Caatinga Appendix

Collection 8

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 offers a summary of the specific methods used in the generation of land cover and land use annual maps for the Caatinga biome in the context of MapBiomas. With each new collection, there was an increase in the number of land cover and land use classes or a revision of the employed method. For instance, from Collection 2.3 onwards, the Random Forest method started to be used in thematic classification, and the parameterization was no longer a trial and error-based process. Instead, the input parameter selection and optimization were achieved via algorithm applications. Another example pertains to the feature space, which is no longer selected by an empirical method, but feature selection algorithms capable of not only reducing dimensionality but also selecting 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 8.0 is described, as well as the improvements applied to the production of these maps. Other methods 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 collected 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 collected based on current classes mapped / New Spatio-temporal filters
5.0	1985 - 2019	Random Forest	Forest Formation, Savanna Formation, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Rocky Outcrop	Stable points, based on 5-years windows/ Feature Importance Analysis/New parameters 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 Sandbank Vegetation.	New class (Wooded Sandbank Vegetation)
7.1	1985 - 2021	Random Forest	Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Rocky Outcrop, Wooded Sandbank Vegetation.	
8.0	1985 - 2021	Random Forest / Gradient Tree Booster	Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Rocky Outcrop, Wooded Sandbank Vegetation.	

2. CLASSIFICATION METHOD

Figure 1 shows the process flow diagram used in the Collection 8 of the Caatinga biome. Since collection 6 some changes have been implemented with the idea of improving the results of the map classification flow. In general, the the process of building the maps of use and cover in the Caatinga Biome are divided into the following steps: Input data, sample collection and feature selections, hyper parameter tuning, classification models, post-classification filters, validation methods and visual inspection, integration of results with Mapbiomas.

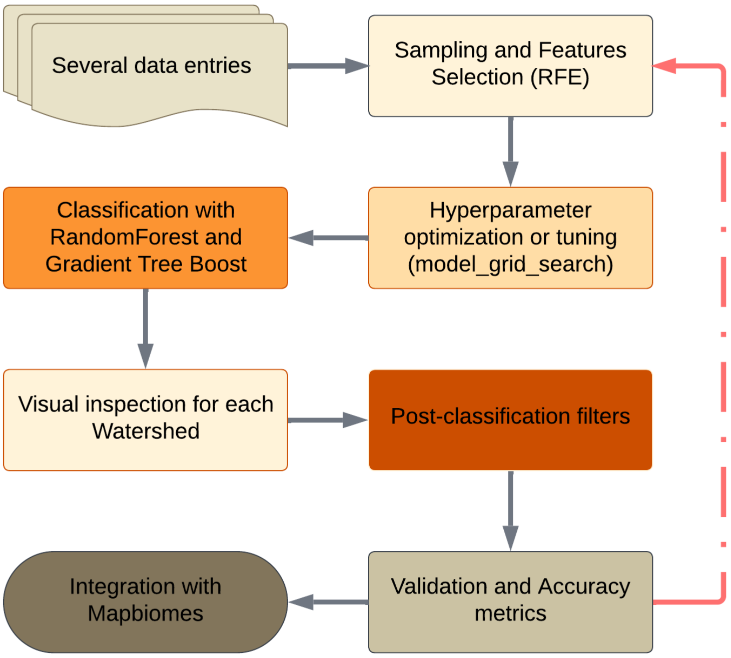

Several data entries
Sampling and Features Selection (RFE)
Classification with RandomForest and Gradient Tree Boost
Hyperparameter optimization or tuning (model_grid_search)
Visual inspection for each Watershed
Post-classification filters
Integration with Mapbiomes
Validation and Accuracy metrics

Figure 1. Simplified general flowchart.

For further details some improvements were added which will be described below (Figure 2).

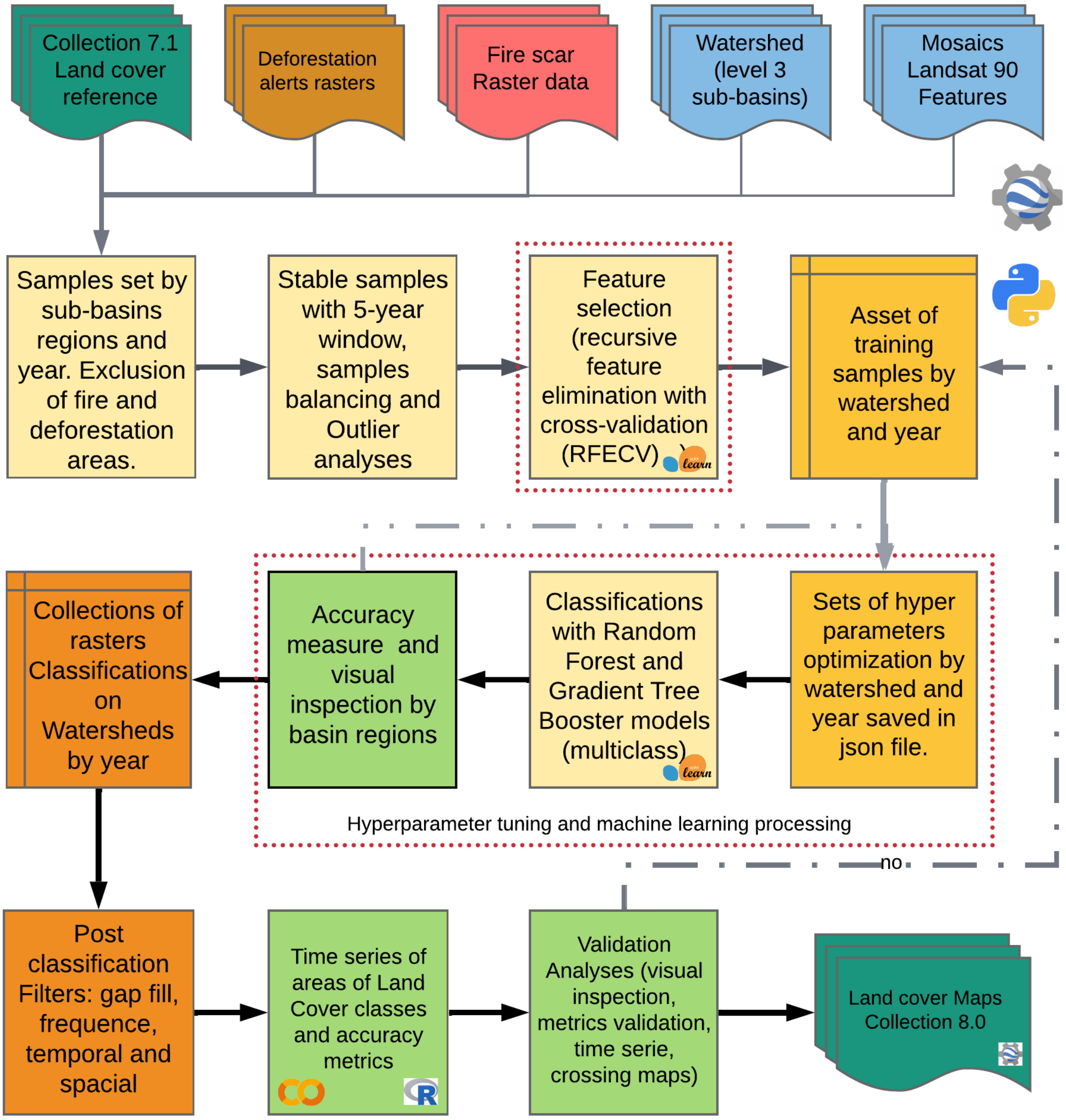

Collection 7.1 Land cover reference
Deforestation alerts rasters
Fire scar Raster data
Watershed (level 3 sub-basins)
Mosaics Landsat 90 Features
Samples set by sub-basins regions and year. Exclusion of fire and deforestation areas.
Stable samples with 5-year window, samples balancing and Outlier analyses
Feature selection (recursive feature elimination with cross-validation (RFECV)
Asset of training samples by watershed and year
Collections of rasters Classifications on Watersheds by year
Accuracy measure and visual inspection by basin regions
Classifications with Random Forest and Gradient Tree Booster models (multiclass)
Sets of hyper parameters optimization by watershed and year saved in json file.
Hyperparameter tuning and machine learning processing
no
Post classification Filters: gap fill, frequence, temporal and spacial
Time series of areas of Land Cover classes and accuracy metrics
Validation Analyses (visual inspection, metrics validation, time serie, crossing maps)
Land cover Maps Collection 8.0

Figure 2. Classification process of MapBiomas Collection 8 (1985-2022) in the Caatinga biome.

2.1 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.

2.2 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 season, so Caatinga savanna formations are expected to show great variation in spectral response throughout 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 3).

