Caatinga Appendix

Collection 10

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 Santos (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 summarizes the specific methodologies employed for generating annual land use and land cover (LULC) maps of the Caatinga biome within the MapBiomas initiative. With each successive collection, the methodology evolved, incorporating either new LULC classes or revisions to existing methods. For instance, a key development from Collection 2.3 to 8.0 was the implementation of the **Random Forest** model for thematic classification, which replaced the trial-and-error parameterization used in the initial collections, BREIMAN (2001).

For Collection 8.0, the **Gradient Tree Boosting (GTB)** model was implemented in parallel with **Random Forest (RF)**, with the final mapping process determined by the model exhibiting superior performance. From Collection 9.0 onwards, including Collection 10.0, GTB became the main classifier to mapping all mosaics.

The *Photovoltaic Power Plants layer* is a critical, cross-cutting dataset created using a **U-Net** deep learning model in collection 10. This model was developed to federally map all registered photovoltaic power plants in Brazil, based on data from the National Electric Energy Agency (ANEEL) website. The precise geographical coordinates of each power plant were stored as an asset within the Google Earth Engine (GEE) platform, serving as the foundational database for generating image "patches" used to train the deep learning model.

The table 1 summarizes the evolution of the mapping methodologies across different collections. This document subsequently details each step developed and implemented for Collection 10.0, highlighting the improvements applied to the map production process. Methodologies employed in previous collections are accessible through the MapBiomas ATBD link (https://mapbiomas.org/download-dos-atbds).

Table 1. Overview of LULC collections of the Caatinga biome.

Collection	Time Interval	Method	Class	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, 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, Herbaceous Sandbank Vegetation.	New class (Herbaceous Sandbank Vegetation)
7.1	1985 - 2021	Random Forest	Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Rocky Outcrop, Herbaceous Sandbank Vegetation.	
8.0	1985 -	Random	Forest, Savanna, Grassland,	

	2022	Forest / Gradient Tree Booster	Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Rocky Outcrop, Herbaceous Sandbank Vegetation.	
9.0	1985 - 2023	Gradient Tree Booster/ cluster	Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Herbaceous Sandbank Vegetation. Rocky Outcrop	Rocky outcrop class was made using a cluster model
10.0	1985 - 2024	Gradient Tree Booster/ cluster / Deep Learning	Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Herbaceous Sandbank Vegetation. Rocky Outcrop, photovoltaic power plant	The photovoltaic power plant class was mapped using the UNet model.

2. METHODOLOGY OVERVIEW

The process flow diagram utilized in Collection 10.0 of the Caatinga biome is depicted in Figure 1. This flowchart combines a few of each node's smaller procedures that have been improved in this most recent collection. A few modifications have been made since collection 6 with the intention of enhancing the map classification flow's outcomes. Generally speaking, the following procedures are involved in creating the land cover and land use maps in the Caatinga Biome: Data input, sample gathering, feature selection, hyperparameter tuning, models of classification, post-classification filters, techniques for validation and visual inspection, and integration of outcomes with MapBiomas.

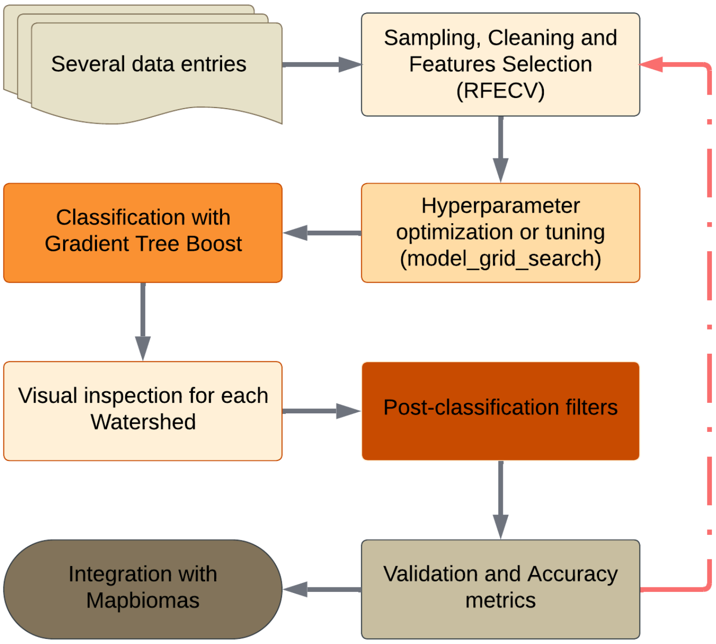

Several data entries
Sampling, Cleaning and Features Selection (RFECV)
Hyperparameter optimization or tuning (model_grid_search)
Classification with Gradient Tree Boost
Visual inspection for each Watershed
Post-classification filters
Validation and Accuracy metrics
Integration with Mapbiomas

Figure 1. Simplified general flowchart.

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

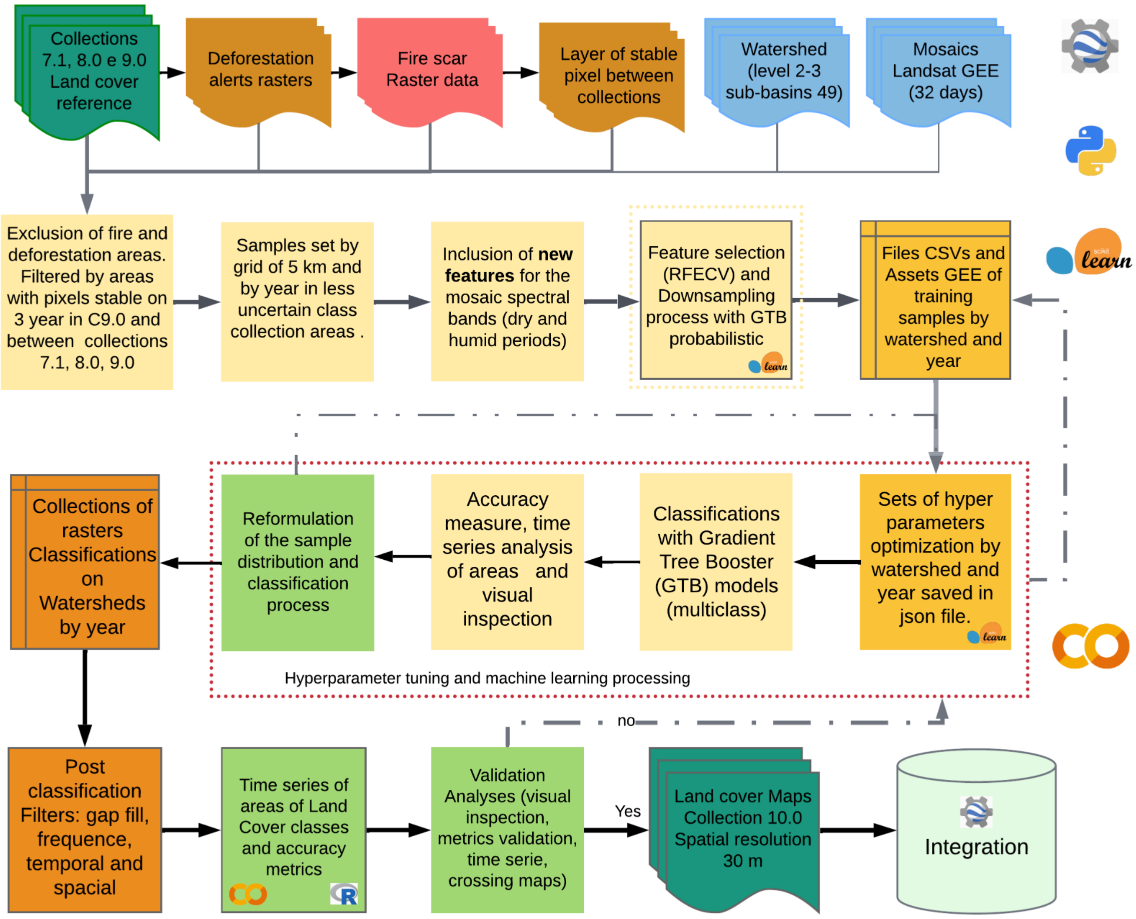

Collections 7.1, 8.0 e 9.0 Land cover reference
Deforestation alerts rasters
Fire scar Raster data
Layer of stable pixel between collections
Watershed (level 2-3 sub-basins 49)
Mosaics Landsat GEE (32 days)
Exclusion of fire and deforestation areas. Filtered by areas with pixels stable on 3 year in C9.0 and between collections 7.1, 8.0, 9.0
Samples set by grid of 5 km and by year in less uncertain class collection areas .
Inclusion of new features for the mosaic spectral bands (dry and humid periods)
Feature selection (RFECV) and Downsampling process with GTB probabilistic
Files CSVs and Assets GEE of training samples by watershed and year
Collections of rasters Classifications on Watersheds by year
Reformulation of the sample distribution and classification process
Accuracy measure, time series analysis of areas and visual inspection
Classifications with Gradient Tree Booster (GTB) 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)
Yes
Land cover Maps Collection 10.0 Spatial resolution 30 m
Integration

Figure 2. Classification process of MapBiomas Collection 10.0 (1985-2024) in the Caatinga biome.

3. IMAGE PRE-PROCESING AND STUDY AREA: The Caatinga Biome

3.1 LANDSAT IMAGE MOSAICS

In the initial collections, classification relied on Landsat 5 (TM), 7 (ETM+), and 8 (OLI) Surface Reflectance (SR) data. With Collection 6.0, we transitioned to using SR data exclusively. From Collection 7.0 through Collection 9.0, Landsat Collection 2, Tier 1 (T1) Surface Temperature (ST) products were incorporated. These Collection 2 Landsat products were generated using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) algorithm (version 3.4.0), available on Google Earth Engine (GEE) via the asset IDs: "LANDSAT/LT05/C02/T1_L2" for Landsat 5, "LANDSAT/LE07/C02/T1_L2" for Landsat 7, and "LANDSAT/LC08/C02/T1_L2" for Landsat 8.

A significant improvement for Collection 10.0 was the on-the-fly generation of annual and seasonal mosaics. For this, we utilized Google's asset "LANDSAT/COMPOSITES/C02/T1_L2_32DAY" to create comprehensive annual, dry, and rainy season composites. This directly supported both sample acquisition and classification, marking a departure from methods used in prior collections.

Despite this advancement, we encountered a notable obstacle: the new mosaics exhibited a greater number of pixel gaps, particularly in areas of bare soil. To overcome this, we implemented a robust correction method. This involved utilizing mosaics from our previous methodology and applying linear regression on a band-by-band basis to approximate and align the two datasets. Consequently, we were able to seamlessly fill the pixel gaps in our new mosaics with accurate data derived from the earlier, more complete imagery (see example below):

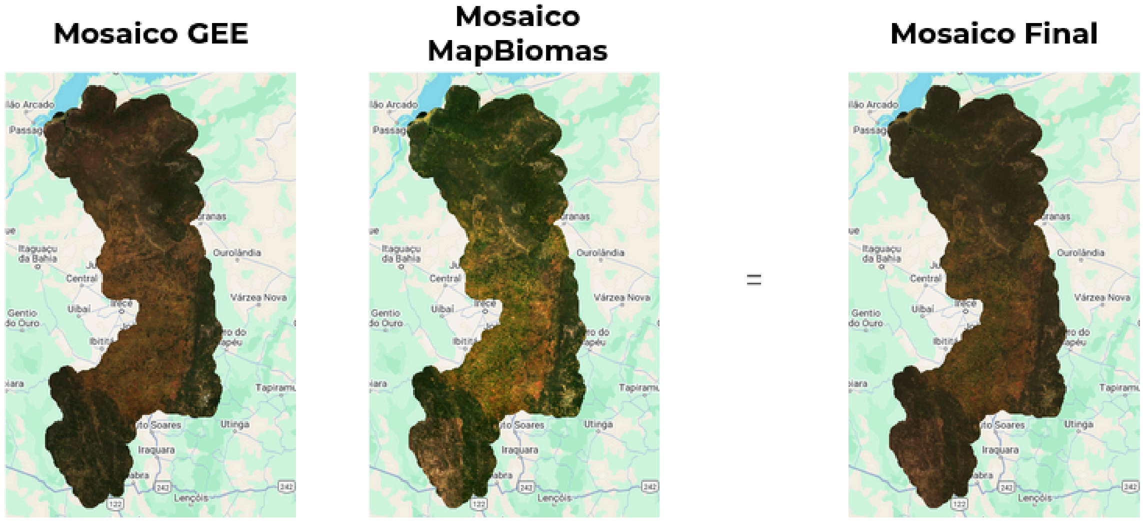


Figure 3: Result of linear regression between the GEE mosaic and the MapBiomas mosaic.

The mosaics generated using the previous methodology can be accessible via GEE path "*projects/nexgenmap/MapBiomas2/LANDSAT/BRAZIL/mosaics-2*", and are saved in the asset project MapBiomas along with all the processing done to clean the data.

The latest mosaic incorporates visible, infrared, and SWIR bands across the three aforementioned periods (annual, dry, and rainy seasons). Additionally, it includes descriptive statistics computed for the dry and wet periods, various spectral indices, and spectral mixing fractions, resulting in a comprehensive dataset of 142 bands.

