Caatinga Appendix

Collection 6

Version 1

General Coordinator
Washington de Jesus Sant'anna da Franca Rocha (UEFS)

Team
Diego Pereira Costa (GEODATIN/UEFS)
Frans Pareyn (APNE)
José Luiz Vieira (APNE)
Rodrigo Nogueira de Vasconcelos (GEODATIN/UEFS)
Soltan Galano Duverger (GEODATIN/UEFS)

Overview

This document represents the summary of the specific methods used in the generation of maps for the Caatinga biome in the context of MapBiomas. The most complete description of the general methods used in the project is present in the general ATBD of Mapbiomas (https://mapbiomas.org/download-dos-atbds)

1. Classification method

Figure 1 shows the process flow used in the Collections 6 of the Caatinga biome. In terms of processing, collection 6 is similar to collections 4 and 5. However, some improvements were added which will be described below

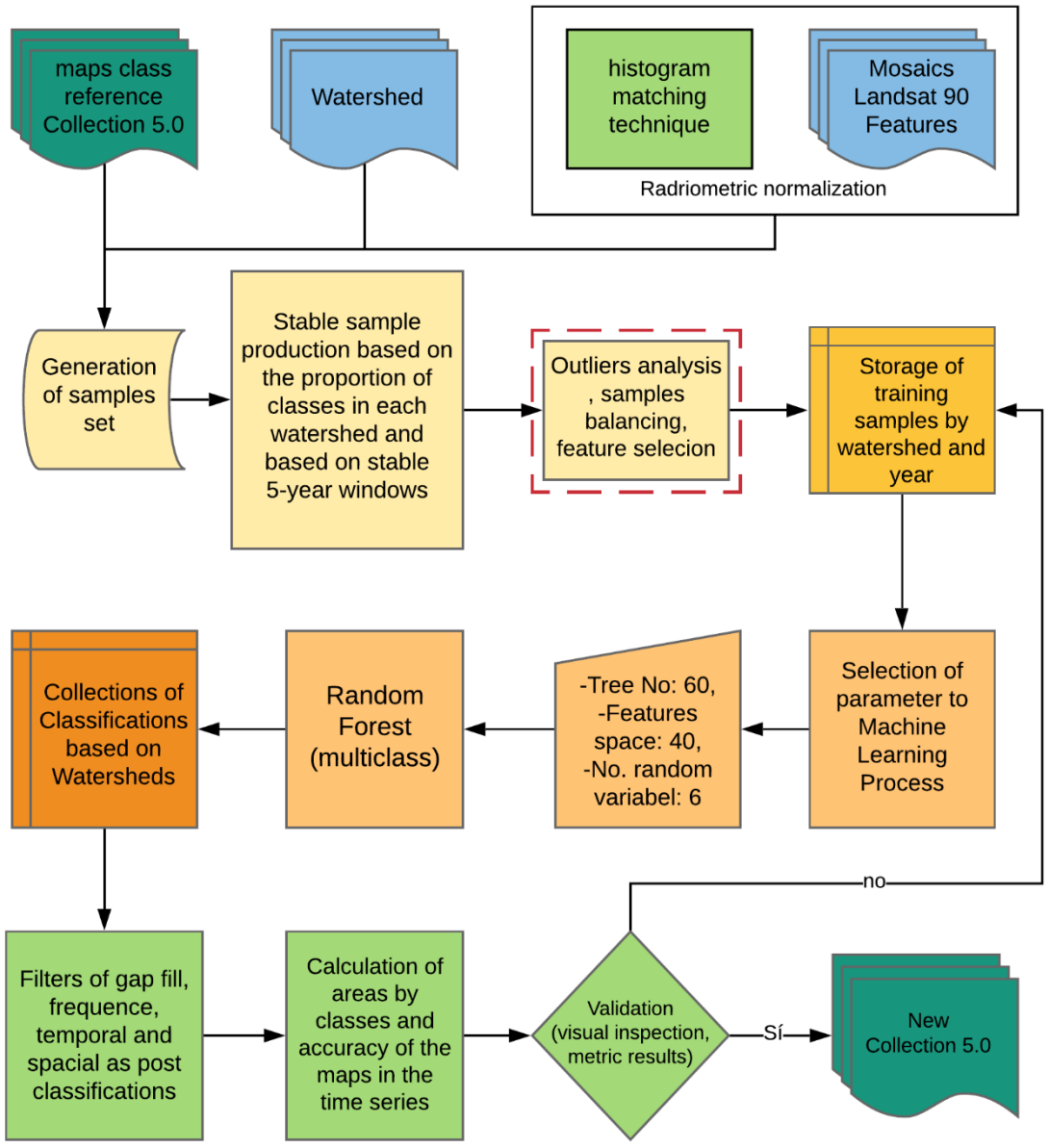

Figure 1. Classification process of MapBiomas Collection 6 (1985-2020) in the Caatinga biome.

Table 1. The evolution of the Caatinga mapping collections in the MapBiomas Project, its periods, mapped classes, brief methodological description, and global accuracy in Level 1, 2, and 3.

Collection	Period	Method	Global Accuracy
4.1	34 years 1985-2018	Random Forest	Level 1: 81.9% Level 2: 79.9% Level 3: 74.3%
5.0	35 years 1985-2019	Random Forest	Level 1: 81.8% Level 2: 80.0% Level 3: 75.4%
6.0	36 years 1985-2020	Random Forest	Level 1: 81.1% Level 2: 75.0% Level 2: 74.9%

2. **Landsat image mosaics**

In previous collections, the classification was performed using Landsat 5 (TM), 7 (ETM +), and 8 (OLI) (Landsat TOA data). In collection 6.0, we used data from the surface reflectance (SR) collection. Furthermore, all mosaics were normalized from a reference image dictionary associated with each L5, L7, L8 image collection. The technique used for this procedure was histogram matching.

