Cerrado - Appendix

Collection 5.0

Version 1

General coordinator
Ane A. Alencar

Team
Bárbara Zimbres
Camila Balzani Marques
Felipe E. B. Lenti
Joaquim J. S. P. Pereira
Julia Z. Shimbo
Luiz Felipe Morais Martenexen
Valderli J. Piontekowski
Vera L. S. Arruda
Wallace Vieira da Silva

1. Overview of the Cerrado classification

The classification strategy of the Cerrado biome within the MapBiomas project consisted of applying decision trees to generate annual maps of the predominant native vegetation (NV) types, which were distinguished in three classes: Forest, Savanna, and Grassland. The method used to generate these annual maps evolved over time, with significant improvements from the first MapBiomas Collection to the present.

The overall approach for the classification of the Cerrado native vegetation consisted of multiple steps: 1) defining the optimum period of the year to build annual Landsat mosaics that is most useful for our purpose; 2) defining a set of remote sensing metrics to be included as potential predictors (feature space); 3) generating reference training samples to calibrate the classification algorithm; 4) applying a post-classification treatment, which includes a series of filters (gap-fill, temporal, frequency, spatial, incidence) to generate a consistent time series and eliminate noise; and finally, 4) integrating the resulting maps with the other cross-cutting themes. A visual inspection and validation analysis of the results were then conducted to evaluate the results of the classification.

In the first two Collections, empirical decision trees were applied as the classification approach, with nodes defined based on expert knowledge of the spectral features of each class. Collection 1.0 covered the period of 2008 to 2015, and was published in 2016. Collections 2.0 and 2.3 covered the period of 2000 to 2016, and were published in 2018. The classification using Random Forest was implemented in Collection 2.3, and from this point onward, the empirical decision tree was used for the purpose of generating stable samples, which were classified as the same NV type over the considered period (2000-2016). These stable samples were used to train the Random Forest models for the classification of the entire time series. Collections 3.0 and 3.1 expanded the period covered to 1985–2017, which a methodological paper was published (Alencar et al. 2020). Collections 4.0 and 4.1 saw a significant improvement in mapping accuracy in comparison to the first collections, and no longer used empirical decision trees to generate training samples. In these two collections, training samples were collected based on the stable samples from the previous collection (3.1). To further minimize eventual bias in our training dataset, for Collection 5 we used reference maps (Collection 4.1) of remaining native vegetation to restrict the area inside which to collect training samples for NV classes.

The basic classification unit in the first four collections was a grid at the scale of 1:250.000, and the classification algorithm was run independently for each grid cell (n=172 tiles). These artificial classification units tend to produce inconsistencies in the contact lines between grids. For Collection 5.0, a new set of classification units were adopted, defined on the basis of regional variation of biophysical and land-use attributes. The Cerrado 19 ecoregions proposed by Sano et al. (2019) were further subdivided, considering Brazil major watersheds and the regional-scale spatial pattern of land-use/land-cover classes in Collection 3.1 (2017). A total of 38 final regions were defined this way to substitute the need for regular grids and to better compartmentalize the environmental heterogeneity typical of the Cerrado biome, which potentially affects the spectral signatures of NV, even within the same NV class.

The general methodological scheme applied to the Cerrado NV classification in Collection 5.0 is presented below (Figure 1) . All the scripts to classify and post-process the Cerrado biome are available at: https://github.com/mapbiomas-brazil/cerrado.