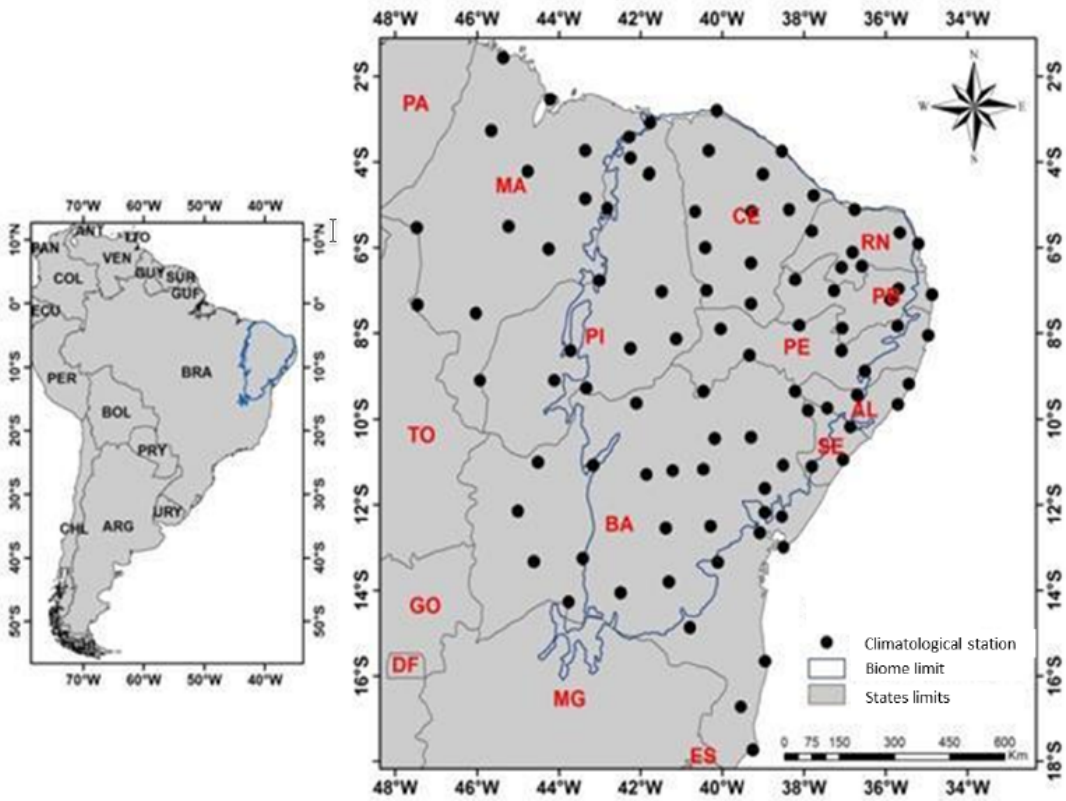


Figure 3. Location of the climatic stations used for the construction of the rainfall series for a selection of the mosaic periods in the Caatinga biome.

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 4) 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.

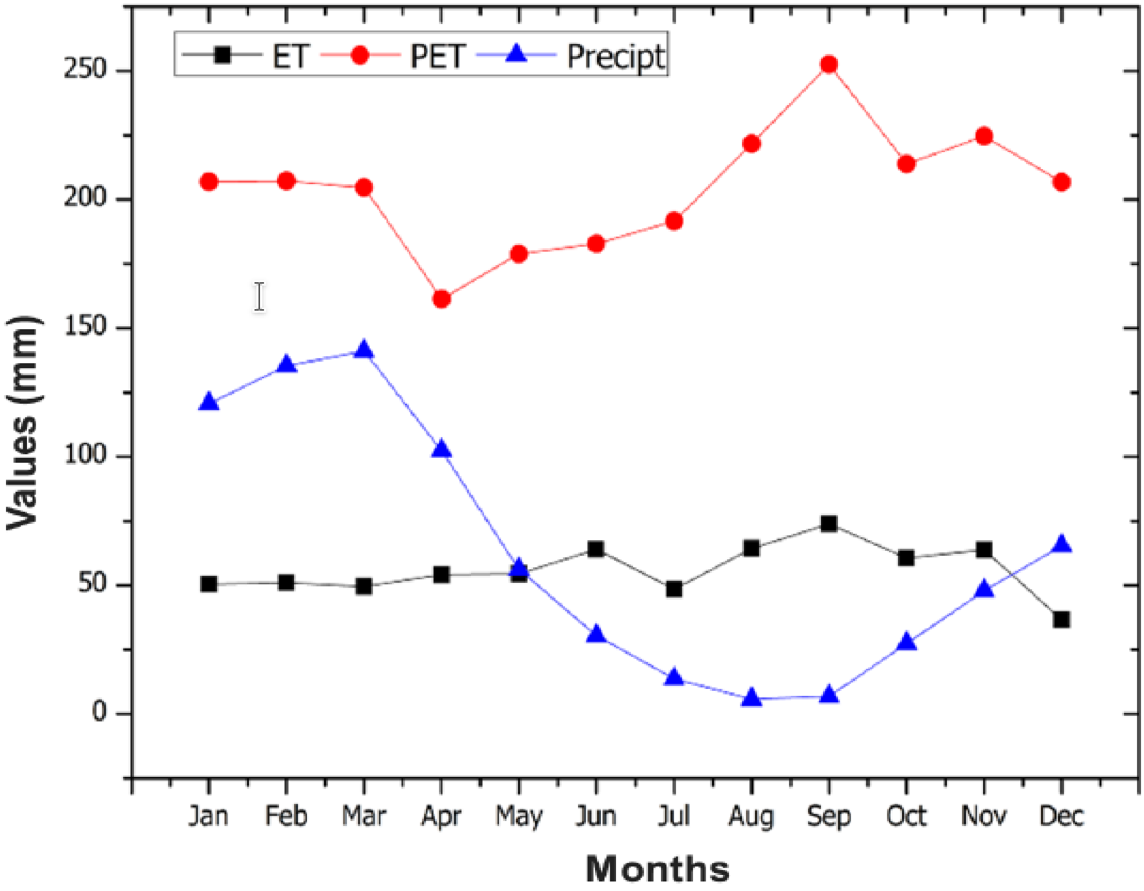


Figure 4. Temporal variation of water balance with monthly mean precipitation, evapotranspiration, and potential evapotranspiration variables in the Caatinga biome.

2.3 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.

2.4 Mosaic quality

The mosaic quality was evaluated using the frequency of each available pixel in the Caatinga biome (Figure 5). 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). There is no change for Collection 8.

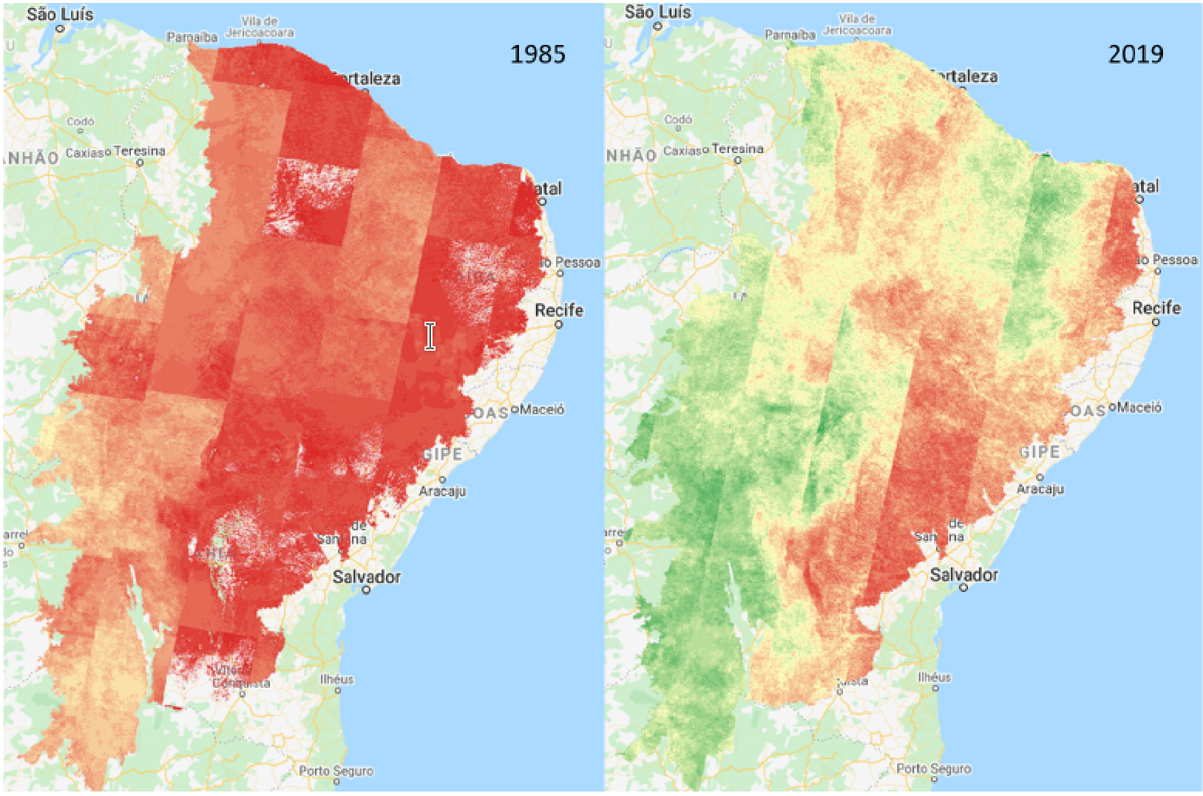


Figure 5. Landsat pixel availability in 1985 and 2019 in the Caatinga biome. Colors refer to data pixel availability, where red is low, yellow is medium, and green is high.

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 6). 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, 7 and 8 it was used the watershed limits 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.