2.1 **Definition of the period**

The image selection period for the Caatinga biome was defined aiming to minimize confusion between different natural vegetation and other land use and land cover (LULC) (*e.g.* cultivated areas) due to extreme phenological changes while trying to maximize the coverage of Landsat images after cloud removing/masking. Unlike most other Brazilian biomes, the climate of the Caatinga biome has a large seasonal variation of precipitation being the main factor determining the physiological behavior of vegetation throughout the year. Caatinga vegetation is classified as seasonal in their majority, expressing great deciduousness over the year. Only a small fraction of tree species do not lose leaves during dry station so that Caatinga savanna formations are expected to show great variation in spectral response through the year. To define the periods for the mosaic construction, we used the rainfall data of the Northeast region of Brazil, considering the strong seasonal component in this region. Initially, an evaluation of the entire available time series (1961-2015) was made. This dataset was obtained from the INMET (www.inmet.gov.br). The data evaluation was performed through visual inspection of the annual graphs and historical averages for each of the climatic stations with data available for the Caatinga biome (Figure 2)

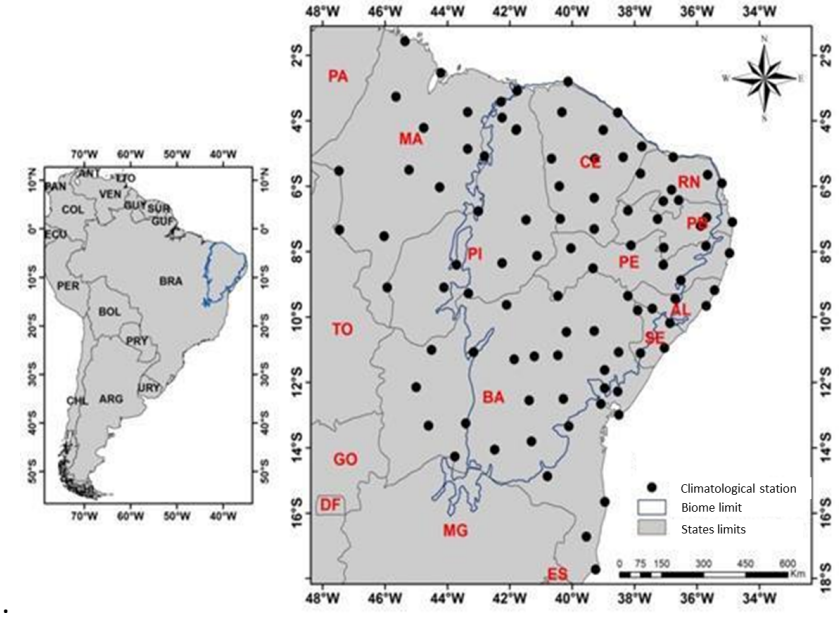

Figure 2. Location of the climatic stations used for the construction of the rainfall series for a selection of the mosaic periods in the Caatinga biome.