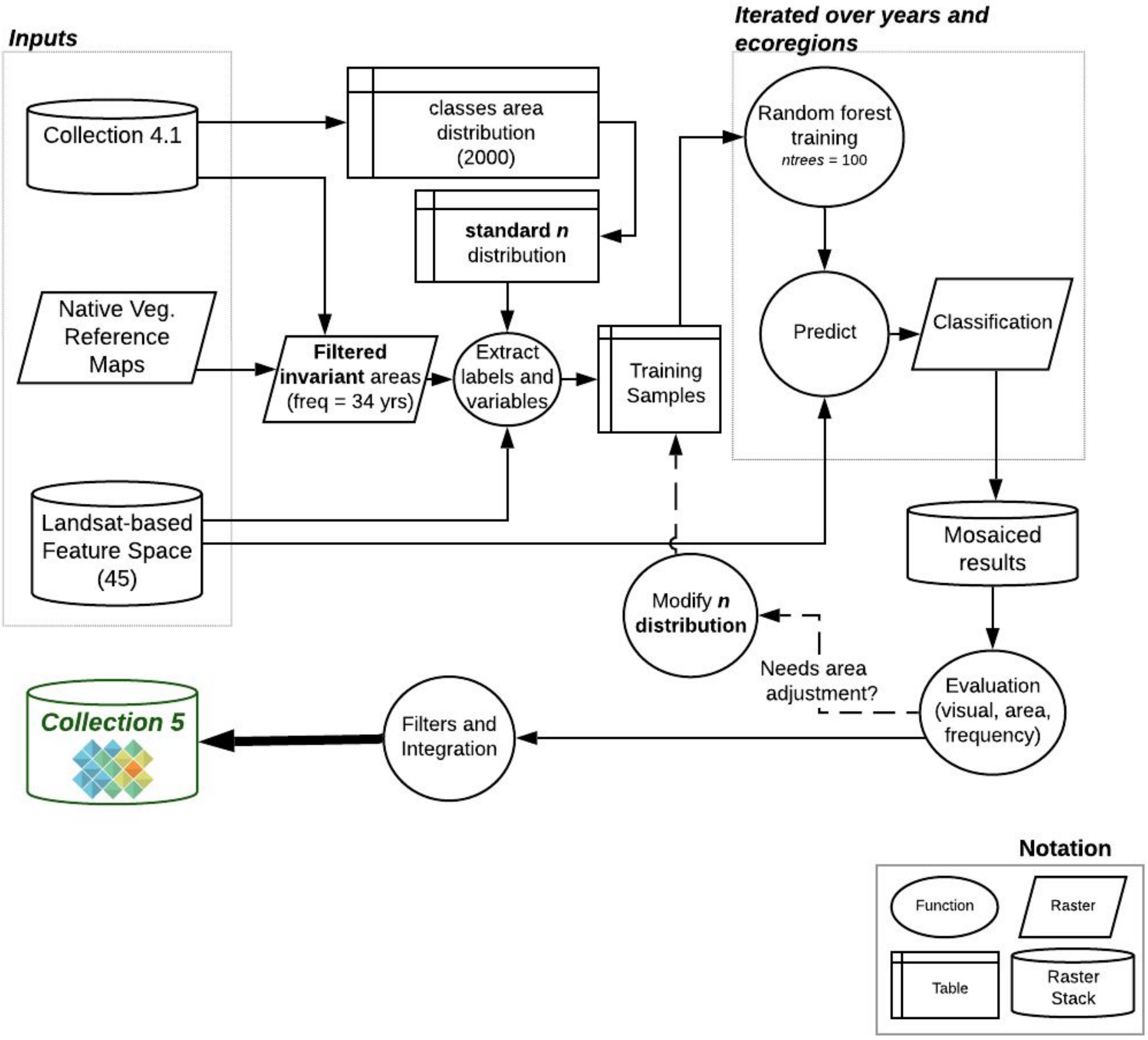


Figure 1. Overview of the method for classifying native vegetation in the Cerrado biome in MapBiomas collection 5.0.

2. Landsat image mosaics

The first step to classify the native vegetation (Forest, Savanna and Grassland) of Cerrado was to generate the mosaic of images that were used in the classification. The mosaic of images consists of a composition of the best pixels that are extracted from all the images available in a defined period within a year. Once the initial and final dates of this period were defined, the median of the images from that period was calculated, generating one median value per pixel. The aggregation of these composed pixels was conducted for each year, producing the annual mosaicked images, which were then submitted to classification.

Several tests were conducted to define the optimum period of images to compose the annual mosaics. Due to the effect of seasonality on the Cerrado vegetation spectral response, compositions of images from the rainy and dry seasons were evaluated. The tests included classifying images from the end of the rainy season when the Cerrado vegetation is still vigorous, and there is a higher probability of getting images with lower cloud cover when compared to the peak of the rainy season. Tests were also done with a composition of images from the end of the dry season which includes the months between July and September. The tests demonstrated that the use of images from the rainy season only would result in an overall greener mosaic, and the chances of increasing commission errors in the classification of the Forest class were higher. On the other hand, if we selected images acquired in the last three months of the dry season only, the mosaic would result in a drier mosaic, underestimating forest cover mainly due to the lower ability for mapping deciduous forests (Figure 2).