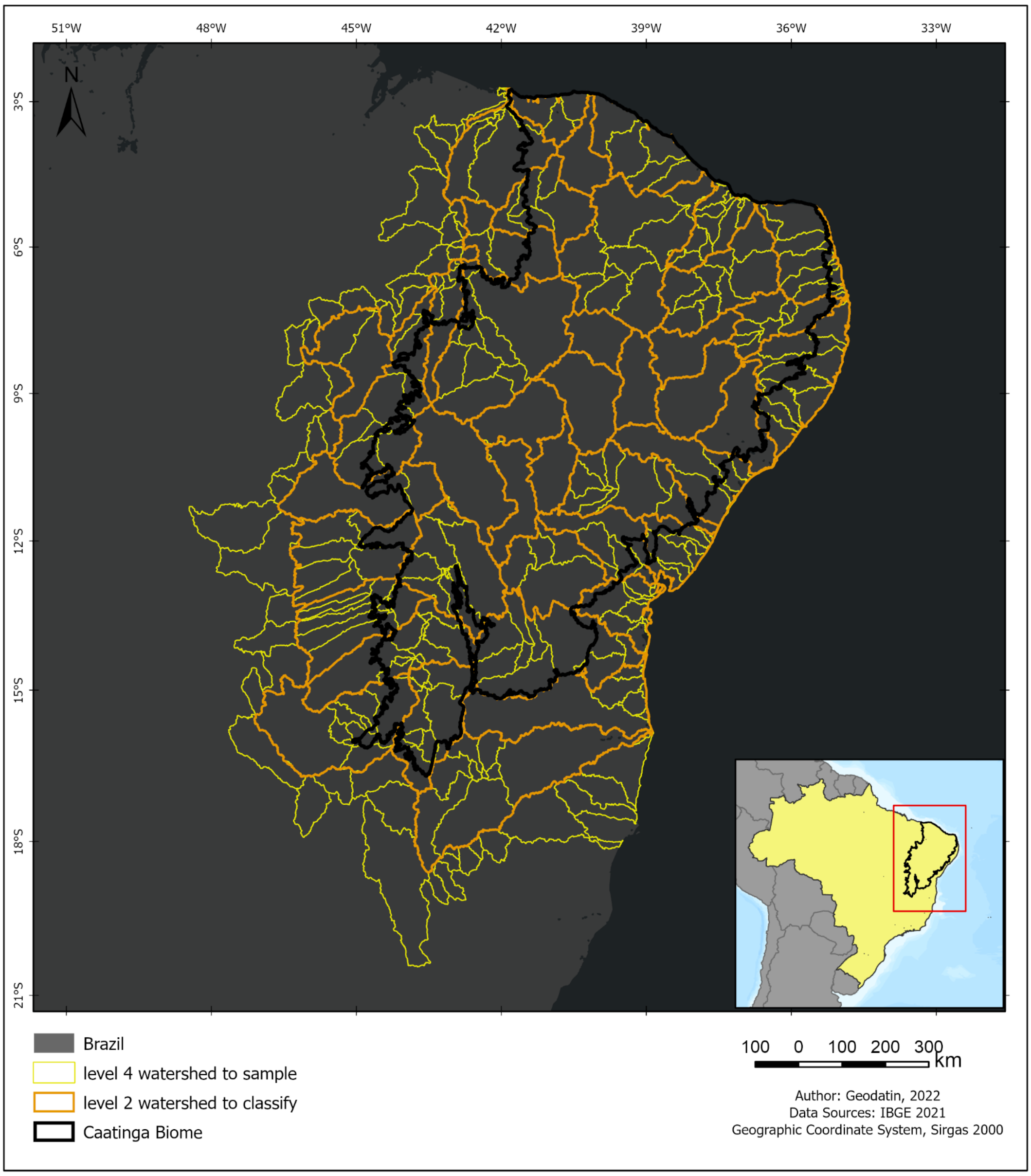


Figure 6. The Caatinga watersheds used in the classification and samples of Collections 7.

4. CLASSIFICATION

4.1 Land cover and land use classes

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 8 (Table

2). Some of these classes were integrated with the cross-cutting themes in a further step. The class Mosaic of Uses in the Caatinga was later superimposed by the Agriculture or Pasture classes, 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 8.

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 Sandbank Vegetation	49	Natural	Land cover	Wooded Sandbank 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 Uses	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	Beach and Dune, Urban Infrastructure and Mining.
4.4. Other non Vegetated Areas	25	Anthropic	Land cover	Non-permeable surface areas (infrastructure, urban expansion or mining) not mapped into their classes and regions of exposed soil in natural or crop areas. Mixed class that includes natural and anthropic areas.
5. Water	33	Natural /	Land cover /	Rivers, lakes, dams, reservoir and other

		Anthropic	Land use	water bodies
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 3 watershed, counting with 42 regions to collect. Each region was sorting at least 1200 samples per class, this condition forced the function ee.Image().stratifiedSample() collect samples in small areas in a specific class.

The first problem with collecting in this way is that classes with little presence in the sub-basin region will not have enough samples in the collection, and so the imbalance of samples by class is a natural process.

The spectral information is essentially derived from the mapbiomas mosaic, but after analyzing the first set of samples, a significant number of other spectral indexes were calculated from the bands 'blue_median', 'green_median', 'red_median', 'nir_median', 'swir1_median', 'swir2_median' present in the mosaic. The new indexes calculated were the following :

"ratio", "rvi", "awei", "iia", "gemi", "cvi", "gli", "afvi", "avi", "bsi", "brba", "dswi5", "lswi", "mbi", "ui", "osavi", "ri", "brightness", "wetness", "nir_contrast", "red_contrast"

The areas from which the points were collected went through 4 conditional layers: a layer indicating the areas of incident pixels, another indicating the areas of stable pixels in a 5-year window, areas where no deforestation occurred in the last 3 years of the series, and areas where no fire scars occurred. In this way, the collected points were stored in a folder in the mapbiomas asset, with each FeatureCollection indicating the region of the basin that was collected and the year.

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 it was 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 (7) shows an example of features from 2020 in Caatinga watersheds and distributions of quantities and percentages sample by class.

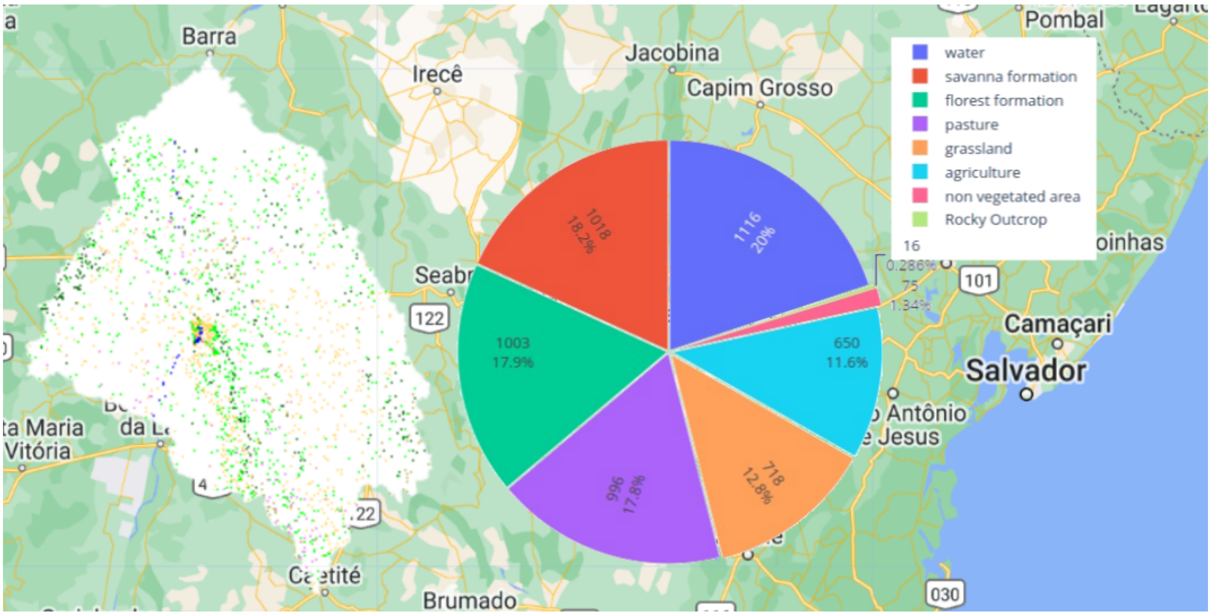


Figure 7. Map with distribution samples by class, and plot pie of distribution of the 2020 for one watershed region.

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). In Collection 8, a larger number of spectral indices were calculated to expand the feature space of the MapBiomas mosaic. The goal was to find a reduced space that offers more separability and contrast between targets. The image below (Figure 8) depicts an instance of the samples corresponding to sub-basin “744” which have an unbalanced distribution due to the nature of the data.

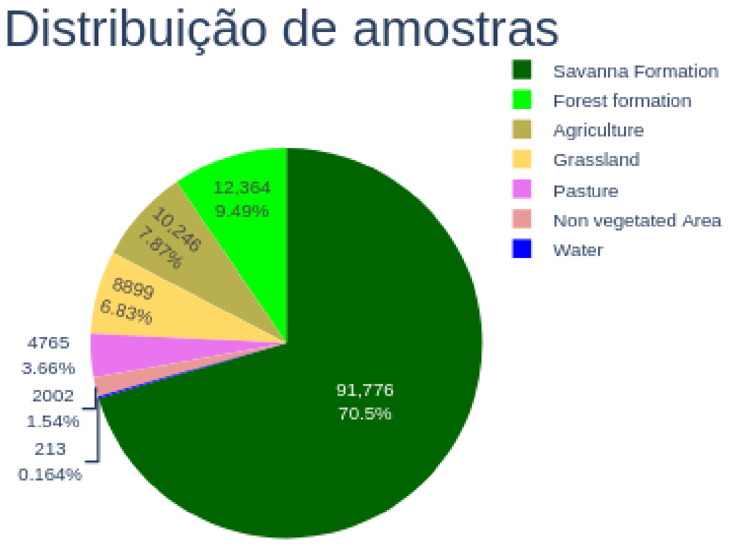


Figure 8: Distribution of samples for sub-basin 744 in the year 2000.

Achieving separability in the feature space is a prevalent challenge when performing remote sensing image classification in the Caatinga Biome. Figure 9 demonstrates that separability within a spectral band is limited for various targets in the image.

