



Caatinga - Appendix

Collection 9

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 summarizes the specific methods used to generate land cover and land use annual maps for the Caatinga biome in the context of MapBiomias. 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 based on trial and error.

In collection 9.0, another model was used, the Gradient Tree Boost. The latter was implemented with the idea of measuring the performance of each classification output together with Random Forest and finalizing the mapping process with the model with the best performance.

For both models, an 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 9.0 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 MapBiomias

(<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 | Class | Main Improvements |
|------------|---------------|--|--|---|
| Beta & 1 | 2008 - 2015 | Empirical Decision Tree | Forest Formation, Non-Forest, Water Mask. | Proof of concept |
| 2.0 | 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 |
| 2.3 | 2000 - 2016 | | | |

| | | | | |
|-----------|-------------|---------------------------------------|--|---|
| 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 - 2022 | Random Forest / Gradient Tree Booster | Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Rocky Outcrop, Wooded Sandbank Vegetation. | |
| 9.0 | 1985 - 2023 | Gradient Tree Booster/ cluster | Forest, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other Non-vegetated Area, Wooded Sandbank Vegetation. Rocky Outcrop | Rocky outcrop class was made using a cluster model |

2. CLASSIFICATION METHOD

A more compact flowchart of the Caatinga biome mapping process in Collection 9.0 is shown in Figure 1. A few methodological modifications have been made since collection 6 with the intention of improving the map classification. The general procedures used for creating the land use and land cover (LULC) maps for the Caatinga Biome comprise: Data input, sample gathering, feature selection, hyperparameter tuning, models of classification, post-classification filters, validation and visual inspection, and integration with MapBiomias crosscutting themes.

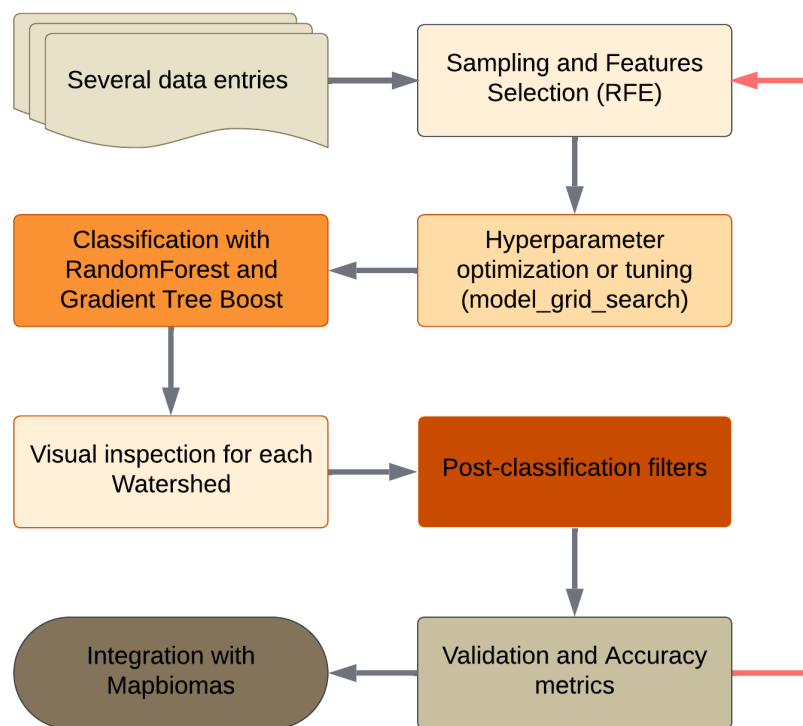


Figure 1. Simplified general flowchart.

Some improvements are illustrated with more detail in Figure 2 and are following described .

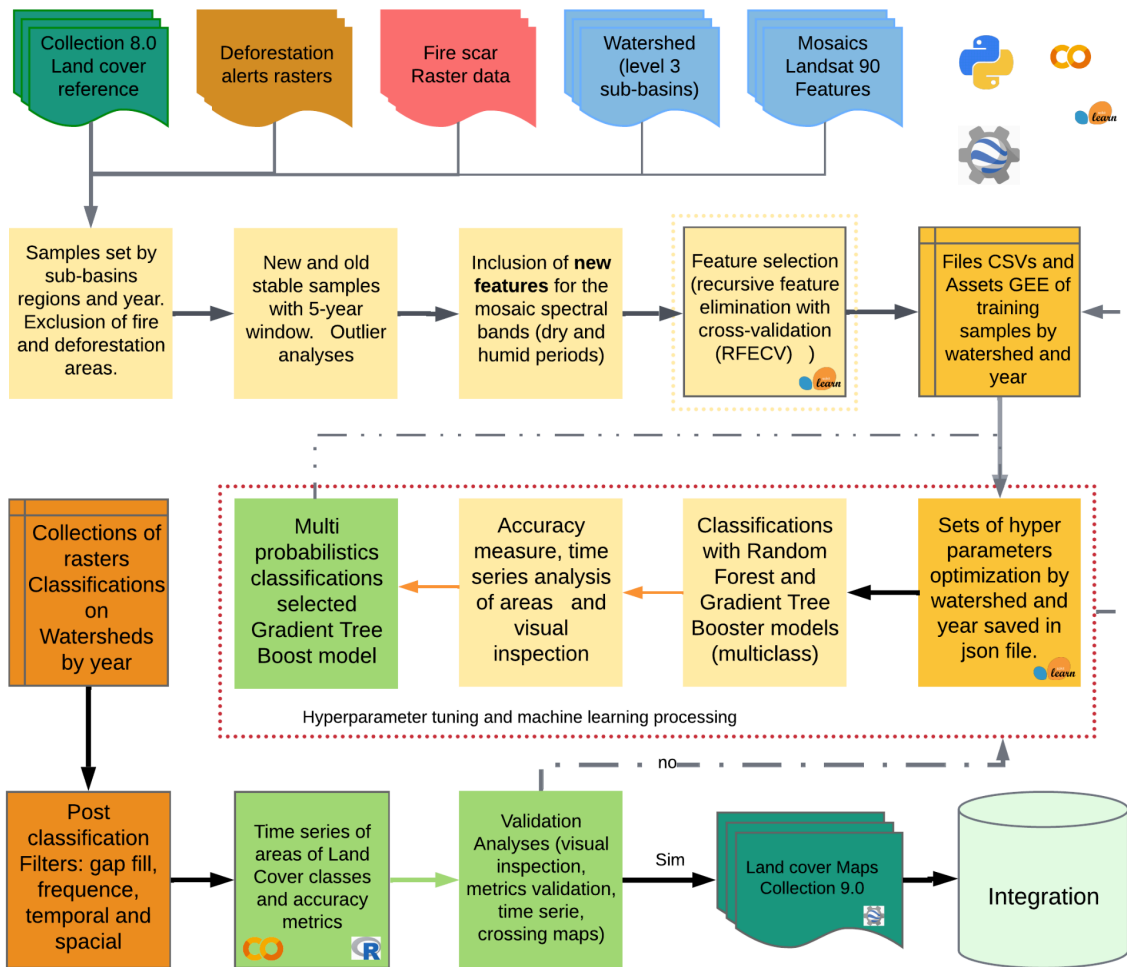


Figure 2. Classification process of MapBiomass Collection 9.0 (1985-2023) in the Caatinga biome.

2.1 Landsat Image Mosaics

In previous collections, the classification was performed by using Landsat 5 (TM), Landsat 7 (ETM+), and Landsat 8 (OLI). Up to Collection 6.0, we used data from the surface reflectance (SR) data, and from Collection 7.0 onwards we used Landsat Collections 2 surface temperature (ST) products. These Collections 2 were created with the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) algorithm (version 3.4.0) available on GEE as id asset "LANDSAT/LT05/C02/T1LL2" for Landsat 5, "LANDSAT/LE07/C02/T1LL2" for Landsat 7, and "LANDSAT/LC08/C02/T1LL2" for Landsat 8.

The mosaic building can be accessible via the path "projects/nexgenmap/MapBiomass2/LANDSAT/BRAZIL/mosaics-2", and it is saved in the asset project MapBiomass along with all the processing done to clean the data.

This mosaic includes fractions from spectrum unmixing, 119 spectral bands between spectral indices, and descriptive statistics computed for dry and wet seasons.

2.2 Definition of the period

In order to minimize confusion caused by extreme phenological changes between different types of natural vegetation and other land use and land cover (LULC) categories, such as cropland, the image selection period for the Caatinga biome was defined with the goal of maximizing the coverage of Landsat images after cloud removal/masking. The Caatinga climate, which varies greatly in seasonal precipitation compared to most other Brazilian biomes, is the principal driver of the physiological activity of plants throughout the year. The bulk of caatinga vegetation is categorized as seasonal, exhibiting significant deciduousness throughout the year. Therefore, 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).

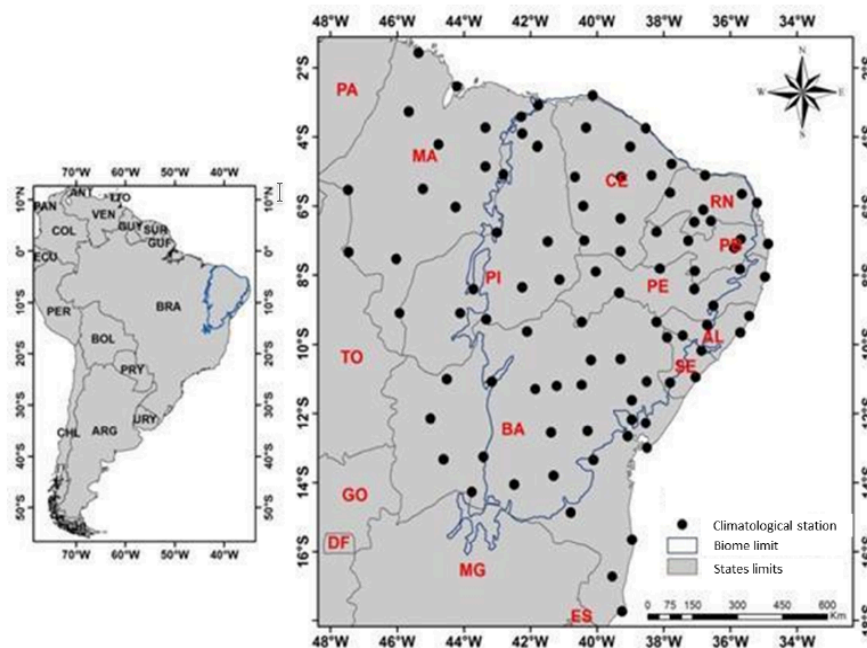


Figure 3. Location of the weather stations which rainfall series were evaluated for the selection of the mosaic periods in the Caatinga biome.