3.2 DEFINITION OF THE PERIOD

To accurately classify LULC in the Caatinga biome, the project focused on selecting Landsat imagery that maximizes cloud-free coverage. This was critical due to the Caatinga's extreme phenological changes (seasonal leaf loss) driven by its unique, highly seasonal rainfall patterns**,** the period between January to July (with higher levels of rainfall). To define the best image selection periods for mosaic construction, they analyzed rainfall data for Brazil's Northeast region from INMET (1961-2015), accounting for the strong seasonality that directly impacts the vegetation's physiological activity. This dataset was obtained from the INMET (*www.inmet.gov.br*).

3.3 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.

3.4 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, characterized by reduced noise from elements such as clouds, relief and cloud shadows. For Collections 4.1 through 9.0, a single change to this calculation refers to the limit of the biome that was updated (IBGE, 2019). There is no change for Collection 10.0.

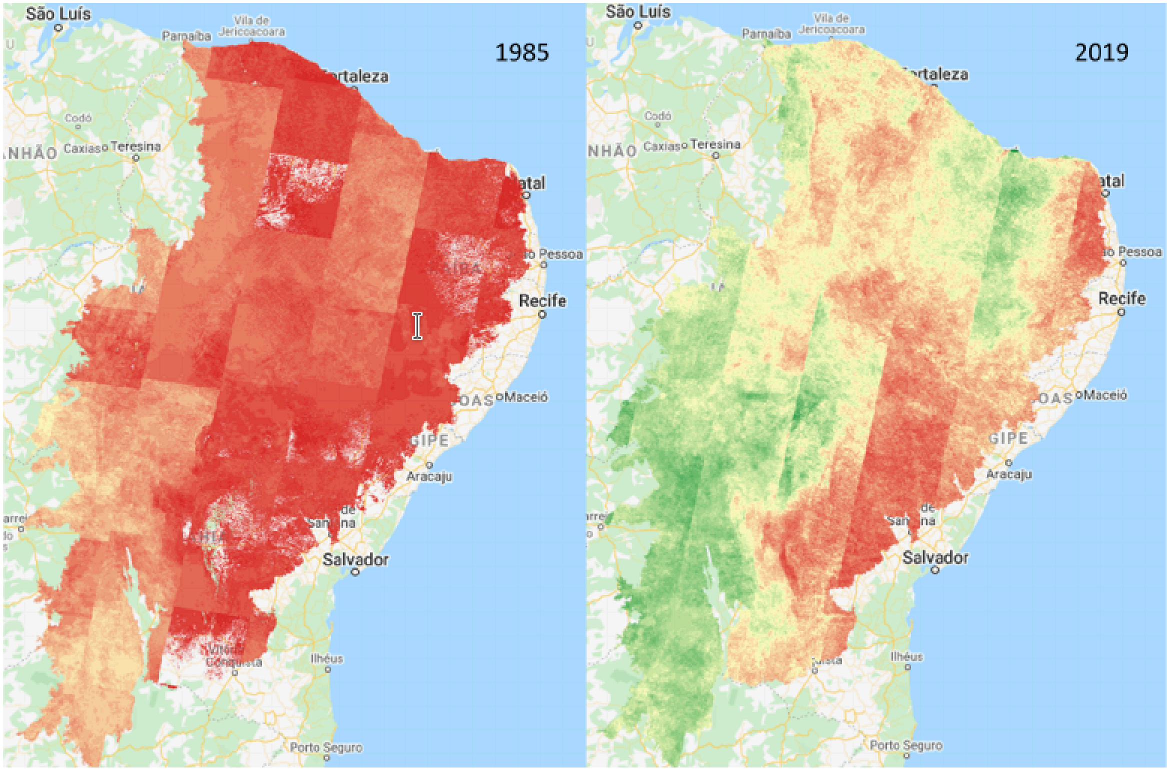
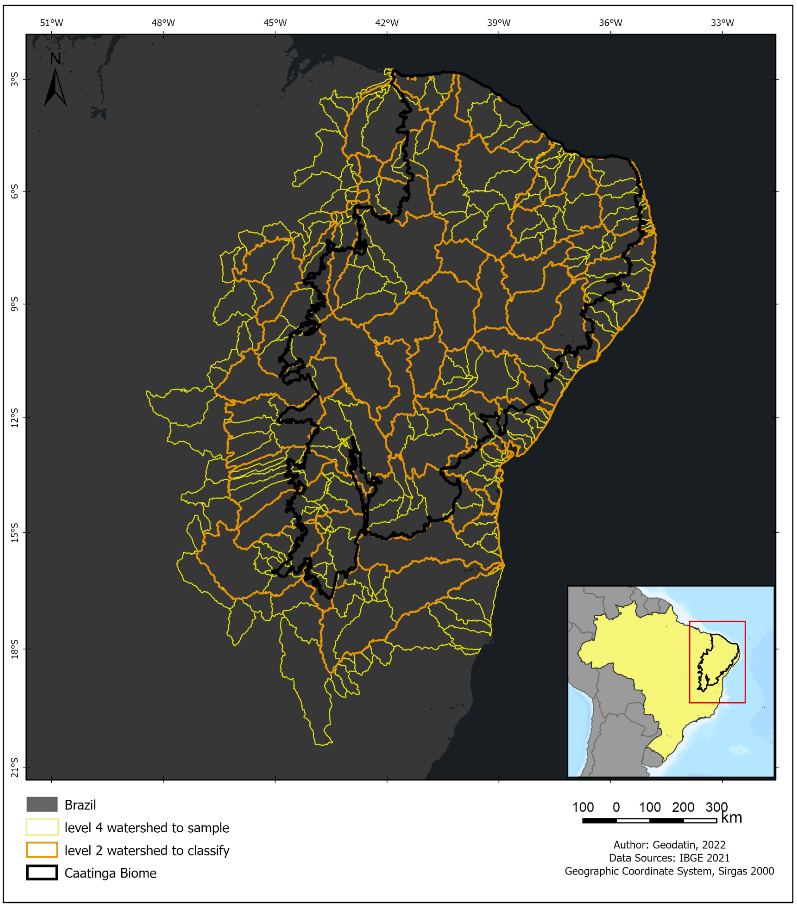


Figure 4. Landsat pixel availability in 1985 and 2019 in the Caatinga biome. Colors refer to data pixel availability, where red is low, yellow is medium, and green is high.

3.5. DEFINITION OF REGIONS FOR CLASSIFICATION

Classifying homogenous regions helps reduce the spectral variability among pixels, both within and between LULC classes and allows the use of a consistent set of samples to classify large areas of the mosaic. However, it is a computationally expensive task. To address this, the Caatinga biome was divided into smaller areas based on watershed boundaries provided by the Agência Nacional de Águas (*www.ana.gov.br*) (Figure 5). The natural borders of the basins helped maintain the homogeneity of the areas and allowed for the automation of the sampling process using GEE's Python API. In earlier versions, level 4 watershed basins were selected, dividing the biome into 320 regions.

Due to changes in biome boundaries (IBGE, 2019), additional basins were incorporated in Collection 5. For Collections 6.0 through 9.0, a merged version combining level 2 and level 4 watershed boundaries was employed, which reduced the Caatinga biome's division to 42 regions. In other words, watershed that are already small at level 2 and were very fractionated at level 4 will remain with the level 2 polygon. In Collection 10.0, this division was further refined to 49 regions.



Author: Geodatin, 2022
Data Sources: IBGE 2021
Geographic Coordinate System, Sirgas 2000

Figure 5. Watershed basins used in the classification and sampling of the MapBiomas LULC collections for Caatinga biome.

4. CLASSIFICATION PROCESS

4.1 LAND COVER AND LAND USE CLASSES

The digital classification of the Landsat mosaics in the Caatinga biome aimed to map ten specific LULC classes from the MapBiomas Collection 10.0 legend (Table 2). Some of these classes were later integrated with the cross-cutting themes.

Specifically, the Mosaic of Uses class in the Caatinga was subsequently refined by overlaying Agriculture or Pasture classifications. This left the Mosaic of Uses class predominantly in areas of temporary crops (which are very common in the Caatinga biome) or where it was not possible to distinguish between Agriculture and Pasture. Additionally, other classes like Rocky Outcrop, Other Non-Vegetated Areas, and Photovoltaic power plants were fine-tuned with specific, targeted classifications.

Table 2. Land cover and land use classes considered for digital classification of Landsat mosaics in the Caatinga biome in the MapBiomas Collection 10.

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 Herbaceous Sandbank Vegetation	49	Natural	Land cover	Herbaceous 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.
4.6 Photovoltaic power plant	75	Anthropic	Land cover	A "photovoltaic power plant" is a medium to large-scale installation designed to generate electricity directly from sunlight, primarily focused on commercializing the energy. In Brazil, plants with a capacity greater than 5 MW are considered large-scale, while those up to 5 MW are classified as mini-generation, according to regulations (Law 14.182/2021; Law 10.438/2002; Decree 5.025/2004; ANEEL Resolution 127/2004). The electricity

				generated is connected to the National Interconnected System (SIN), which distributes power throughout the country. In terms of land use, these plants occupy significant areas: it's estimated that an installation in tropical regions requires about 1 ha per MW using fixed modules, potentially varying to 2–3 ha/MW depending on technology (trackers) and panel arrangement. National examples confirm this range: the Nova Olinda Solar Park (292 MW across 690 ha ≈ 2.4 ha/MW), and the Pirapora Solar Complex (321 MW across ≈ 1,500 ha, about 4.7 ha/MW).
5. Water	33	Natural / Anthropic	Land cover / Land use	Rivers, lakes, dams, reservoir and other water bodies
6. Non Observed	27	non Observed	non Observed data	non Observed data

4.2 SAMPLE PROCESS

The most recent methodology of the sampling process was initiated in collection 8.0 and it has been refined continually to collection 10.0, and aims to establish pixel collection areas with the least uncertainty in the label, for this purpose, exclusion and inclusion criteria for collection areas were established.

The exclusion criteria consider areas where there was some intra-annual change and could corrupt the annual spectral information. The changes considered are burned areas, deforested areas, areas within a buffer of gaps from clouds or cloud shadows and areas that show variability between consecutive years.

The inclusion criteria consider areas where as a likely sample pixel only in areas that were stable over a 3-year window. Another inclusion criterion was to consider those pixels with the same labels in both collection 8.0 and 9.0. To achieve these criteria for each region grid, sorting at least 500 samples per class was required, which compelled the use of the function ee.Image().stratifiedSample() to collect samples from small areas inside a class.

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", “gvmi”,”gcvi”,”gsavi”, "cvi",”gli",”ndvi”,”ndti”,"afvi","avi","bsi","brba","dswi5","lswi","mbi","ui","osavi","ri","brightness", "wetness", "nir_contrast", "red_contrast".

Until Collection 8.0, the sampling process concluded with an outlier elimination step. For this, the Learning Vector Quantization (LVQ) algorithm, specifically the *ee.Clusterer.wekaLVQ()* function based on Kohonen (2003), was employed. This clustering algorithm allowed for the grouping of all samples within their respective categories. Subsequently, the two largest clusters (in terms of pixel count) for each class were selected for analysis. Finally, each feature was retained based on a percentage (x%) of the class's total number, aiming for approximately 1000 pixels per feature. Figure 6 illustrates an example of characteristics from 2020 within Caatinga watershed basins, along with the distribution of sampled quantities and percentages per class.

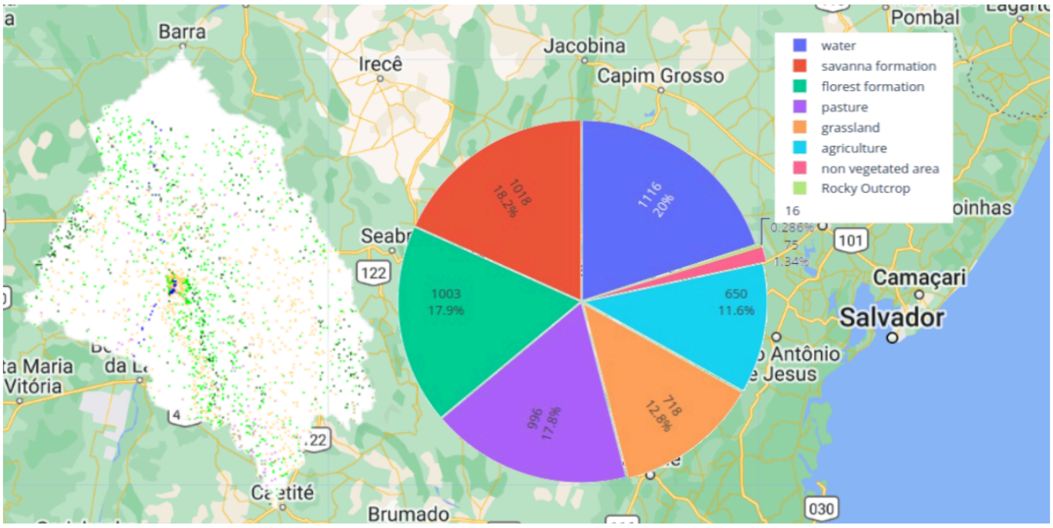


Figure 6. Map with distribution samples by class, and plot pie of distribution of the 2020 for one watershed region.

