Cerrado - Appendix

Collection 4

Version 1

General coordinator
Ane A. Alencar

Team
Bárbara Zimbres
Camila Balzani Marques
Felipe E. B. Lenti
Isabel de Castro Silva
João Ribeiro
Julia Z. Shimbo
Valderli J. Piontekowski
Vera L. S. Arruda

1. 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 scheme. The mosaic of images represents a mosaic composition of the best pixels that are extracted from all the images available between a defined period within a year. Once the initial and final dates of this period were defined, the median of these pixels was calculated generating one value for each pixel. The aggregation of these composed pixels generated then the annual mosaic that was submitted to classification.

Several tests were done to define the optimum period of images to compose the annual mosaics. Due to the high impact of the seasonality on Cerrado vegetation spectral response compositions of images from the rainy and dry seasons were evaluated. The tests included classification using the end of the rainy season when the Cerrado vegetation is still vigorous and there is higher probability of getting images with lower cloud observation if compared with the peak of the rainy season. Tests were done also with a composition of images generated in the end of the dry season which includes the months between July and September. The tests demonstrated that if only images from the rainy season were used, the result was a greener mosaic and the chances of increasing the commission errors in the classification were higher, since more areas that were not forest (*i.e.* dense savannas) can be classified as forest. On the other hand, if only the images acquired during the last three months of the rainy season were selected, the mosaic results in a drier aspect, underestimate the forest cover mainly due to the lost ability for mapping deciduous forests (Figure 1).

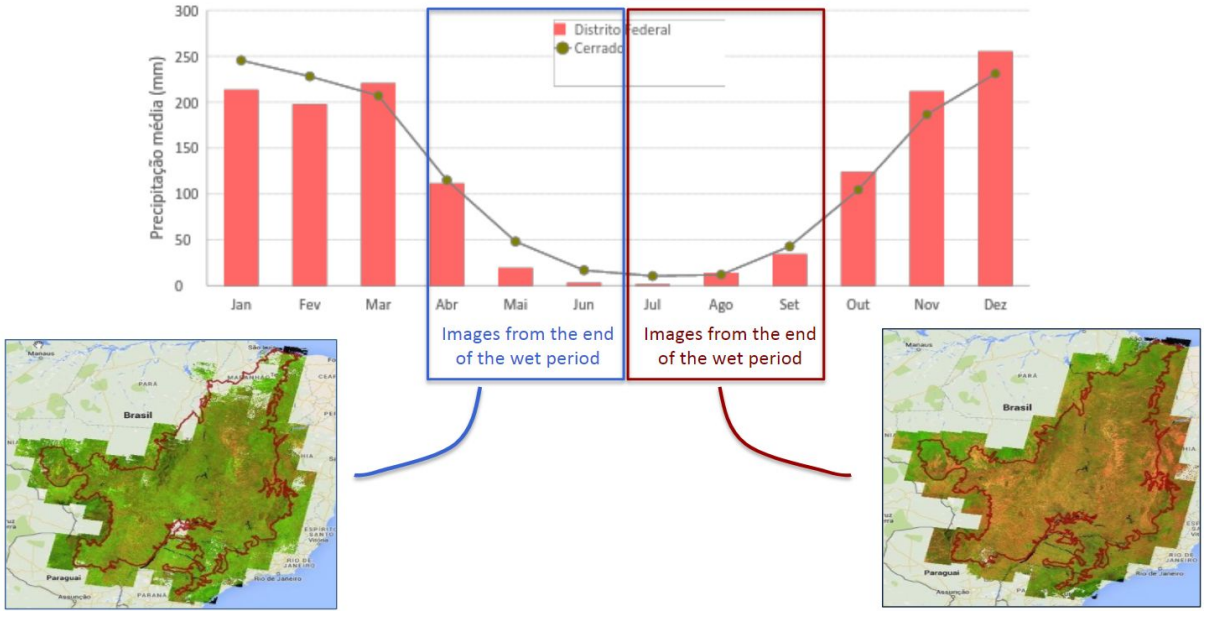

Figure 1. Pixel composite mosaics in the end of the rainy season and the end of the dry season in the Cerrado biome.

Based on the tests described above, a large window was decided for the selection of the initial and final dates for the mosaics. 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 months window between the months of April to September (Figure 2). The median value of the pixels selected during this wider period demonstrated to accommodate better the difficulties in mapping forest presented in the shorter window tests. In fact, this strategy averaged the commission and omission errors of the strategies presented above generating 34 mosaics (Figure 3), by adding the mosaic of 2018 in the Collection 4.