Then, a periodic window scan was carried out for the entire Caatinga biome, indicating that the period between January and July (with higher rainfall in the Caatinga biome) (Figure 4) is more likely to provide images with spectral contrast capable of separating different LULC categories for the biome. The choice of these sets of parameters helped to define the mosaics with better spectral quality and fewer amounts of noise and clouds in the images for the biome.

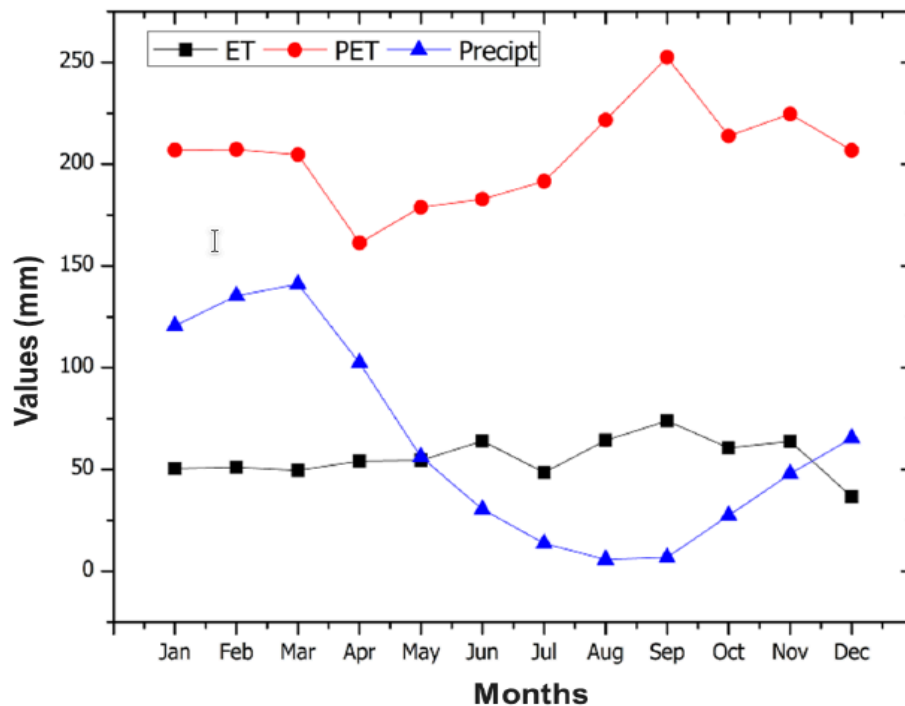


Figure 4. Temporal variation of monthly mean precipitation (Precipt), evapotranspiration (ET), and potential evapotranspiration (ET) in the Caatinga biome.

2.3 Image selection

For the selection of Landsat scenes to build the annual mosaics by map sheet, 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.

2.4 Mosaic quality

The mosaic quality was evaluated by computing the frequency of each available pixel during the selected period in the Caatinga biome (Figure 5). As a result of the selection criteria, all of them presented satisfactory quality (i.e. less noise such as clouds, relief and clouds shadows.). In Collections 4.1, 5.0, 6.0, 7.0 and 8.0, a single change to this calculation refers to the Caatinga biome boundary, that was updated (IBGE, 2019). There is no change for Collection 9.0.

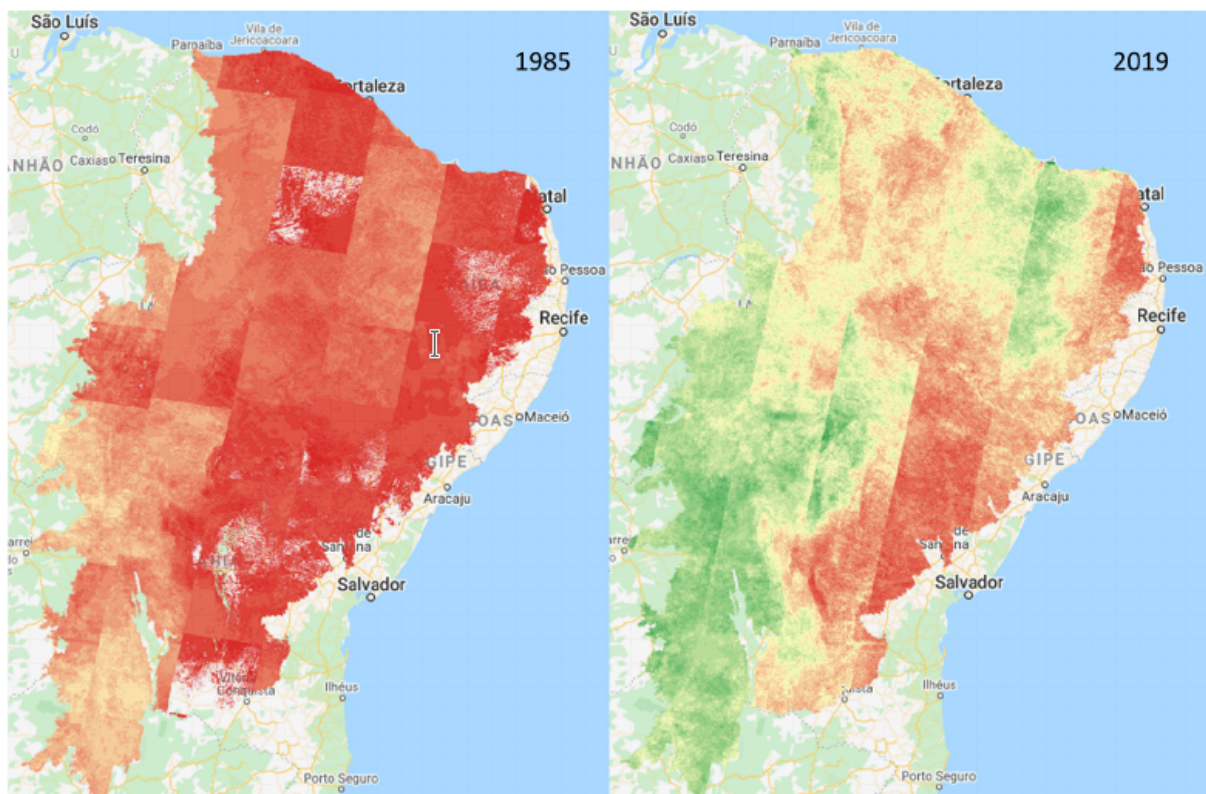


Figure 5. Example of Landsat pixel availability for the mosaics of 1985 and 2019 in the Caatinga biome. Colors refer to pixel availability, where red is low, yellow is medium, and green is high.

3. DEFINITION OF REGIONS FOR CLASSIFICATION

A classification done in homogenous regions reduces the variability between the spectral values of the pixels, both within each LULC category as between categories, and allows the use of the same set of samples to classify large areas of

the mosaic. However, given that this is a computationally expensive task, we divided the biome into smaller areas based on watershed boundaries available by the Agência Nacional de Águas (www.ana.gov.br) (Figure 6). By taking these natural borders, we aimed higher homogeneity within each area, allowing for the automation of the sampling process using GEE's Python API. In earlier collections, the level 4 of watershed boundaries was used and the biome was divided in 320 regions.

Due to the the update of the boundaries of the biomes (IBGE, 2019) in Collection 5, some additional basins were included. In Collections 6.0, 7.1, 8.0 and 9.0 we used another division merging between level 2 and level 4 watershed boundaries, resulting in a final division of the Caatinga biome into 42 regions, figure 6.

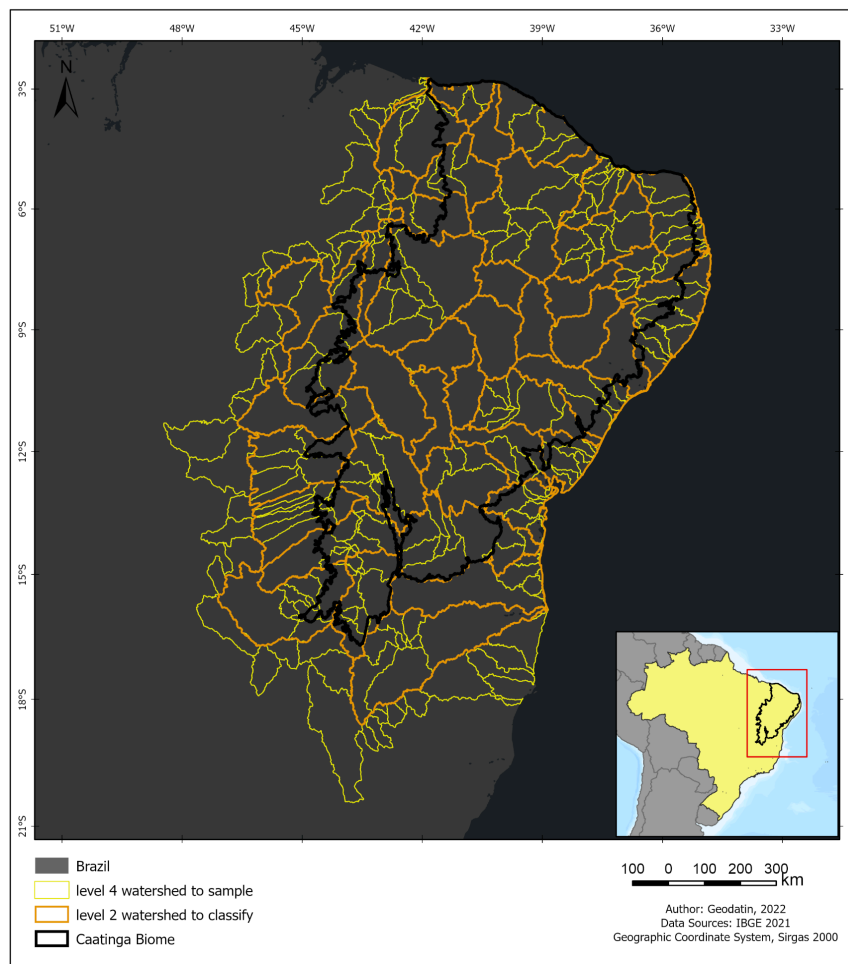


Figure 6. The Caatinga watersheds used in the classification and sampling of the latter MapBiomas collections.

