Caatinga Appendix

Collection 7

Version 1

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

Team
Diego Pereira Costa (GEODATIN/UEFS)
Rodrigo Nogueira de Vasconcelos (GEODATIN/UEFS)
Nerivaldo Afonso (GEODATIN/UEFS)
Rafael Oliveira Franca Rocha (GEODATIN/UEFS)
Soltan Galano Duverger (GEODATIN/UEFS)
Deorgia Tayane Mendes de Souza (UEFS/PPGM)
Jocimara Souza Lobão (UEFS/PPGM)

1. 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. For each new collection, there was an increase in the number of land cover classes or a change in the methodology used. For example, from collection 2.3 onwards, the Random Forest method started to be used in thematic classification and the parameterization was no longer done by trial and error, but by the application of algorithms for the selection and optimization of input parameters. Another example comes from features space, which is no longer selected by an empirical method and started to use feature selection algorithms that allow both to reduce dimensionality and to select the best features for the classification model. Table 1 summarizes the evolution of the methods used in the preparation of maps by collection and throughout the document each step developed and used in the collection is described, as well as the improvements applied to the production of these maps. Other methodologies used in previous collections can be accessed at ATBD of Mapbiomas (https://mapbiomas.org/download-dos-atbds).

Table 1. A brief review of the evolution of Caatinga collections, their intervals, methods, mapped classes, and the main improvements.

Collection	Time Interval	Method	Classes	Mainly Improvements
Beta & 1	2008 - 2015	Empirical Decision Tree	Forest Formation, Non-Forest, Water Mask.	Proof of concept
2.0 2.3	2000 - 2016 2000 - 2016	Empirical Decision Tree Random Forest	Forest Formation, Savanna Formation, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Areas.	Land use and land cover samples collect / Spatio-temporal filters
3.0 & 3.1	1985 - 2017	Random Forest	Same as collection 2.3.	Land use and land cover samples collect based on current classes mapped / Added Mosaic of Agriculture and Pasture class / New Spatio-temporal filters
4.0 & 4.1	1985 - 2018	Random Forest	Same as collection 2.3	Land use and land cover samples collect based on current classes mapped / New Spatio-temporal filters
5.0	1985 - 2019	Random Forest	Forest Formation, Savanna Formation, Grassland, Mosaic of Agriculture and	Stable points, based on 5-years windows/ Feature Importance Analysis/New parameters

			Pasture, Water, Other Non-vegetated Area, Rocky Outcrop	for the RF implementation/ Division of processing by watershed/ New class (Rocky Outcrop) / Spatio-temporal filters
6.0	1985 - 2020	Random Forest	Same as Collection 5.0.	New Mosaic Collection
7.0	1985 - 2021	Random Forest	Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Rocky Outcrop, Wooded Restinga.	New class (Wooded Restinga)

2. Classification method

Figure 1 shows the process flow diagram used in the Collections 7 of the Caatinga biome. In terms of processing, Collection 7 is something similar to Collections 4, 5, and 6. However, some improvements were added which will be described below.

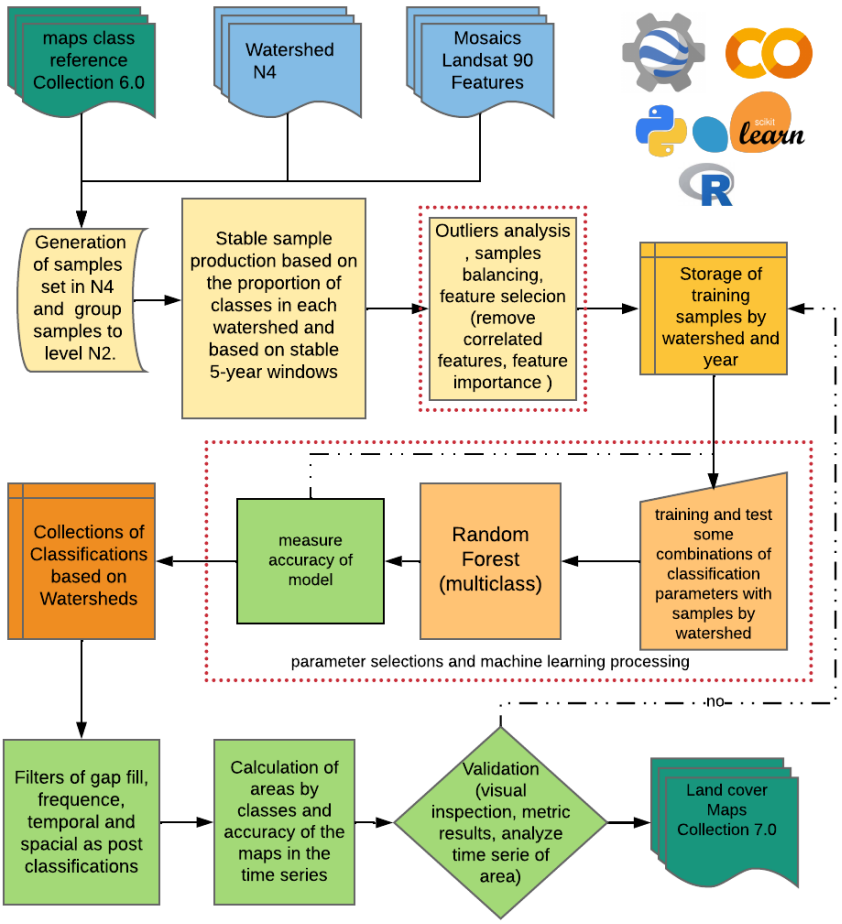

Figure 1. Classification process of MapBiomas Collection 7 (1985-2021) in the Caatinga biome.

2. Landsat image mosaics

In previous collections, the classification was performed using Landsat 5 (TM), 7 (ETM+), and 8 (OLI) (Landsat SR data). In collection 6.0, we used data from the surface reflectance (SR) collection. Collection 7.0 was created by the Landsat images Collections 2 ST products. These Collections 2 of Landsat was created with the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) algorithm (version 3.4.0) available on GEE as id asset *"LANDSAT/LT05/C02/T1_L2"* for Landsat 5, as id asset *"LANDSAT/LE07/C02/T1_L2" for* Landsat 7 and *"LANDSAT/LC08/C02/T1_L2"* for Landsat 8. The mosaic building is saved in the asset project Mapbiomas with all processing to get the data cleaned, it is accessed by path *"projects/nexgenmap/MapBiomas2/LANDSAT/BRAZIL/mosaics-2".* This mosaic has 119 spectral bands between spectral indexes, fractions from spectral unmixing, and descriptive statistics calculated by period dry and wet.

