase, the temporal filter targets unrealistic trajectories that involve NV-to-NV transitions. For instance, it is not ecologically plausible for a pixel to change from Forest formation to Grassland and then revert to Forest formation within a short timeframe, such as a five-year period. The filter evaluates the temporal patterns of such NV-to-NV transitions and replaces them with more plausible classifications based on the surrounding temporal context.

The temporal filter has two phases. In the first phase, it compares the trajectories of all pixels with reference trajectories that represent the expected behavior for NV-loss or NV-regrowth events. This process is carried out iteratively, starting with the most recent years in the time series. For NV-loss, the reference trajectory is a 4-year window. The first two years are classified as NV, followed by a transition to Mosaic of Use for at least one additional year. If the focal pixel (i.e., the candidate for change) does not follow this trajectory, the conversion in the focal year is considered invalid. To handle focal pixels with spurious losses, the reclassification rule prioritizes the most recent information available in the time series. The focal pixel is then reclassified to the same class as its temporal neighbor in the following year.

A similar procedure is carried out for NV-regrowth filtering, but applying a 5-year temporal window, and the reference trajectory involves the first two years classified as Mosaic of Use. This is followed by a transition to NV for the focal pixel, and persistence as NV for two additional years. Any deviations from this reference trajectory are considered spurious changes, and the transition is invalidated by reclassifying the focal pixel to the same class as its temporal neighbor in the most recent year. It's important to note that the temporal filter is not applied to the years 1985, 1986, and 2022. These years are specifically used for evaluating persistence criteria in the temporal neighborhood of a given focal pixel. Additionally, the year 2021 is not filtered for NV-regrowth because an additional year is required in the temporal window defined for this reference trajectory.

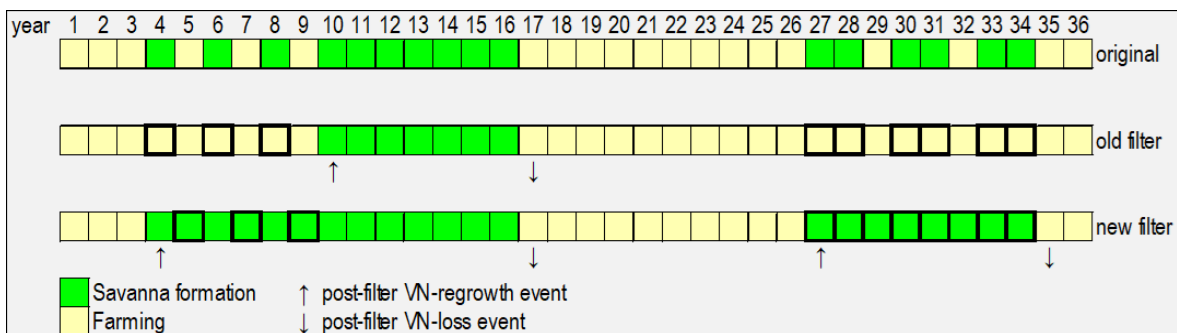


Figure 10. The effect of the Temporal Filter on a theoretical classification trajectory (original). The sequence labeled 'old filter' illustrates the filtered trajectory employing the Collection 7.1 implementation, while the sequence labeled 'new filter' corresponds to the trajectory utilizing the Collection 8 implementation (initial phase).

The second and final phase in the Temporal Filter implemented in Collection 8.0 aims at eliminating trajectories that are not coherent and ecologically realistic over time and that remain after the first phase is completed. This filter follows a series of sequential steps:

1. The filter evaluates all pixels in a 5-year (from 1986 to 2018) and 4-year (from 1986 to 2019) moving window. It corrects any pixel value that shows a specific class in the previous year (year -1), undergoes a change in the current year, and then returns to the initial class in the subsequent years (year +2 or +3). This process was applied for each class in this order: Savanna formation (4), Forest Formation (3), Grassland Formation (12), Wetland (11), Mosaic of Uses (21), River, Lake, and Ocean (33), and Other Non-vegetated Area (25).
2. Similar to the first step, the filter uses a 3-year moving window (from 1986 to 2020) to correct middle years concerning -1 and +1 years for each class in the same order as in the first step.

4.4. Frequency filter

The Frequency Filter was exclusively applied to pixels classified as native vegetation for at least 90% of the time series. For pixels meeting this criterion, if a particular class, such as Forest Formation, was assigned to the pixel over more than 75% of the time, that class was confirmed as the classification for the entire period. The same rule was applied to the Savanna Formation, Wetland, and Grassland Formation classes, but with a frequency criterion of 50% of the time series. Regarding the Rocky Outcrop class, a frequency criterion of 70% was applied in the first round of classification, and this was increased to 90% in the second round.