Then, a periodic window scan was carried out for the entire Caatinga biome, indicating that the period between January to July (with higher levels of rainfall in the Caatinga biome) (Figure 3) is more likely to obtain images with spectral contrast capable of separating different classes of LULC for the biome. The choice of these sets of parameters helped to define the mosaics with better spectral quality and less amount of noise and clouds in the images for the biome.

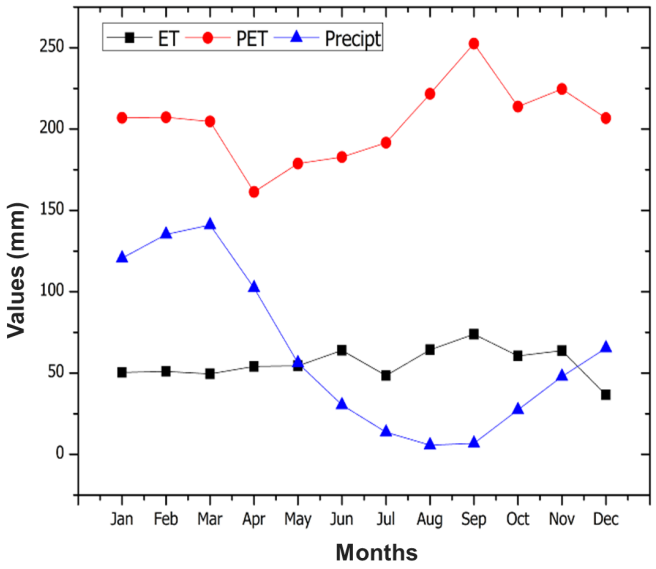

Figure 3. Temporal variation of water balance with monthly mean precipitation, evapotranspiration and potential evapotranspiration variables in the Caatinga biome.

2.2 **Image selection**

For the selection of Landsat scenes to build the mosaics by map sheet for the year, within the acceptable period, a threshold of 90% of cloud cover was applied (i.e. any available scene with up to 90% of cloud cover was accepted). When needed, due to excessive cloud cover and/or lack of data, the acceptable period was extended to encompass a larger number of scenes to allow the generation of a mosaic without missing data. Whenever possible, this was made by including months at the beginning of the period, in the winter season.

For the generation of the mosaics by map sheet we used the parameters described (period and cloud cover). The selected Landsat scenes were processed to generate the temporal mosaic that covers the area of the chart.

2.3 **Final quality**

The mosaic quality was evaluated using the frequency of each available pixel in the Caatinga biome (Figure 4). As a result of the selection criteria, all of them presented satisfactory quality. In Collections 4.1, 5, and 6, a single change to this calculation refers to the limit of the biome that was updated (IBGE, 2019).