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 considerable seasonal variation of precipitation, the main factor determining the physiological behavior of vegetation throughout the year. Caatinga vegetation is classified as seasonal in its majority, expressing great deciduousness over the year. Only a small fraction of tree species do not lose leaves during dry station so 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).

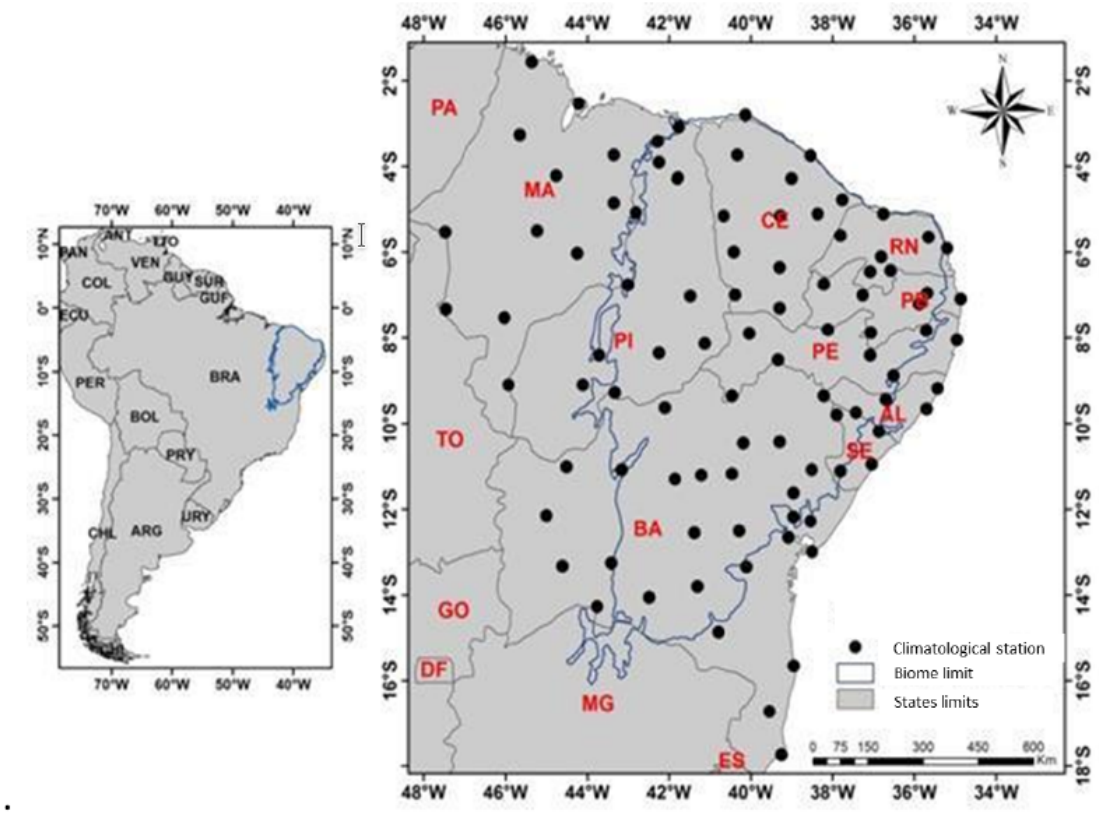

Figure 2. Location of the climatic stations used for the construction of the rainfall series for a selection of the mosaic periods in the Caatinga biome.

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.

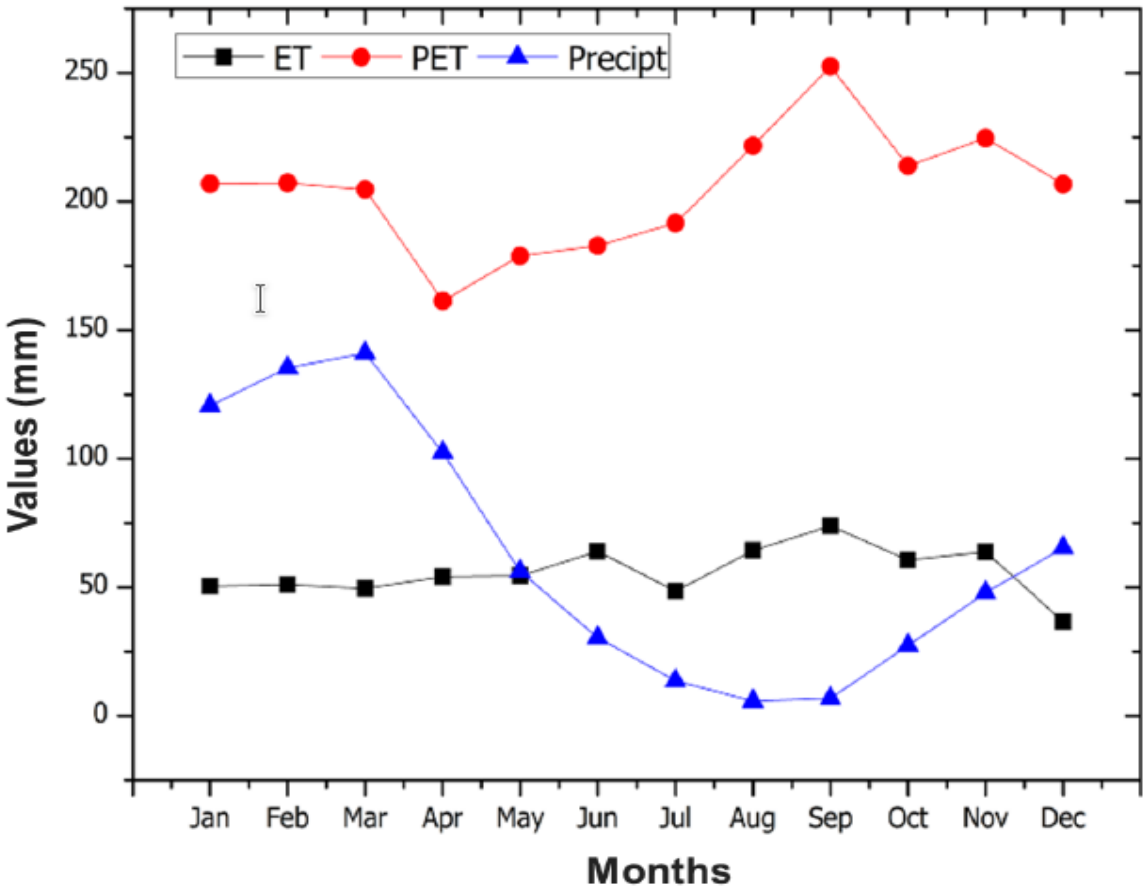

Figure 3. Temporal variation of water balance with monthly mean precipitation, evapotranspiration, and potential evapotranspiration variables in the Caatinga biome.