This application of the frequency filter resulted in a more stable classification of native vegetation classes. Notably, the 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. By considering the long-term frequency of each class assignment, the frequency filter ensures a more reliable representation of the dominant land cover types in the Cerrado biome throughout the 38-year time series. It provides an additional layer of consistency, refining the classification and minimizing uncertainties related to occasional temporal fluctuations in the pixel classification over the years.

4.5. Spatial filter

The spatial filter implemented in Collection 8.0 plays a crucial role in refining the classification accuracy by addressing misclassifications at the edges of pixel groups. It utilizes the "connectedPixelCount" function, inherent to the Google Earth Engine platform, which identifies connected components (neighbors) sharing the same pixel value. Through this approach, isolated pixels that lack the minimum required number of

connected identical neighbors are considered for further assessment. The spatial filter sets a minimum connection value of six connected pixels, corresponding to an area of approximately 0.54 hectares. This implies that for a pixel to preserve its classification, it is necessary for it to possess a minimum of six adjacent pixels that share an identical value. By setting a minimum mapping unit, the spatial filter helps eliminate spurious noise and artifacts caused by isolated pixels that do not conform to the prevailing land cover patterns within the Cerrado biome.

4.6. Geomorphological filter

The geomorphology filter plays a critical role in refining the classification of wetlands within the Cerrado biome, taking into account their spectral similarities with water bodies. Geomorphological and geological factors significantly influence the wetlands in the Cerrado region. For the filter, the data of geomorphological units from IBGE (Brazilian Institute of Geography and Statistics) was used to delineate relief units associated with wetlands and increase the accuracy of the classification. For each year, the classification process is filtered based on specific geomorphological conditions. If the land use class is “12” (Wetlands) and the geomorphological unit is “23” (“floodplain”) or “29” (“fluviolacustre plain”), the land use class is reclassified to “11” (Campestre formation). This reclassification is applied to regions where wetlands share similarities with floodplains or fluviolacustre plains. On the other hand, in areas characterized by other geomorphological units, the Wetlands classification remains unchanged.

4.7. Integration with cross-cutting themes

In the integration of cross-cutting themes and biomes' maps for each year from 1985 to 2022. This procedure was governed by a set of well-defined prevalence rules, as outlined in Table 6. Notably, there was one singular case in which the general prevalence rules for the Cerrado biome did not apply. In this instance, Pasture (15), Citrus (47), and Cotton (62) had a higher prevalence than Savanna Formation (4) and Grassland Formation (12), except within protected areas. Importantly, these prevalence rules do not apply within Environmental Protection Areas (APA). In APAs, the general rules are followed for the integration of cross-cutting themes and biomes' maps.

Table 6. General prevalence rules - Mapbiomas Collection 8.0

Class	Pixel value	Prevalence
Mining	30	1

Beach and Dune	23	2
Mangrove	5	3
Aquaculture	31	4
Salt Flat	32	5
Urban Infrastructure	24	6
Sugar Cane	20	7
Soybean	39	8
Rice	40	9
Other Temporary Crops	41	10
Perennial Crop	36	11
Coffee	46	12
Citrus	47	13
Other Perennial Crops	48	14
Temporary Crop	19	15
Forest Plantation	9	16
Rocky Outcrop	29	17
Other Non Vegetated Areas	25	18
River, Lake and Ocean	33	19
Forest Formation	3	20
Savanna Formation	4	21
Wetland	11	22
Grassland Formation	12	23
Pasture	15	24
Mosaic of Uses	21	25

5. VALIDATION

The accuracy analysis of Collection 8.0 was conducted using a dataset provided by LAPIG/UFG, consisting of approximately 25,000 reference sample pixels for the Cerrado biome. Each class of the MapBiomas legend was assigned to these samples for each year between 1985 and 2022, by interpreters trained in Cerrado vegetation, ensuring expert knowledge in the classification process. The analysis included calculations of global and per-class accuracy, as well as omission and commission errors, and quantity and allocation

disagreements, using the confusion matrix that compared the reference dataset to the sample pixels from the integrated (public) version of Collection 8.0.

The results showed that the global accuracy of Collection 8.0 was 84.7% in Level-1 (a 0.47% improvement compared to Collection 7.1) and 76.1% in levels 2 and 3, slightly lower than Collection 7.1. This makes Collection 8.0 in Level-1 the most accurate compared to the other collections (Figure 11).