4. CLASSIFICATION

4.1 Land cover and land use classes

The digital classification of the Landsat mosaics in the Caatinga biome sought to map a subset of ten LULC classes from the MapBiomias legend in Collection 9.0 (Table 2). Some of these classes were later integrated with the cross-cutting themes. The Mosaic of Uses class in the Caatinga was overlaid by the Agriculture or Pasture classes, remaining only 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 MapBiomias Collection 9.0.

| 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 | 22 | Anthropic | Land use | Beach and Dune, Urban Infrastructure |

| | | | | |
|--------------------------------|----|---------------------|-----------------------|--|
| Area | | | | 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 / 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 and feature selection

The most recent methodology of the sampling process was initiated in collection 8.0. The spectral information of samples this collection were updated and joined with the samples data of collection 9.0. The first criterion was to exclude areas where there have been changes in coverage in the year of the mosaic. As changes we considered burned areas, deforested areas, areas within a buffer of gaps from clouds or cloud shadows, and areas that show variability between consecutive years. The second was to consider only areas that were stable over a 3-year window and match its class with the same class in the previous year's collection. To achieve these criteria for each region, sorting at least 1500 samples per class was required, which compelled the use of the function `ee.Image().stratifiedSample()` to collect samples from small areas inside a class. Also, to correct the imbalance of samples occasionally generated due to insufficient samples in a region, sample pixels were collected manually in the mosaic for the years 2016 and 2021, filtered by the same criteria mentioned above. Additionally, samples from collection 8.0 that met the criteria of the collection 9.0 were selected using the cluster noise removal filters.

The spectral information is essentially derived from the MapBiomass 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 samples were collected were subjected to four conditional layers: one indicating the areas of coincidences pixels, figure 21, another indicating the areas of stable pixels over a 5-year period, areas where no deforestation occurred in the last three years of the series, and areas with no fire scars. In this fashion, the gathered points were placed in a folder within the MapBiomias file, with each feature collection denoting the region and year.

The final step with the sampling process was to eliminate outliers by class. The Learning Vector Quantization algorithm was then implemented in the function `ee.Clusterer.wekaLVQ()` from Kohonen (2003). This cluster algorithm enables the grouping of all samples in the new category. The first two sets of clusters with more pixels in each class were then chosen for analysis. Later, each feature was kept as 50 % of the number class, with a quantity of around 1000 pixels.

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, wet 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 region. With this, the Gradient Tree Boost classifier, is minimized using the gradient descent technique to minimize errors, achieves better performance with normalized data. The Random Forest also improves its performance as it excludes outliers from the mosaics with normalization, fitting the pixel values within the same range of the sample pixel values.

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 MapBiomias Collection 7.0 (General ATBD MapBiomias, 2020). In Collection 8.0, a larger number of spectral indices were calculated to expand the feature space of the MapBiomias mosaic. The goal was to find a reduced space that offers more separability and contrast between targets. The figure 7, depicts an

instance of the samples corresponding to watershed “744” which have an unbalanced distribution due to the nature distribution the class cover on real regions.

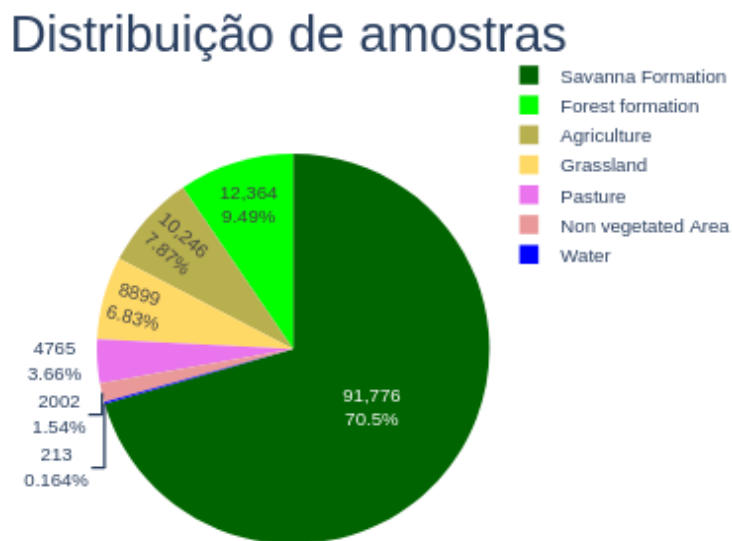


Figure 7: Distribution of samples for sub-basin 744 in the year 2000.

Achieving separability in the feature space is an important challenge when performing remote sensing image classification in the Caatinga Biome. Figure 8 demonstrates that separability within a spectral band is limited for various targets in the image.

Another way of visualizing this can be seen in Figure 9, which plots the "blue_median", "green_median", "red_median", "nir_median" bands of the mosaic for six coverage classes.

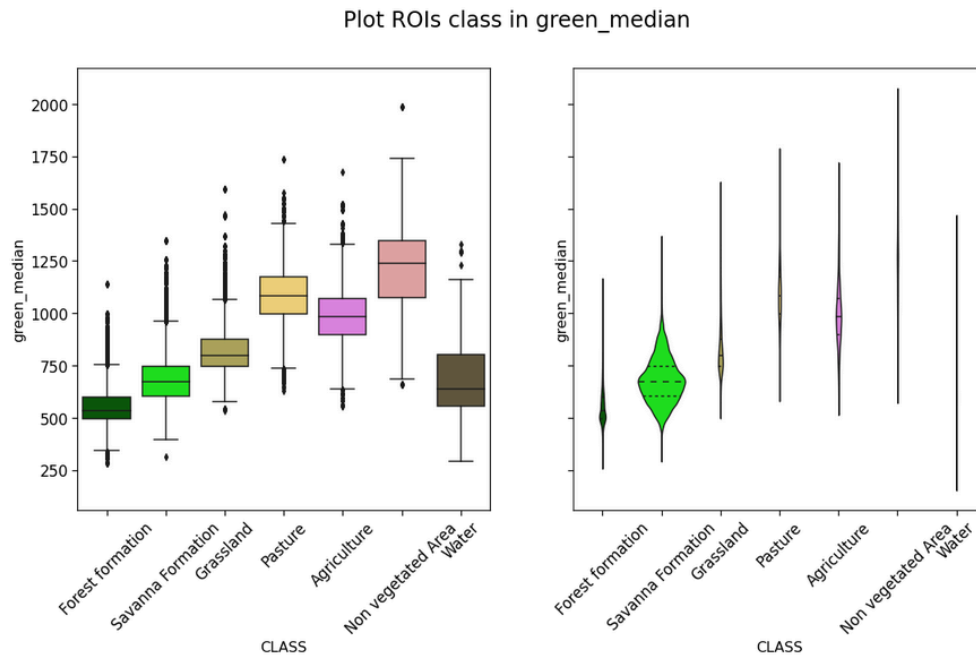


Figure 8: Box and violin plots from samples of spectral band “GREEN” in the main land cover classes mapped by the Caatinga team. In axis X have class cover and axis Y spectral values from band green.

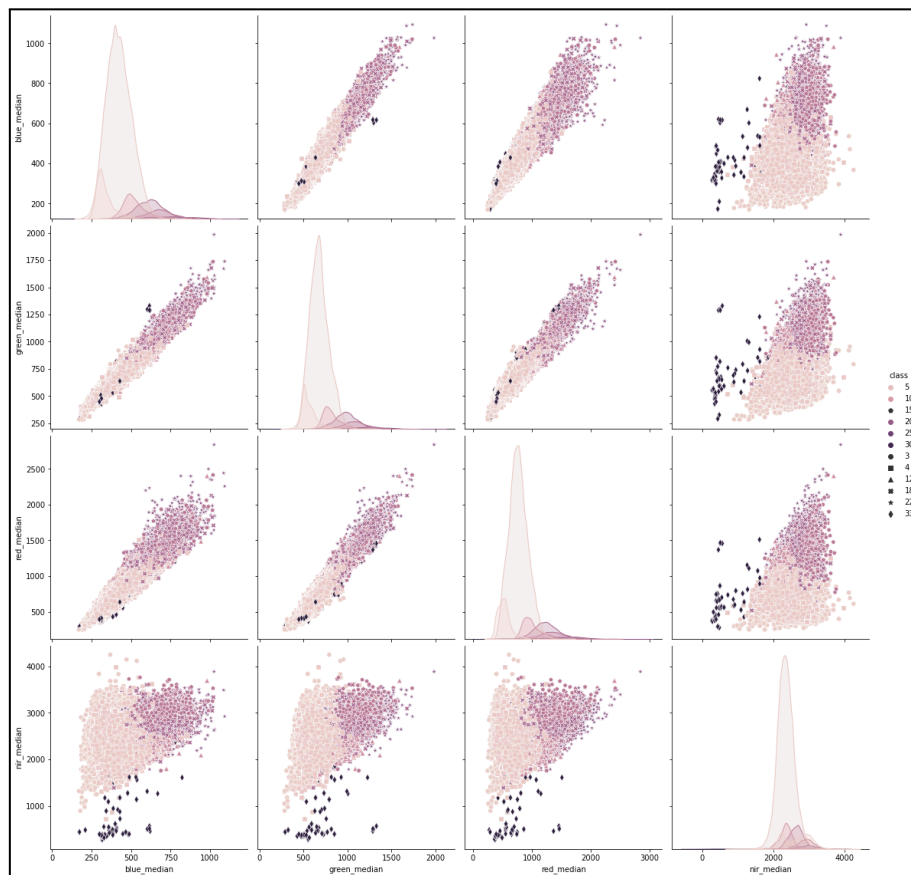


Figure 9: Spatial distribution of samples for variables, “blue_median”, “green_median”, “red_median”, “nir_median”.