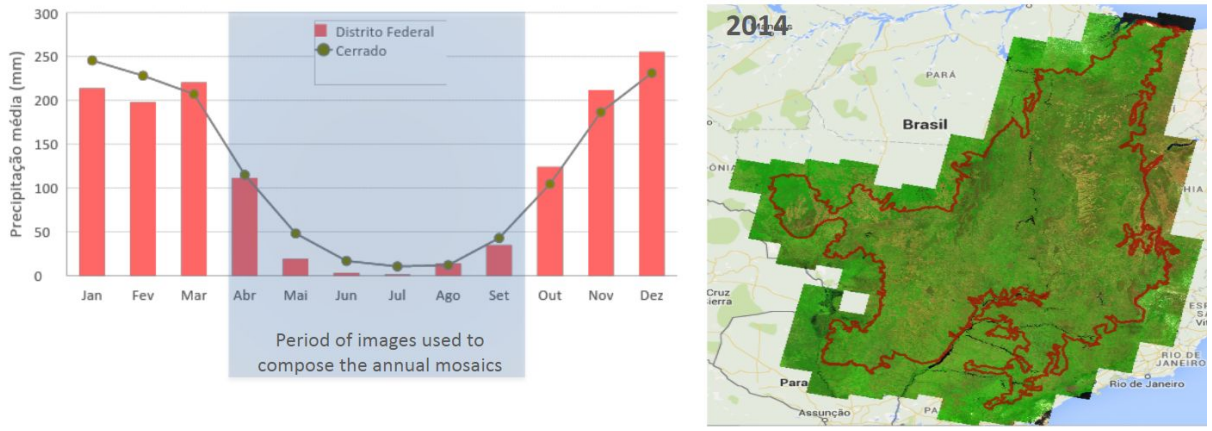

Figure 2. Window period used to define the final pixel composite annual mosaics used in the classification of MapBiomas Collection 4 in the Cerrado biome.