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 Mosaic 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 better quality (i.e Less noise such as clouds, relief and clouds shadows.). In Collections 4.1, 5, 6, and 7, a single change to this calculation refers to the limit of the biome that was updated (IBGE, 2019).

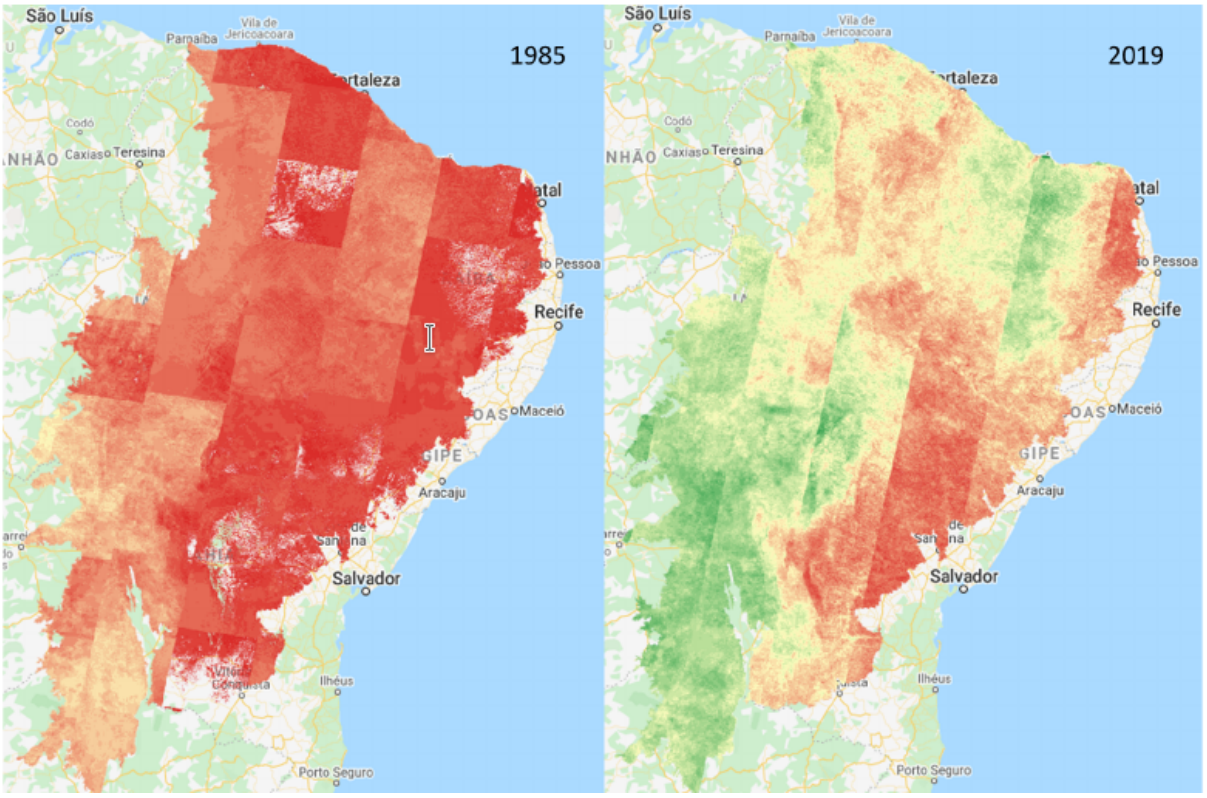

Figure 4. Landsat pixel availability in 1985 and 2019 in the Caatinga biome. Colors refer to data pixel availability, where red is low, yellow is medium, and green is high.

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, but in Collections 6 and 7 was used the limited watershed with 42 regions.

The classification in homogenous regions reduces the variability between the spectral values of the pixels outside and inside the coverage classes, as well as

allows the same samples to classify large areas of the mosaic. The sampling process for areas large in the google earth engine (GEE) is a computationally expensive task, that is why in this work small areas were selected at level 4 watershed. The level 4 watershed has 320 regions, then this sampling process was automated using the api python of GEE.