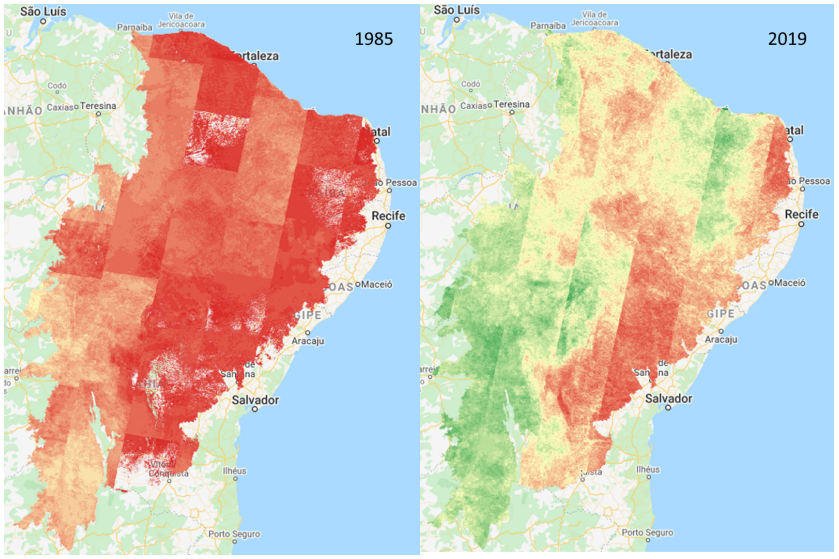

Figure 4. Landsat pixel availability in 1985 and 2019 in the Caatinga biome. Colors refer to data pixel availability, where red is low, yellow is medium, and green is high.

3. **Definition of regions for classification**

The Caatinga Biome was divided into 42 regions based on watershed boundaries available by the Agência Nacional de Águas (www.ana.gov.br) (Figure 5). In this case, we merged watersheds, level 3 and level 4. Due to the changes in the limits of the biomes (IBGE, 2019) in Collection 5, another region was added, reaching 39 in total.

The classification in homogenous regions reduces the confusion of samples and classes, as well as allows a better balance of samples.

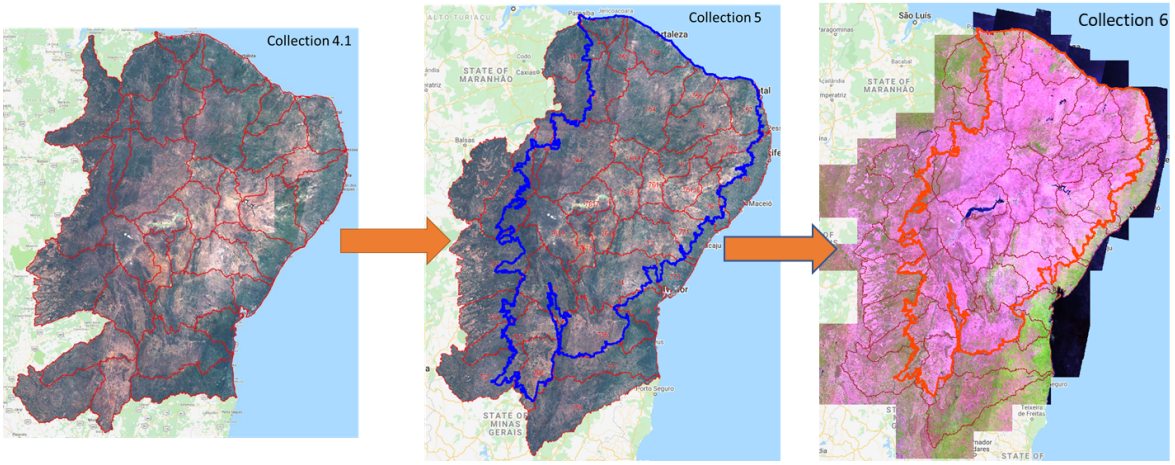

Figure 5. The evolution of the Caatinga watersheds used in the classification of Collections 4.1, 5, and 6.

4. **Classification**

4.1 **Classification scheme**

The digital classification of the Landsat mosaics in the Caatinga biome aimed to map a subset of seven LULC classes of the MapBiomas legend in Collection 6 (Table 1), which were integrated with the cross-cutting themes in a further step. The class Mosaic of Agriculture and Pasture in the Caatinga was later superimposed by Agriculture or Pasture class, remaining in areas of temporary crops (very common in the Caatinga biome) or where it was not possible to distinguish between these two classes.

Table 1. Land cover and land use classes considered for digital classification of Landsat mosaics in the Caatinga biome in the MapBiomas Collection 6.