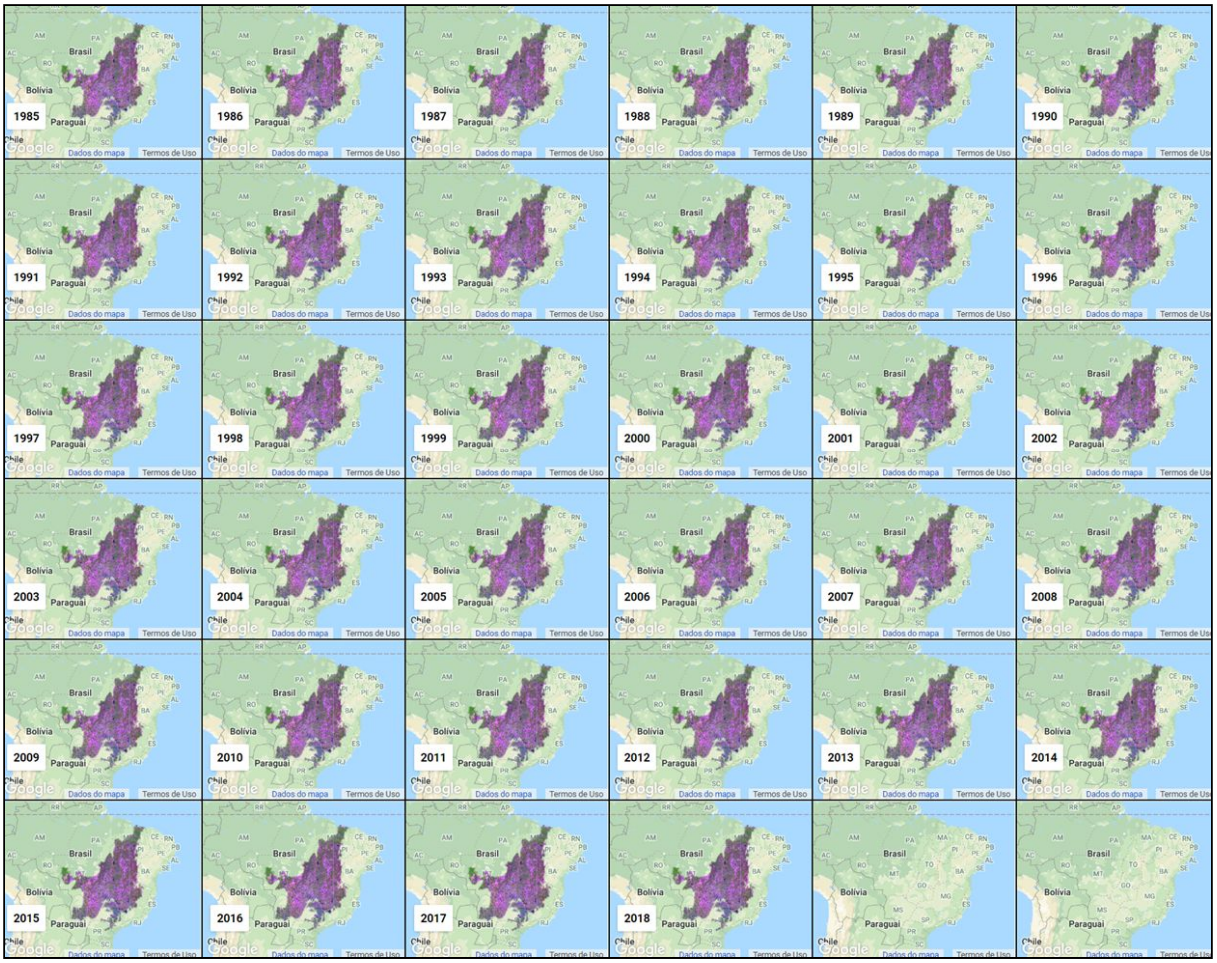

Figure 3. Annual Landsat image mosaics of the Cerrado biome from 1985 to 2018 in the MapBiomas Collection 4.

2. Classification

The Collection 4 was improved with the Random Forest approach by collecting training samples only in areas with stable classification during the 33 years in collection 3.1. Sample sufficiency per tile was attained by using a larger window to sort samples and use them to train the per tile classifier: 3x3 tiles in collection 4 *versus* 1 buffered tile (50 km) in collection 3.1.

2.1. Classification scheme

The classification of the Landsat mosaics for the Cerrado biome considered eight land use and land cover (LULC) classes of the MapBiomas Collection 4 Legend (Table 1), which was later integrated with the cross-cutting theme classes in a subsequent step.

Table 1. Land cover and land use categories considered for digital classification of Landsat mosaics for the Cerrado biome in the MapBiomas Collection 4.

Legend class of Collection 3	Numeric ID	Color
1.1.1. Forest Formations	3	
1.1.2. Savanna Formations	4	
2.2. Grassland	12	
3.1 Pasture	15	
3.2.1 Annual and Perennial Crop	18	
4.4. Other Non-Vegetated Areas	22	
5.2 River, Lake and Ocean	26	
6. Not Observed	27	

The development of the Collection 4 with annual maps of Cerrado main natural vegetation formations from 1985 to 2018 followed the steps:

(1) Definition of areas of stable classes considering the Collection 3.1 time series (2000 - 2017). Urban Area pixels were used as a proxy for collection samples for the Non-vegetated Area Samples;

(2) Preliminary Random Forest (fitted model) with sub-sample set in order to evaluate remote sensed variable importance (*i.e.* feature space);

(3) Assessing area proportion per class/per year in order to balance sample set for each prediction/classification (per tile per year)

(4) Training of each classifier (tile/year) using balanced samples and selected feature space. Minimum sample size per class was set to 300 and maximum sample size per class was set to 1500.

(5) Classification (prediction) using Random Forest as implemented in Google Earth Engine (parameters: mtry = 5; ntrees = 60; nodeSize = 2).

2.2. Feature space

The feature space used in the Cerrado biome in the Collection 4 was defined with a statistical approach by fitting several Random Forest classification models, each using a subset of 400 unique samples per class and considering 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 (mean decrease in accuracy for each native vegetation class: Forest, Savanna and Grassland formations). All cases were compiled in a dataframe and 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 abovementioned classes. As expected, these sets

shared several variables, i.e. information that was important to the global accuracy and one or more of the specific classes. We selected 48 variables as the feature space for training/classification (Table 2).

Table 2. Feature space subset considered in the classification of the Cerrado biome Landsat image mosaics in the MapBiomas Collection 4.

Variable	Description	Statistic	Temporal range	Script acronym
Green	Landsat band	minimum	year	min_green
Green	Landsat band	median	year	median_green
Greendr 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 Clorophyll Vegetation Index	median	year	median_gcvi
GCVI	Green Clorophyll 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

2.3. Classification algorithm, training samples and parameters

For the final classification, a Random Forest classifier was trained for each tile using 300 to 1500 samples per class. The actual sample size for each model (tile/year) is a function of the distribution of class areas in each case. Samples for each class were randomly selected both from within the target tile and from its immediate neighbors (*i.e.* one target tile plus eight adjacent tiles). The usage of samples from neighboring tiles was adopted as it has improved the spatial consistency of the maps since collection 3.1. All tiles were classified using 60 decision trees and a subset of five randomly selected variables per node (mtry).