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 | |
| median | median | | median dry | | | |
| median dry | media dry | | median wet | | | |
| median wel | stdDev | | min | | | |
| green | median texture | GCVI | median | soil | median | |
| | stdDev | | median dry | | median dry | |
| red | median | NDVI | median wet | | ndfi | median wet |
| | median dry | | amp | | | stdDev |
| | median wet | | median | median | | |
| | min | | median dry | median dry | | |
| nir | median | NDWI | median wet | sefi | median wet | |
| | median dry | | amp | | min | |
| | median wet | | median | | median dry | |
| | min | | median dry | | median wet | |
| SWIR1 | median | SAVI | median wet | shade | stdDev | |
| | median wet | | median | | median | |
| | min | | median dry | | median dry | |
| | stdDev | | median wet | | median wet | |
| SWIR1 | median | PRI | stdDev | | min | |
| | median wel | | median | | amp | |
| | min | | median dry | | | |
| | stdDev | | median wet | | | |

The feature space of collection 9.0 has been expanded to be more robust and to follow good data augmentation practices used in data science (Table 4).

Table 4: Feature space subset indexes calculated from the estimated bands of the Landsat mosaic of mapBiomass in the Caatinga biome in the MapBiomass Collection 9.0.

| 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 regions 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 10), in order to eliminate some variables 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.

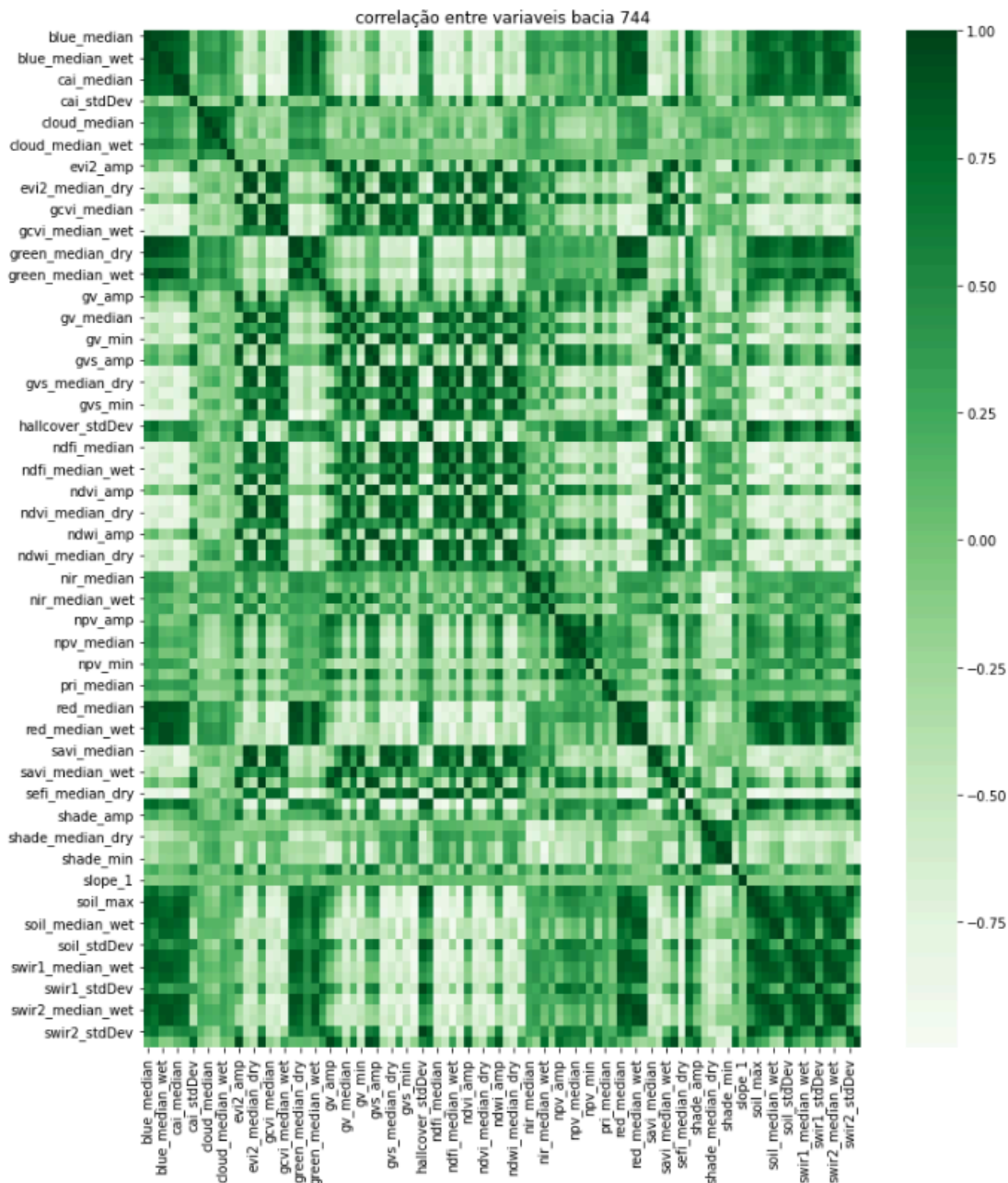


Figure 10. Example plot of correlation between variables of samples from year 2020.

Since Collection 8.0, the model includes 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 can be found at:

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 11). 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 11: Example of the implemented feature selection function (RFECV) and a list of selected variables.

A script was implemented for the **Hyperparameter Tuning** process after selecting the variable sets by regions 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 12, where the optimal pair of parameters would be (40, 0.175).

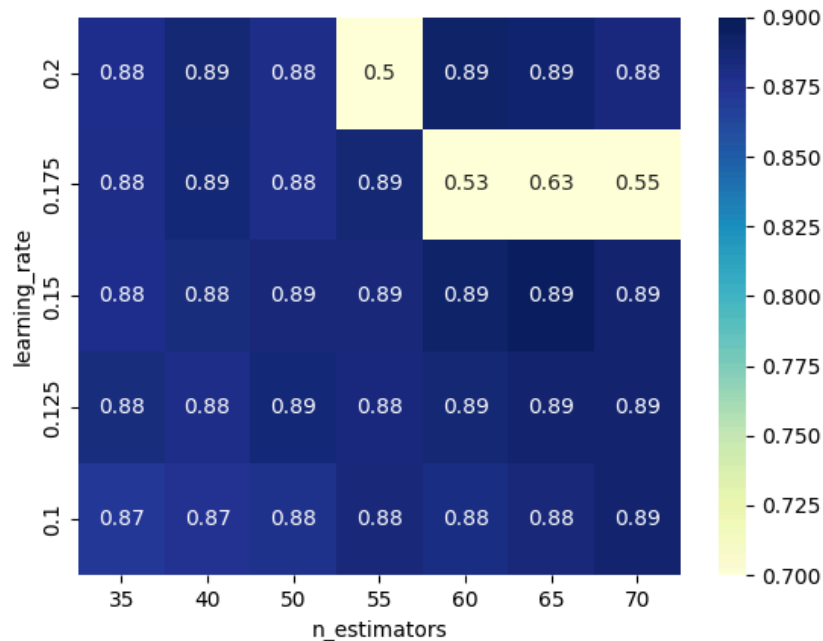


Figure 12. Example plot of combination of "learning rate" parameters and "n estimators".

Part of the code implemented for selecting optimal parameters is shown in Figure 13. Each pair of optimal parameters for year and 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 13: Part of the code implemented for the Hyperparameter tuning process.

For each set of 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 MapBiomass's Github (<https://github.com/mapbiomas-brazil/caatinga>).