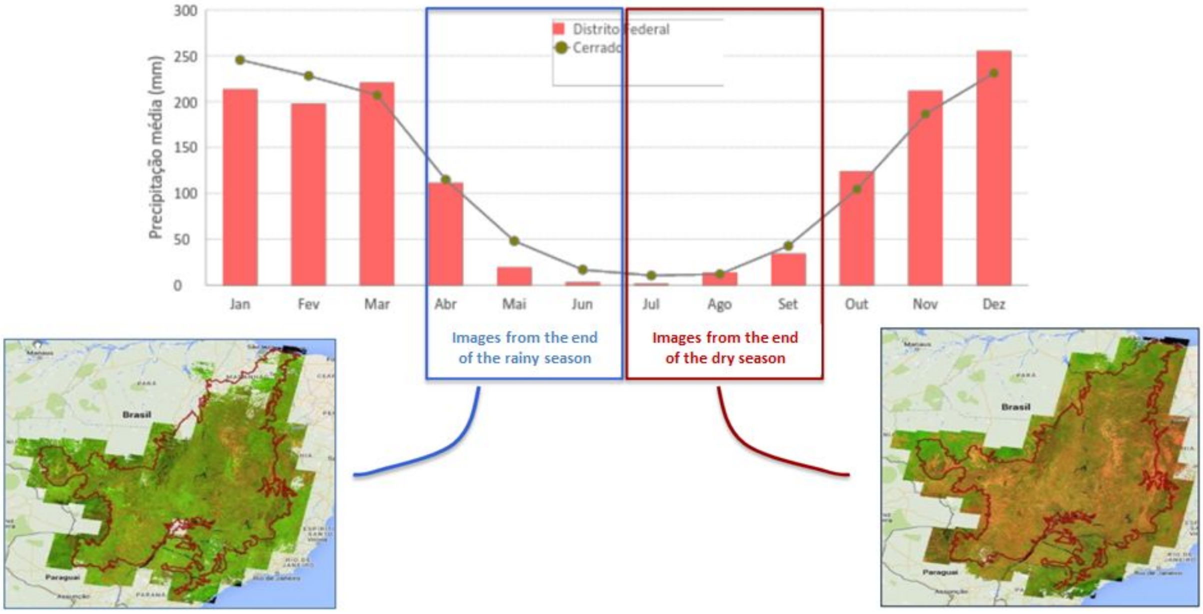


Figure 2. Pixel composite mosaics at the end of the rainy season and at the end of the dry season in the Cerrado biome.

Based on the tests described above, a large window was decided for the selection of the initial and final dates for the mosaic generation. These dates were individually selected for each of the 172 tiles and for each year. The criteria for the selection of these dates included the use of a maximum six month window between April and September (Figure 3). The median value of the pixels selected during this wide period was shown to better resolve the mapping issues which resulted from the narrow window tests. In fact, this strategy averaged the commission and omission errors between the narrow window tests. We ended up with 35 mosaics (Figure 4), by adding one year (the mosaic of 2019) to Collection 4.1 Landsat image mosaics.

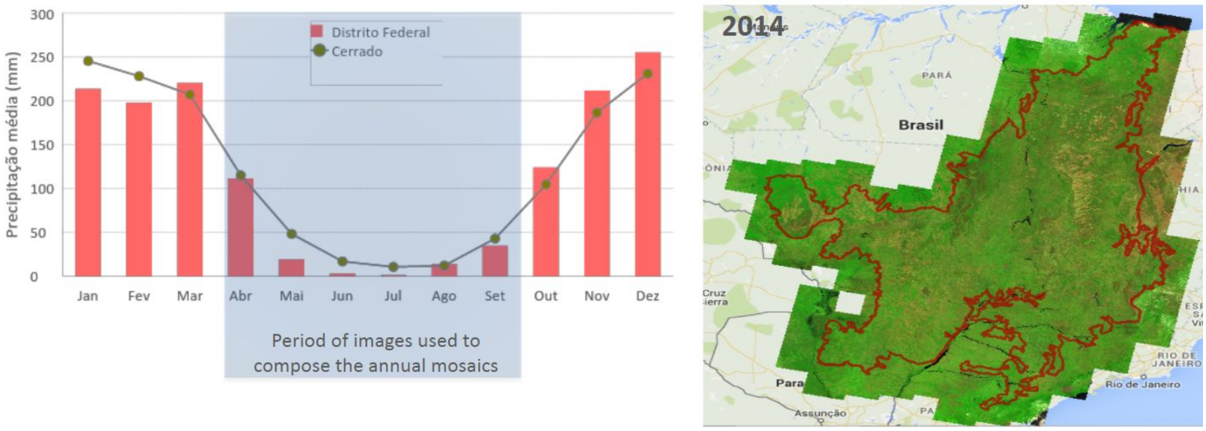


Figure 3. Window period used to define the final pixel composite mosaics used in the classification of MapBiomas Collection 5.0 in the Cerrado biome.