Legend class	NEW ID	Natural/Antrópico	Land Cover / land Use	General description
1.1.1. Forest Formation	3	natural	Land cover	Vegetation with predominance of continuous canopy - Savanna-Estépica Florestalada, Seasonal Semi-Deciduous and Deciduous Forest.
1.1.2. Savanna Formation	4	natural	Land cover	Vegetation with predominance of semi-continuous canopy species - Savanna-Estépica Arborizada, Savana Arborizada.
2.2. Grassland	12	natural	Land cover	Vegetation with predominance of herbaceous species (Savanna-Estépica Parque, Savana-Estépica Grassy-Woody, Savanna Park, Savanna Grassy-Woody)
2.4. Rocky Outcrop	29	natural	Land cover	Rocks naturally exposed on the earth's surface without soil cover, often with partial presence of rupicolous vegetation and high slope
3.3. Mosaic of Agriculture and Pasture	21	antropic	Land use	Use agricultural areas where it was not possible to distinguish between pasture and agriculture.
6. Non Observed	27	non observed	non observed data	non observed data

4.2 **Feature space**

The feature space for digital classification of the LULC classes in the Caatinga biome comprised a subset of 66 features (Table 2), taken from the complete feature space of MapBiomas Collection 6 (General ATBD MapBiomas, 2020). All watersheds were analyzed individually in terms of feature importance. These variables included the original Landsat reflectance bands, as well as vegetation indexes and spectral mixture modeling-derived variables. The definition of this subset and the classifier parameters were made based on n tests conducted through machine learning-based libraries. All codes used are available in the repository of MapBiomas's Github (https://github.com/mapbiomas-brazil/caatinga).

Table 2. Feature space subset considered in the classification of Landsat image mosaics in the Caatinga biome in the MapBiomas Collection 6.

'awei'	'awei_median'	'awei_median_dry'	'blue_stdDev'	'brightness_median'
'cvi_median_dry'	evi'	'evi_median'	'evi_median_dry'	'gcvi'
'green_stdDev'	'lai'	'lai_median'	'lai_median_dry'	ndfia'
'ndfia_median'	'ndfia_median_dry	'ndvi'	'ndwi'	'ndwi_median_dry'
'nir_1'	'nir_median'	nir_median_dry'	'nir_stdDev'	'osavi'
'osavi_median_dry'	'red_median'	'red_stdDev'	'soil'	soil_median'
'soil_median_dry'	'spri'	'spri_median'	'spri_median_dry'	'swir1_median'
'gv'	swir1_median_dry'	'gv_median'	'gv_median_dry'	'ratio'
'ndfia_median'	ndfia'	'ndfia_median_dry'	'ndvi'	'ndwi'
'ndwi_median_dry'	'nir_1'	'nir_median'	nir_median_dry'	'nir_stdDev'
'osavi'	'osavi_median_dry	'red_median'	'red_stdDev'	'soil'
soil_median'	'soil_median_dry'	'spri'	'spri_median'	'spri_median_dry'
'swir1_median'	'gv'	swir1_median_dry'	'gv_median'	'gv_median_dry'
'ratio'				

4.3 Classification algorithm, training samples, and parameters

The digital classification was performed by watershed, year by year, using a *Random Forest* algorithm (Breiman, 2001) available in the *Google Earth Engine*. Training samples for each watershed were defined following a strategy of using pixels in which the vegetation cover/land use remained the same in the five years windows of Collection 5 named as “stable samples”. The parameters used in the classifier were: 'numberOfTrees': 60, 'variablesPerSplit': 6, 'minLeafPopulation': 3 'maxNodes': 10, 'seed': 0. The sequence of temporal space filters used in collection 5 was 1 - GapFill, 2 - Temporal, 3 - Frequency, 4 - Spatial

4.3.1 **Stable sample from Collection 4.1**

The extraction of stable samples from the previous map Collection 4.1 followed several steps aiming to ensure their confidence as training areas.

First, based on the class cover percentage in each watershed and the five years windows, we generated random stable samples. To sample size determination, based on statistical representativeness, we used Tortora (1978) as a reference. A minimum of 300 samples was used for rare classes that do not cover at least 10% of the region area. All codes used are available in the repository of MapBiomas’ Github (https://github.com/mapbiomas-brazil/caatinga).

4.3.2 **Final classification**

Final classification was performed for all regions and years with samples. It was used the same subset of samples for all the years, and it was trained in the same mosaic of the year that was classified.