In collection 9.0, one of the strategies used to improve the performance of the classifiers was to normalize the data, the mosaic for the landsat median bands and for each median period. Specifically for the dry period, rainy period and annual period in the “blue_median”, “green_median”, “red_median”, “nir_median”, “swir1_median”, “swir2_median” bands. The statistics for normalization were saved for each year and each watershed basin. With this, the Gradient Tree Boost classifier, which uses the gradient descent technique to minimize errors, achieves better performance with normalized data.

In collection 10, a new methodology was required due to the large volume of collected samples. The first step involves collecting samples from pixel areas with lower label uncertainty. The second step combines all samples within the 49 classification regions. The third step applies a downsampling method within each sample set.

4.3 COLLECT SAMPLES PROCESS AND DOWNSAMPLING METHOD

Sample collection prioritizes areas where pixels are least likely to have misclassified labels. Several aforementioned criteria serve as filters for these collection areas. These areas were subsequently divided into 761 grids, covering the 49 basins (Figure 7). For Collection 10.0, 500 pixels were collected per grid for each class. This strategy resulted in an average of approximately 500,000 points per sample set (basin/year).

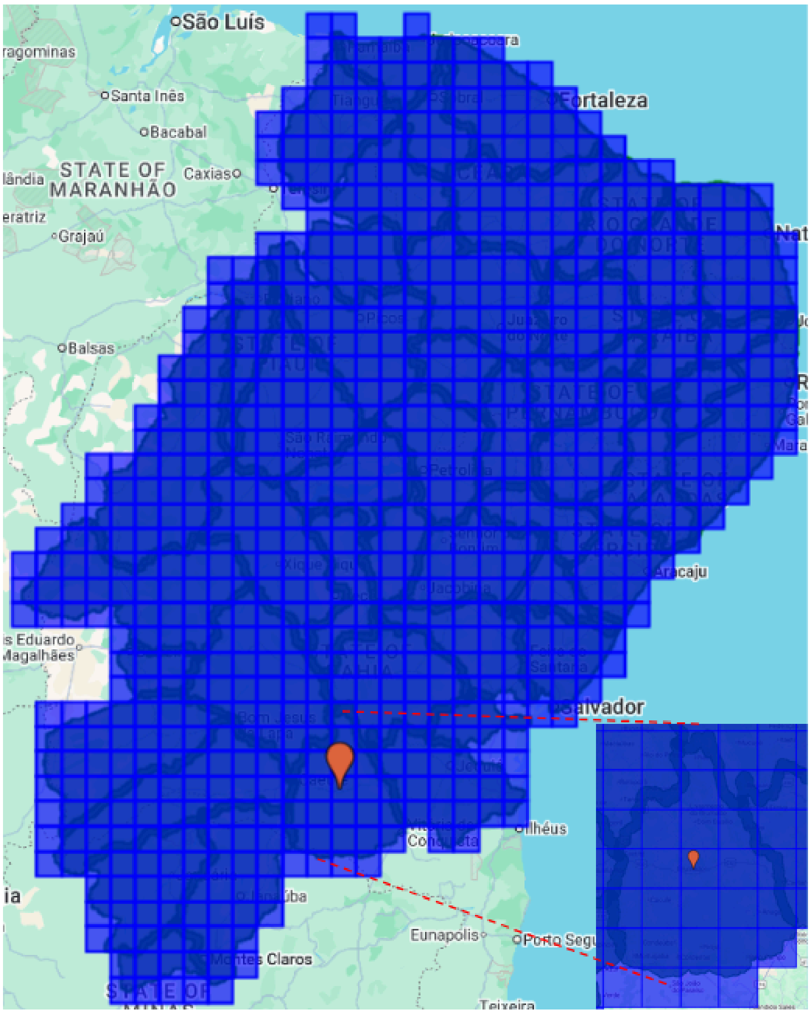


Figure 7. Collection areas by grid and their watershed will be grouped.

As mentioned earlier, the GTB classifier demands more computational power than RF, making it challenging to use very large sample volumes for training. This is precisely why we employ *downsampling*. This methodology allows us to extract a significantly smaller, yet representative, subset of the initial data, while also removing potential outlier pixels from the overall sample.

4.3.1 Enhanced Probabilistic Gradient Tree Boosting Methodology

Inspired by the "Isolation Forest" algorithm (LIU et al., 2008), we developed a new methodology using *probabilistic Gradient Tree Boosting (GTB)*. This approach enhances the selection of training samples, replacing the previous method that relied on the Google Earth Engine (GEE) API.

4.3.1 Sample Selection Process

Our process begins by selecting subsets of samples for each year and for each basin within our dataset. The core idea is to refine the training data by identifying the most confident and representative pixels.

Here's how it works:

1. Initial Separation*:* We first separate samples belonging to the three primary natural vegetation classes:
    - Forest Formation
    - Savanna Formation
    - Grassland Formation
2. Probabilistic Output*:* For every pixel classified by the GTB model, the output isn't just a single class label. Instead, the model provides a *probability vector* indicating the likelihood of that pixel belonging to each class.

    Example: A probability vector like (0.2,0.85,0.12) means:
    - 20% chance of being Forest Formation
    - 85% chance of being Savanna Formation
    - 12% chance of being Grassland Formation

3. *Identifying High-Confidence Candidates:* Even though the pixel above would be ultimately labeled as "Savanna" (since 0.85 is the highest probability), the crucial insight comes from its **high confidence score** for that class. In the *feature space* (the multi-dimensional representation of the pixel's characteristics), the classifier effectively "sees" this pixel as having an 85% probability of being Savanna. This makes it an ideal candidate for inclusion in the training set for the Savanna class.

4.3.2 Reducing Training Set Size

To significantly optimize and reduce the size of the overall training dataset, we applied a strategic selection process:

- For each of the natural vegetation classes, we specifically chose 100 pixels whose classification probabilities fell within high-confidence intervals: (0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0). This ensures that only the most confidently classified pixels are retained for training.

This same refined approach was also extended to agricultural classes, including:

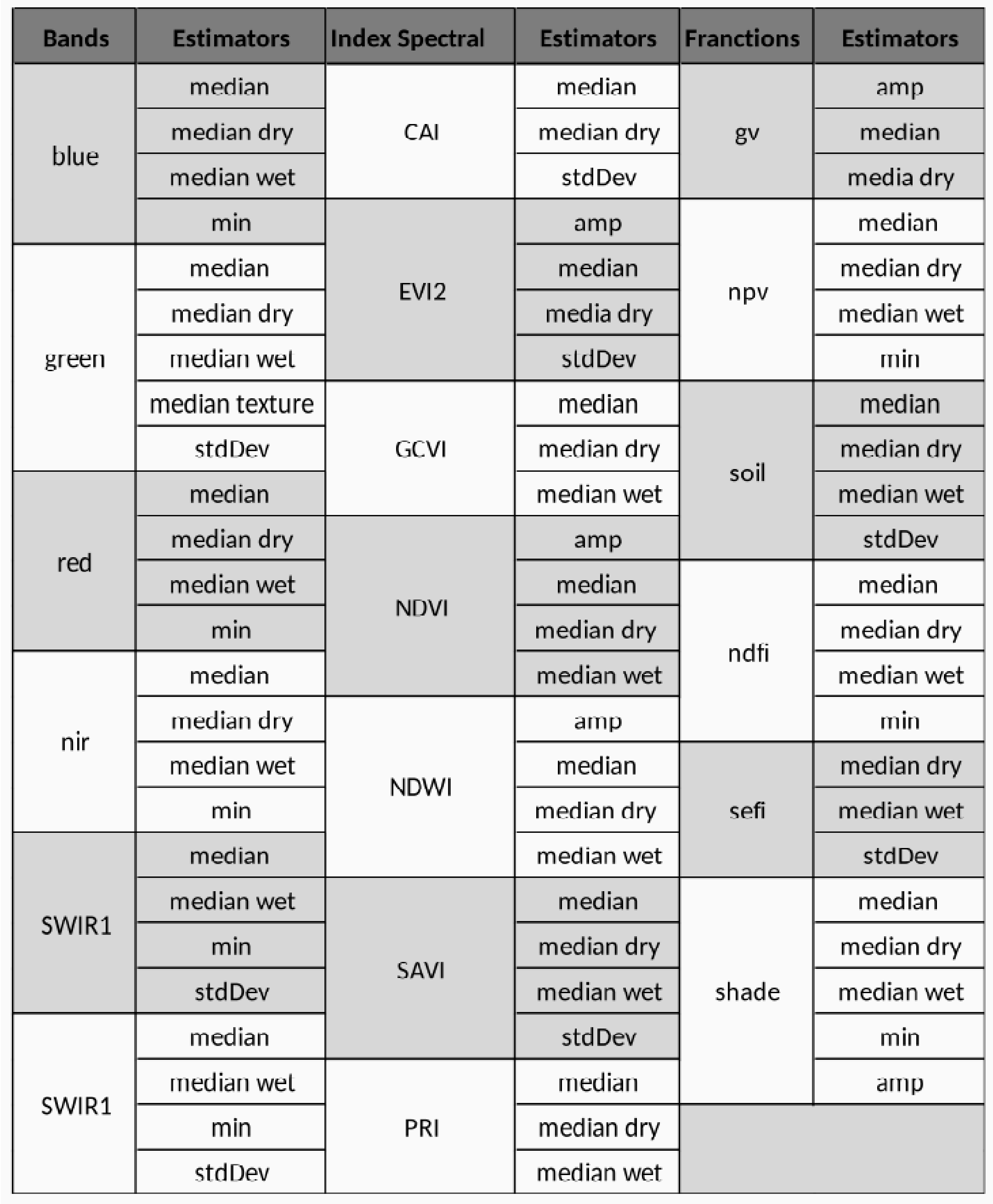
- Pasture
- Agriculture
- Mosaic of Uses (mixed land use)

Classes with low representation in the initial dataset, such as Water, Other Non-Vegetated Areas and Rocky Outcrop, were not included in this specific sampling strategy. They had too few samples to meaningfully apply this high-confidence selection method.

4.4 FEATURE SPACE AND FEATURE SELECTION PROCESS

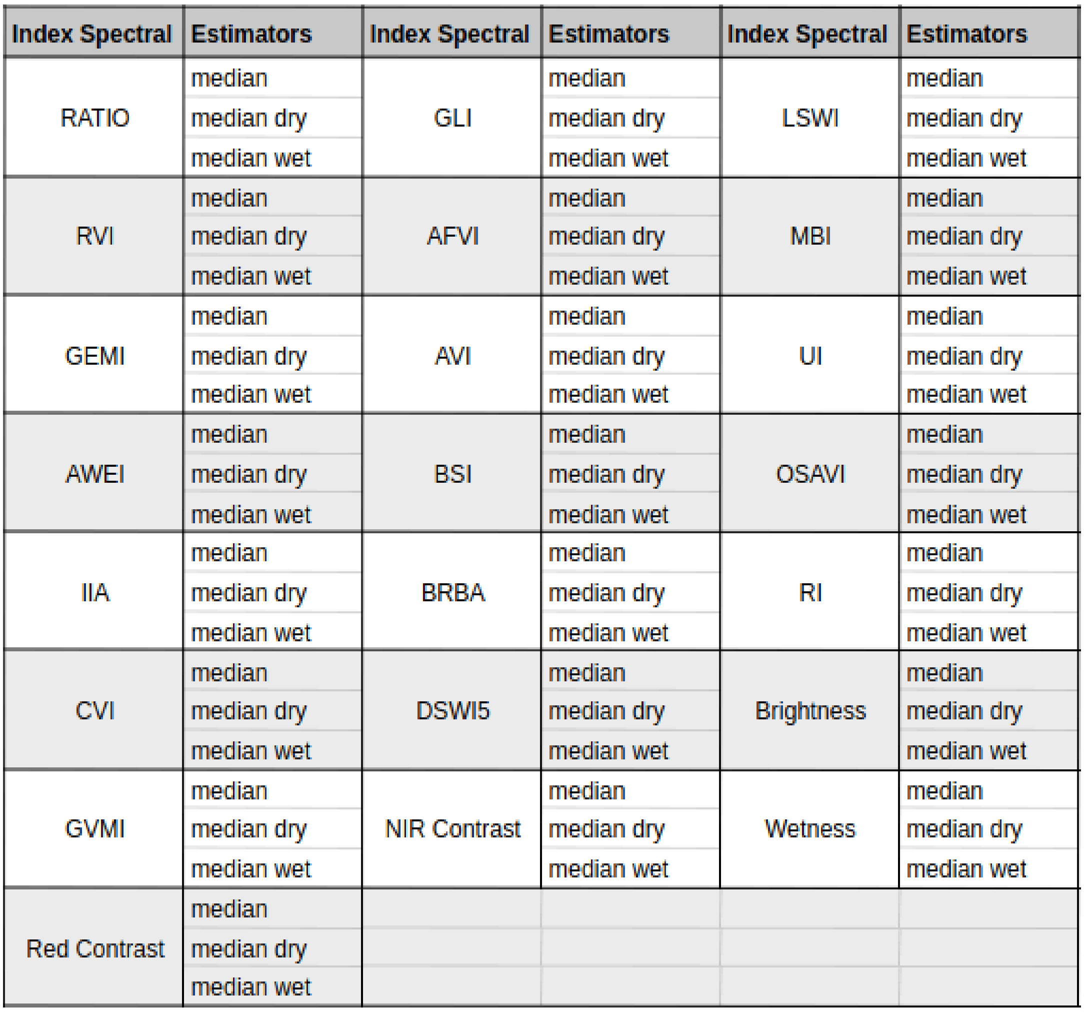
The feature space for digital classification of the LULC classes in the Caatinga biome comprised a subset of 75 features (Table 3), taken from the complete feature space of MapBiomas Collection 7.0 (General ATBD MapBiomas, 2020). In Collection 8.0, a larger number of spectral indices were calculated 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.