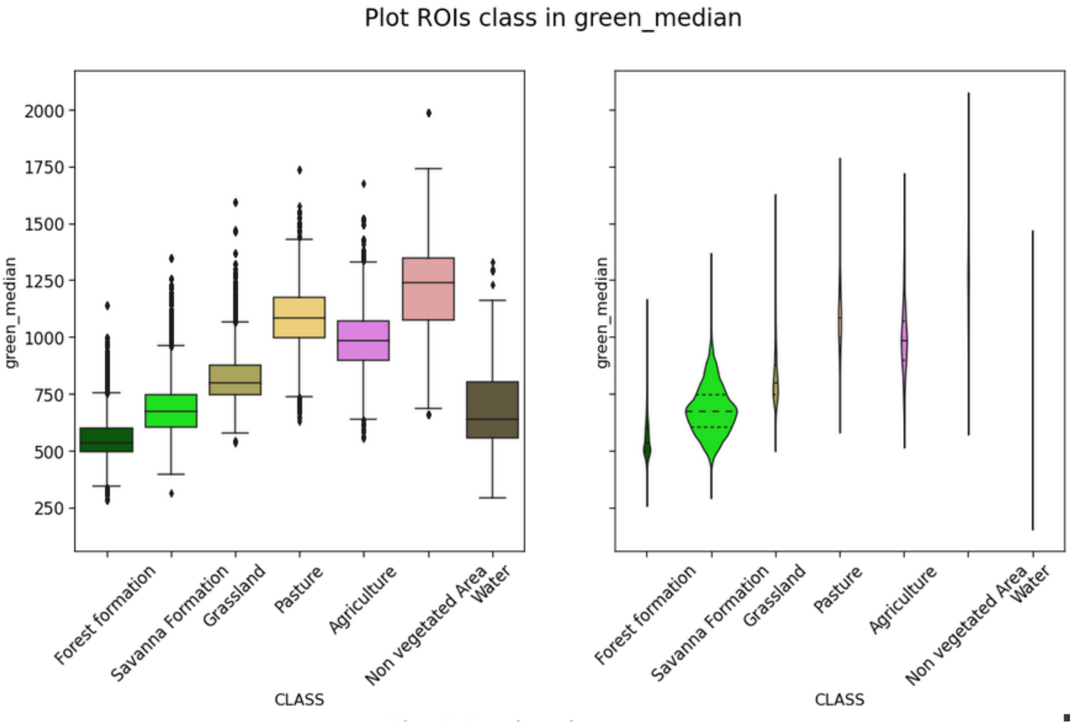


Figure 9: Box and violin plots from samples of spectral band “GREEN” in the main land cover classes mapped by the Caatinga team.

Another way of visualizing this can be seen in the image below (Figure 10), which plots the "blue_median", "green_median", "red_median", "nir_median" bands of the mosaic for six coverage classes.

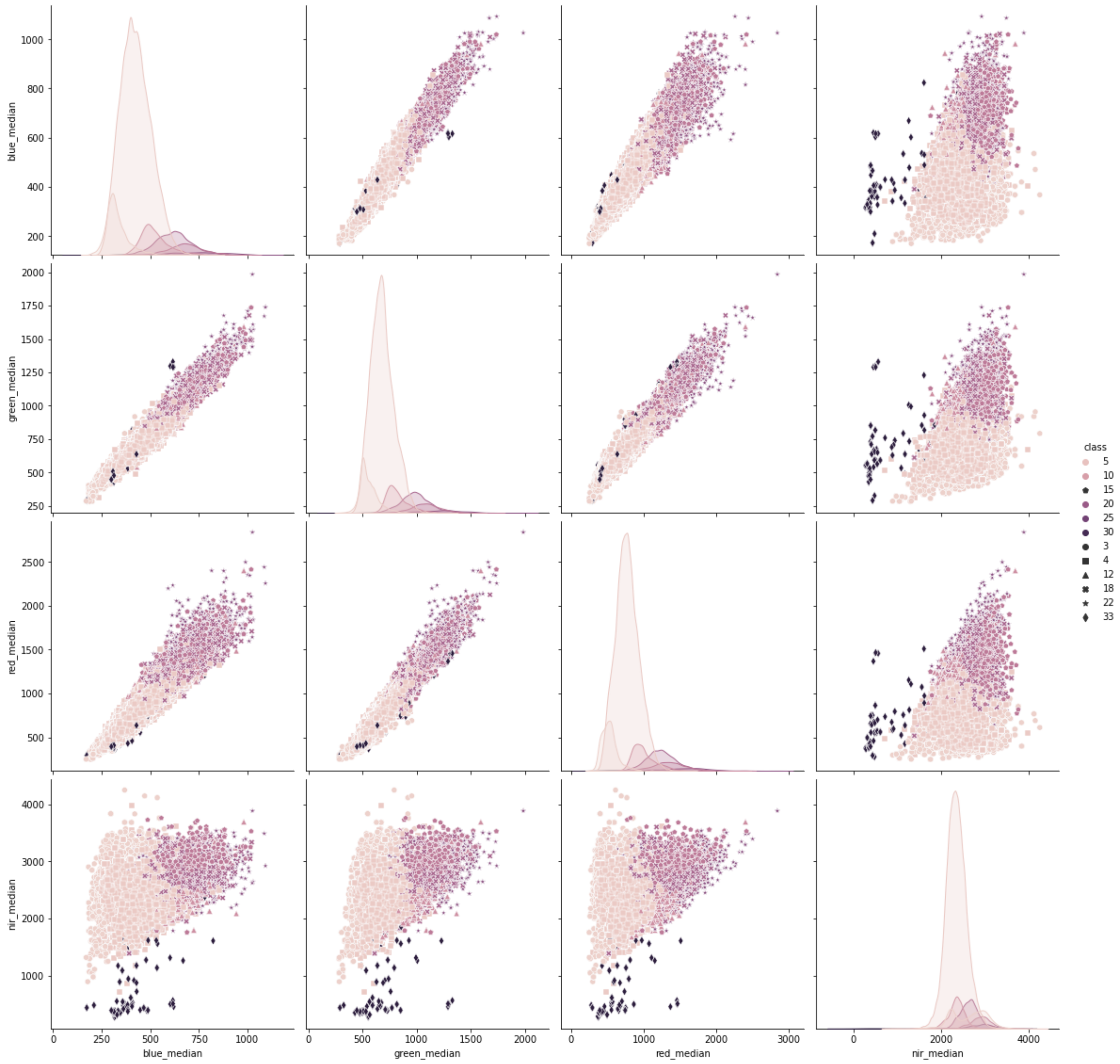
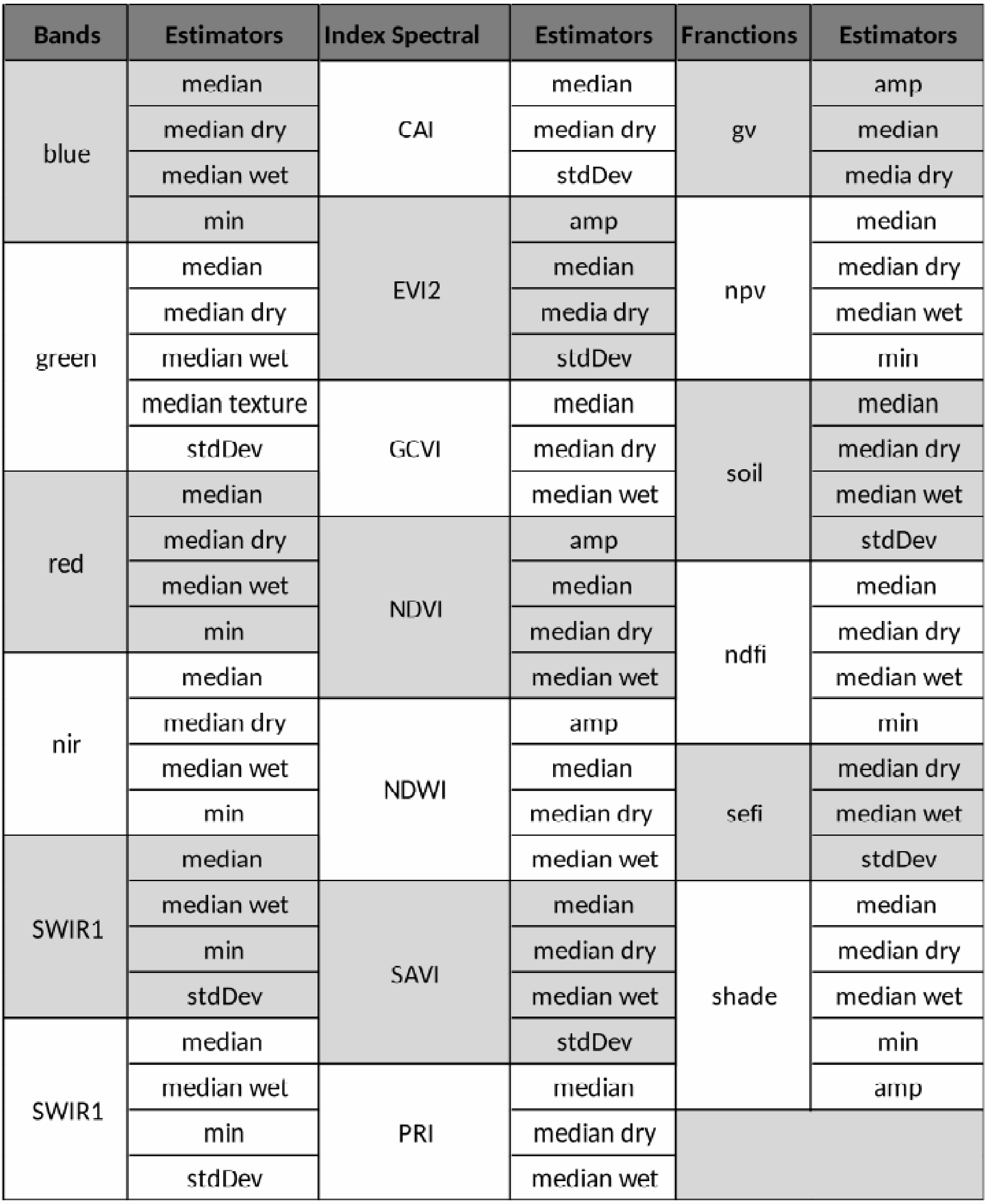


Figura 10: Distribuição espacial de amostras para as variáveis, “blue_median”, “green_median”, “red_median”, “nir_median”.

