Atlantic Forest - Appendix

Collection 10

Version 1

General coordinator

Natalia Crusco

Luis Guedes Pinto

Team

Fernando Frizeira Paternost

Yasmin Gelli Kloosterman

Marcos Reis Rosa

Jacqueline Freitas

1. Overview of classification method

The initial classification of the Atlantic Forest biome within the MapBiomas project consisted of applying decision trees to generate annual maps of the predominant native vegetation (NV) types, which were distinguished in three classes: Forest, Savanna, and Grassland. The method used to generate these annual maps evolved over time, with significant improvements from the first MapBiomas Collection to the present.

Collection 1.0 covered the period from 2008 to 2015 and was published in 2016. Collections 2.0 and 2.3 covered the period from 2000 to 2016 and were published in 2018. The classification using Random Forest algorithm was implemented in Collection 2.3, and from this point onward, the empirical decision tree was used to generate stable samples, which were classified as the same NV type over the considered period (2000-2016). These stable samples were used to train the Random Forest models to classify the entire time series by using Landsat imagery. Collections 3.0 and 3.1 expanded the period covered to 1985–2017. Collections 4, 5, 6, and 7 used training samples collected based on the stable samples from the previous collection with adjustments in the sample balance and new samples collected to improve specific regions. In Collection 8, multiple classifications with different SEEDs from the RF were used. The value of the seed (whether positive or negative) was also used to alter the sample balancing across different classes. A total of 10 different classifications were performed per region and the MODE of all classifications was then used to determine the final class assigned to each pixel. Additionally, some post-classification filters were developed with the primary goal of reducing areas of false regeneration and false deforestation in the biome. The production of Collection 9, with land cover and land use annual maps for the period of 1985-2023, followed a sequence of steps in the Atlantic Forest biome, similar to those used in the previous Collections 4 to 8 (Figure 1). However, some improvements were added up, particularly in the mosaics, balance of samples, post classification filters and auxiliary maps (Table 1). In collection 9, the approach of generating classifications using different SEED values was replaced by employing the MultiProbability approach, which calculates the probability of a pixel belonging to multiple classes and assigns the final classification value to the class with the highest probability, while storing the highest probability value to create an uncertainty map. Collection 10 included several improvements since the initial processing phase. Some highlights are related to the use of cloud mosaics, classification of the herbaceous sandbank vegetation class starting from the extraction of stable samples, a new classification process for the rocky outcrop class, and updated post-classification filters.

Table 1. The evolution of the Atlantic Forest mapping collections in the MapBiomas Project, its periods, level and number of classes, brief methodological description, and global accuracy in Levels 1 and 2.

Collection	Period	Levels /N. Classes	Method	Global Accuracy
Beta & 1	8 years 2008-2015	1 / 7	Empirical Decision Tree	
2.0 & 2.3	16 years 2000-2016	3 / 13	Empirical Decision Tree & Random Forest (2.3)	
3.0 & 3.1	33 years 1985-2017	3 / 19	Random Forest	Level 1: 87.3% Level 3: 82.4% *
4.0 & 4.1	34 years 1985-2018	3 / 19	Random Forest	Level 1: 89.0% Level 3: 84.2% *
5.0	35 years 1985-2019	4 / 21	Random Forest	Level 1: 91.4% Level 2: 87.4% *
6.0	36 years 1985-2020	4 / 25	Random Forest	Level 1: 91.9% Level 2: 87.3%
7.0	37 years 1985-2021	4 / 27	Random Forest	Level 1: 91.9% Level 2: 87.1%
8.0	38 years 1985-2022	4 / 28	Random Forest	Level 1: 92.3% Level 2: 87.2%
9.0	39 years 1985-2023	4 / 29	Random Forest	Level 1: 91.7% Level 2: 86.6%
10	40 years 1985-2024	4 / 30	Random Forest	Level 1: 91.5% Level 2: 86.1%

** Due to hierarchy changes in the forest classes, level 2 of collection 6 and 7 is being compared to level 3 of previous collections.*