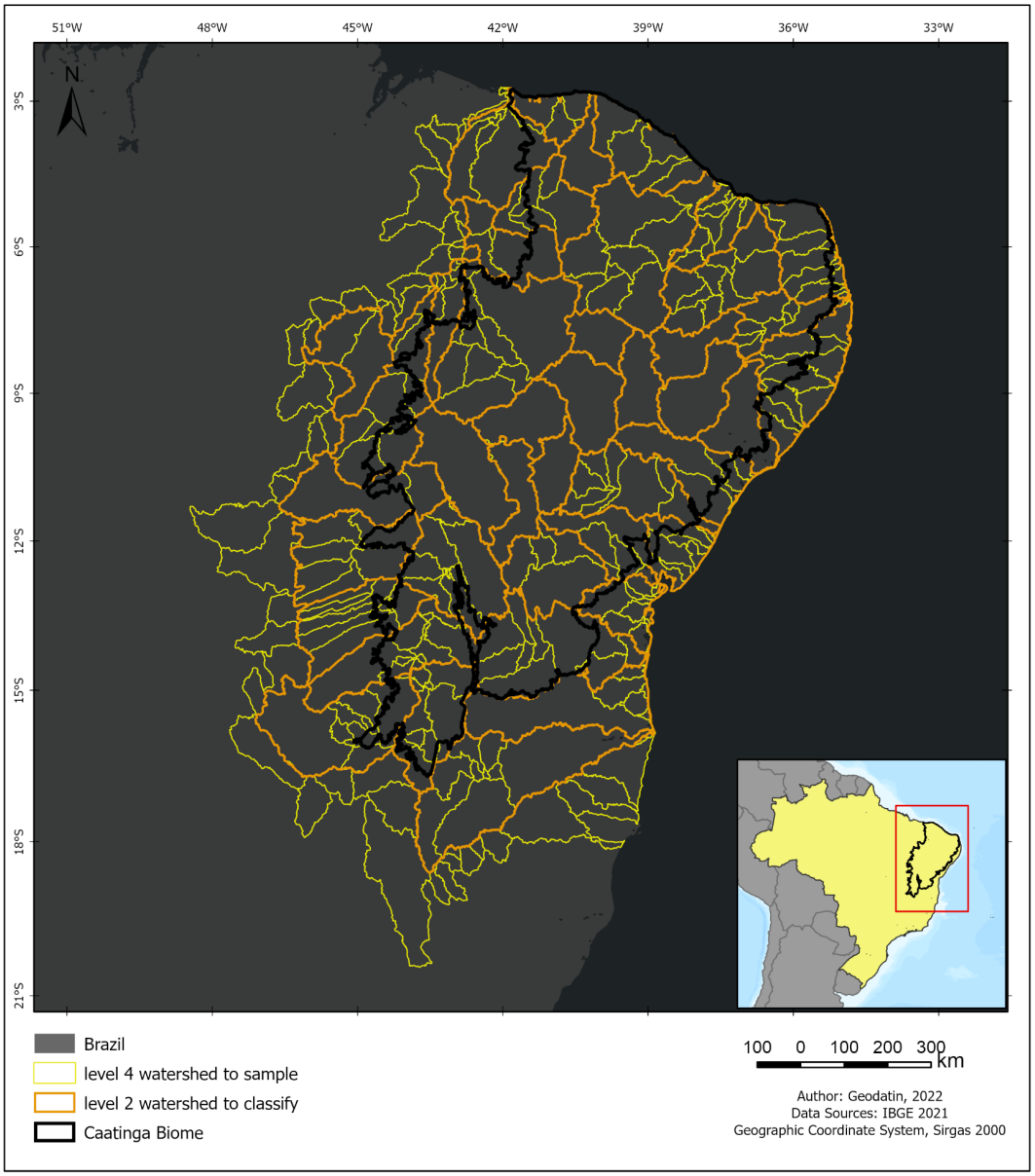

Figure 5. The Caatinga watersheds used in the classification and samples of Collections 7.

4. CLASSIFICATION

4.1 Class of cover to map

The digital classification of the Landsat mosaics in the Caatinga biome aimed to map a subset of ten LULC classes of the MapBiomas legend in Collection 7 (Table

2), some of these classes 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 the 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. Other classes were tuned with specific classifications, such as Rocky Outcrop and Other non Vegetated Areas.

Table 2. Land cover and land use classes considered for digital classification of Landsat mosaics in the Caatinga biome in the MapBiomas Collection 7.

Legend class	**ID**	**Natural / Anthropic**	**Land cover / Land use**	**General description**
1.1 Forest Formation	3	Natural	Land cover	Vegetation with predominance of continuous canopy-Savana- Estépica, Florestalada, Seasonal Semi-Deciduous and Deciduous Forest.
1.2 Savanna Formation	4	Natural	Land cover	Vegetation with predominance of semi-continuous canopy species - savanna- shrub savanna- savanna woodland.
1.4 Wooded Restinga	49	Natural	Land cover	Restinga vegetation includes herbaceous plant communities dominated by shrubs or small trees. These species are frequently wide-spread and occur in coastal areas of Southeastern Brazil
2.2 Grassland	12	Natural	Land cover	Vegetation with predominance of herbaceous species (steppe Savannah 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 the partial presence of rupicolous vegetation and high slope.
3.3 Mosaic of Agriculture and Pasture	21	Anthropic	Land use	Use agriculture areas where it was not possible to distinguish between pasture and agriculture.
4. Non vegetated Area	22	Anthropic	Land use	
4.4. Other non Vegetaded Areas	25	Anthropic	Land cover	
5. Water	33	Natural / Anthropic	Land cover / Land use	
6. Non Observed	27	non Observed	non Observed data	non Observed data

4.2 Sample process and feature selection

A sampling task is an expensive process for large areas in the GEE platform. The strategy of the sampling process was to select regions in the level 4 watershed, counting 320 regions to collect. For each region was sorting at least 500 samples per class, this condition forced the function ee.Image().stratifiedSample() collect samples in small areas in a specific class. These classes were. When all data samples are saved in the asset folder, they are regrouped at level 2 watershed, using merges of feature collections.

The last process with the samples is to remove outliers by class. Then the algorithm Learning Vector Quantization was implemented in the function ee.Clusterer.wekaLVQ() from Kohonen, 2003. This cluster algorithm allows a group of all samples in the new category. Then for each class is selected the first two groups of clusters with more pixels that belong to the same class in analysis. Later each feature is saved with x percent of the number class that the quantity be approximately 1000 pixels. Figure (6) shows an example of features from 2020 in Caatinga watersheds and distributions of quantities and percentages sample by class.