4.4 Classification algorithm, training samples, and parameters

During the classification process, the input data is adjusted to allow the MapBiomass mosaics to be classified by regions and year. The data is then displayed using a GEE script and visual inspection by the team's analysts to assess the classification results by regions 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. During each round of classification, two versions are 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 both classifiers is shown in Figure 14. The final version of the classification was done for all regions and years with samples from the same subset used for all years, and it was trained in the same mosaic of the classified year.

```
'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 14: Example parameters for the Random Forest and Gradient Tree Boost classifiers.

5. POST-CLASSIFICATION

The rules of the post-classification filters were adapted to the cover classes used in the Caatinga biome. Like other biomes, it follows the sequence of fill-gap filter, spatial filter, temporal filter, spatial filter and temporal filter. How these filters work is described below.

5.1 Gap Fill filter

This filter aims to fill in data (pixels) in the image that do not have observations. For this filter, the time series of raster maps from the classification is used. Landsat images from the last years of the series have more satellites and consequently the annual mosaics are created with more images and fewer pixel gaps. The filter works by searching, for each pixel gap, the first pixel with a value in the historical series.

5.2 Spatial filter

The 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 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 reclassified 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 reclassified 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 changed 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 2023).

6. ANALYSIS OF SECONDARY VEGETATION

Another data produced from the land use and coverage maps is secondary vegetation data. The methodology for creating these layers can be found in: https://brasil.mapbiomas.org/wp-content/uploads/sites/4/2024/04/Deforestation_-_Secondary_Vegetation-Appendix-ATBD-Collection-8.docx.pdf, Figure 15.