3. Post-classification

Due to the pixel-based classification method and the long temporal series, a list of post-classification spatial and temporal filters was applied. The post-classification process includes the application of gap-fill, temporal, spatial and frequency filters.The temporal filter rules were adapted for the land cover and land use classes used in the Cerrado biome and were complemented by specific rules to adjust for cases where a pixel appeared.

3.1. Complementary samples

After classification, we observed a confusion between cast shadows and water bodies in Cerrado regions with steep terrain associated with mountains (nine tiles). To address this issue we collected complementary samples of shade pixels and trained a Random Forest classifier using only water samples, shade samples and a reduced feature space, focused on variables that help to distinguish these two targets. We then generated a layer containing only pixels that were originally mapped as water (water mask) and ran this supplementary classification to assign each pixel as “confirmed” water or shade pixel, in which case it was remapped to the “Not observed” class. Finally, pixels originally classified as water in each of the nine tiles were substituted for the layer generated in this post-classification step.

3.2. Gap fill

In this filter, no-data values (“gaps”) are theoretically not allowed and are replaced by the temporally nearest valid classification. In this procedure, if no “future” valid position is available, then the no-data value is replaced by its previous valid class. Therefore, gaps

should only exist if a given pixel has been permanently classified as no-data throughout the entire temporal domain.

3.3. Temporal filter

The temporal filter uses the subsequent years to replace pixels that have invalid transitions.

3.4. Frequency filter

Frequency filters were applied only in pixels that were considered “stable native vegetation” . If a “stable native vegetation” pixel is at least 60% of the years as the same class, all years are changed to this class. The result of these frequency filters is a classification with more stable classification between native classes (e.g. Forest and Savanna). Another important result is the removal of noises in the first and last year of the classification.

3.5. Spatial filter

The spatial filter avoids unwanted modifications to the edges of the pixel groups (blobs), a spatial filter 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. In this filter, at least six connected pixels are needed to reach the minimum connection value. Consequently, the minimum mapping unit is directly affected by the spatial filter applied, and it was defined as 6 pixels (~0,5 ha).

3.6. Integration with cross-cutting themes

The cross-cutting themes and the biomes data were integrated for each of the 34 years in the period 1985-2018.. This integration was guided by a set of specific hierarchical prevalence rules (Table 3). As output of this step, a final land cover and land use map for each chart of the MapBiomas project was produced. An exception in the Cerrado biome was that Pasture (derived from Collection 3.1) had prevalence over Grassland Formation, excepted in the Conservation Units. In the Amazon Cerrado transition (charts) the Cerrado classification had prevalence over Non Forest Natural Formation areas classified by the Amazon.

Table 3. Prevalence rules for combining the output of digital classification with the cross-cutting themes in Collection 4.

Collection 4	Prevalence Rule
4.1. Beach and Dune	1
1.1.3. Mangrove	2
5.2. Aquaculture	3
2.3. Apicum	4
5. Water	5

5.1. River, Lake and Ocean	5
1.2. Forest Plantation	6
4.4. Mining	7
4.2. Urban Infrastructure	8
3.2. Farming	9
3.2.1. Annual and Perennial Crop	9
3.2.1. Semi-perennial Crop	9
1.1.1. Forest Formation	10
1.1.2. Savana Formation	11
4.5. Rocky Outcrop	12
2.1. Wetlands	13
2.2. Grassland	14
2.4. Other Non-Forest Natural Formation	14
3.1. Pasture	15
4.3. Other Non-Vegetated Area	16
3.3 Mosaic of Agriculture and Pasture	17
6. Not Observed	18

4. References

MCTI - MINISTRY OF SCIENCE, TECHNOLOGY AND INNOVATION. 2015. III Brazilian inventory of anthropogenic emissions and removals of greenhouse gases not controlled by the Montreal Protocol. Brasília, DF: MCTI.

MMA - MINISTRY OF THE ENVIRONMENT. 2015. TerraClass Cerrado - Mapping of the Use and Coverage of Cerrado Land in the year 2013. Brasília, DF: MMA.

FBDS - BRAZILIAN FOUNDATION FOR SUSTAINABLE DEVELOPMENT. 2015. Land use mapping for the Cerrado. Rio de Janeiro, RJ: FBDS.