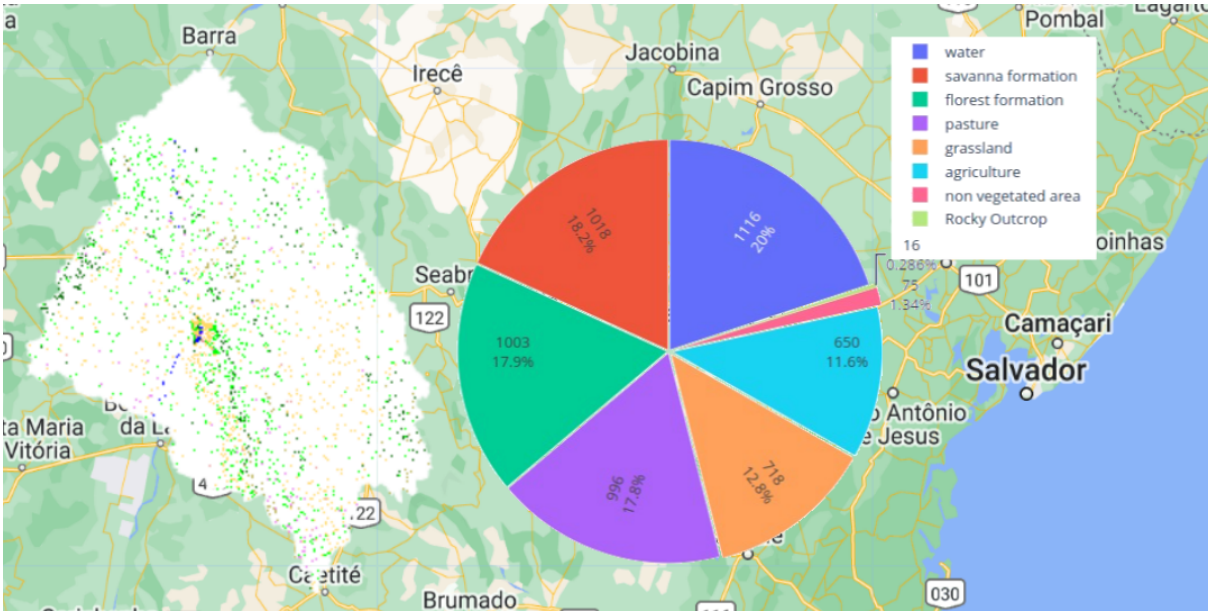

Figure 6. Map with distribution samples by class, and plot pie of distribution of the 2020 for one watershed region.

4.3 Feature space

The feature space for digital classification of the LULC classes in the Caatinga biome comprised a subset of 75 features (Table 3), taken from the complete feature space of MapBiomas Collection 7 (General ATBD MapBiomas, 2020).

Table 3: Feature space subset considered in the classification of Landsat image mosaics in the Caatinga biome in the MapBiomas Collection 7.

Bands	Estimators	Index Spectral	Estimators	Franctions	Estimators
blue	median	CAI	median	gv	amp
	median dry		median dry		median
	median wet		stdDev		media dry
	min	EVI2	amp	npv	median
green	median		median		median dry
	median dry		media dry		median wet
	median wet		stdDev		min
	median texture	GCVI	median	soil	median
	stdDev		median dry		median dry
red	median		median wet		median wet
	median dry	NDVI	amp		stdDev
	median wet		median	ndfi	median
	min		median dry		median dry
nir	median		median wet		median wet
	median dry	NDWI	amp		min
	median wet		median	sefi	median dry
	min		median dry		median wet
SWIR1	median		median wet		stdDev
	median wet	SAVI	median	shade	median
	min		median dry		median dry
	stdDev		median wet		median wet
SWIR1	median		stdDev		min
	median wet	PRI	median		amp
	min		median dry		
	stdDev		median wet		

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 first step was measuring the correlation between feature Collection, Figure 7, and correlated variables would be eliminated from the least important following the score. For to calculate correlation was used the function corr() from pandas library of python.

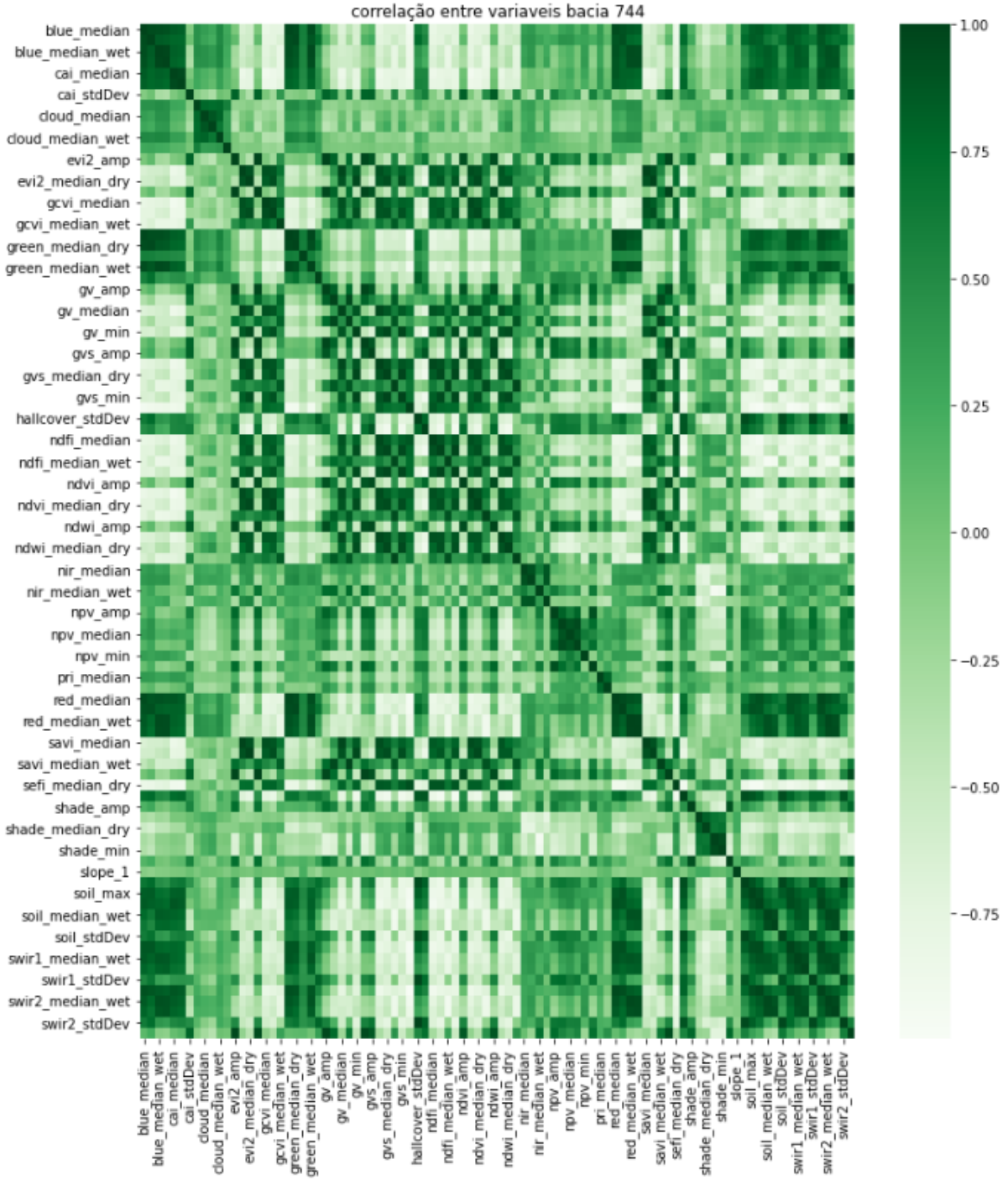

Figure 7. Example the plot correlation of watersheds samples from 2020 year.