5. **Post-classification**

The temporal filter rules were adapted for the classes used in the Caatinga biome and were complemented by specific rules to adjust for cases where a pixel appeared.

5.1 **Gap Fill filter**

This filter aims to fill data (pixels) in images that do not have observations. In practice, if no valid “future” position is available, the value with no data is replaced by its previous valid class. In this way, only gaps that have no observation permanently as no data in the entire time domain will be worthless.

5.2 **Spatial filter**

The applied spatial filter uses a mask to change only pixels connected to five or fewer pixels of the same class. These pixels were replaced by the MODA value of its eight neighbor’s pixels.

5.3 **Temporal filter**

The applied temporal filter uses the subsequent years to replace pixels that have invalid transitions. In the first process, the filter looked for any natural class (3-FOREST FORMATION, 4-SAVANNA FORMATION, 12-GRASSLAND, 13- OTHERS NO FOREST FORMATION) that was not this class in 85 and was equal to these classes in 86 and 87 and then corrected 85 class to avoid any regeneration in the first year. In the second process, the filter looked at the pixel value in last year that was not 21MOSAIC OF AGRICULTURAL OR PASTURE and was equal to 21- MOSAIC OF AGRICULTURAL OR PASTURE in the previous two years. The value in last year was then converted to 21-MOSAIC OF AGRICULTURAL OR PASTURE to avoid any regeneration in the last year. The third process looked in a 3-year

moving window to correct any value that was changed in the middle year and return to the same class next year. This process was applied in this order: [33-RIVER, LAKE, OCEAN, 13-OTHERS NO FOREST FORMATION, 4-SAVANNA FORMATION, 29-ROCKY OUTCROP, 21-MOSAIC OF AGRICULTURAL OR PASTURE, 3-FOREST FORMATION, 12-GRASSLAND]. The last process was similar to the third process but it was a 4- and 5-years moving window that corrected all middle years.

5.3 **Frequency filter**

A frequency filter was applied only in pixels that were considered "stable natural vegetation" (at least all series of years as [3-FOREST FORMATION, 4-SAVANNA FORMATION, 12-GRASSLAND]). If a "stable natural vegetation" pixel was at least 80% of years of the same class, all years were changed to this class. The result of this frequency filter was a more stable classification between natural classes (ex: forest and savanna). Another important result was the removal of noises in the first and last year in classification (i.e. 1985 and 2019).

5.4 **Incident filter**

An incident filter was applied to remove pixels that change too many times in the series of years. All pixels that changed more than eight times and were connected to less than six pixels that also changed more than eight times were replaced by the MODE value. This avoided changes in the border of the classes.

All forest pixels that changed more than ten times and were connected to less than 66 pixels that also changed more than ten times were replaced to 21 (Mosaic of Pasture and Agriculture). Forests are stable and these areas that change this much are not forests.

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 replaced to 21 (Mosaic of Pasture and Agriculture). These areas were also some type of land use.

6. **Validation strategies**

The validation of each process was produced using independent validation points provided by Lapig/UFG. We used all points that both interpreters considered the same class, resulting in more than 12,000 validation points. The figure below shows the result of the accuracy analysis for the level 1 legend of the MapBiomas Collection 6 (1985-2020) (Figure 6).