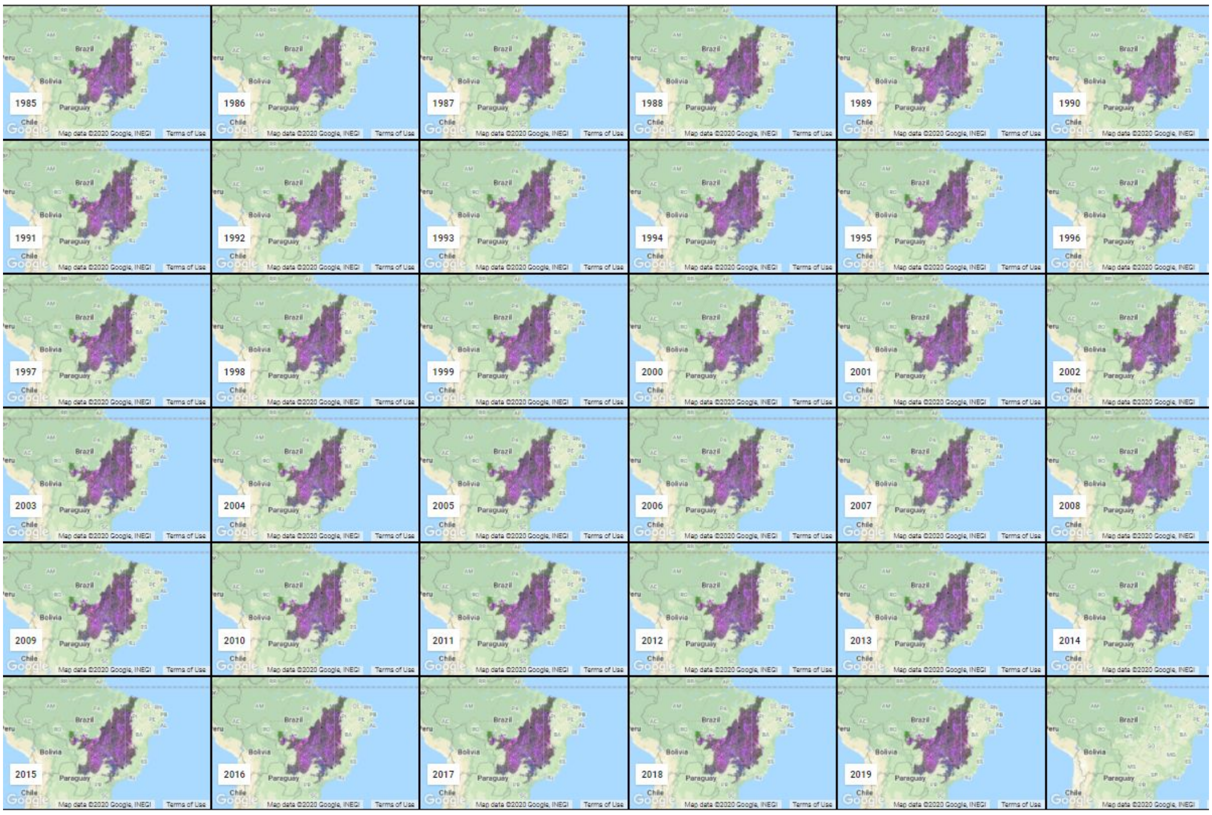


Figure 4. Annual Landsat image mosaics of the Cerrado biome from 1985 to 2019 in MapBiomas Collection 5.0.

3. Classification

Collection 5.0 was built using Random Forest models (one for each region and each year) calibrated with training samples collected based on areas with stable classification during the 34 years in Collection 4.1, and on reference maps for NV. Sample size was first set to 7,000 per classification unit (region), and distributed among classes proportionally to each class area in Collection 4.1 (year 2000). The minimum number of samples was defined as 700, in order to provide sufficient samples for those classes comprising less than 10% of a given region. Preliminary versions revealed that this threshold was not suitable for the class "River, Lake and Ocean", as it resulted in great commission errors for this class. Therefore, we defined a specific minimum number of samples for this class of 250.

Classifications were produced independently for each region and year, and the resulting time-series was post-processed using filters to enhance temporal and spatial coherence. Intermediate versions were evaluated based on visual inspection to identify regions presenting spatial discontinuities with neighbouring regions, or marked omission/commission errors for a given class. Those cases were reclassified considering a modified distribution of sample size per class, which was set by an interpreter to compensate for the proportional excess/lack of area of a given class in the evaluated version. Finally, maps representing cross-cutting themes were integrated with the final version of the Cerrado biome time-series to produce MapBiomas Collection 5.0.

3.1. Classification scheme

The classification of the Landsat mosaics for the Cerrado biome considered eight land use and land cover (LULC) classes in the MapBiomas Collection 5.0 legend (Table 1), which were later integrated with the cross-cutting theme classes as a subsequent step.

Table 1. Land cover and land use categories considered for digital classification of the Landsat mosaics for the Cerrado biome in MapBiomas Collection 5.0.

Legend class of Collection 5	Numeric ID	Color
1.1.1. Forest Formation	3	🟩
1.1.2. Savanna Formation	4	🟩
2.2. Grassland	12	🟫
3.1 Pasture	15	🟨
3.2 Agriculture	18	🟪
4.4. Other Non-Vegetated Areas	25	🟪
5.2 River, Lake and Ocean	33	🟦
6. Not Observed	27	⬜

The development of Collection 5.0 annual maps of the Cerrado biome, from 1985 to 2019, was conducted in the following steps:

(1) Definition of stable areas considering the Collection 4.1 time series (1985 - 2018) and reference maps for NV. Urban area pixels were used as a proxy for collecting samples of the Non-Vegetated Area class;

(2) Assessing area proportion of all classes in order to balance the sample set for each run of the classification model per region and per year;

(3) Training of each classifier (per region and per year) using balanced samples and the selected feature space. Minimum sample size per class was set to 700 (250 for “River, Lake and Ocean” class covering less than 10% of the region) and maximum sample size per class was set to 7000.

(4) Final classification (prediction) using Random Forest as implemented in the Google Earth Engine platform (ntrees = 100).

3.2. Feature space

The feature space used in the Cerrado biome in the Collection 5.0 was the same used in Collection 4.1. It was defined with a statistical approach by fitting several preliminary Random Forest classification models, each using a subset of 400 unique samples per class, and considering Landsat mosaics from five years (1989, 1994, 2007, 2011, and 2016). Variable importance for each case was evaluated in terms of mean decrease in accuracy when a given variable was absent in the model. We evaluated mean decrease in accuracy in general terms (global accuracy), and in specific terms (for each native vegetation class: Forest, Savanna and Grassland). The final feature space was selected among the 30 variables with the highest average importance for global accuracy, as well as the 30 top variables considering the accuracy of each of the above mentioned classes. As expected, these sets shared several variables, so that, according to these criteria, we ended up with 48 variables forming the feature space for the final training/classification (Table 2).