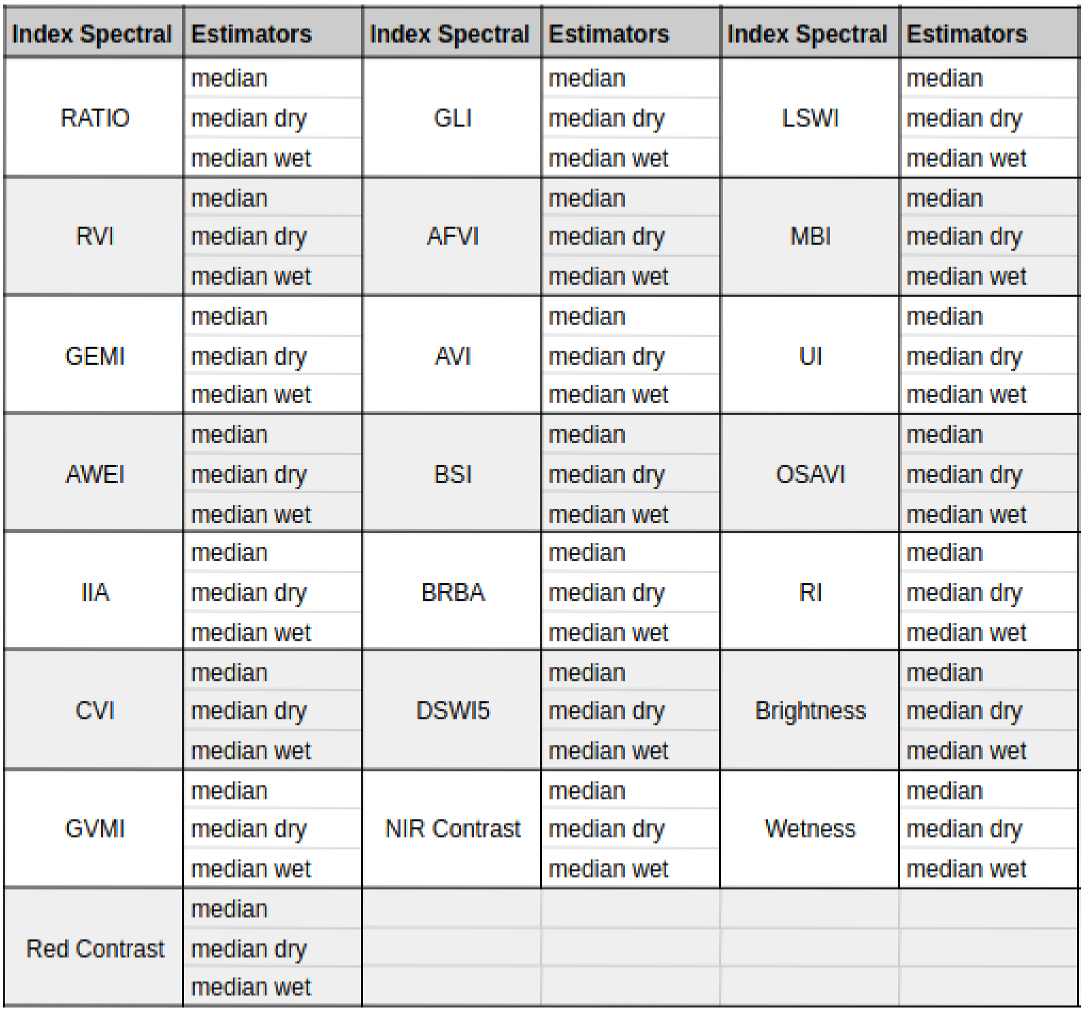
Table 3: Feature space subset considered in the classification of Landsat image mosaics in the Caatinga biome in the MapBiomas Collection 8.



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		

The feature space of this collection has been expanded to be more robust and to follow good data augmentation practices used in data science, see Table 4.

Table 4: Feature space subset indexes calculated from the estimated bands of the Landsat mosaic of mapBiomas in the Caatinga biome in the MapBiomas Collection 8.



Index Spectral	Estimators	Index Spectral	Estimators	Index Spectral	Estimators
RATIO	median	GLI	median	LSWI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
RVI	median	AFVI	median	MBI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
GEMI	median	AVI	median	UI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
AWEI	median	BSI	median	OSAVI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
IIA	median	BRBA	median	RI	median
	median dry		median dry		median dry
	median wet		median wet		median wet
CVI	median	DSWI5	median	Brightness	median
	median dry		median dry		median dry
	median wet		median wet		median wet
GVMI	median	NIR Contrast	median	Wetness	median
	median dry		median dry		median dry
	median wet		median wet		median wet
Red Contrast	median				
	median dry				
	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 variables (Figure 1)1, and some variables would be eliminated from the least important criteria following the score. To calculate correlation it was used the function corr() from Pandas Library of Python.

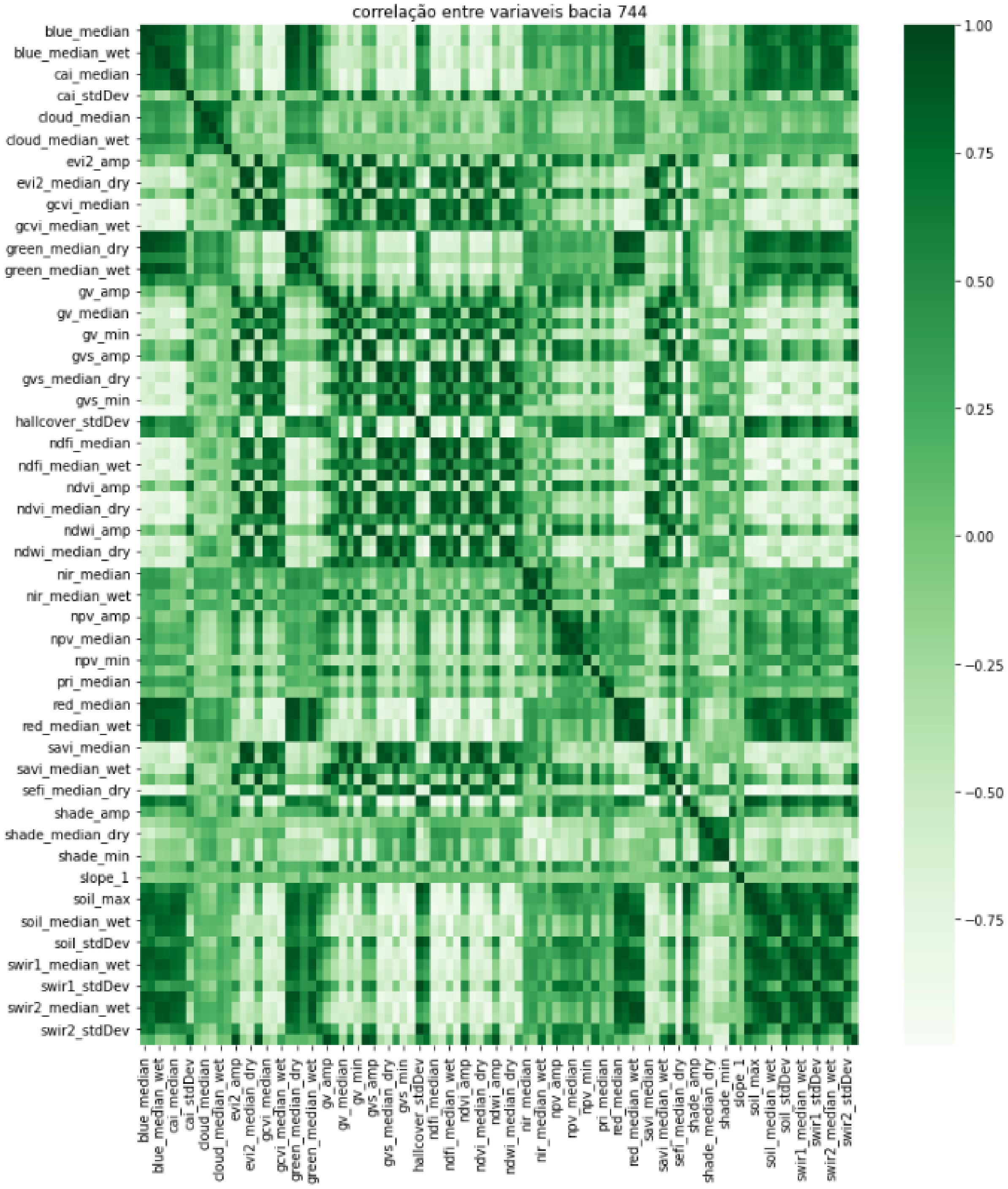
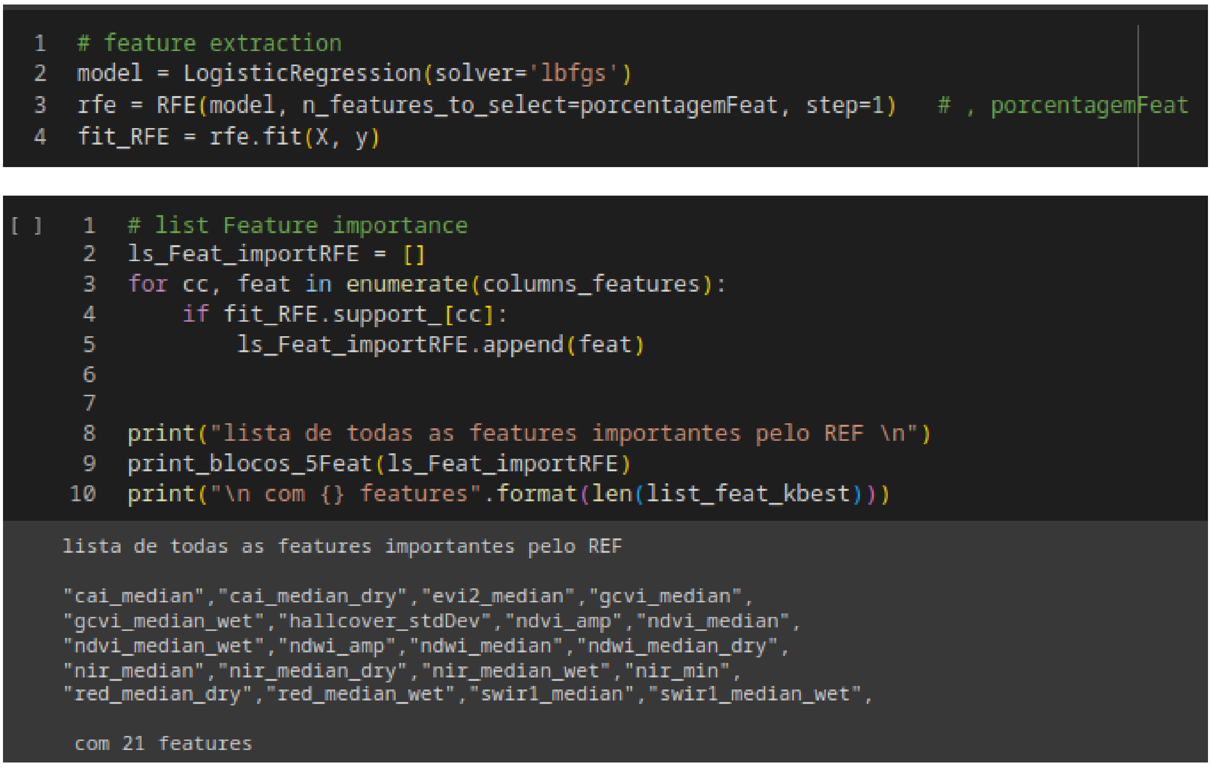