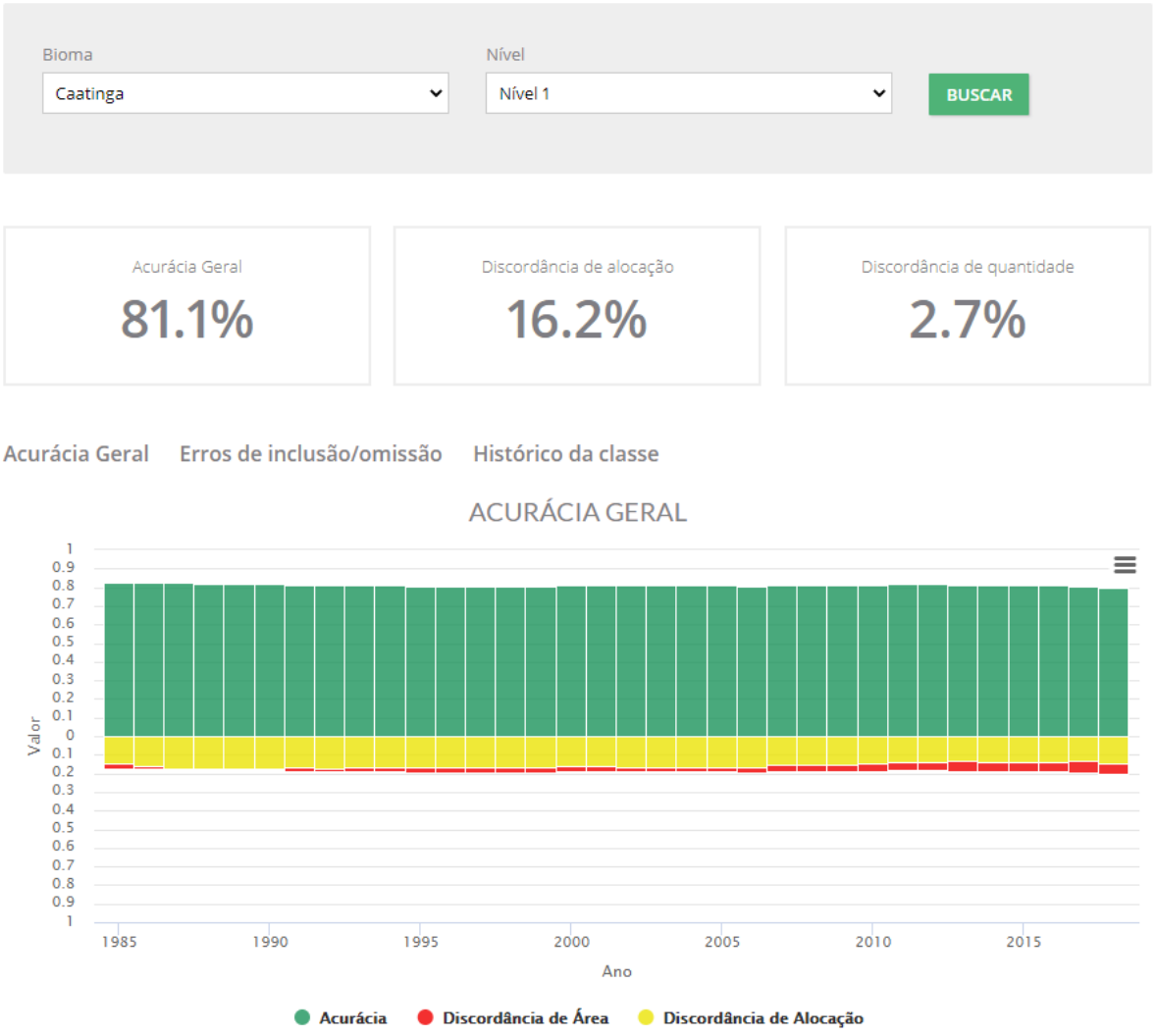

Figure 6. Accuracy of level 1 of MapBiomas Collection 6 in the Caatinga biome (1985-2020).

7. References

Breiman, 2001. Classification and regression based on a forest of trees using random inputs. doi:10.1023/A:1010933404324.

ARCOVA, F. C. S.; CICCO, V. DE; ROCHA, P. A. B. Precipitação efetiva e interceptação das chuvas por floresta de Mata Atlântica em uma microbacia experimental em Cunha - São Paulo. **Revista Árvore**, v. 27, n. 2, p. 257–262, 2003.

IBGE. **Vegetação RADAM**. Disponível em: <ftp://geoftp.ibge.gov.br/informacoes_ambientais/acervo_radambrasil/vetores/>. Acesso em: 30 maio. 2018.

IBGE Mapa de Biomas do Brasil – primeira aproximação. Rio de Janeiro, 2020, disponível em https://www.ibge.gov.br/geociencias/informacoes-ambientais/15842-biomas.html?=&t=downloads, acessado em julho de 2020;

Tortora, R.D. 'A Note on Sample Size Estimation for Multinomial Populations." The American Statistician 32:3 (August 1978), 100-102.