Table 2. Feature space subset considered in the classification of the Cerrado biome Landsat image mosaics in the MapBiomas Collection 5.0.

Variable	Description	Statistic	Temporal range	Script acronym
Green	Landsat band	minimum	year	min_green
Green	Landsat band	median	year	median_green
Green dry season	Landsat band	median	seasonal ; NDVI below first quartile	median_green_dry
Red	Landsat band	minimum	year	min_red
Red	Landsat band	median	year	median_red
Red dry season	Landsat band	median	seasonal ; NDVI below first quartile	median_red_dry
Red wet season	Landsat band	median	seasonal ; NDVI above first quartile	median_red_wet
Near Infrared (NIR)	Landsat band	median	year	median_nir

Near Infrared (NIR)	Landsat band	minimum	year	min_nir
Near Infrared (NIR)	Landsat band	Standard deviation	year	stdDev_nir
Near Infrared (NIR) dry season	Landsat band	median	seasonal ; NDVI below first quartile	median_nir_dry
Near Infrared (NIR) wet season	Landsat band	median	seasonal ; NDVI above first quartile	median_nir_wet
Shortwave Infrared 1 (SWIR 1)	Landsat band	median	year	median_swir1
Shortwave Infrared 1 (SWIR 1) dry season	Landsat band	median	seasonal ; NDVI below first quartile	median_swir1_dry
Shortwave Infrared 1 (SWIR 1) wet season	Landsat band	median	seasonal ; NDVI above first quartile	median_swir1_wet
Shortwave Infrared 2 (SWIR 2)	Landsat band	median	year	median_swir2
Shortwave Infrared 2 (SWIR 2) dry season	Landsat band	median	seasonal ; NDVI below first quartile	median_swir2_dry
EVI2	Enhanced vegetation index 2	Standard deviation	year	stdDev_evi2
EVI2	Enhanced vegetation index 2	amplitude	year	amp_evi2
EVI2 dry season	Enhanced vegetation index 2	median	seasonal ; NDVI below first quartile	median_evi2_dry
EVI2 wet season	Enhanced vegetation index 2	median	seasonal ; NDVI above first quartile	median_evi2_wet
GV	Green vegetation fraction	Standard deviation	year	stdDev_gv
GV	Green vegetation fraction	amplitude	year	amp_gv
GVS	GV / (100 - shade)	median	year	median_gvs
GVS dry season	GV / (100 - shade)	median	seasonal ; NDVI below first quartile	median_gvs_dry
Shade	Shade fraction	median	year	median_shade
NDFI	Normalized Difference Fraction Index	median	year	median_ndfi

NDFI dry season	Normalized Difference Fraction Index	median	seasonal ; NDVI below first quartile	median_ndfi_dry
NDFI wet season	Normalized Difference Fraction Index	median	seasonal ; NDVI above first quartile	median_ndfi_wet
NDFI	Normalized Difference Fraction Index	amplitude	year	amp_ndfi
NDVI	Normalized Difference Vegetation Index	median	year	median_ndvi
NDVI dry season	Normalized Difference Vegetation Index	median	seasonal ; NDVI below first quartile	median_ndvi_dry
NDVI wet season	Normalized Difference Vegetation Index	median	seasonal ; NDVI above first quartile	median_ndvi_wet
NDVI	Normalized Difference Vegetation Index	amplitude	year	amp_ndvi
NDVI	Normalized Difference Vegetation Index	Standard deviation	year	stdDev_ndvi
NDWI	Normalized Difference Water Index	median	year	median_ndwi
SAVI	Soil-adjusted vegetation index	Standard deviation	year	stdDev_savi
SAVI dry season	Soil-adjusted vegetation index	median	seasonal ; NDVI below first quartile	median_savi_dry
SAVI wet season	Soil-adjusted vegetation index	median	seasonal ; NDVI above first quartile	median_savi_wet
WEFI	Woodland ecosystem fraction index	standard deviation	year	stdDev_wefi
WEFI	Woodland ecosystem fraction index	amplitude	year	stdDev_wefi
WEFI wet season	Woodland ecosystem fraction index	median	seasonal ; NDVI above first quartile	median_wefi_wet
GCVI	Green Chlorophyll Vegetation Index	median	year	median_gcvi
GCVI	Green Chlorophyll Vegetation Index	median	seasonal ; NDVI above first quartile	median_gcvi_wet
Hall cover	Hall cover vegetation index	median	year	median_hallcover

PRI	Photochemical reflectance index	median	year	median_pri
PRI	Photochemical reflectance index	median	seasonal ; NDVI below first quartile	median_pri_dry
Slope*	Terrain slope	identity	fixed	slope

3.3. Classification algorithm, training samples, and parameters

As mentioned before, a Random Forest classifier was trained for each region using 700 to 7000 samples per class. The actual sample size for each model (region/year) was a function of the distribution of class areas in each case. Samples for each class were randomly selected both from within the target region and from its adjacent regions (*i.e.* within a buffer of 15 km). All regions were then classified using 100 decision trees.