Figure 11. Example the plot correlation of watersheds samples from 2020 year.

Unlike the previous collections, this collection was implemented with the Recursive Feature Elimination (RFE), an alternative feature selection method that automatically tunes the number of features selected with cross-validation. Therefore, for each set of samples (basin / year), was saved a list of features selected from the feature elimination process (ZHANG AND JIANWEN, 2009; RAMEZAN, 2022). A simple example can be accessed at:

https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html

The RFE() class can be used by calling the python sklearn library (Figure 12). There are two methods of the class that can be used to filter the selected variables: the "support_()" method and the "ranking_" method. with the former we can select the surviving variables from a list of "TRUE" or "FALSE", and with the latter we can extract the ranking of the "TRUE" variables.



```
# feature extraction
model = LogisticRegression(solver='lbfgs')
rfe = RFE(model, n_features_to_select=porcentagemFeat, step=1)   # , porcentagemFeat
fit_RFE = rfe.fit(X, y)
```

```
# list Feature importance
ls_Feat_importRFE = []
for cc, feat in enumerate(columns_features):
    if fit_RFE.support_[cc]:
        ls_Feat_importRFE.append(feat)


print("lista de todas as features importantes pelo REF \n")
print_blocos_5Feat(ls_Feat_importRFE)
print("\n com {} features".format(len(list_feat_kbest)))
```

```
lista de todas as features importantes pelo REF

"cai_median","cai_median_dry","evi2_median","gcvi_median",
"gcvi_median_wet","hallcover_stdDev","ndvi_amp","ndvi_median",
"ndvi_median_wet","ndwi_amp","ndwi_median","ndwi_median_dry",
"nir_median","nir_median_dry","nir_median_wet","nir_min",
"red_median_dry","red_median_wet","swir1_median","swir1_median_wet",

 com 21 features
```

Figure 12: Example of the implemented feature selection function (RFE) and a list of selected variables.

A script was implemented for the **Hyperparameter Tuning** process after selecting the variable sets by drainage basin and year. The GridSearchCV() function, along with the Pipeline() function, is capable of testing various parameter combinations for the model. It is then possible to establish which combination of parameters represents the best score or accuracy. The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid. An example of the "learning rate" parameters and "n estimators" is shown in figure 13, where the optimal pair of parameters would be (40, 0.175).

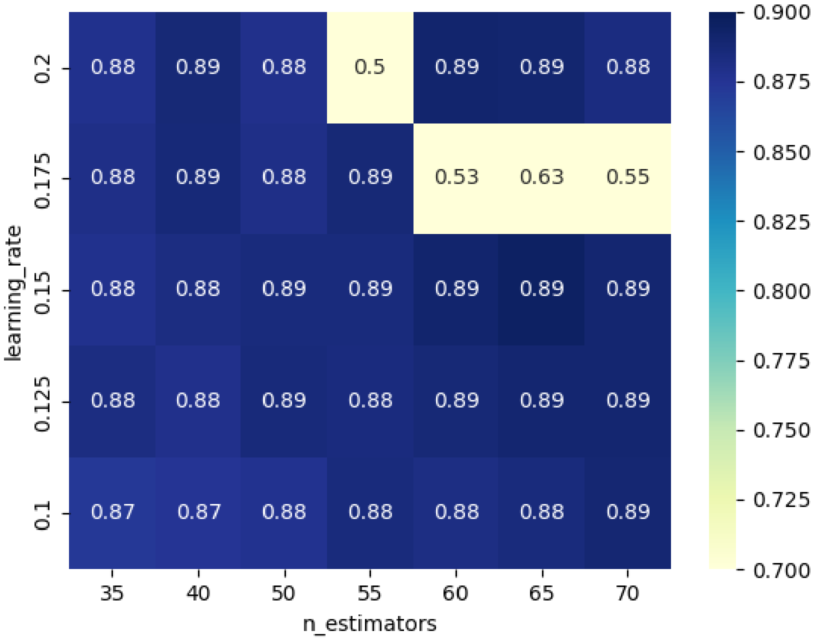
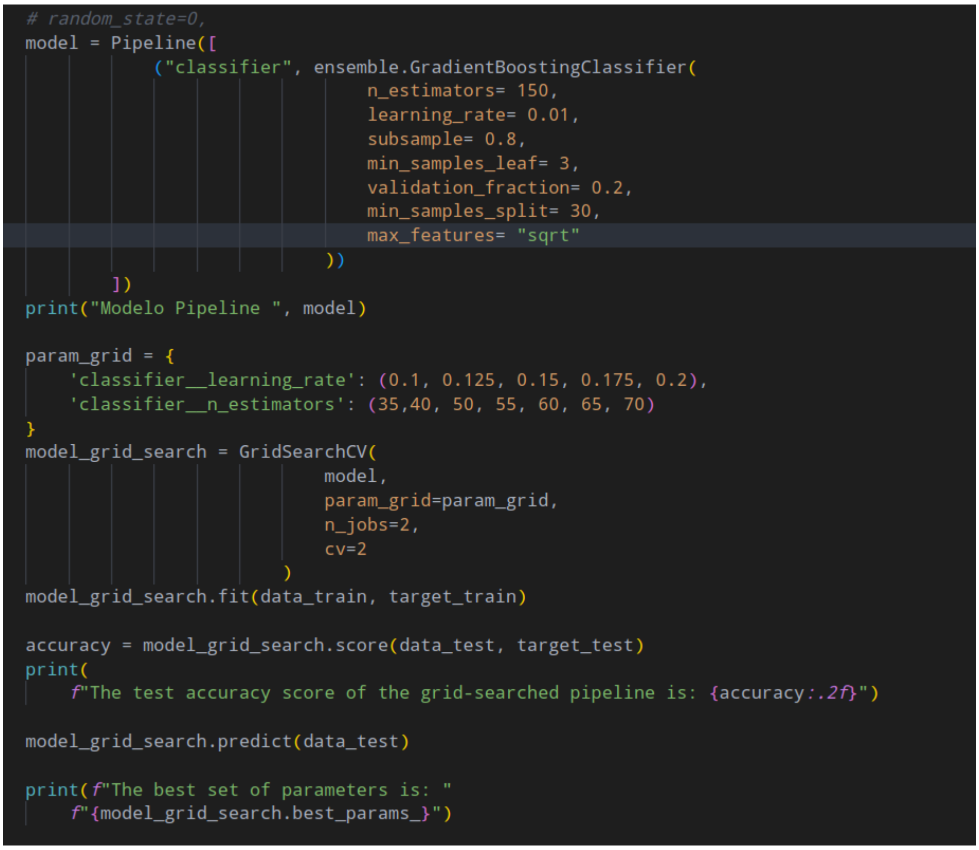


Figure 13. Example of the plot of combination of "learning rate" parameters and "n estimators".

Part of the code implemented for selecting optimal parameters is shown in the following image (Figure 14). Each pair of optimal parameters for year and hydrographic region is saved in a single json file.