Table 3: Feature space subset considered in the classification of Landsat image mosaics in the Caatinga biome in the MapBiomas Collection 10.



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		

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 10.



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				

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.

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.

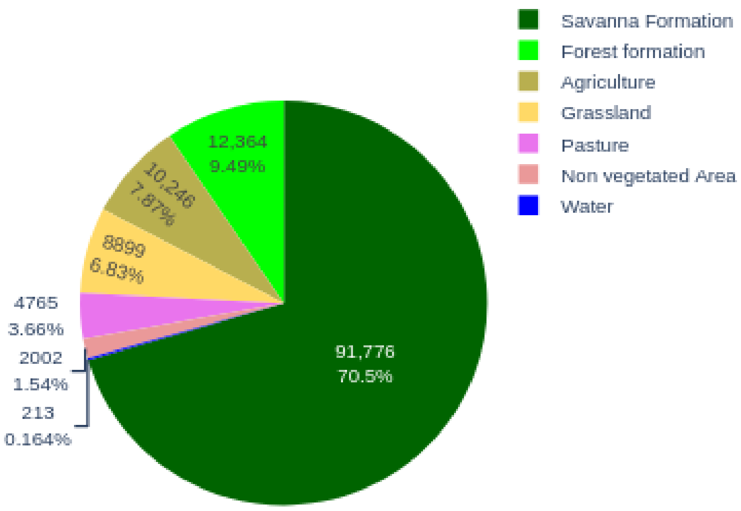


Figure 8: Distribution of samples for sub-basin 744 in the year 2000.

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

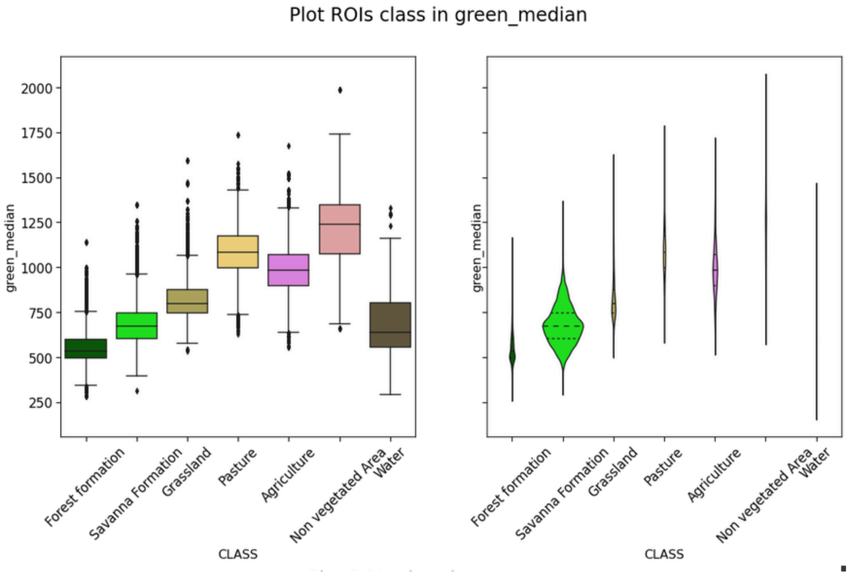


Figure 9: Box and violin plots from samples of spectral band “GREEN” in the main land cover classes mapped by the Caatinga team.

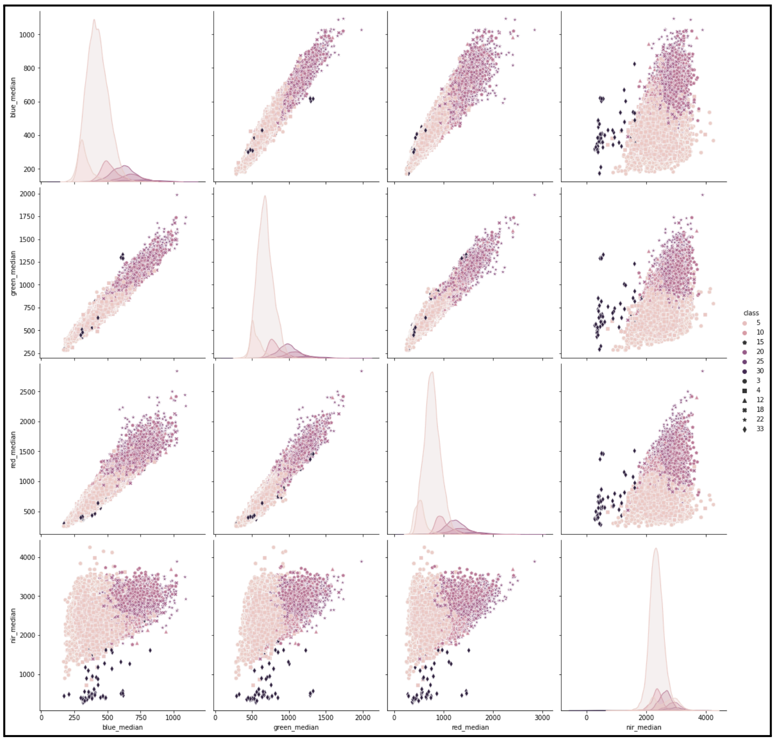


Figure 10: Distribuição espacial de amostras para as variáveis, “blue_median”, “green_median”, “red_median”, “nir_median”.

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 11), and some variables would be eliminated from the least important criteria following the score.

To calculate the correlation between the indices, the *corr()* function was used for each set of samples. The *corr()* function is implemented in the Pandas library of the python language. The python scripts were implemented in colab.

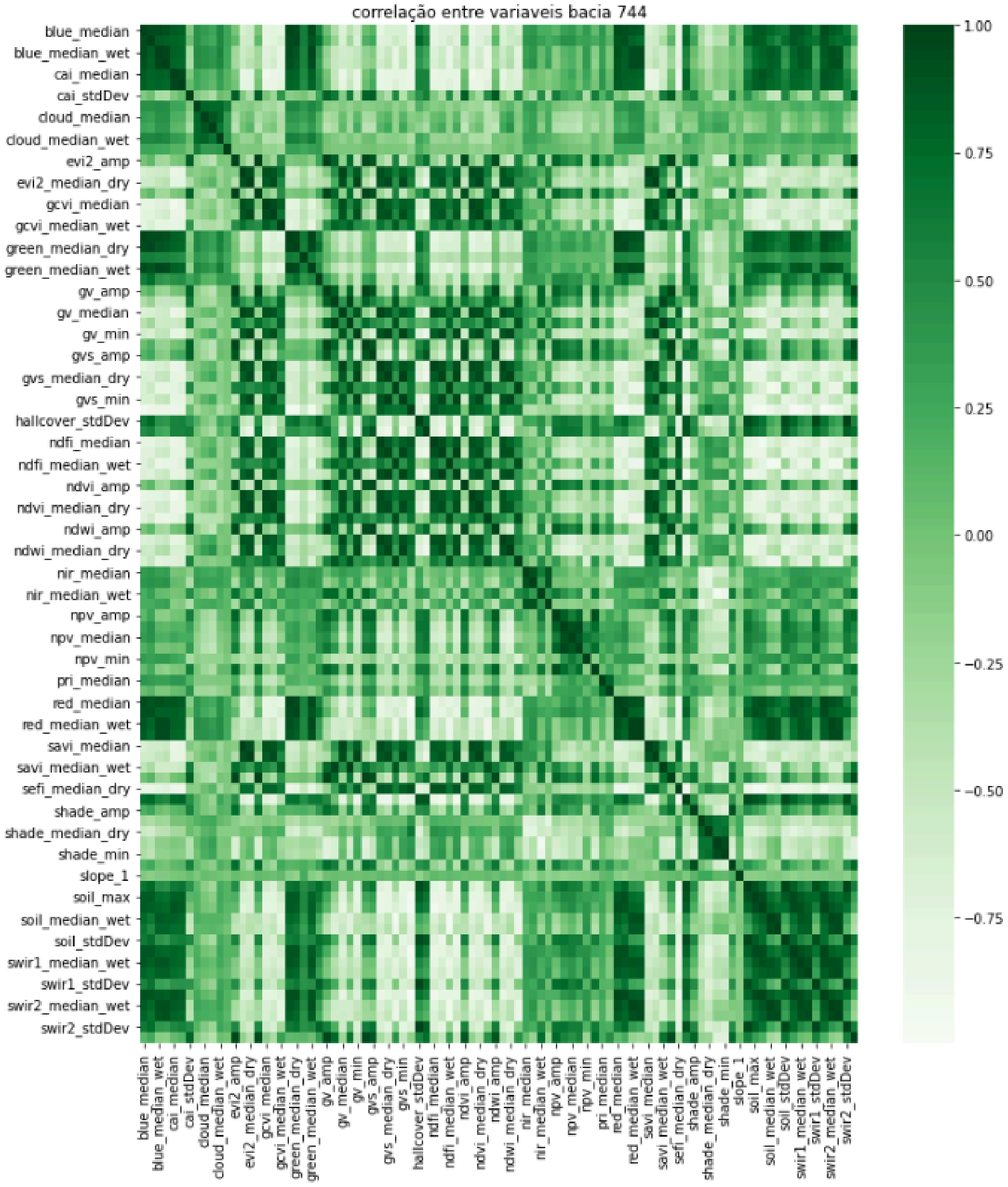


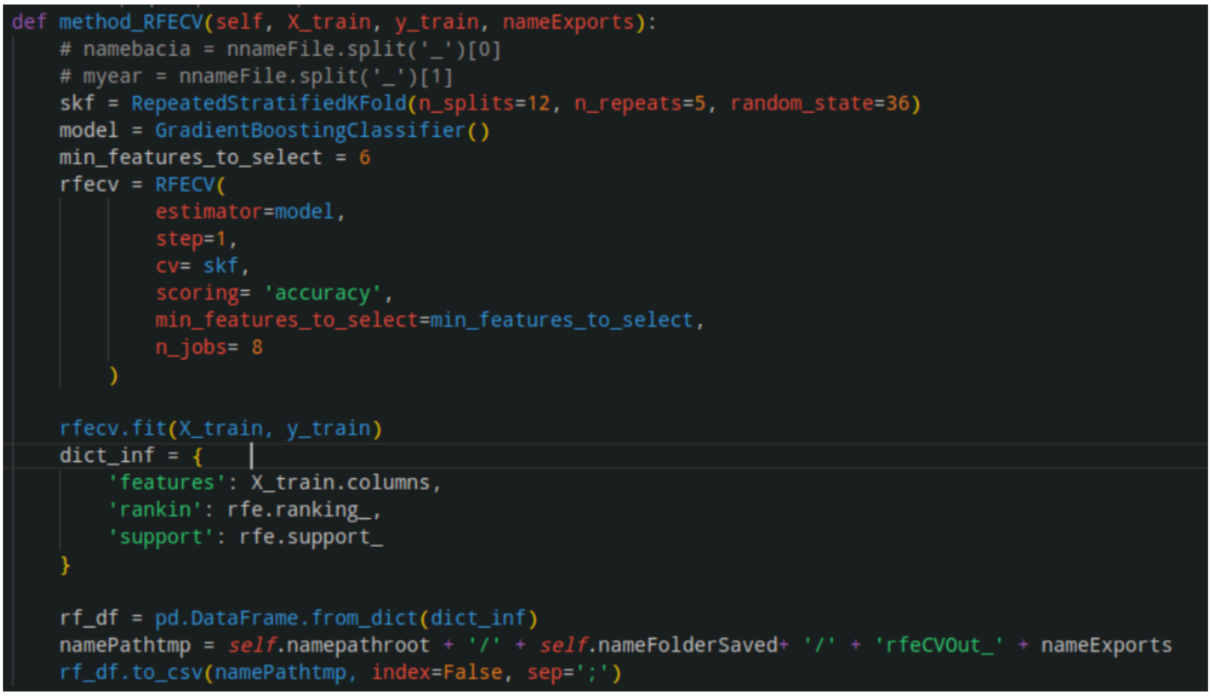
Figure 11. Example the plot correlation of watersheds samples from the year 2020.

Since Collection 8.0, the model has included Recursive Feature Elimination with Cross Validation (RFECV), an alternate feature selection method that uses cross-validation to automatically optimize the amount of features picked. As a result, for each set of data (basin / year), a list of characteristics chosen during the feature removal procedure was saved (*ZHANG AND JIANWEN, 2009; RAMEZAN, 2022*). A basic example may be found at the link below:

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

The *RFECV()* function can be accessed by using the python Sklearn library (Figure 12). There are two methods in which the class can be used to filter the selected variables: the "support_()" method and the "ranking_" method. With the former we can choose the surviving variables from a list of "TRUE" or "FALSE", and with the latter we can extract the ranking of the "TRUE" variables.