4. Post-classification

The pixel-based classification method and the fact that the algorithm was run for each year independently in a long temporal series, a series of post-classification spatial and temporal filters was applied to increase consistency and eliminate classification mistakes. The post-classification process included a gap-fill procedure, as well as temporal, spatial, frequency, and incidence filters, each presented below.

4.1. Temporal Gap Fill Filter

No-data values (gaps) produced by cloud covered (or cloud shadow) pixels in a given image, were filled by the temporally nearest future valid classification. If no future valid classification was available, then the no-data value was replaced by its previous valid classification. Therefore, gaps should only remain in the final classified map when a given pixel was consistently classified as no-data throughout the entire temporal series.

4.2. Temporal filter

The temporal filter uses the subsequent years to replace pixels that have invalid transitions in a given year. It follows sequential steps:

1) As a first step, the filter searches for any native vegetation class (Forest, Savanna, and Grassland) that was not classified as such in 1985, and was correctly classified in 1986 and 1987, and then corrects the 1985 value.
2) In the second step, the filter searches for pixel values that were not Pasture, Agriculture, or Other Non Vegetated Areas (classes representing anthropogenic use) in 2019, but were classified as such in 2017 and 2018. The value in 2019 is then corrected to match the previous years to avoid any regeneration detection in the last year (which can not be corroborated).
3) In the third step, the filter evaluates all pixels in a 3-year moving window to

correct any value that changes in the second year (midpoint of the window) but returns to the same class in the third year. This process is applied observing prevalence rules, in this order: Pasture (15), Agriculture (18), Other non Vegetated Areas (25), River, Lake and Ocean (33), Savanna (4), Rocky Outcrop (29), Grassland (12), Forest (3).

4) The last step is similar to the third process, but consists of a 4- and 5-year moving window that corrects all middle years running in the same order of class prevalence.

4.3. Frequency filter

The frequency filter was applied only on pixels that were classified as native vegetation (no conversion transitions) throughout the time series. If such a pixel was classified as the same class over more than 60% of the period, that class was assigned to that pixel over the whole period. The results of this frequency filter was a more stable classification of native vegetation classes. Another important result was the removal of noise in the first and last year of the classification, which can not be adequately assessed by the temporal filter.

4.4. Spatial filter

The spatial filter avoids misclassifications at the edge of pixel groups, and was built based on the "connectedPixelCount" function. Native to the GEE platform, this function locates connected components (neighbours) that share the same pixel value. Thus, only pixels that do not share connections to a predefined number of identical neighbours are considered isolated. At least six connected pixels are required to reach the minimum connection value. Consequently, the minimum mapping unit is directly affected by the spatial filter applied, and it was defined as six pixels (~0,5 ha).

4.5. Incidence filter

An incident filter was applied to remove pixels that changed too many times over the 35 years. All pixels that changed more than eight times, and were connected to less than 6 same-class pixels that also changed more than eight times, were replaced by the MODE value. This avoids spurious transitions at the border of the same-class pixel group.

Savanna Formation and Grassland pixels that changed more than ten times, and were connected to less than 66 pixels that also changed more than ten times, were classified as Other Non-Vegetated Areas (Class 25). Natural classes tend to be stable, and these areas that change too often are likely not native vegetation.

As a final rule, all forest pixels that changed more than eight times, and were connected to more than 66 pixels that also changed more than eight times, were also classified as Other Non-Vegetated Areas (Class 25). This rule aims to filter out areas of commercial tree plantations mapped as Forest Formation; as the growth period for *Eucalyptus* sp. and *Pinus* sp. commercial forest stands is approximately seven to eight years.

4.6. Integration with cross-cutting themes

The cross-cutting themes and the biomes' classified maps were integrated for each of the 35 years in the period 1985-2019. This integration was guided by a set of specific hierarchical prevalence rules (Table 3). A final land cover and land use map of the MapBiomas project Collection 5.0 is the output of this last step. An exception to the prevalence rules in the case of the Cerrado was that the Pasture class had prevalence over Grassland, except within Conservation Units.

Table 3. Prevalence rules for combining the output of the Cerrado classification with the cross-cutting themes in Collection 5.0.

Collection 5.0	Prevalence Rule
4.3. Mining	1
4.1. Beach and Dune	2
1.1.3. Mangrove	3
5.2. Aquaculture	4
2.3. Salt flat	5
4.2. Urban Infrastructure	6
1.2. Forest Plantation	7
3.2.1.2. Sugar Cane	8
3.2.1.1. Soybean	9
3.2.1.3. Other annual crops	10
3.2.2. Perennial Crop	11
3.2.1. Annual Crop	12
3.1. Pasture	14
4.4. Other non Vegetated Area	15
5.1. River, Lake and Ocean	16
1.1.1. Forest Formation	17
1.1.2. Savanna Formation	18
2.2. Grassland	20
6. Not Observed	18

5. Validation

Accuracy analysis was performed based on the dataset produced by LAPIG comprising about 25k reference sample-pixels for the Cerrado. One of the classes of the MapBiomas legend was assigned to each sample in each year (1985 - 2019) by an interpreter trained by experts in Cerrado vegetation (for details on the sampling design please consult the ATBD and Accuracy Methodological Report). Global and per class accuracy, omission and commission errors, as well as quantity and allocation disagreements were calculated based on the confusion matrix that confronts the reference dataset to sample-pixels from the integrated (public) version of Collection 5.