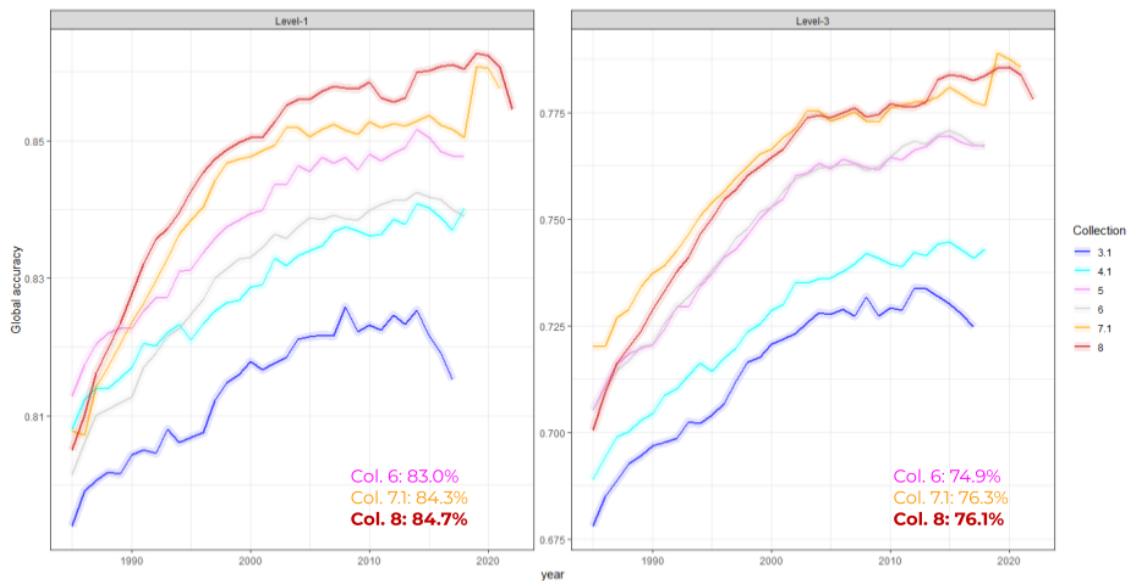


Figure 11. Global accuracy for the Cerrado biome at legend level 1 and level 3. The x-axis represents the years (from 1985 to 2022), while the y-axis represents the global accuracy value (from 0 = low accuracy to 1 = high accuracy). The colored lines indicate the accuracy per year of the current collection (8.0 - red line) and the previous collections (7.1, 6, 5, 4.1 and 3.1 - orange to dark blue lines). The overall average accuracies over the whole period for the last three collections are indicated next to the respective lines.

The improvements in accuracy, when compared to Collection 7.1, were mainly attributed to the decrease in commission errors across some of the native vegetation classes. The Forest Formation and the Savanna Formation showed commission error reductions of 13.19% and 4.51%, respectively. The Grassland exhibited a significant increase, with a commission error of 18.92%. For omission errors, the Forest Formation had a significant reduction of approximately 16.85%. For the Savannah class, a reduction of about 3.87% was observed. However, the Grassland class showed an increase in the omission error, around 7.21%. The observed accuracy metrics highlight the complexity of the Cerrado biome classification, especially in the Grassland class, but also demonstrate that the improvements implemented in Collection 8.0 reflect significant advances in the overall accuracy of the mapping and a reduction in commission and omission errors in the

Forest and Savanna classes. All accuracy metrics are available at <https://mapbiomas.org/accuracy-statistics>.

6. 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.A.C.; Arruda, V.L.S.; Silva, W.V.d.; Conciani, D.E.; Costa, D.P.; Crusco, N.; Duverger, S.G.; Ferreira, N.C.; Franca-Rocha, W.; Hasenack, H.; Martenexen, L.F.M.; Piontekowski, V.J.; Ribeiro, N.V.; Rosa, E.R.; Rosa, M.R.; dos Santos, S.M.B.; Shimbo, J.Z.; Vélez-Martin, E. Long-Term Landsat-Based Monthly Burned Area Dataset for the Brazilian Biomes Using Deep Learning. *Remote Sens.* 2022, 14, 2510. <https://doi.org/10.3390/rs14112510>

Burke, M.; Lobell, D.B. Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proc. Natl. Acad. Sci. USA* 2017, 114, 2189–2194.

Donchyts, Gennadii, Hessel Winsemius, Jaap Schellekens, Tyler Erickson, Hongkai Gao, Hubert Savenije, and Nick van de Giesen. "Global 30m Height Above the Nearest Drainage (HAND)", *Geophysical Research Abstracts*, Vol. 18, EGU2016-17445-3, 2016, EGU General Assembly (2016).

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.; Penuelas, J.; Field, C. B. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. *Remote Sensing of Environment*, v.41, n.1, p.35-44, 1992.

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. 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 Sens.* 2018, 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*.

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. Assessing the Spatial and Occupation Dynamics of the Brazilian Pasturelands Based on the Automated Classification of MODIS Images from 2000 to 2016. *Remote Sens.* 2018, 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(2018), 818–828.

Souza, C.M.; Roberts, D.A.; Cochrane, M.A. Combining spectral and spatial information to map canopy damage from selective logging and forest fires. *Remote Sens. Environ.* 2005, 98, 329–343.

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