The definition of this subset and the classifier parameters were made based son the weights on the function feature_importance from the score on Random Forest Classification implemented in Scikit-learn library python, Figure 8.

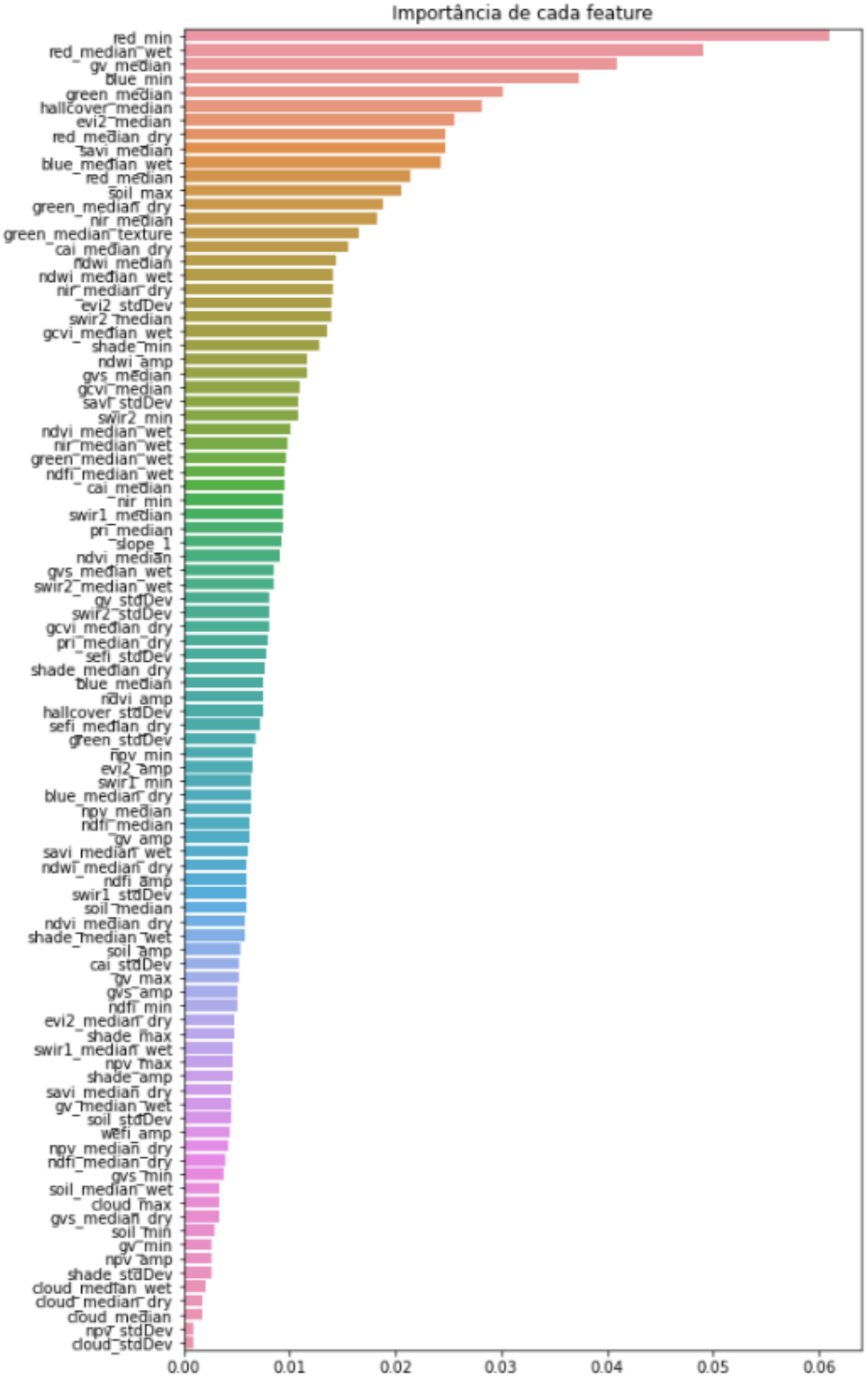

Figure 8. Example of the plot of one list of feature importance.

Later for each watershed sample a list of variables was saved to later be called in the classification stage. All codes used are available in the repository of MapBiomas's Github (https://github.com/mapbiomas-brazil/caatinga).

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.

Final classification was performed for all regions and years with samples. It was using 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 with no observation remain with no data.

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 21-MOSAIC 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 significant improvement was the fluctuation decrease in the extreme years of the mapped series (i.e. 1985 and 2019).

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 85,000 validation points. The figure below shows the result of the accuracy analysis for the level 3 legend of the MapBiomas

Collection 7 (1985-2018) (Figure 9). The metrics showing are historical and globals accuracy, allocation disagreement and quantity disagreement.