Primary vegetation is formed by the following cover classes: Savannah Formation, Forest Formation and Grassland Formation. The areas of these three classes are distributed as follows: 96% Savannah Formation, 3% Forest Formation and 1% Grassland Formation. Secondary vegetation is considered to be vegetation that was mapped as vegetation after being deforested after the first year of the 1985 series. The distribution of this secondary vegetation is as follows: 96% Savannah Formation, 3% Forest Formation and 1% Grassland Formation. As highlights, we have that the loss of primary vegetation will reach 20 million hectares by 2023, corresponding to 34% of all primary vegetation in 1985. Recovery per year has an average of between 4 and 8 thousand hectares.

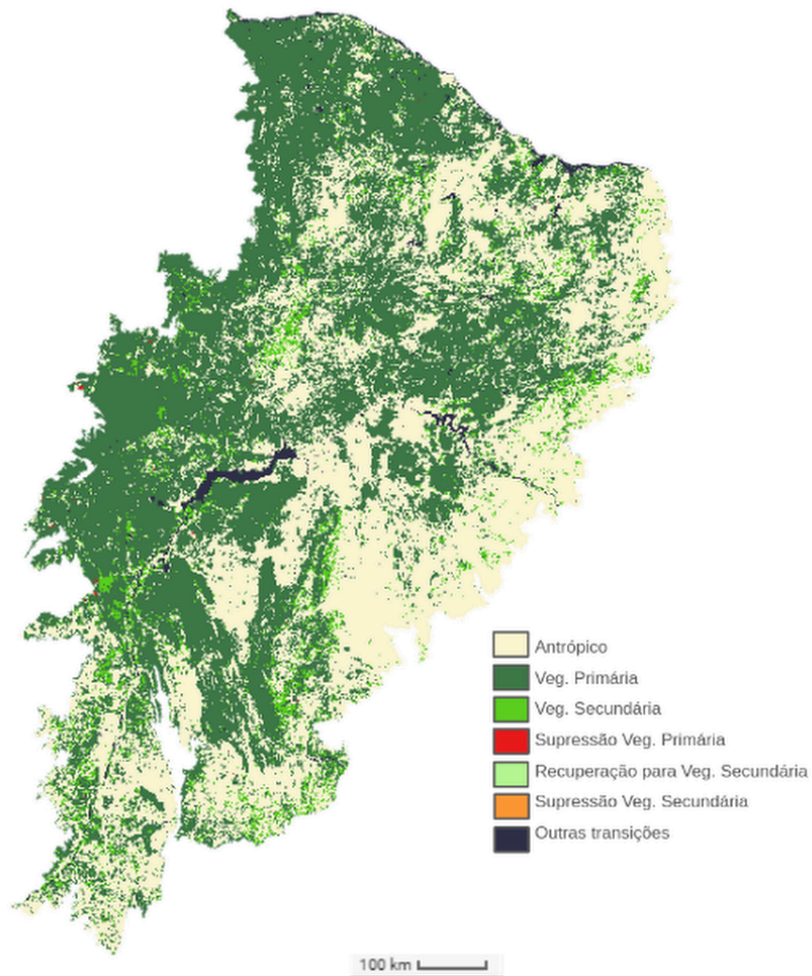


Figure 15: Map of the Primary vegetation and anthropic areas of year 2023.

To understand how the loss of primary vegetation by cover class behaved, see Figure 16, as well as the increase in secondary vegetation by cover class in the historical series from Figure 17. A relationship between the suppression of primary vegetation and the suppression of secondary vegetation can be seen in Figure 18.

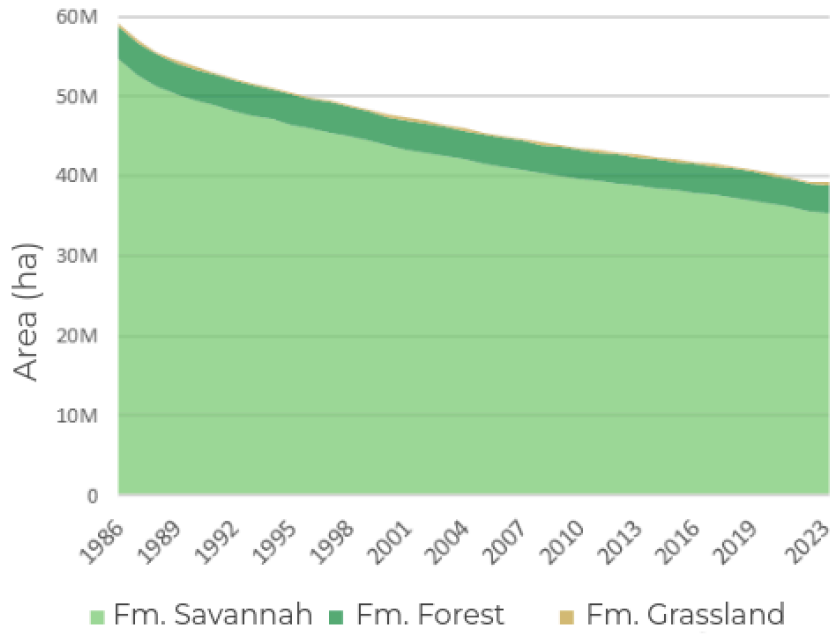


Figure 16: Primary Vegetation Areas by Cover Class for the Historical Series 1986-2023.

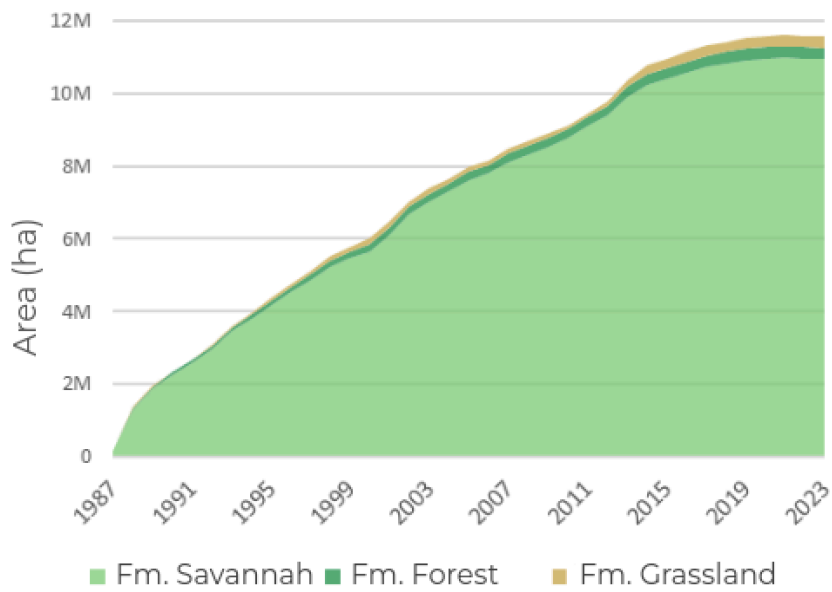


Figure 17: Secondary Vegetation Areas by Cover Class for the historical series 1986-2023.

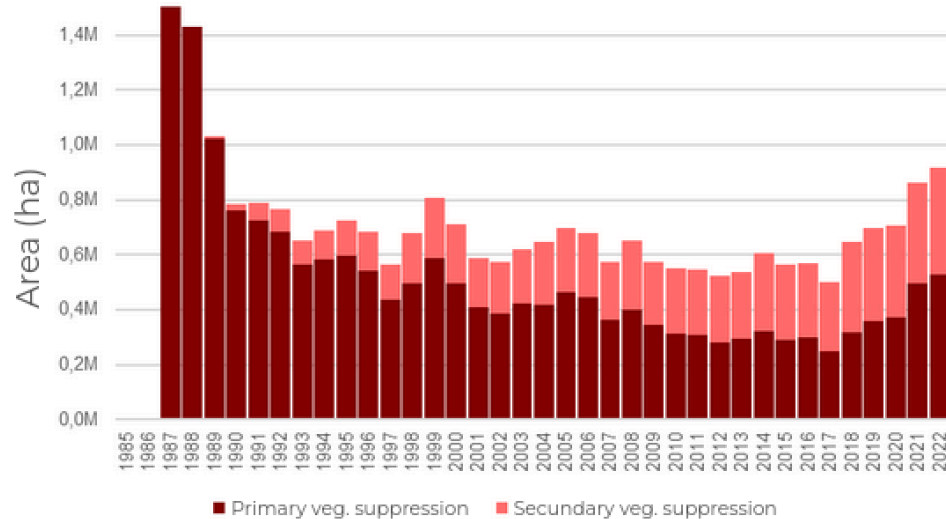


Figure 18: Primary and Secondary Vegetation of suppression areas for the historical aeries 1987-2023.

7. VALIDATION STRATEGIES

The validation stage of each process was created using independent validation points provided of the team of Laboratório de Processamento de Imagens e Geoprocessamento (**Lapig**) from Universidade Federal de Goiás. We used all points that at least two interpreters considered the same class, resulting in more than 85,000 validation points. Figure 19 shows the result of the accuracy assessment for the level 3 legend of the MapBiomias Collection 9.0 (1985-2023). For each year, overall accuracy, allocation disagreement and quantity disagreement (Pontius & Millones, 2011) were computed, and from them the correspondent average values for the whole series.

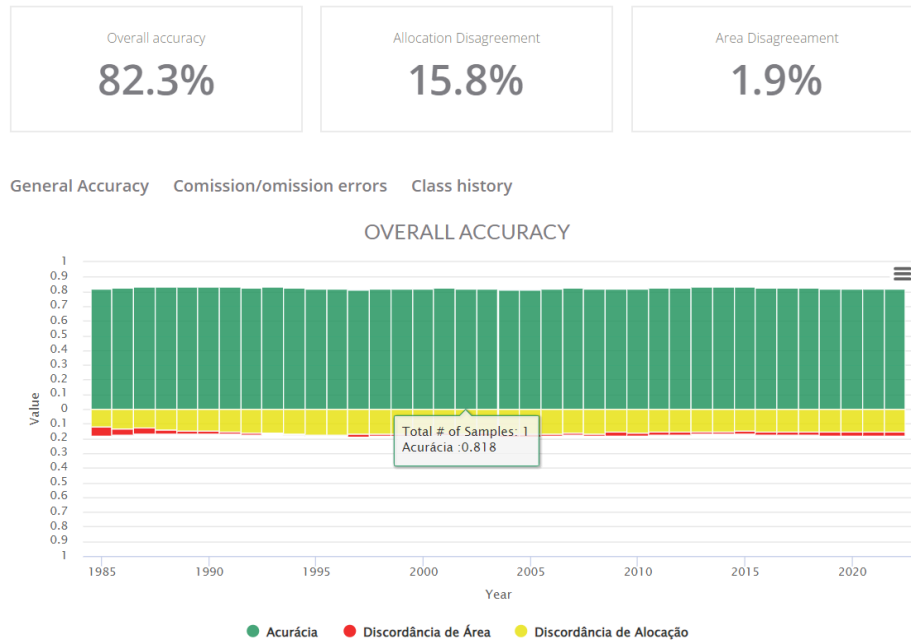


Figure 19. Accuracy of level 3 of MapBiomias Collection 9.0 in the Caatinga biome (1985-2023).

The approach used in this collection allowed to obtain higher accuracy than previous collections (Table 5). Figures 20 and 21 illustrate omission and commission errors. By analyzing these data, it is possible to determine which classes are confused with others and then devise a new strategy to reduce commission and omission errors.

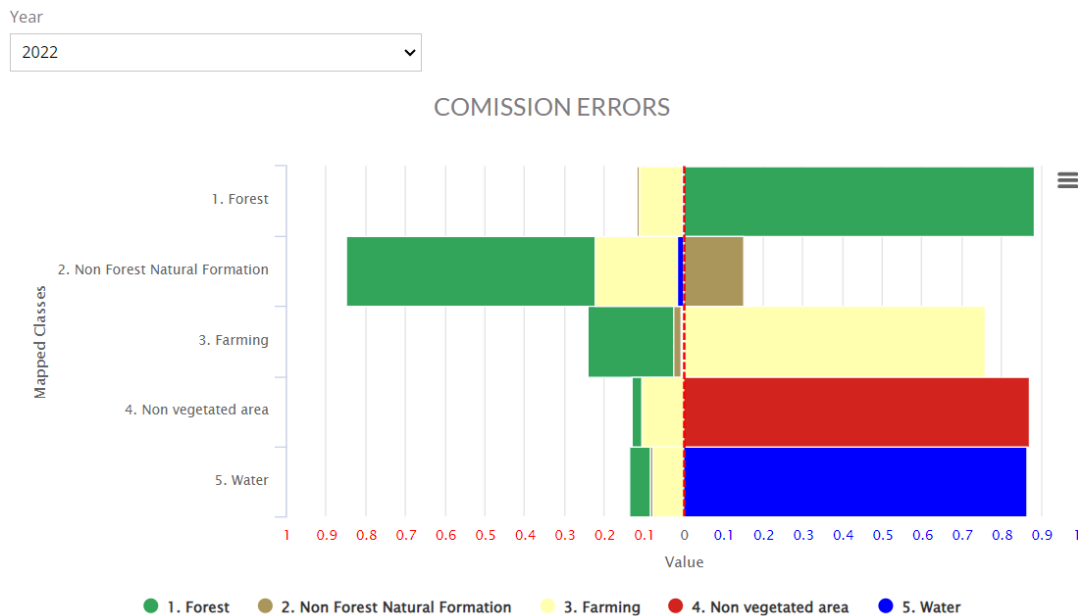


Figure 20. Commission errors of the land cover and land use mapping in the Caatinga biome from year 2022.

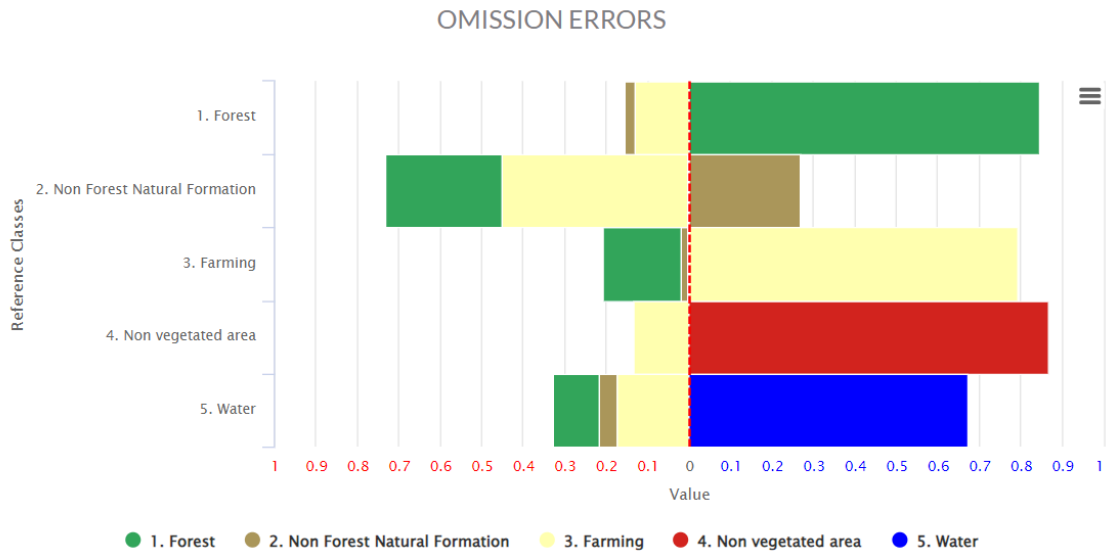