Global accuracy (considering all years) was 83.8%, 81.6% and 74.8% in levels 1, 2 and 3 of the legend, respectively. Allocation disagreement ranged from 11.3% to 18.7% across levels while quantity disagreement range was 4.9% - 6.5%. Across levels, accuracy metrics were slightly higher than Collection 4.1, with improvements in terms of reduced omission of Savanna Formation and River, Lake and Ocean (Figure 5) and reduced commission of Grassland (Figure 6).

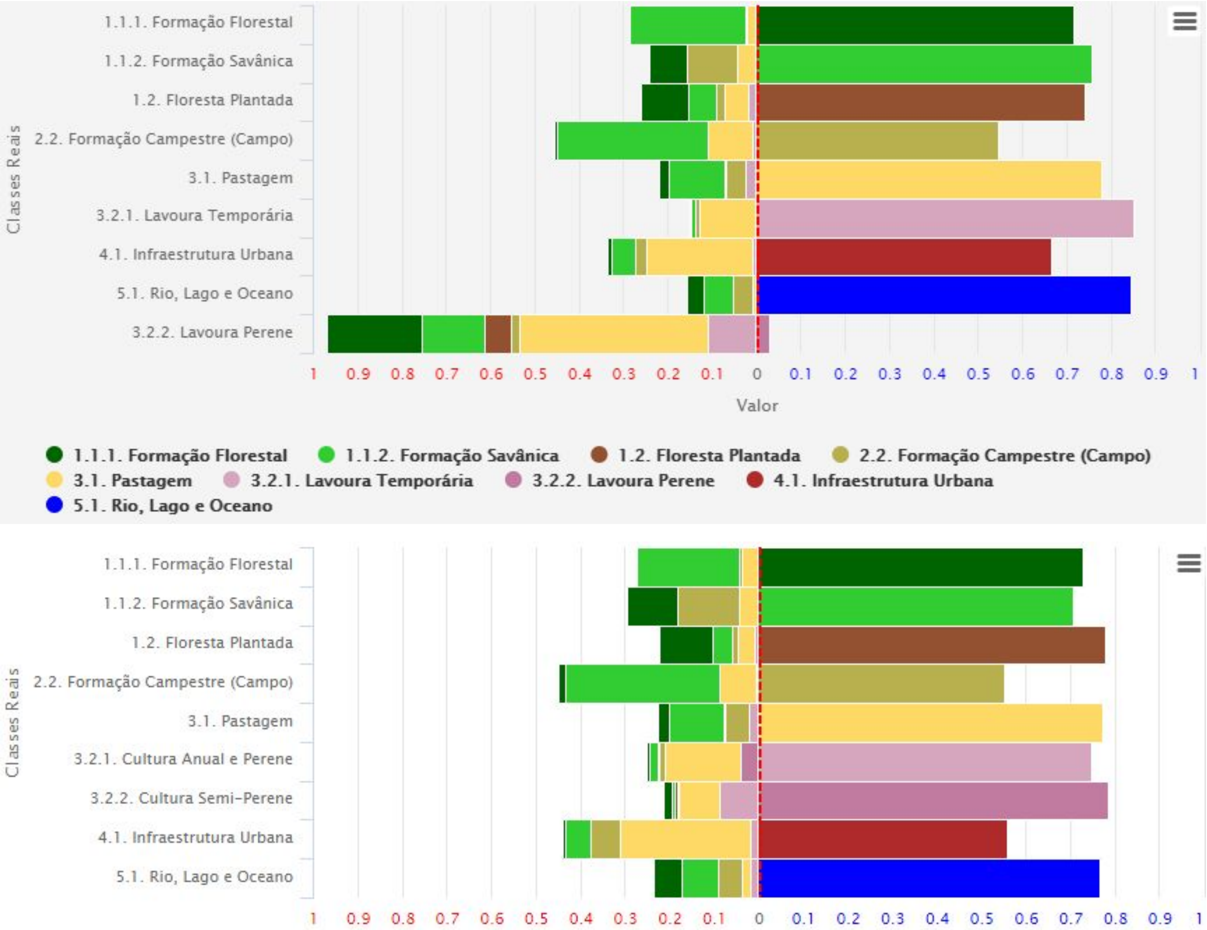
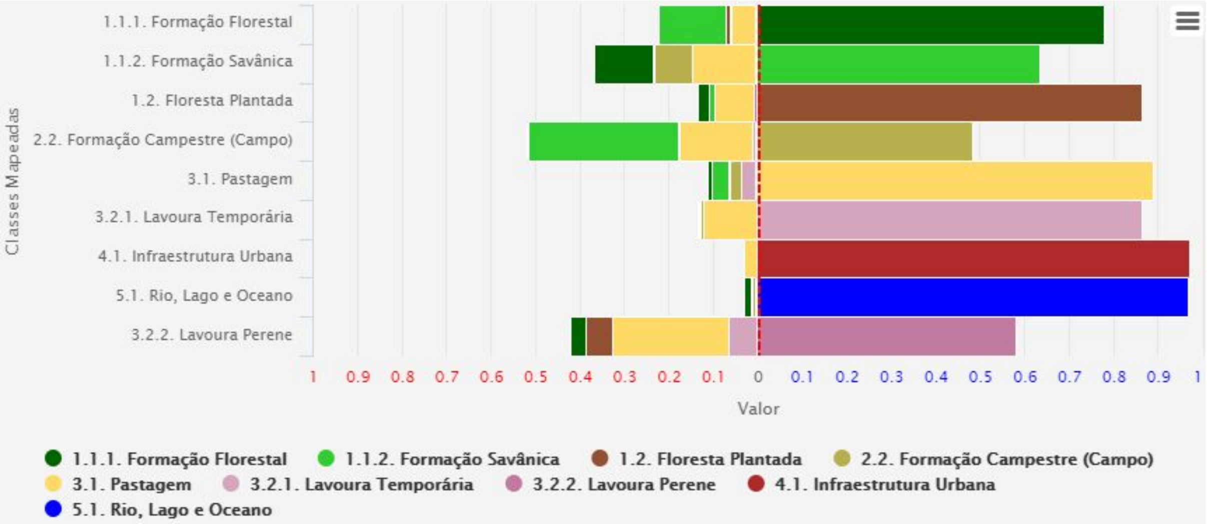