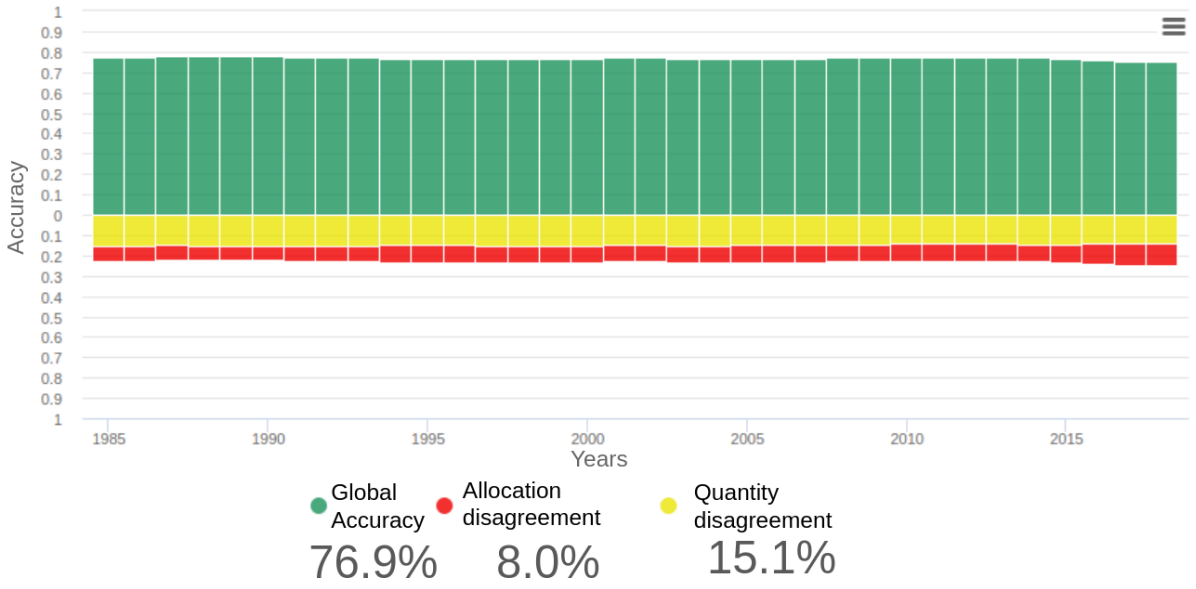

Figure 9. Accuracy of level 3 of MapBiomas Collection 7 in the Caatinga biome (1985-2018).

The methodology applied in this collection had higher accuracy than other collections before 7. The numbers that show these results are in Table 3. Another analysis used is to review the errors of omission and commission, Figures 10 and 11. With these errors we can understand which classes are confused with other classes in the classification. And from that analysis, draw up a new strategy to reduce those errors of commission and omission.

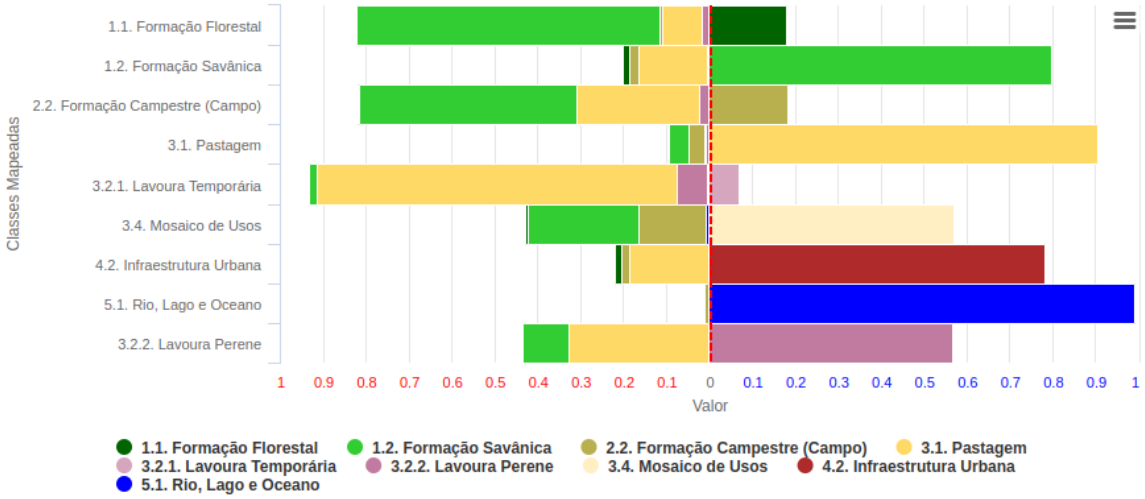

Figure 10. Commission errors of the land cover and land use mapping in the Caatinga.

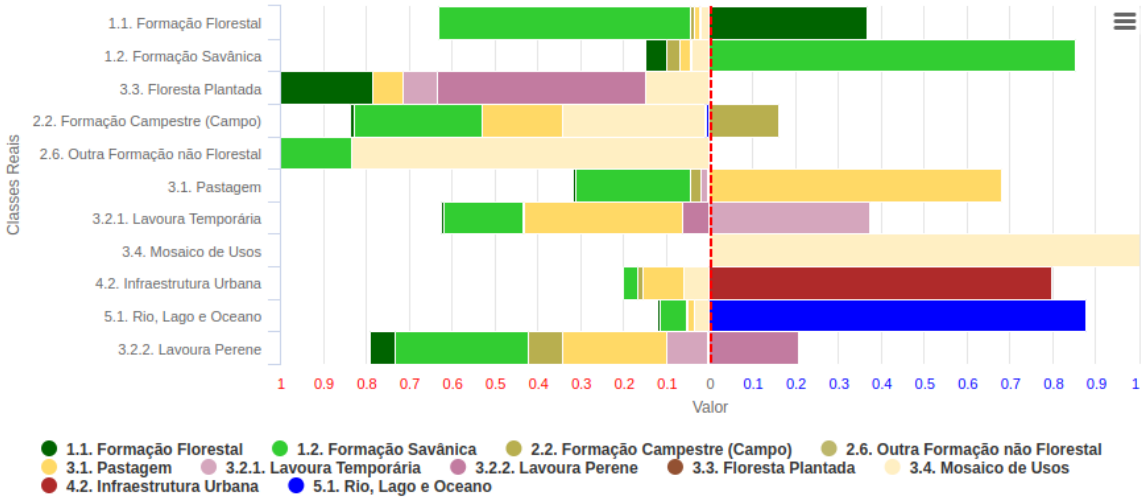

Figure 11. Omission errors of the land cover and land use mapping in the Caatinga.

Table 3. 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, with 34 years the points of references.

Collection	Method	Global Accuracy
3.1	Random Forest	Level 1: 80.0 % Level 2: 78.2 % Level 1: 71.3 %
4.1	Random Forest	Level 1: 81.9 % Level 2: 79.9 % Level 1: 74.3 %
5.0	Random Forest	Level 1: 81.8 % Level 2: 80.0 % Level 1: 75.4 %
6.0	Random Forest	Level 1: 81.1% Level 2: 75.0 % Level 1: 74.9 %
7.0	Random Forest	Level 1: 81.6 % Level 2: 76.9 % Level 1: 76.9 %

If we plot all values in the accuracy series then we can compare better to see all results of the other collections, Figure 12.