Figure 21. Omission errors of the land cover and land use mapping in the Caatinga biome from year 2022.

Table 15 shows the evolution of overall accuracy results across collections 3.1 and 9.0. Also the different classification algorithms used.

Table 15. The evolution of the Caatinga mapping collections in the MapBiomias Project, with classification methods and overall accuracy in Level 1, 2, and 3, based on 34 years with reference points.

| Collection | Method | Overall 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 % |

If we plot all values in the accuracy series, we can better compare and observe all of the data from the different collections, Figure 22.

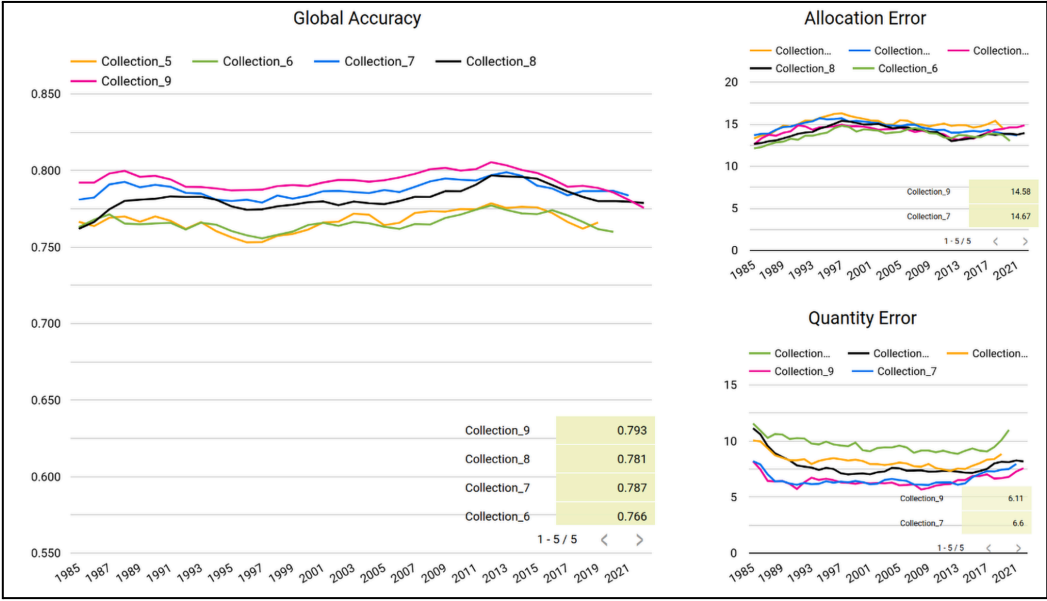


Figure 22. Plot of Accuracy metrics of level 3 of MapBiomass Collections 5.0, 6.0, 7.0, 8.0 and 9.0 in the Caatinga biome (1985-2023).

Another technique to assess the quality of a time series of LULC maps is to examine the area of each class across time as shown in Figure 23. This type of analysis makes it possible to compare the areas of the classes throughout the time 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.

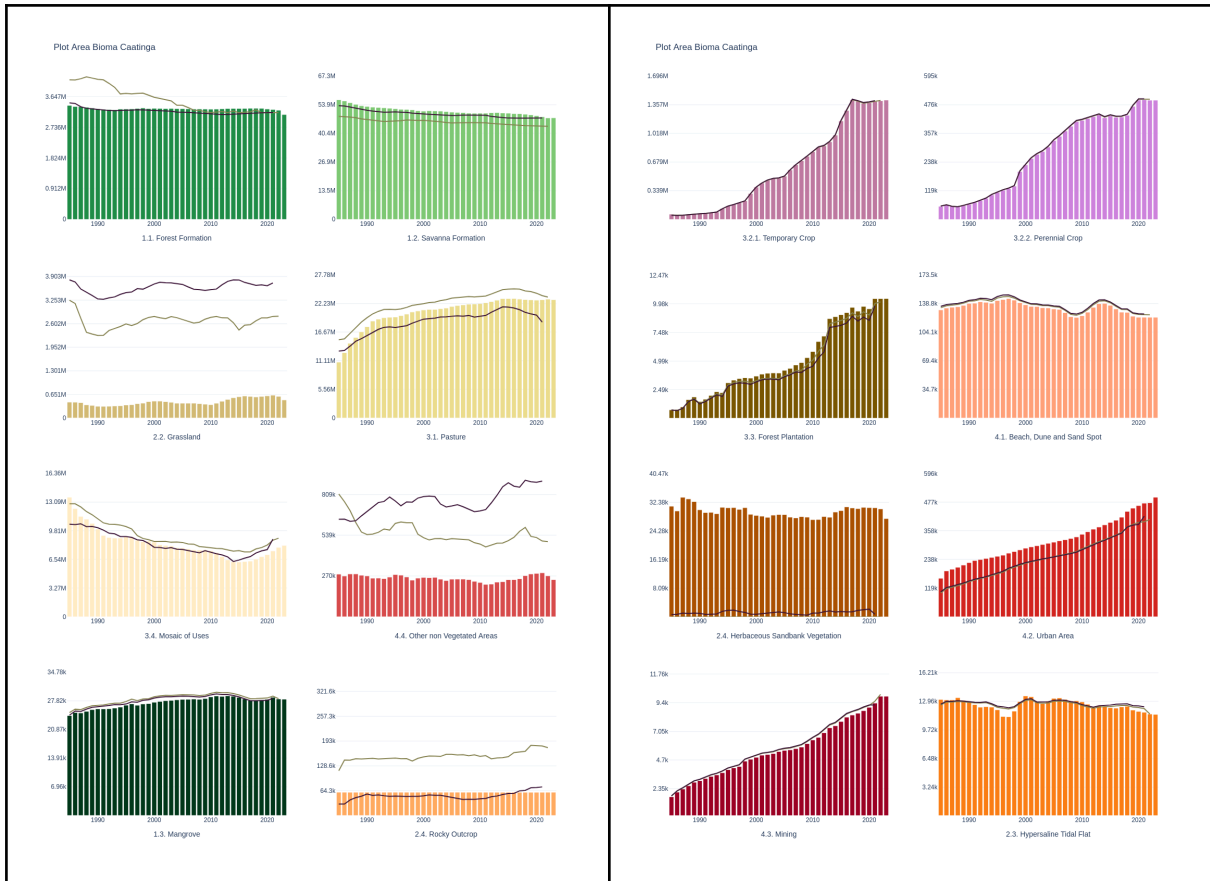


Figure 23. Time series of level 3 classes of MapBiomas Collection 9.0 in the Caatinga biome (1985-2023) per area (ha). The green and brown lines represent respectively collections 7.1 and 8.0.

Another way of validating the coverage data was to compare the concordance of the LULC classes between collections. For this analysis we used a rule implemented by the Pampa team that accounts for coincidences (Figure 24).






| | | | | | |
|--------------------------|---|---|---|---|---|
| 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 |
| | <i>Classificações com maior incerteza</i> → | | | | |

Figure 24: 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 biome and the Chapada Diamantina, with coordinates [-12.8821188, -41.6785173], show the greatest disagreements (Figure 25). These places of greatest discordance indicate where new samples should be collected in order to discriminate the correct class. They also indicate areas where the pixels in the mosaic may have clouds, noise or shadows.

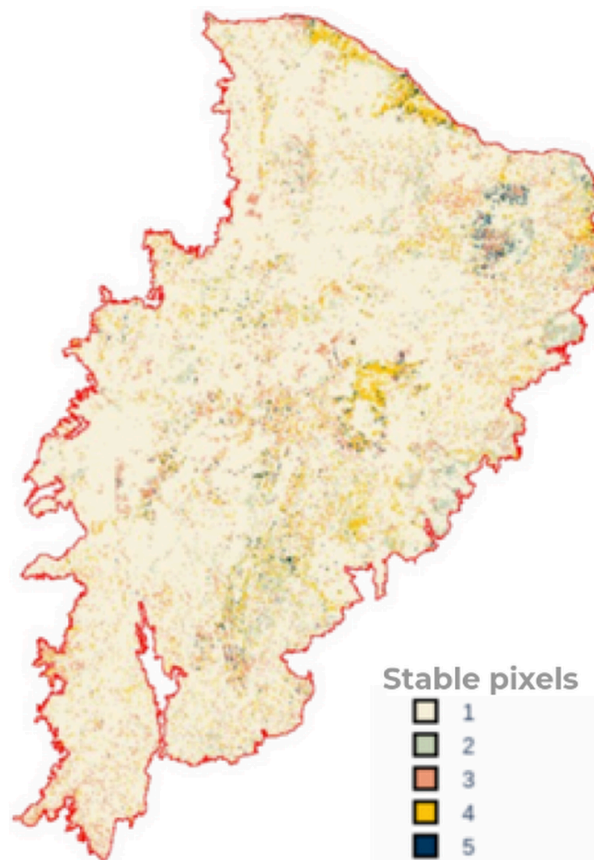


Figure 25: Pixel coincidence between collections 6.0, 7.1 and 8.0.

The concordance statistics over the time series show that on average 76 % of the pixels are coincident between the 3 collections (Figure 26). 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.

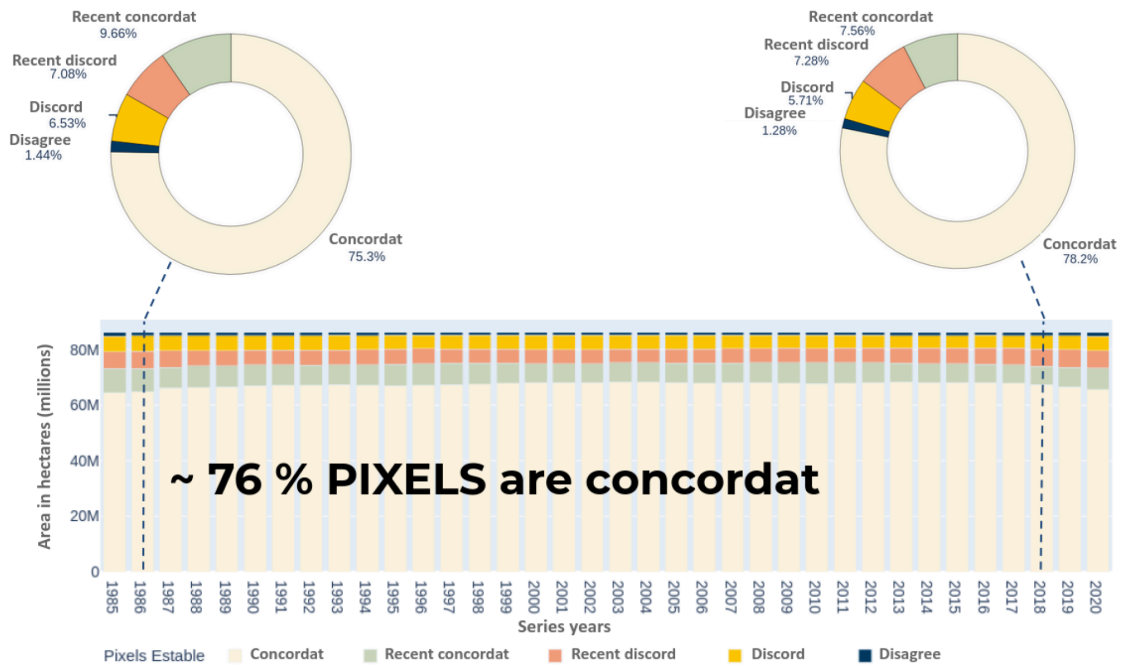


Figure 26: Statistics of the areas of agreement for collections 6.0, 7.1 and 8.0.

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 27a and 27b).