If the number of variables in "TRUE" is less than 10, then banking consecutive to 1 is taken as a condition (e.g. 2,3,4,5 etc.).



```
def method_RFECV(self, X_train, y_train, nameExports):
    # namebacia = nnameFile.split('_')[0]
    # myear = nnameFile.split('_')[1]
    skf = RepeatedStratifiedKFold(n_splits=12, n_repeats=5, random_state=36)
    model = GradientBoostingClassifier()
    min_features_to_select = 6
    rfecv = RFECV(
                estimator=model,
                step=1,
                cv= skf,
                scoring= 'accuracy',
                min_features_to_select=min_features_to_select,
                n_jobs= 8
            )

    rfecv.fit(X_train, y_train)
    dict_inf = {
        'features': X_train.columns,
        'rankin': rfe.ranking_,
        'support': rfe.support_
    }

    rf_df = pd.DataFrame.from_dict(dict_inf)
    namePathtmp = self.namepathroot + '/' + self.nameFolderSaved+ '/' + 'rfeCVOut_' + nameExports
    rf_df.to_csv(namePathtmp, index=False, sep=';')
```

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

4.5 HYPERPARAMETER TUNING PROCESS

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).

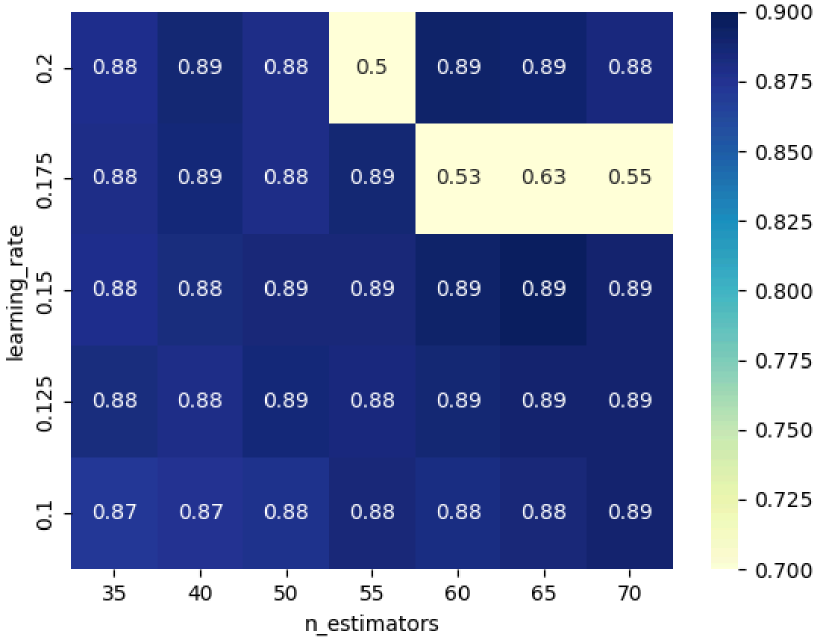


Figure 13. Example of the plot of combination of "learning rate" parameters and "n estimators".

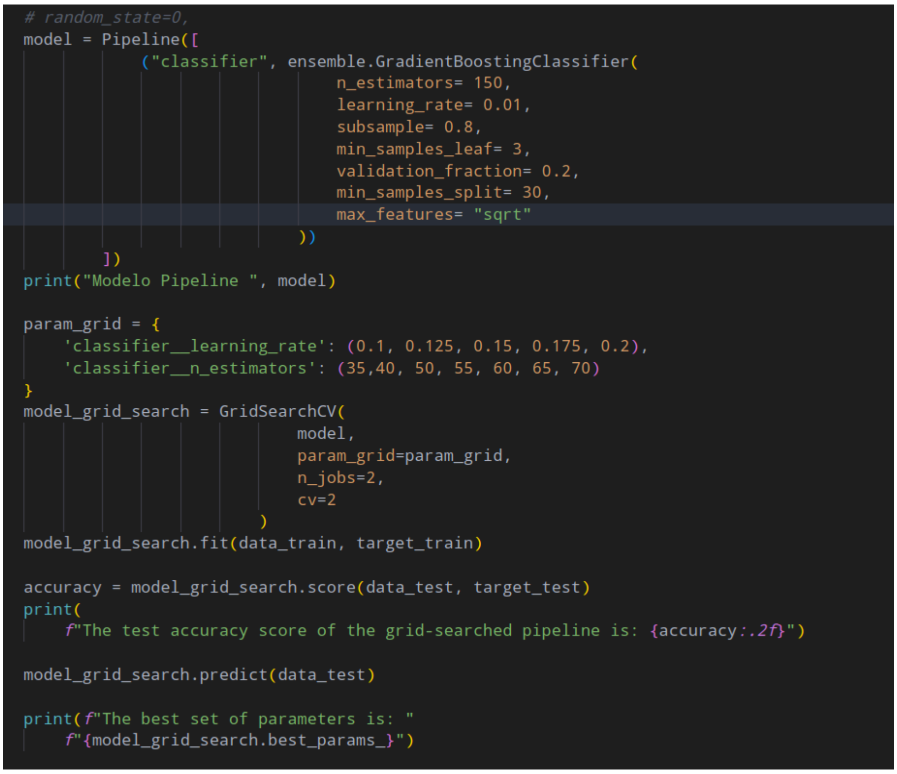
The GridSearchCV() (Grid Search Cross-Validation) is an exhaustive, or "brute-force," hyperparameter optimization technique, which means it can be computationally expensive. It operates as follows:

1. Defining a Hyperparameter "Grid": You define a dictionary where keys are the hyperparameter names for your model, and values are lists of all possible values you want to test for each hyperparameter. This creates a "grid" of every possible combination. (See Figure 14 for an illustration).
2. Exhaustive Training and Cross-Validation: For each hyperparameter combination within your defined grid, `GridSearchCV()` performs the following steps:
   - It trains the model using cross-validation. This involves splitting the training dataset into k "folds" (subsets). The model is then trained k times; each time, it uses k-1 folds for training and one fold for validation.
   - The model's performance is evaluated on each validation fold using a pre-defined scoring metric (e.g., accuracy, F1-score, Mean Squared Error (MSE), etc.).

- The final score for that specific hyperparameter combination is the average of the scores obtained across all k folds.

3. Selecting the Best Combination: After evaluating all possible combinations in the grid via cross-validation, `GridSearchCV()` identifies the combination of hyperparameters that yielded the best average validation score.
4. Final Model Refitting: Once the optimal hyperparameter combination is found, `GridSearchCV()` (by default, if `refit=True`) retrains the model using the entire original training dataset with these winning hyperparameters. This ensures you have a robust final model trained on all available data.

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.



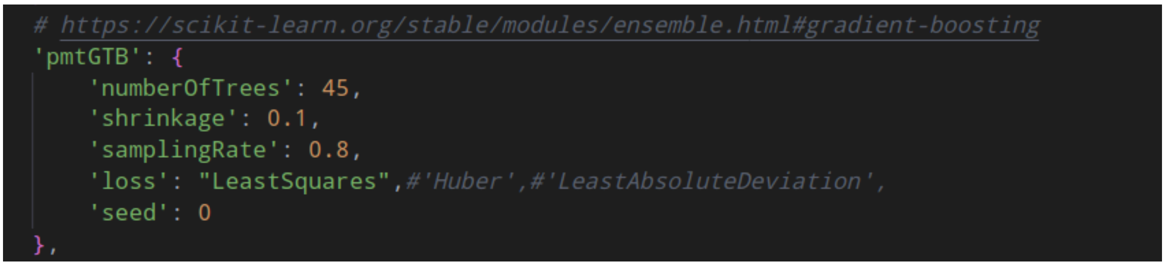
```
# random_state=0,
model = Pipeline([
            ("classifier", ensemble.GradientBoostingClassifier(
                                n_estimators= 150,
                                learning_rate= 0.01,
                                subsample= 0.8,
                                min_samples_leaf= 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.

For each watershed sample, a list of variables was kept for eventual use in the classification process. All the codes used in this stage are available in the repository of MapBiomas's Github (https://github.com/mapbiomas-brazil/caatinga).

4.5 Classification algorithm

During the classification process, the input data is adjusted to allow the MapBiomas mosaics to be classified by hydrographic basin and year. The data is then displayed using a GEE script and reviewed by the team's analysts to assess the classification results by basin and year. The primary objective of this step is to identify regions that require additional samples or classification parameter changes. Once identified, these areas are included in the map correction cycle. As explained before in collection 8.0, two algorithms were simultaneously reviewed. One is generated using the Random Forest classification (BREIMAN 2001), and the other is the result of the Gradient Tree Booster classification (LAWRENCE et al. 2004). An example of the parameters for GTB classifiers is shown in figure 15.



```
# 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 Gradient Tree Boost classifiers.

For the classes of *Forest Formation, Savanna Formation, Grassland Formation, Pasture, Agriculture, Mosaic of Uses, Other Non-Vegetated Areas, and Water Bodies,* the **Gradient Tree Boosting (GTB) classifier** is applied. This process uses a specific sample set, a predefined list of spectral bands, a set of classifier parameters, and a dictionary indicating the percentage of samples per class for each region/year processed in the construction of the map series that compose the collection.

Each classification version undergoes a set of stringent criteria, including accuracy, smoothness of area curves across the entire series, and spatial coherence. These criteria determine whether the map series for a given region demonstrates superior quality compared to the same region in previous collections.

The **Rocky Outcrop** class presents significant mapping challenges due to its inherent spectral mixture, often combining exposed soil with sparse grasses or small plants growing amidst the rocks. To address this complexity, our mapping effort leveraged data from the Geological Survey of Brazil (SGB), which provides presence points for these outcrops.

We demarcated these outcrop areas using polygons that encompassed over 80% of the outcrop's extent. Within these polygons, we applied an unsupervised clustering algorithm, specifically `ee.Clusterer.wekaXMeans` implemented in Google Earth Engine (GEE), configured to produce a maximum of three clusters, PELLEG AND MOORE (2000). From the resulting raster output, we identified the cluster value corresponding to the rocky outcrop class and then selected all such areas across the Caatinga Biome.

While this clustering process for identifying rocky outcrops is time-consuming, it proved fundamental. It provided a robust foundation for constructing a high-quality dataset of labeled image patches (or "chips") essential for training subsequent Deep Learning models.

4.5.1 CLASSIFIER RANDOM FOREST

The **Random Forest** algorithm is a bagging (Bootstrap Aggregating) ensemble method. It builds multiple independent decision trees, each trained on a random sample of the data (with replacement) and utilizes a random subset of features. The final prediction is determined by aggregating the individual tree predictions, typically through majority voting for classification tasks or by averaging for regression tasks (BREIMAN, 2001).

Advantages of Random Forest:

- High Accuracy and Robustness: Generally offers high performance and is less prone to overfitting compared to individual decision trees, as aggregating predictions from many trees reduces variance.
- Handles High-Dimensional Data Well: Can efficiently manage datasets with many features.
- Handles Missing Data and Outliers: It's relatively robust to missing data and outliers, as tree splits are less affected by extreme values.

- Feature Importance: Provides a measure of feature importance, helping identify which variables are most relevant for prediction.
- Parallelizable: Trees are built independently, allowing for parallel training and speeding up the process on large datasets. Therefore, within GEE it is possible to build more than 100 trees with a large training set and not have memory errors or time outs.
- Fewer Parameters for Tuning: Generally requires less hyperparameter tuning compared to Gradient Boosting.

Disadvantages of Random Forest:

- *Less Interpretable:* It's considered a "black-box model" because combining many trees makes interpreting the final model more difficult than a single decision tree.
- *May Not Be Best for Imbalanced Data:* Can have a bias towards the majority class in imbalanced datasets. Techniques like resampling or class weights can mitigate this. This is considered the biggest weakness for the satellite image classification process, because both the samples and the presence of classes within the images are unbalanced.

4.5.2 CLASSIFIER GRADIENT TREE BOOSTING

Gradient Tree Boost (GTB) is a boosting ensemble method that builds decision trees sequentially. Each new tree built in sequence is used to correct the errors (residuals) of the preceding tree, with the goal of minimizing a specific loss function.

Advantages of Gradient Tree Boosting:

- *High Accuracy:* Frequently achieves state-of-the-art accuracy in many problems, outperforming Random Forest in some cases, especially when well-tuned.
- *Ability to Capture Complex Patterns:* By iteratively correcting errors, it's very good at capturing non-linear relationships and complex interactions in the data.
- Handles Imbalanced Data Well: Can handle imbalanced datasets more effectively by focusing on difficult-to-predict examples (those with larger residuals). This property fits perfectly into the challenge of classifying 8 cover classes within a large volume of Landsat images.
- Flexibility: Can be optimized for a variety of loss functions, making it applicable to various problem types (regression, classification, ranking, etc.).

- Feature Importance: Similar to Random Forest, it also provides measures of feature importance. This property is widely used in the selection of bands and spectral indices to be used during the final classification process.