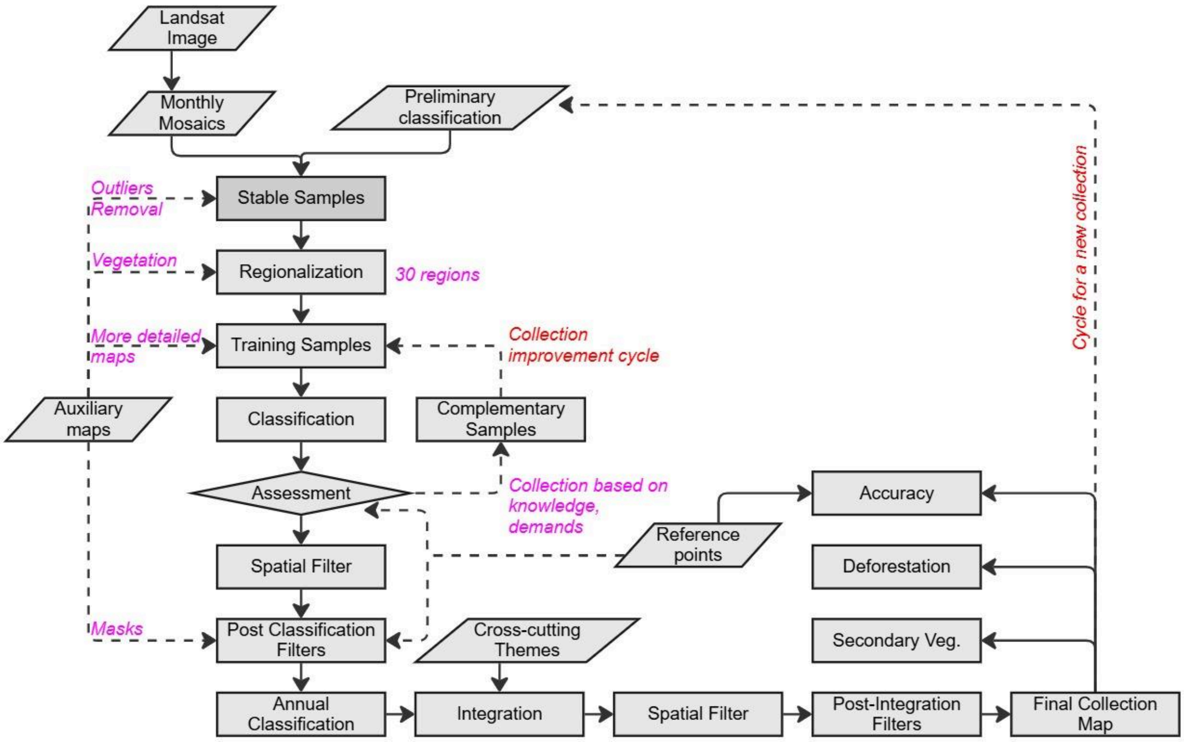


Figure 1. Classification process of Collection 10 in the Atlantic Forest biome. Inclined gray rectangles represent databases, while linear rectangles point key-steps in the workflow. Solid arrows indicate main flow, and dashed arrows point to assessment and evaluation cycles. Pink and red text shows additional information for specific steps.

2. Landsat image mosaics

The methodology for selecting Landsat images and mosaics is progressively refined across collections and is detailed in the following sections:

2.1 Collection 9 and Prior

2.1.1. Definition of the temporal period

Until collection 5, the classification was performed by using Landsat 5 (TM), 7 (ETM+) and 8, (OLI) top of atmosphere (TOA) data. In the collection 6, we adopted the use of surface reflectance (SR) data, and the use of TOA was discontinued. Since Collection 7 we adopted USGS Landsat 8 Level 2, Collection 2, Tier 1.

The mosaic of images consists of a composition of the pixels that are extracted from all the images available in a defined period within a year. Once this period's initial and final dates were defined, the per-pixel statistical parameters (i.e median, min, max, amplitude) were calculated, generating one median image with several bands. The aggregation of these composed pixels was conducted for each year, producing the annual Landsat mosaics, which were then submitted for classification.

Despite the diversity of ecosystems and the great extent of the biome both in latitudinal amplitude and in coast extension (Figure 2), the image selection period for the Atlantic Forest biome was defined from April to September (Figure 3) aiming to maximize the coverage of Landsat images after cloud removal/masking.