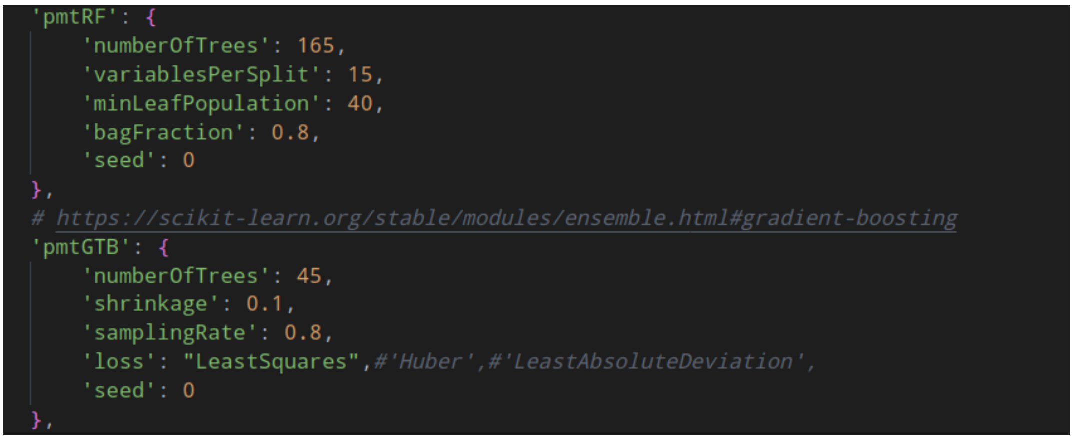
```
# random_state=0,
model = Pipeline([
            ("classifier", ensemble.GradientBoostingClassifier(
                                n_estimators= 150,
                                learning_rate= 0.01,
                                subsample= 0.8,
                                min_samples_leaf= 3,
                                validation_fraction= 0.2,
                                min_samples_split= 30,
                                max_features= "sqrt"
                            ))
        ])
print("Modelo Pipeline ", model)

param_grid = {
    'classifier__learning_rate': (0.1, 0.125, 0.15, 0.175, 0.2),
    'classifier__n_estimators': (35,40, 50, 55, 60, 65, 70)
}
model_grid_search = GridSearchCV(
                            model,
                            param_grid=param_grid,
                            n_jobs=2,
                            cv=2
                        )
model_grid_search.fit(data_train, target_train)

accuracy = model_grid_search.score(data_test, target_test)
print(
    f"The test accuracy score of the grid-searched pipeline is: {accuracy:.2f}")

model_grid_search.predict(data_test)

print(f"The best set of parameters is: "
    f"{model_grid_search.best_params_}")
```

Figure 14: Part of the code implemented for the Hyperparameter tuning process.

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).

4.4 Classification algorithm, training samples, and parameters

During the classification stage, the input data is adjusted to allow for the classification of the mapBiomas mosaics by hydrographic basin and year. Next, the data is visualized using a GEE script and examined by the team's analysts to assess the classification results by basin and year. The primary objective of this stage is to identify basins that require additional samples or classification parameter changes. Once identified, the team considers these areas for inclusion in the map correction cycle. During each round of classification, two map versions are simultaneously reviewed. One version is produced using the Random Forest classification (BREIMAN 2001), and the other version is the result of the Gradient Tree Booster classification (LAWRENCE et al. 2004). An example of the parameters for both classifiers is shown in figure 15.



```
'pmtRF': {
    'numberOfTrees': 165,
    'variablesPerSplit': 15,
    'minLeafPopulation': 40,
    'bagFraction': 0.8,
    'seed': 0
},
# https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting
'pmtGTB': {
    'numberOfTrees': 45,
    'shrinkage': 0.1,
    'samplingRate': 0.8,
    'loss': "LeastSquares",#'Huber',#'LeastAbsoluteDeviation',
    'seed': 0
},
```

Figure 15: Example parameters for the Random Forest and Gradient Tree Boost classifiers.

Final classification was performed for all regions and years with samples. The same subset of samples was used 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 MODE value of its eight neighbor’s pixels.

5.3 Temporal filter

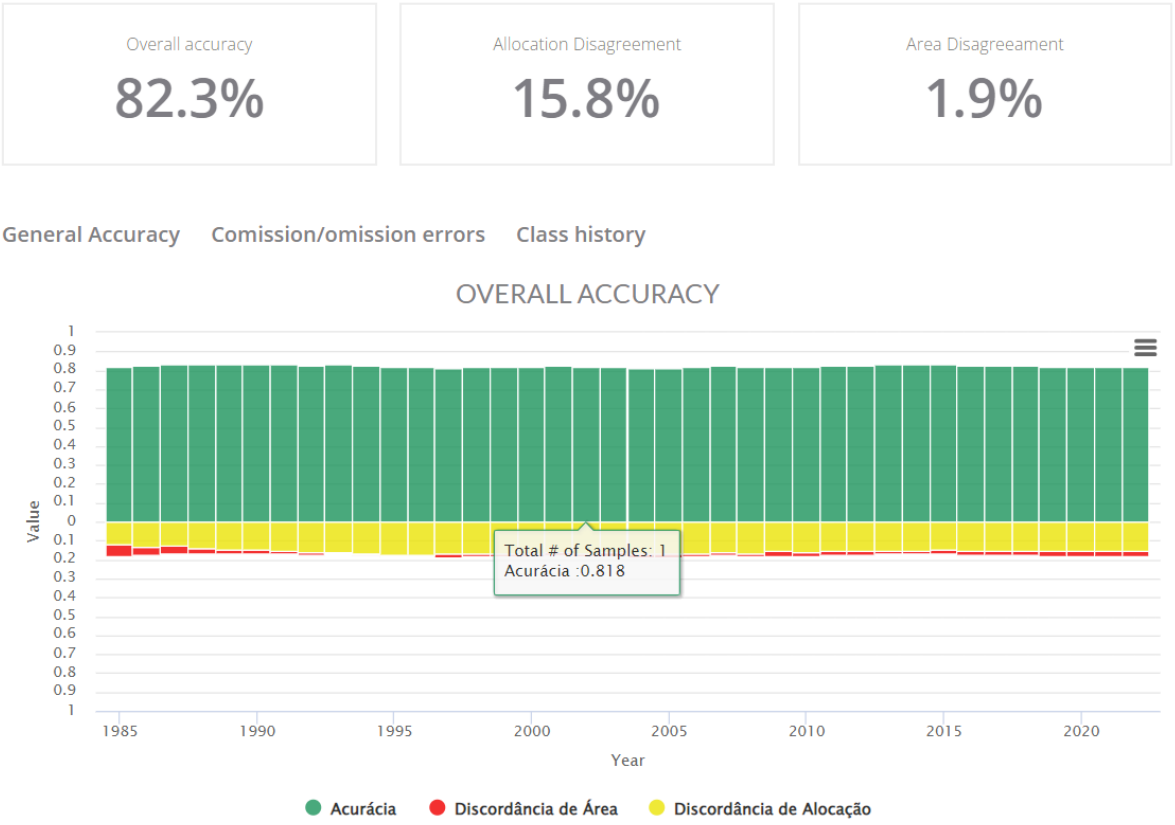
The applied temporal filter uses the subsequent years to replace pixels that have invalid transitions. In the first process, the filter looked for any natural class (3-FOREST FORMATION, 4-SAVANNA FORMATION, 12-GRASSLAND, 13-OTHERS NO FOREST FORMATION) that was not this class in 85 and was equal to these 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.

5.4 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 the 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 16). The metrics showing are historical and global accuracy, allocation disagreement and quantity disagreement.



Overall accuracy
82.3%

Allocation Disagreement
15.8%

Area Disagreeament
1.9%

Figure 16. Accuracy of level 3 of MapBiomas Collection 8 in the Caatinga biome (1985-2022).

The methodology applied in this collection had higher accuracy than other collections before 7. The numbers that show these results are in Table 4. Another analysis used is to review the errors of omission and commission, Figures 17 and 18. 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.

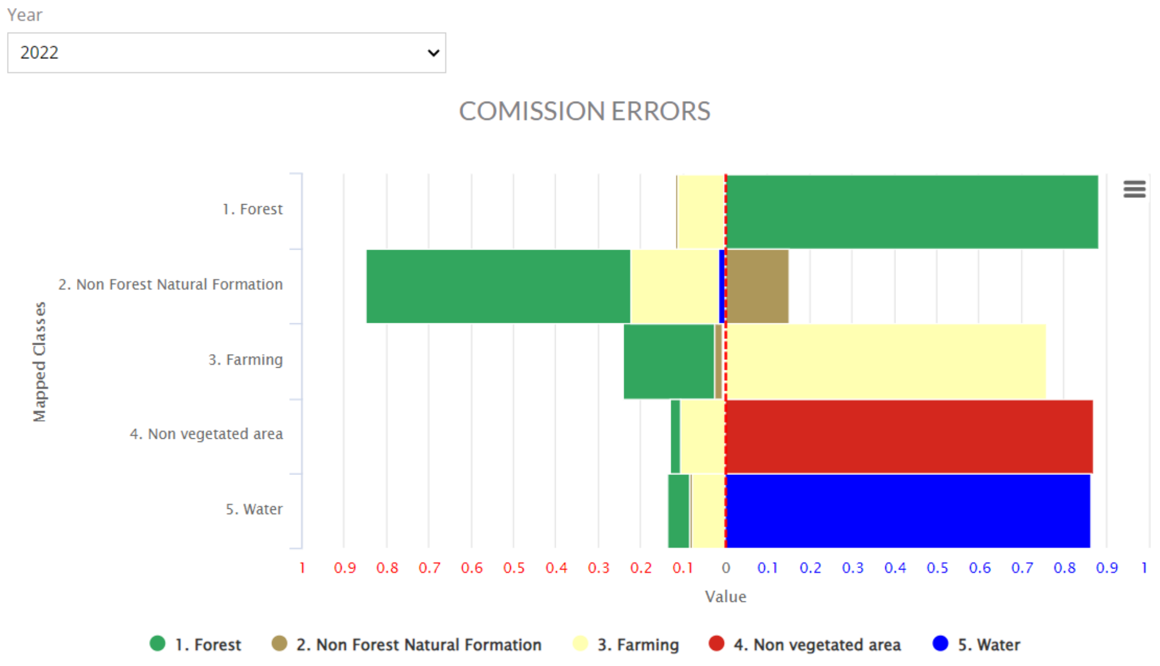


Figure 17. Commission errors of the land cover and land use mapping in the Caatinga.