Disadvantages of Gradient Tree Boosting:

- *More Prone to Overfitting*: Because it builds trees sequentially and focuses on errors, GTB is more susceptible to overfitting, especially on noisy data or if hyperparameters aren't tuned carefully. This property makes the process of acquiring training samples, the feature selection process and the hyperparameter tuning process rigorous in this work.
- Slower to Train: Training is sequential, meaning it cannot be parallelized as easily as Random Forest, resulting in longer training times. Therefore, the construction of many trees within the GTB causes the processing on the GEE platform to time out memory.
- Sensitive to Hyperparameters: Requires more careful and extensive tuning of hyperparameters (such as learning rate, number of trees, and maximum depth) to achieve optimal performance.
- Sensitive to Outliers: By focusing on reducing residuals, outliers can have a disproportionate impact, leading the model to "learn" the noise.

Since Collection 9.0, we prioritized using Gradient Tree Boosting because:

- Maximum accuracy is crucial due to working with highly seasonal data, specifically annual mosaics for the Caatinga biome. Therefore, even though it requires more time to fine-tune the hyperparameters, we selected the model that achieves the best accuracy.
- New methodologies for cleaning the training datasets ensure that the model performs better than previous models.
- Imbalanced datasets are a natural property of remote sensing data, and GTB can be more effective at learning minority classes by focusing on errors.

5. POST-CLASSIFICATION

The temporal filter rules were specifically adapted to account for the phenological and spectral dynamics of the land cover classes characteristic of the Caatinga biome. In addition to the standard temporal consistency checks, custom rules were incorporated to handle anomalous transitions—particularly cases in which a class appears abruptly in a pixel’s time series. These adjustments aim to improve classification stability and reduce spurious temporal fluctuations in areas with high seasonal variability and low vegetation cover.

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, figure 16.

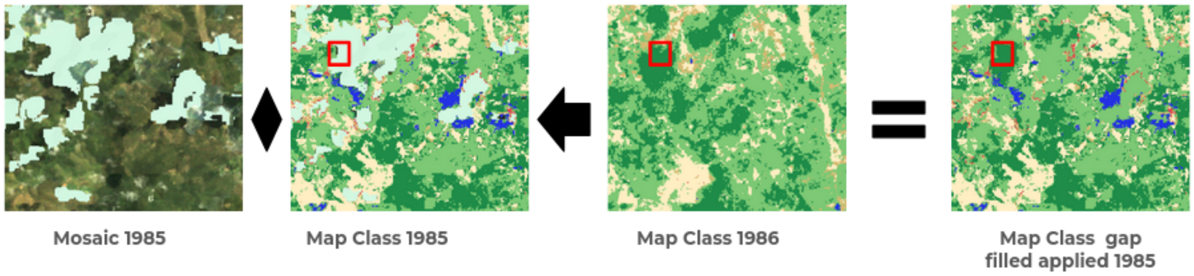


Figure 16: Example of the process gap fill.

5.2 Spatial filter

The applied spatial filter targets isolated or weakly connected pixels by using a connectivity-based mask. Specifically, it identifies pixels that are connected to five or fewer adjacent pixels of the same class. These pixels are then replaced with the statistical mode of their 8-connected neighborhood, effectively smoothing local noise while preserving spatial coherence.

5.3 Temporal filter

The applied temporal filter replaces pixels with faulty transitions with those from succeeding years. In the first step, the filter searched for any natural class (3-FOREST FORMATION, 4-SAVANNA FORMATION, 12-GRASSLAND, 13-OTHERS NO FOREST FORMATION) that was not this class in 1985 but was equivalent to these classes in 1986 and 1987, and then rectified the 1985 class to avoid regeneration in the first year. In the second step, the filter looked at the pixel value from last year that was not 21-MOSAIC OF AGRICULTURAL OR PASTURE but was equal to 21-MOSAIC OF AGRICULTURAL OR PASTURE in the preceding 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 stage looked at a 3-year moving window to fix any values that had altered in the middle year and return to the same class the following year. This method was used in the following 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 fourth and final stage was identical to the previous one, but it employed a four- and five-year moving window to modify 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. INTEGRATION

The final classification results, incorporating post-classification filters, are integrated with the data from the cross-cutting themes across the entire historical series (1985–2024). Ultimately, the final integrated map for the Caatinga biome features 20 classes at level 3 of the Collection 10.0 legend (Figure 16).

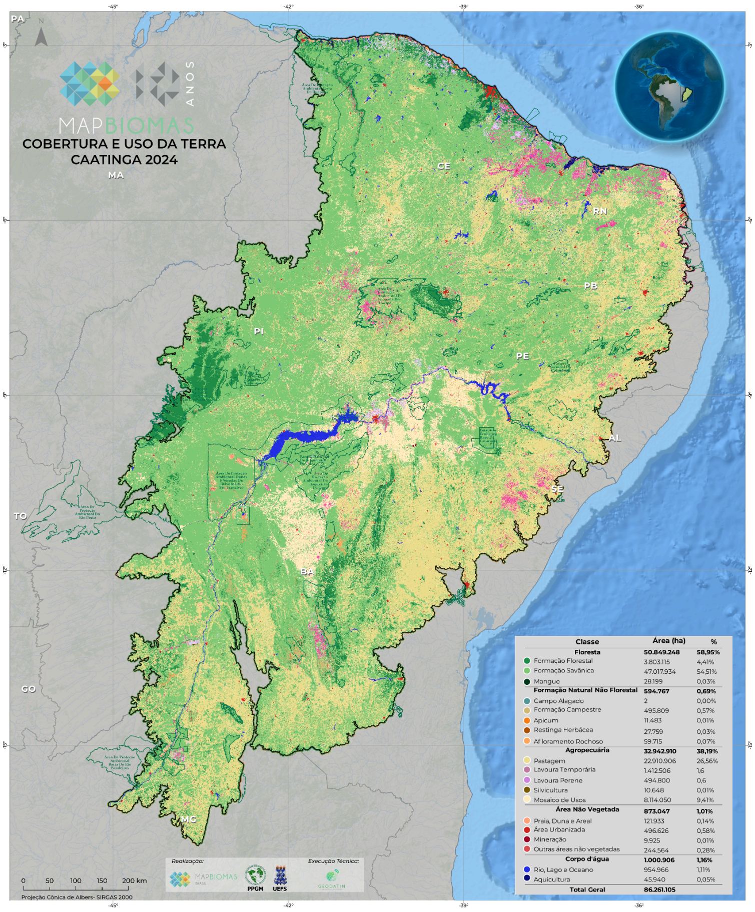
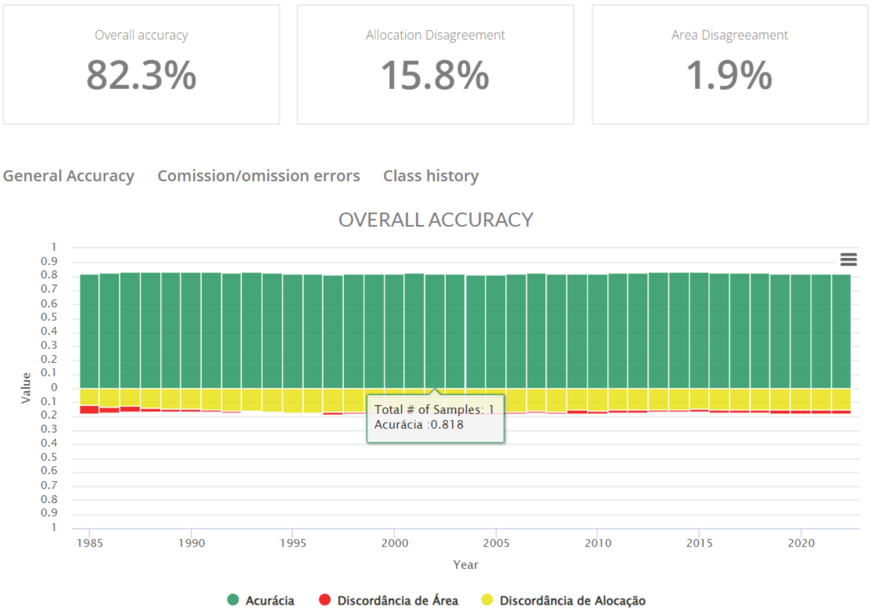

MAPBIOMAS
10 ANOS
COBERTURA E USO DA TERRA
CAATINGA 2024
PA
MA
CE
RN
PB
PE
PI
AL
SE
BA
TO
GO
MG
Classe | Área (ha) | %
Floresta | 50.849.248 | 58,95%
Formação Florestal | 3.803.115 | 4,41%
Formação Savânica | 47.017.934 | 54,51%
Mangue | 28.199 | 0,03%
Formação Natural Não Florestal | 594.767 | 0,69%
Campo Alagado | 2 | 0,00%
Formação Campestre | 495.809 | 0,57%
Apicum | 11.483 | 0,01%
Restinga Herbácea | 27.759 | 0,03%
Af loramento Rochoso | 59.715 | 0,07%
Agropecuária | 32.942.910 | 38,19%
Pastagem | 22.910.906 | 26,56%
Lavoura Temporária | 1.412.506 | 1,6
Lavoura Perene | 494.800 | 0,6
Silvicultura | 10.648 | 0,01%
Mosaico de Usos | 8.114.050 | 9,41%
Área Não Vegetada | 873.047 | 1,01%
Praia, Duna e Areal | 121.933 | 0,14%
Área Urbanizada | 496.626 | 0,58%
Mineração | 9.925 | 0,01%
Outras áreas não vegetadas | 244.564 | 0,28%
Corpo d'água | 1.000.906 | 1,16%
Rio, Lago e Oceano | 954.966 | 1,11%
Aquicultura | 45.940 | 0,05%
Total Geral | 86.261.105
0 50 100 150 200 km
Projeção Cônica de Albers - SIRGAS 2000
Realização:
Execução Técnica:
PPGM
UEFS
-45°
-42°
-39°
-36°

Figure 16. Final land use and land cover map of the Caatinga biome (2024).

7. VALIDATION STRATEGIES

The validation stage of each process was created 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 10.0 (1985-2024) (Figure 17). The metrics showing are historical and global accuracy, allocation disagreement and quantity disagreement.



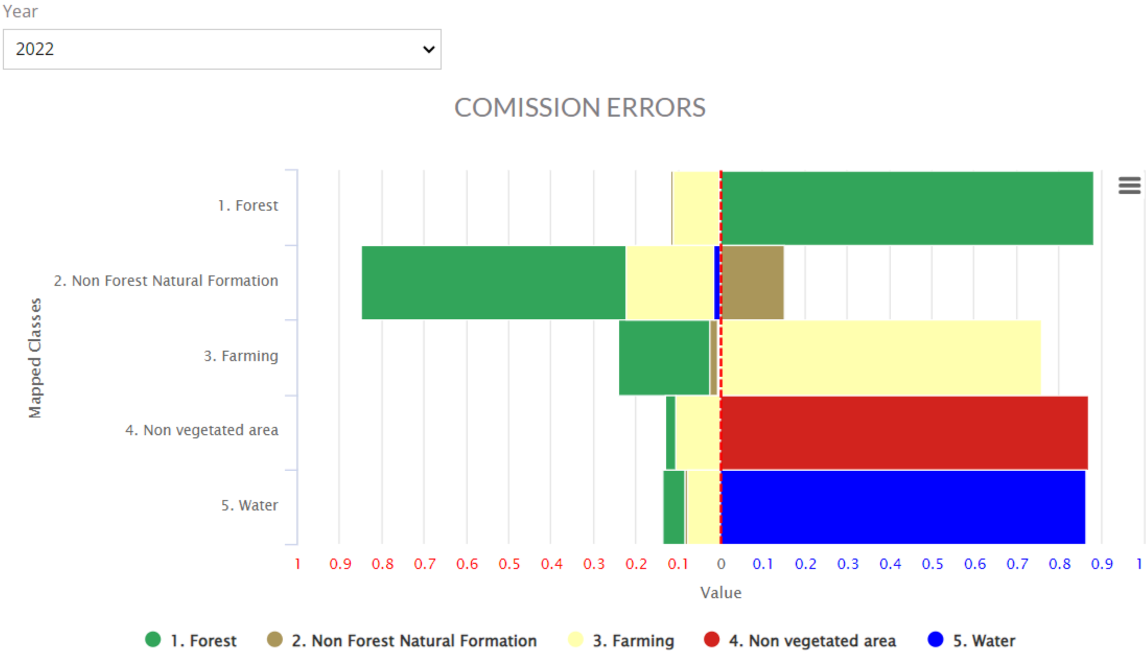
Overall accuracy 82.3%

Allocation Disagreement 15.8%

Area Disagreement 1.9%

Figure 17. Accuracy of level 3 of MapBiomas Collection 10.0 in the Caatinga biome (1985-2024).

The approach used in this collection was more accurate than previous collections. Table 5 has the numbers that demonstrate these outcomes. Figures 18 and 19 illustrate the errors of omission and commission. By analyzing these data, we may determine which classes are confused with others in the categorization. And based on these results, we can devise a new strategy to reduce those errors of commission and omission.