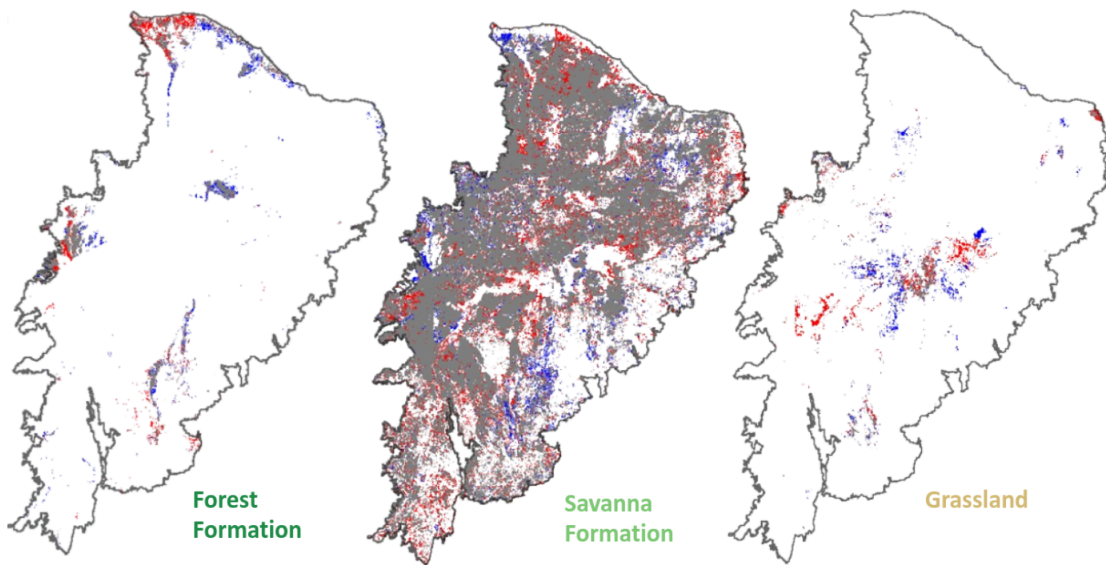


Figure 27a: Concordance for Forest Formation, Savannah Formation and Grassland Formation between collections 7.1 and 8.0. In gray concordat, in blue present only in collection 7.1, and in red present only in collection 8.0.

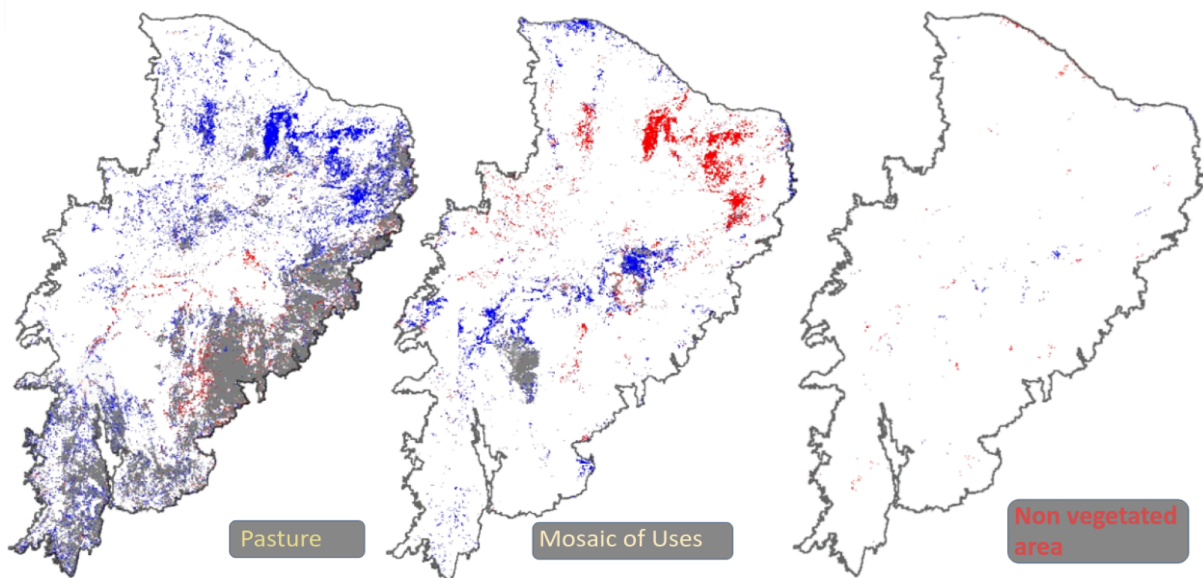


Figure 27b: Concordance for Pasture, Mosaic of uses and Non vegetated areas between collections 7.1 and 8.0. In gray concordat, in blue present only in collection 7.1 , and in red present only in collection 8.0.

Land cover categories with a greater presence in the Caatinga biome, such as Savannah Formation, presented more than 80% of agreement between the last two collections (Figure 28).

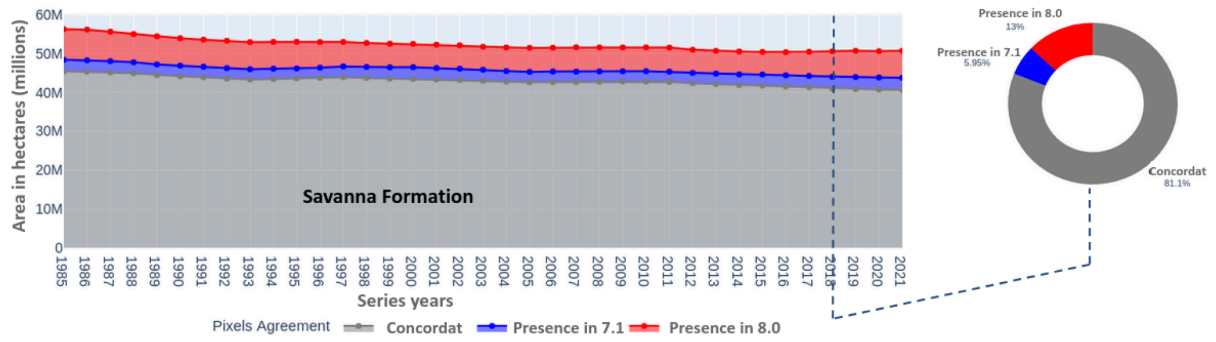


Figure 28: Areas of concordant pixels between collections 7.1 and 8.0.

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 MapBiomass maps (Figure 29).

| Code_Class | Class_mapbiomas | Class Names |
|------------|-----------------|---------------------------------------|
| 1 | 33 | Artificial area |
| 2 | 21 | Agricultural Area |
| 3 | 21 | Managed Pasture |
| 4 | 21 | Mosaic of Occupations in Forest Areas |
| 5 | 21 | Forestry |
| 6 | 3 | Forest Vegetation |
| 9 | 10 | Wet Area |
| 10 | 4 | Countryside Vegetation |
| 11 | 21 | Mosaic of Occupations in Rural Areas |
| 12 | 33 | Continental water body |
| 13 | 33 | Coastal body of water |
| 14 | 22 | Discovered Area |

Figure 29: Unified legend between IBGE classes and Mabiomas classes.

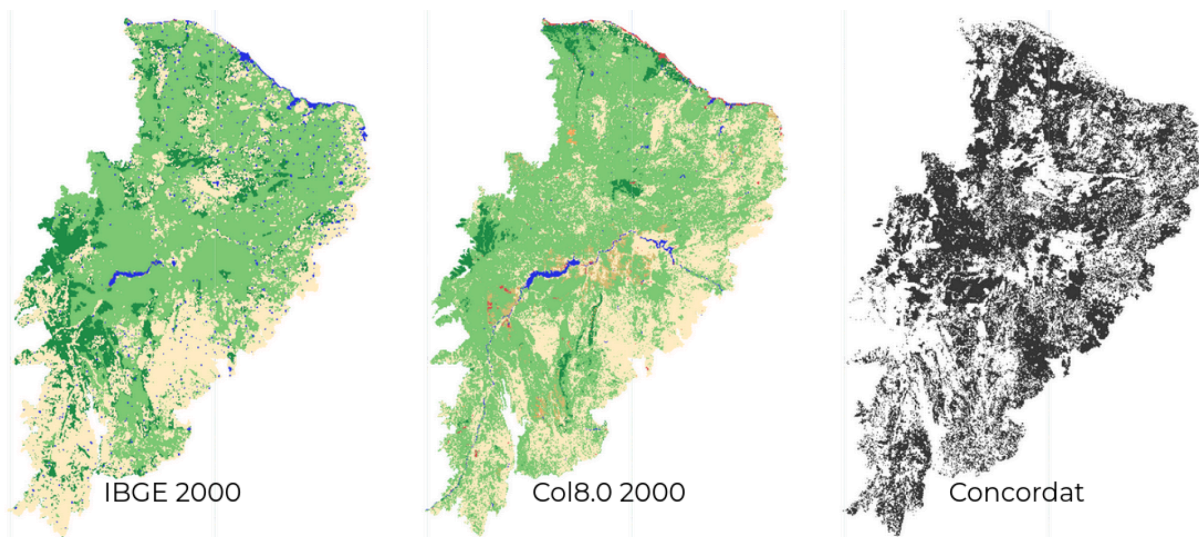


Figure 30: Coincidence Map of Caatinga using the IBGE LULC and MapBiomass LULC for the year 2000.

The greatest coincidences are found in areas to the west and north of the Caatinga biome, where there are large expanses of savannah, and in the south-east, where there is a high presence of pasture (Figure 31).

In seven years of IBGE maps, the differences with the MapBiomas land cover and land use maps was approximately 50% for all years (Figure 31).

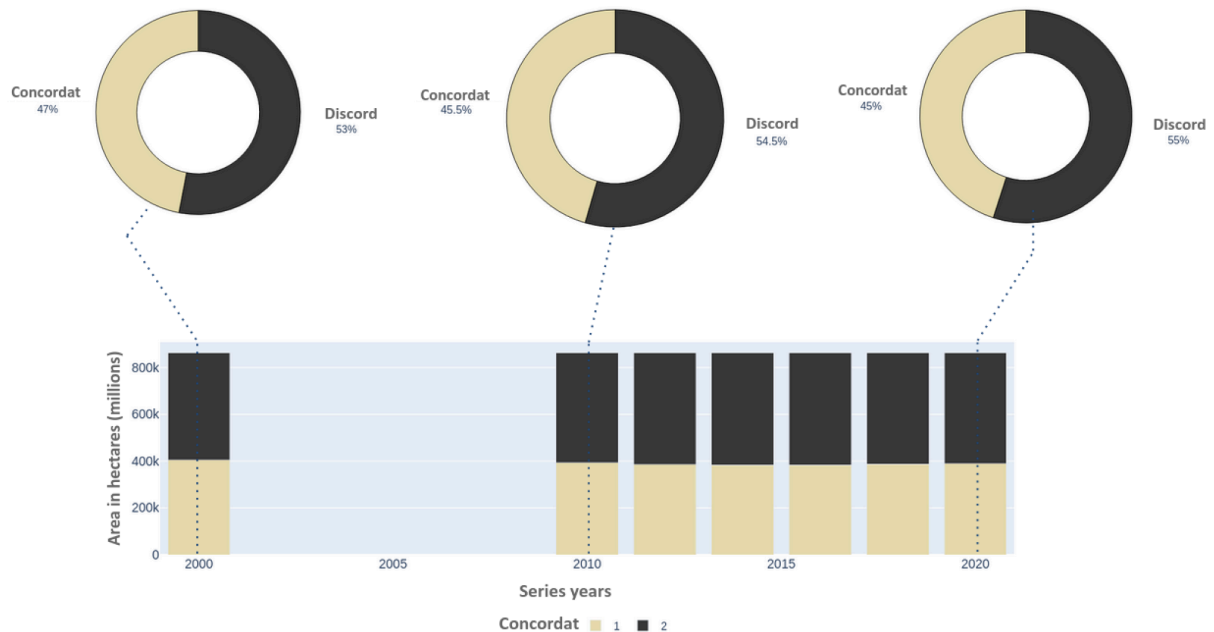


Figure 31: Areas of concordant pixels between the IBGE and MapBiomas maps for the years 2000, 2010, 2015, 2020.

8. REFERENCES

ARCOVA, F. C. S.; CICCIO, 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: <https://bdiaweb.ibge.gov.br/#/consulta/vegetacao>

IBGE Mapa de Biomas do Brasil – primeira aproximação. Rio de Janeiro, 2020, disponível em:

<https://www.ibge.gov.br/geociencias/informacoes-ambientais/vegetacao/15842-biomas.html>

Pontius, R.G., Millones, M., 2011. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int. J. Remote Sens.* 32, 4407–4429. 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.