Figure 5. Omission error per class at the third level of the legend, for Collection 5 (top) and Collection 4.1 (bottom). Bars length from the red dashed line represent either error (to the

left, red values) or precision (to the right, blue values) as a percentage of class area in the reference dataset (sample-based). Colors mainly represent the proportion of each committed class (left).



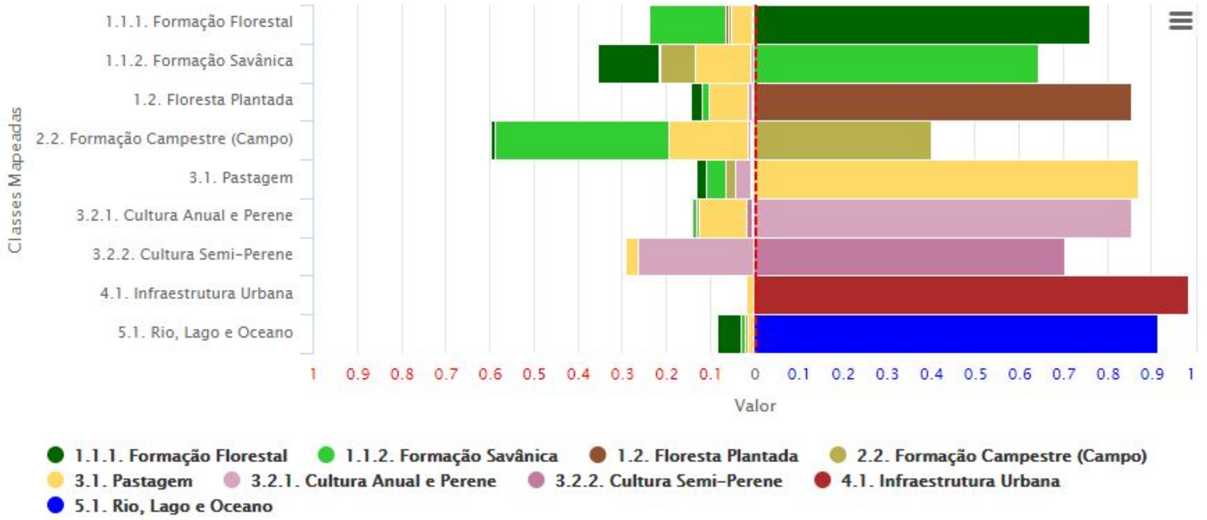


Figure 6. Commission error per class at the third level of the legend, for Collection 5 (top) and Collection 4.1 (bottom). Bars length from the red dashed line represent either error (to the left, red values) or precision (to the right, blue values) as a percentage of class area (pixel count). Colors mainly represent the proportion of each committed class (left).

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. https://doi.org/10.3390/rs12060924.

Sano, E. E., Rodrigues, A. A., Martins, E. S., Bettiol, G. M., Bustamante, M. M. C., Bezerra, A. S., Couto, A. F., Vasconcelos, V., Schüler, J., & Bolfe, E. L. 2019. Cerrado Ecoregions : A spatial framework to assess and prioritize Brazilian savanna environmental diversity for conservation. Journal of Environmental Management, 232(July 2018), 818–828. https://doi.org/10.1016/j.jenvman.2018.11.108