Year: 2022

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

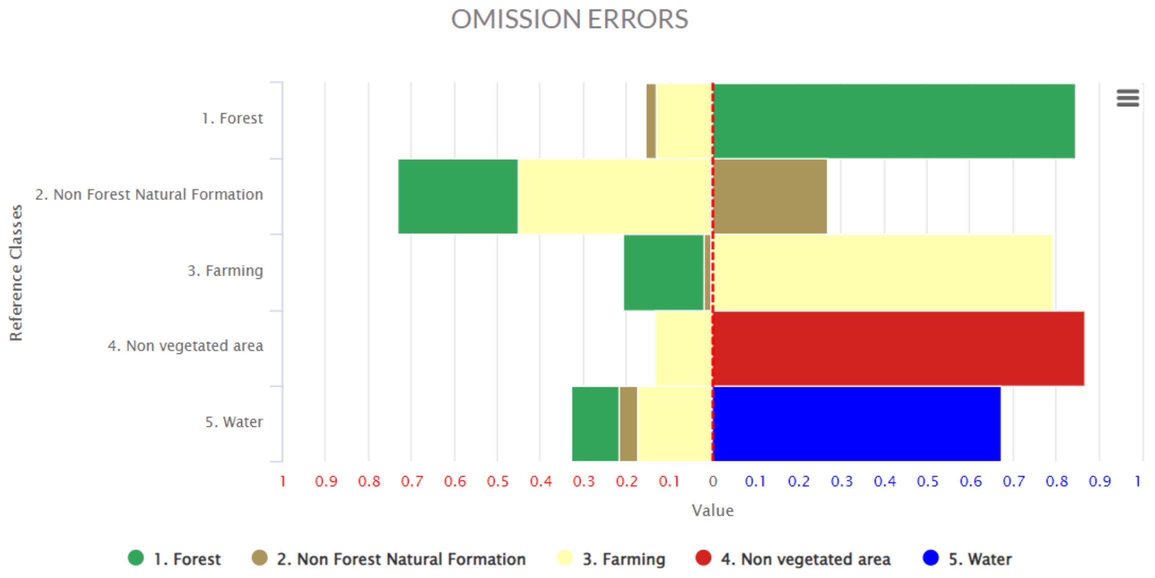


Figure 19. Omission errors of the land cover and land use mapping in the Caatinga.

Table 5. 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 3: 71.3 %
4.1	Random Forest	Level 1: 81.9 % Level 2: 79.9 % Level 3: 74.3 %
5.0	Random Forest	Level 1: 81.8 % Level 2: 80.0 % Level 3: 75.4 %
6.0	Random Forest	Level 1: 82.8% Level 2: 76.6 % Level 3: 74.9 %
7.1	Random Forest	Level 1: 83.7 % Level 2: 78.8 % Level 3: 76.9 %
8.0	Random Forest / Gradient Tree Booster	Level 1: 83.6 % Level 2: 78.2 % Level 3: 76.9 %
9.0	Gradient Tree Booster	Level 1: 84.6 % Level 2: 79.4 % Level 3: 79.3 %
10.0	Gradient Tree Booster	Level 1: 84.6 % Level 2: 79.4 %

		Level 3: 79.3 %

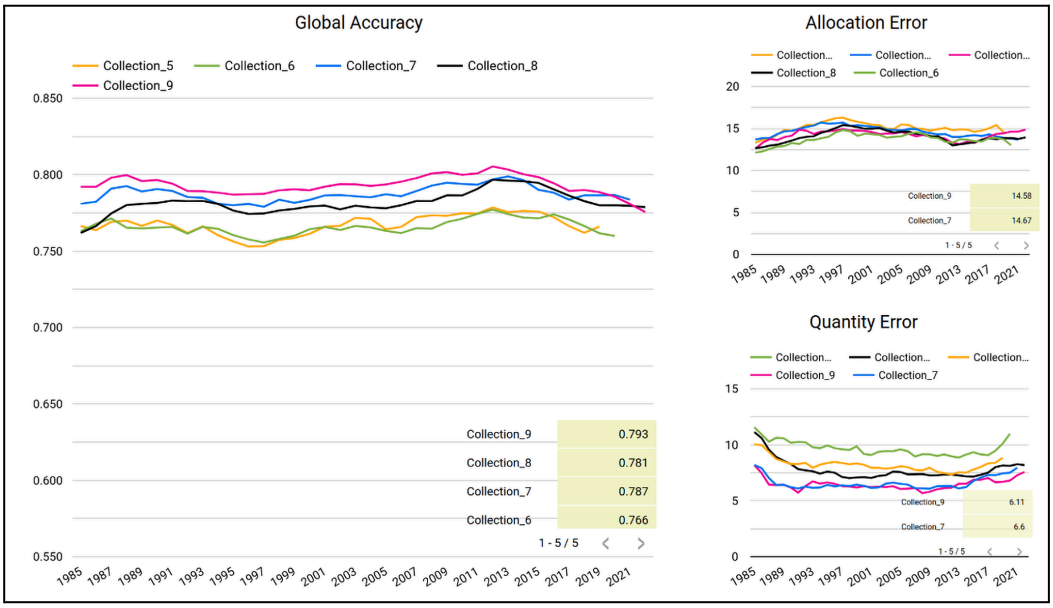


Figure 20. Plot of Accuracy of level 3 of MapBiomas Collections 5.0, 6.0, 7.0, 8.0, 9.0 and 10.0 in the Caatinga biome (1985-2024 years).

If we plot all values in the accuracy series, we can better compare and observe all of the data from the different collections (Figure 20). Another technique to assess the quality of a map series is to examine the area's behavior in relation to each class of land cover across time. Figure 21 depicts time series plots of area by cover class. In addition, the areas of the two collections prior to the 10.0 collection are plotted. This type of analysis makes it possible to compare the areas of the classes throughout the series with other previously published collections. At this point, it is possible to understand where and why there have been changes from one collection to another.

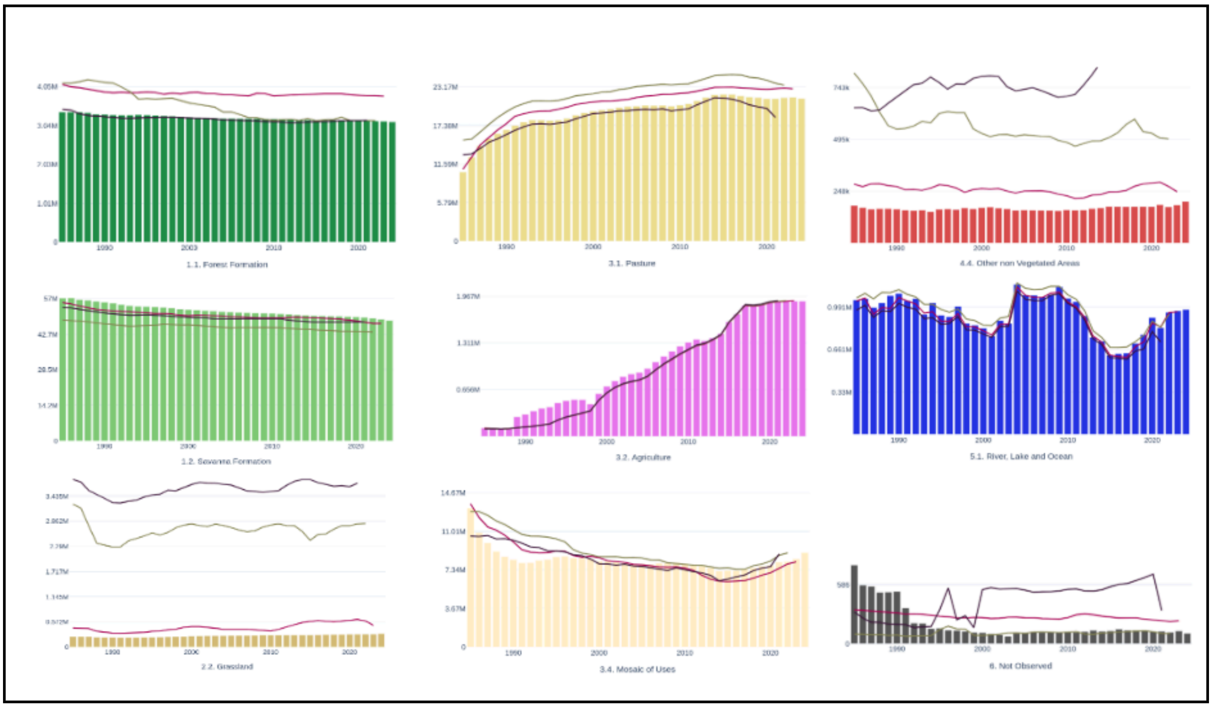
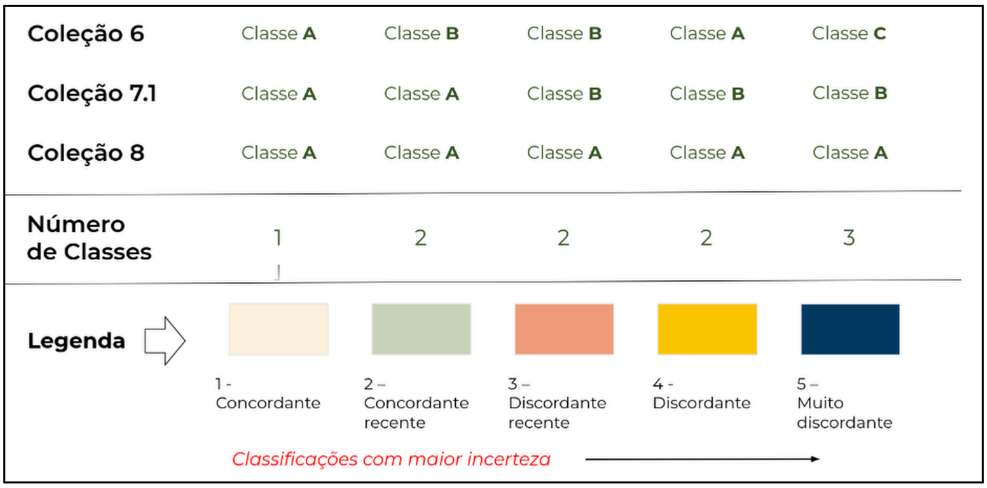


Figure 21. Time series of level 3 classes of MapBiomas Collection 10.0 in the Caatinga biome (1985-2024 years) per area (ha). The green, brown and red lines represent the corresponding areas of collections 7.1, 8.0 and 9.0.

Another way of validating the coverage data was to compare the areas of the classes between collections. For this analysis we used a rule implemented by the Pampa team that accounts for coincidences (Figure 22).



Coleção 6	Classe **A**	Classe **B**	Classe **B**	Classe **A**	Classe **C**
Coleção 7.1	Classe **A**	Classe **A**	Classe **B**	Classe **B**	Classe **B**
Coleção 8	Classe **A**	Classe **A**	Classe **A**	Classe **A**	Classe **A**
Número de Classes	1	2	2	2	3
Legenda ⇨	1 - Concordante	2 – Concordante recente	3 – Discordante recente	4 - Discordante	5 – Muito discordante

Classificações com maior incerteza →

Figure 22: Concordance table between collections 6.0, 7.1 and 8.0.

With this analysis we can infer how much area is being mapped with the same class, how much area is varying between classes from one collection to another and when these disagreements occurred last year. Regions in the north and center of the Caatinga and the Chapada Diamantina show the greatest disagreements (Figure

23). These places of greatest discordance indicate where new samples should be collected and used to find the correct class. They also indicate areas where the pixels in the mosaic have clouds, noise or shadows.

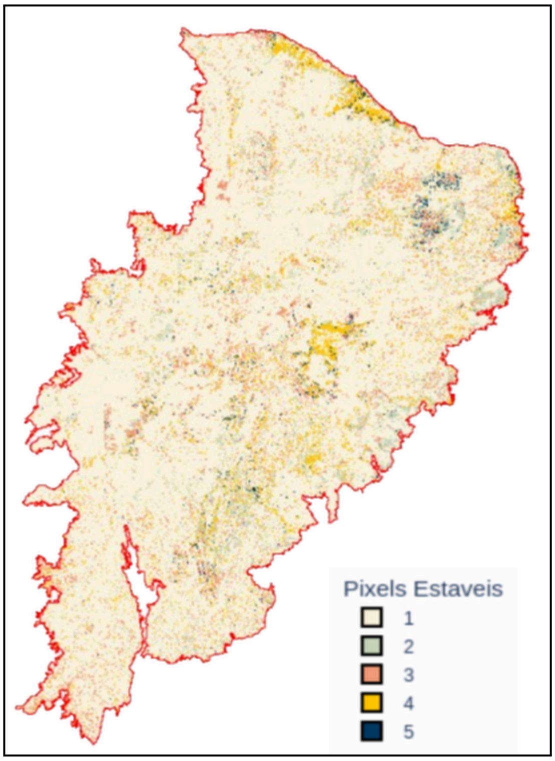


Figure 23: Pixel coincidence model between collections 6.0, 7.1 and 8.0.

The statistics over the series show that on average 76 % of the pixels are coincident between the 3 collections (Figure 24). This percentage is higher than the accuracy of the last three collections, which indicates that this measure does not indicate the quality of the maps, but rather areas with high stability between collections.