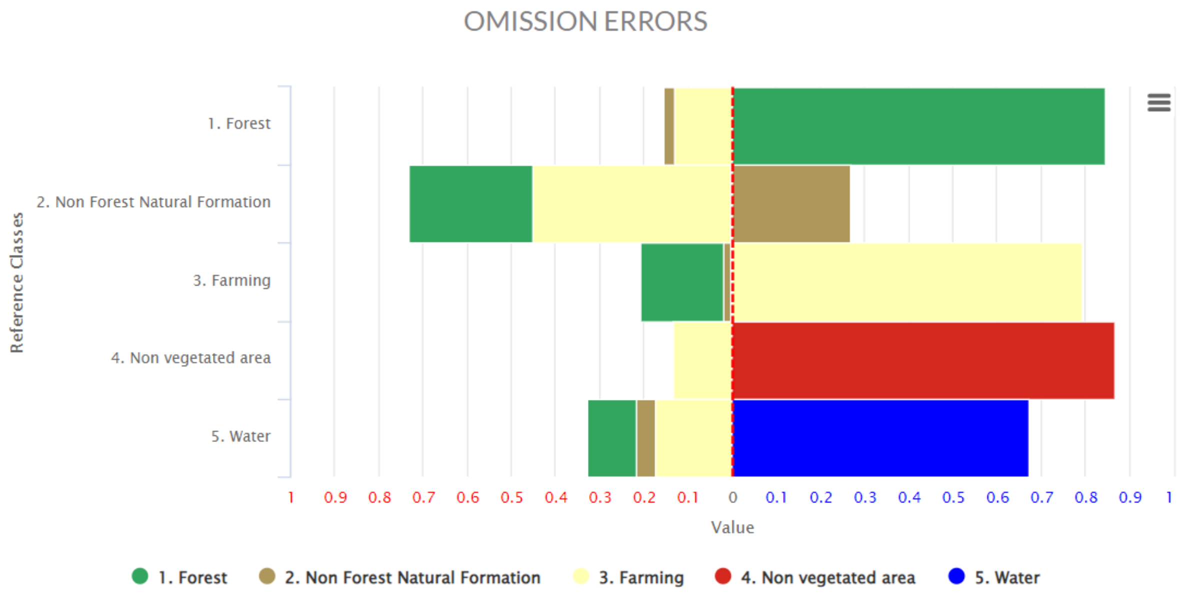


Figure 18. Omission errors of the land cover and land use mapping in the Caatinga.

Table 4. 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 %
7.1	Random Forest	Level 1: 81.6 % Level 2: 76.9 % Level 1: 76.9 %
8.0	Random Forest / Gradient Tree Booster	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 19.

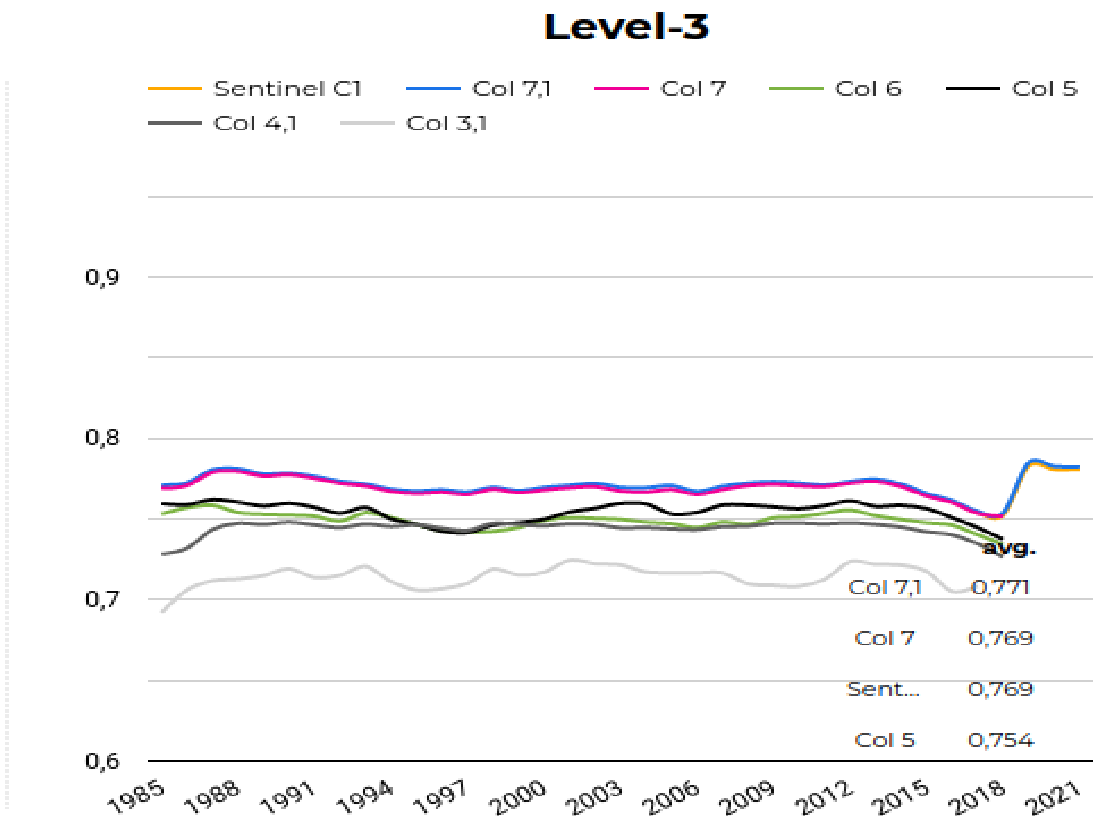


Figure 19. Plot of Accuracy of level 3 of MapBiomas Collections 3.1, 4.1, 5.0, 6.0, 7.0, 7.1 and 8.0 in the Caatinga biome (1985-2018 years).

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 20 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.

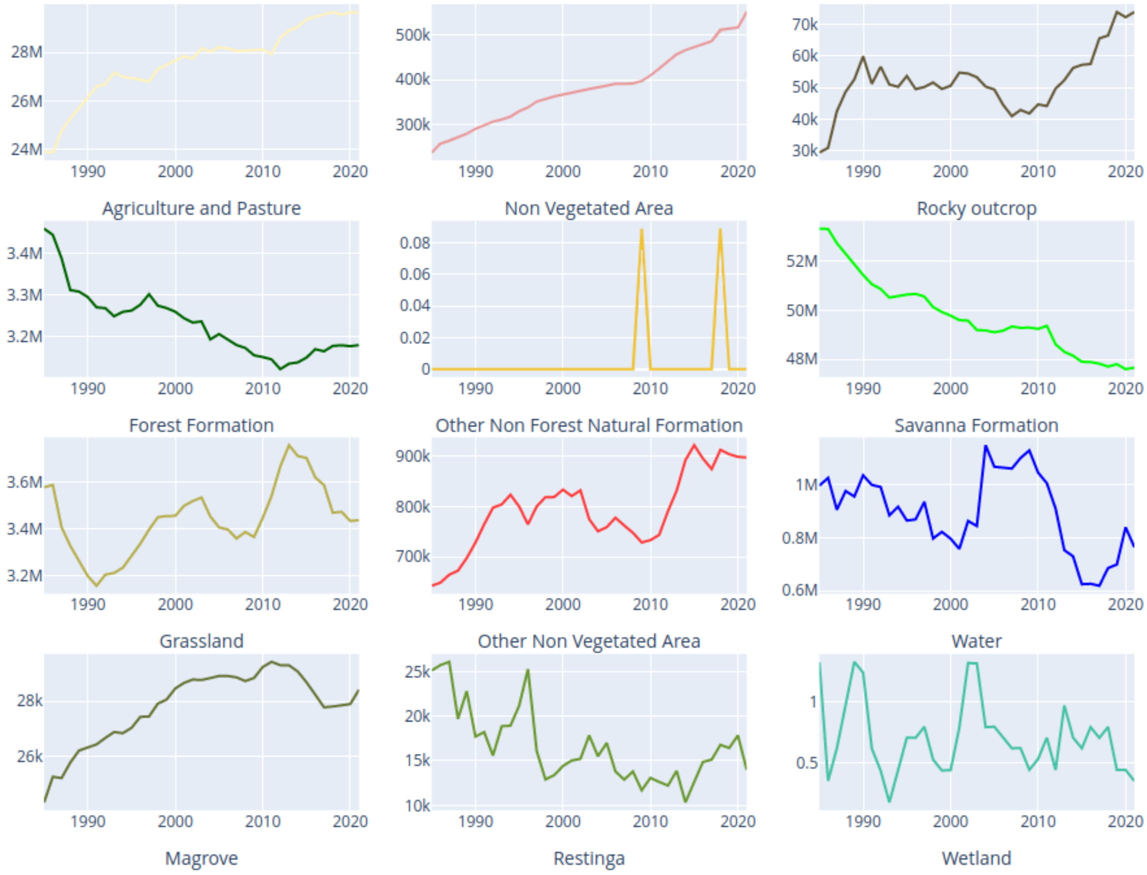


Figure 20. Plot of Area time series of level 3 of MapBiomas Collection 8.0 in the Caatinga biome (1985-2022 years).

7. REFERENCES

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.

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

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;

Lawrence, R., Bunn, A., Powell, S., & Zambon, M. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote sensing of environment*, *90*(3), 331-336.

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.

Zhang, Rui, and Jianwen Ma. "Feature selection for hyperspectral data based on recursive support vector machines." International Journal of Remote Sensing 30.14 (2009): 3669-3677.

Ramezan, Christopher A. "Transferability of Recursive Feature Elimination (RFE)-Derived Feature Sets for Support Vector Machine Land Cover Classification." Remote Sensing 14.24 (2022): 6218.