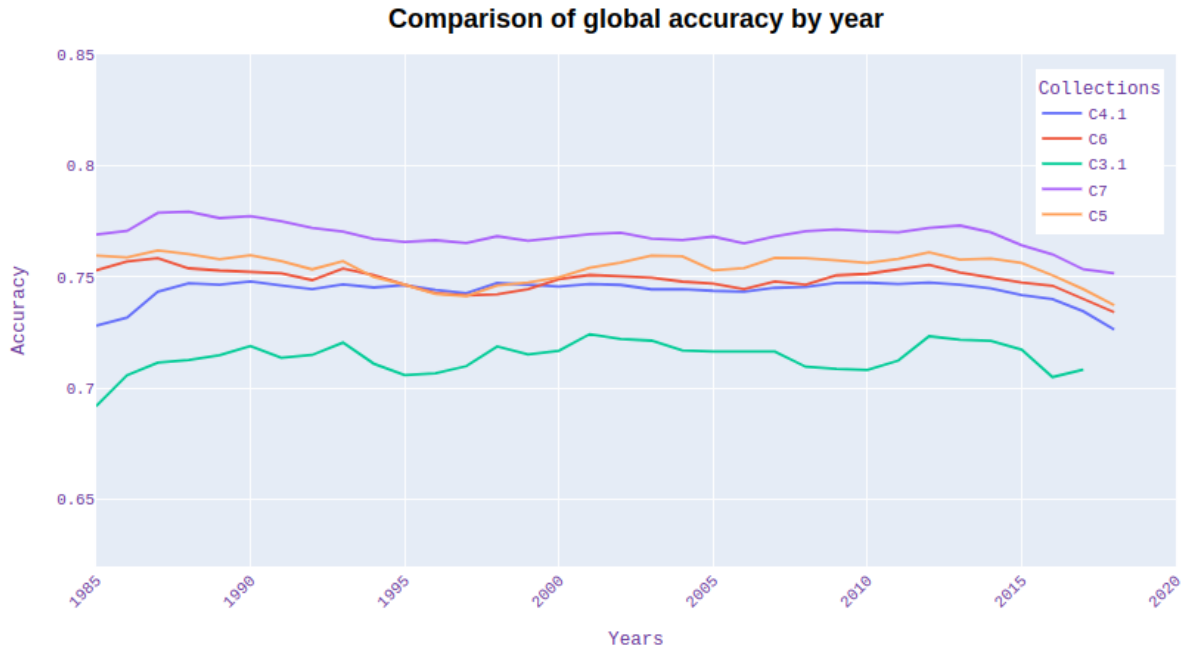

Figure 12. Plot of Accuracy of level 3 of MapBiomas Collections 3.1, 4.1, 5.0, 6.0 and 7.0 in the Caatinga biome (1985-2018 years).

Another way to measure the quality of map series is to analyze the behavior of the area by each class of land cover in the time series. The plots in Figure 13 show the area time series by class of cover. Some cover classes should not have a sudden change from one year to another, so knowing the behavior of the class we can identify these possible errors between the maps of consecutive years. When these errors are identified, it is a matter of correcting them with post-classification filters as explained above.

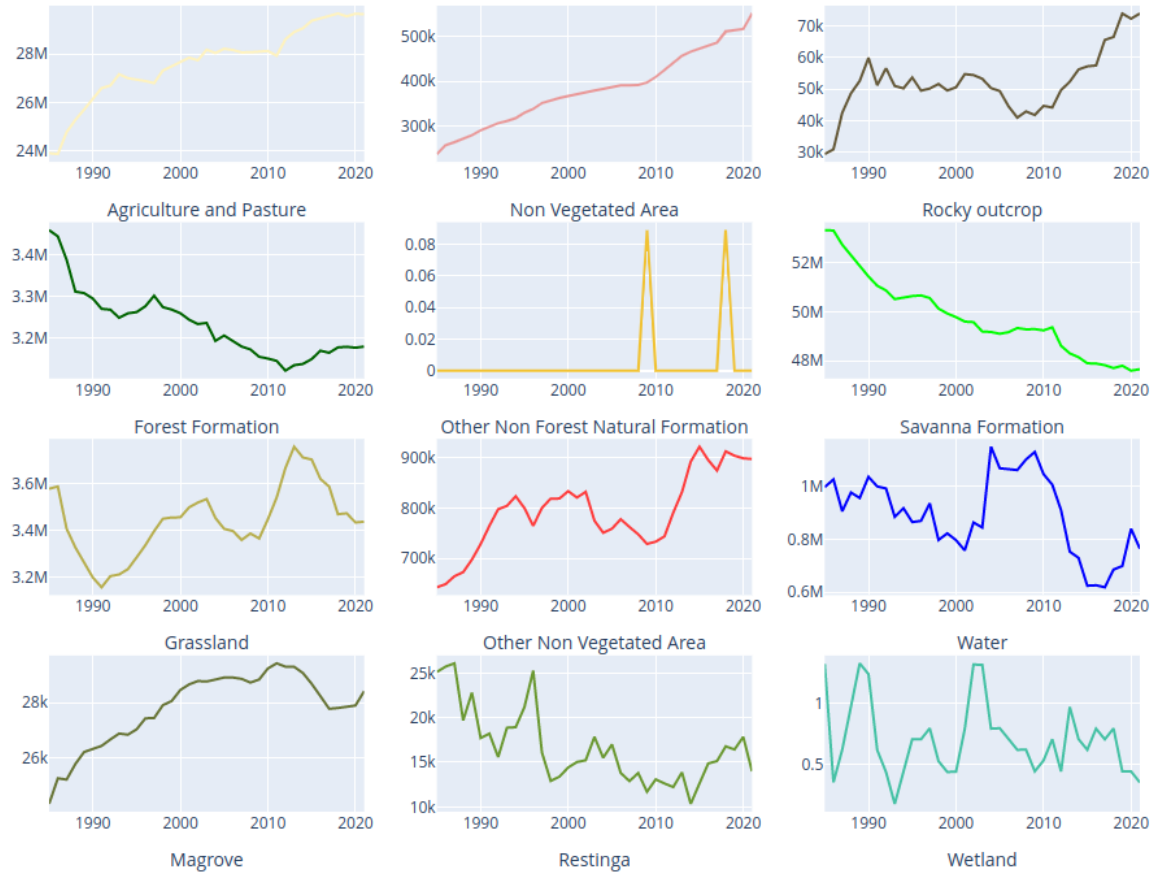

Figure 13. Plot of Area time series of level 3 of MapBiomas Collection 7.0 in the Caatinga biome (1985-2021 years).

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=download s, 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.

T. Kohonen, "Learning Vector Quantization", The Handbook of Brain Theory and Neural Networks, 2nd Edition, MIT Press, 2003, pp. 631-634.