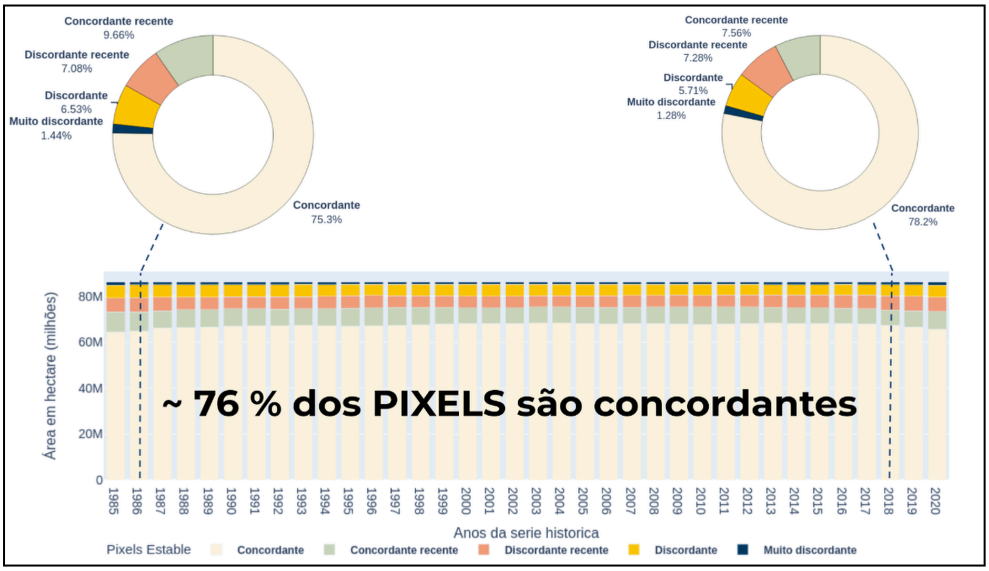


Figure 24: Statistics of the areas of agreement for collections 6.0, 7.1 and 8.0.

Based on this analysis, the question arises as to which classes are affected by large areas of disagreement. Thus, if we analyze the last two collections by cover, then the models would indicate where the pixels are that were classified as savannah in collection 7.1, for example, and are not in collection 8.0, as well as those that are now in collection 8.0 and were not in collection 7.1, figures 25a and 25b.

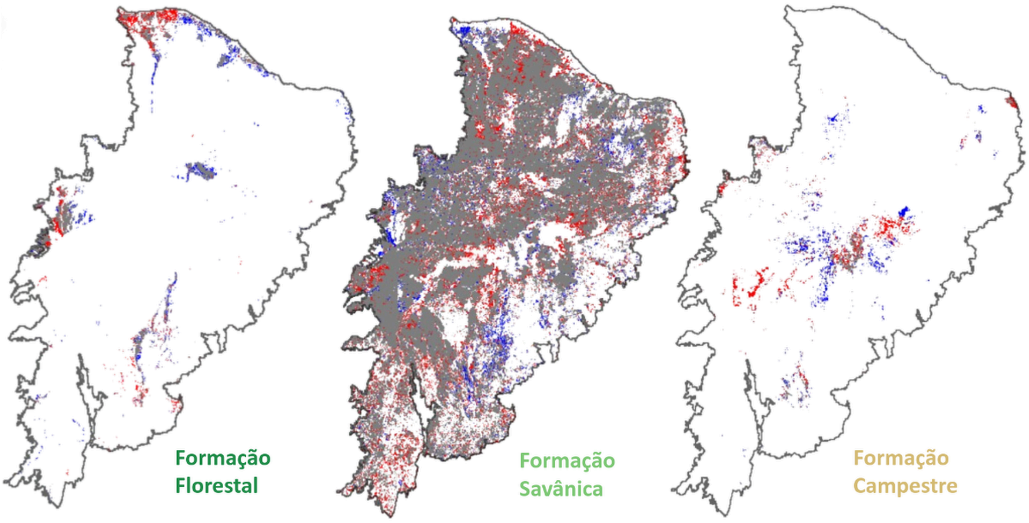


Figure 25a: Cover models for Forest Formation, Savannah Formation and Grassland Formation, for collections 7.1 and 8.0. In gray areas of coincidence, in blue areas only in the 7.1 collection and in red areas that were only mapped for the 8.0 collection.

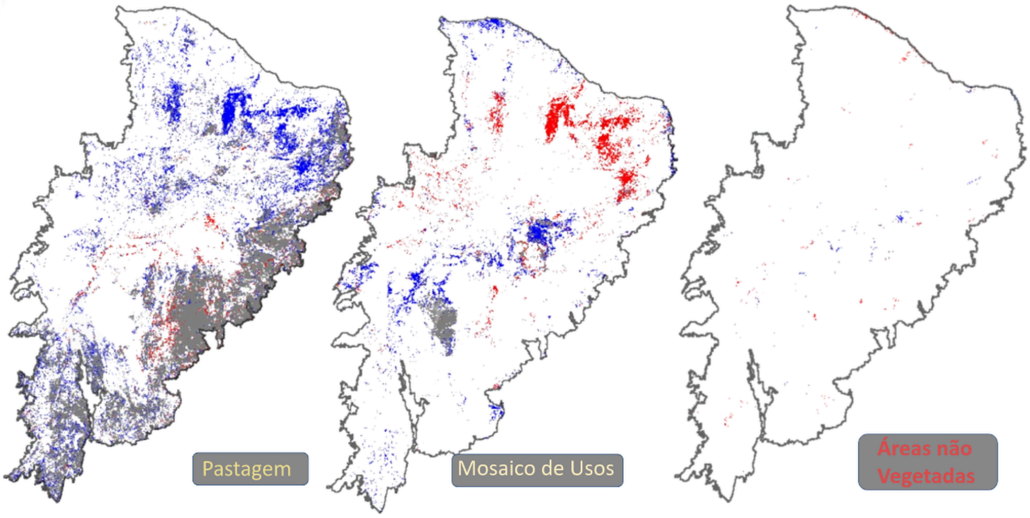


Figure 25b: Cover models for Forest Formation, Savannah Formation and Grassland Formation, for collections 7.1 and 8.0. In gray areas of coincidence, in blue areas only in the 7.1 collection and in red areas that were only mapped for the 8.0 collection.

Land covers with a greater presence in the caatinga, such as savannah Formation, have more than 80% agreement between the last two collections, see figure 26.

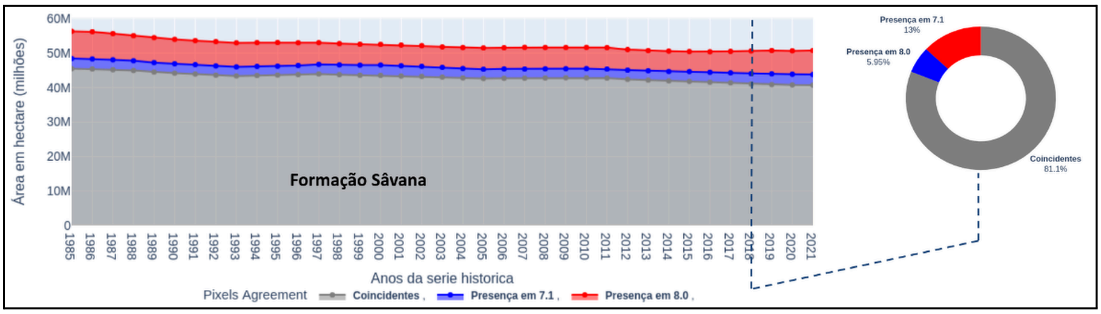
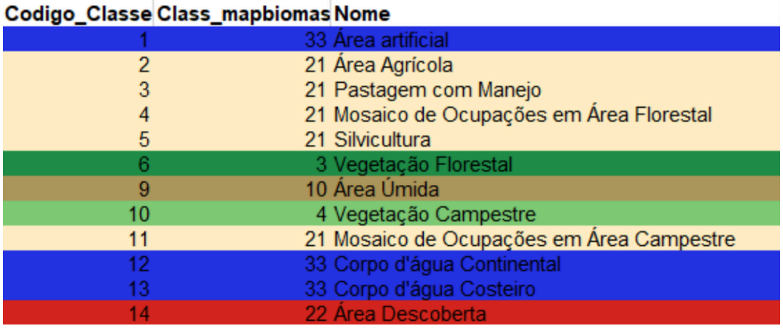


Figure 26: Areas of concordant pixels between collections 7.1 and 8.0.

An analysis of coincidences can be made using maps from other sources. To do this, we used the map from the Brazilian Institute of Geography and Statistics (IBGE), available on the download page of the institute's platform. The comparison was made to homogenize each of the land cover classes from the IBGE product with the MapBiomas maps (Figure 27).



Codigo_Classe	Class_mapbiomas	Nome
1	33	Área artificial
2	21	Área Agrícola
3	21	Pastagem com Manejo
4	21	Mosaico de Ocupações em Área Florestal
5	21	Silvicultura
6	3	Vegetação Florestal
9	10	Área Úmida
10	4	Vegetação Campestre
11	21	Mosaico de Ocupações em Área Campestre
12	33	Corpo d'água Continental
13	33	Corpo d'água Costeiro
14	22	Área Descoberta

Figure 27: Unified legend between IBGE classes and Mabiomas classes.

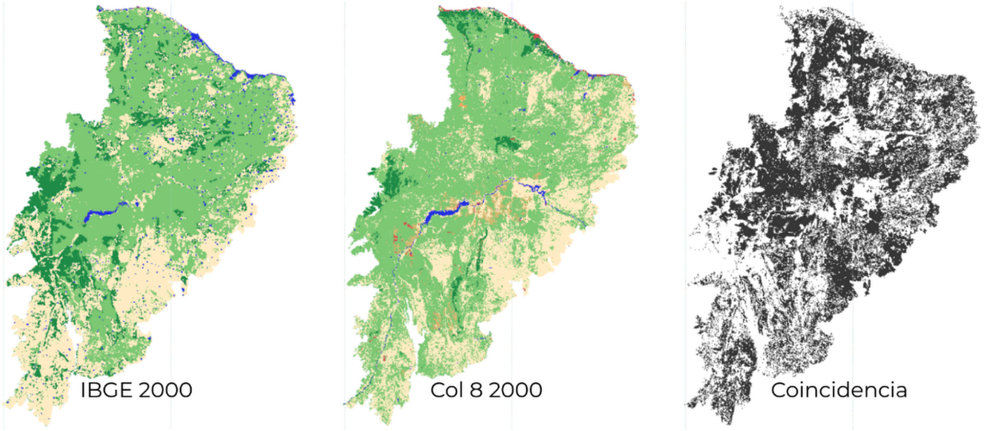


Figure 28: Coincidence Map of Caatinga using the IBGE LULC and MapBiomas LULC for the year 2000.

The greatest coincidences are found in areas to the west and north of the Caatinga where there are large expanses of savannah and in the south-east of the Caatinga where there is a high presence of pasture (Figure 28).

In seven years of IBGE maps, the differences with the MapBiomas cover maps was approximately 50% for all years (Figure 29).

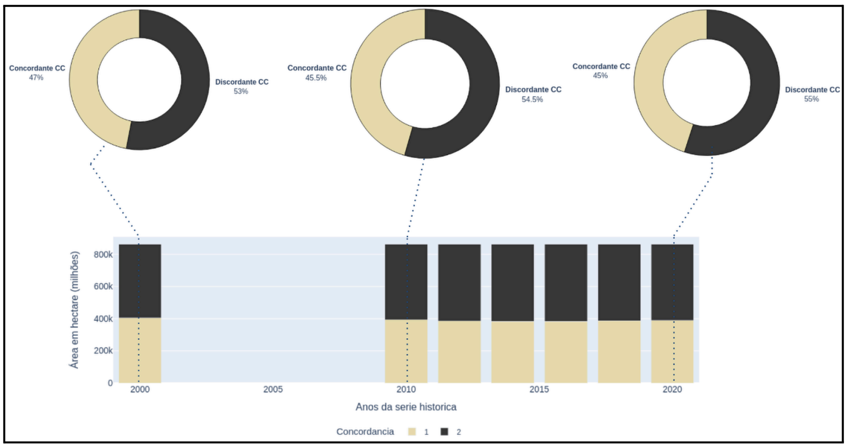


Figure 29: Areas of concordant pixels between the IBGE and mapbiomas maps for the years 2000, 2010, 2015, 2020.

8. 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.

LIU, F.T., TING, K.M. AND ZHOU, Z.H., 2008, December. Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

PATIL, D., PATIL, K., NALE, R. AND CHAUDHARI, S., 2022, July. Semantic segmentation of satellite images using modified U-Net. In *2022 IEEE Region 10 Symposium (TENSYMP)* (pp. 1-6). IEEE.

PELLEG, D. AND MOORE, A.W., 2000, June. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml* (Vol. 1, pp. 727-734).

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.

RONNEBERGER, O., FISCHER, P., & BROX, T. 2015, October. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.

ZHU, X.X., TUIA, D., MOU, L., XIA, G.S., ZHANG, L., XU, F. AND FRAUNDORFER, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, *5*(4), pp.8-36.