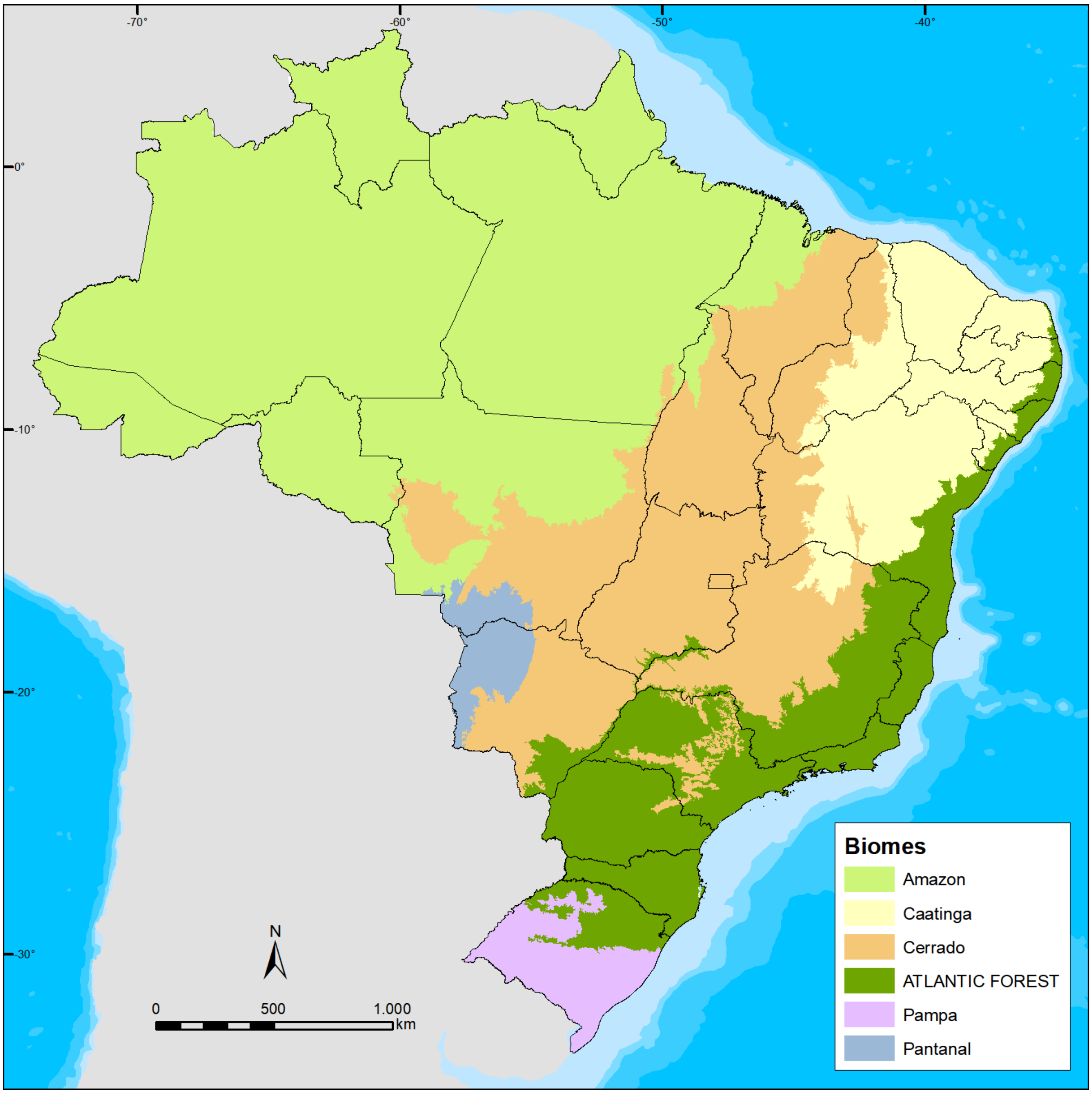


Figure 2. Biomes in Brazil (IBGE 1:250.000).

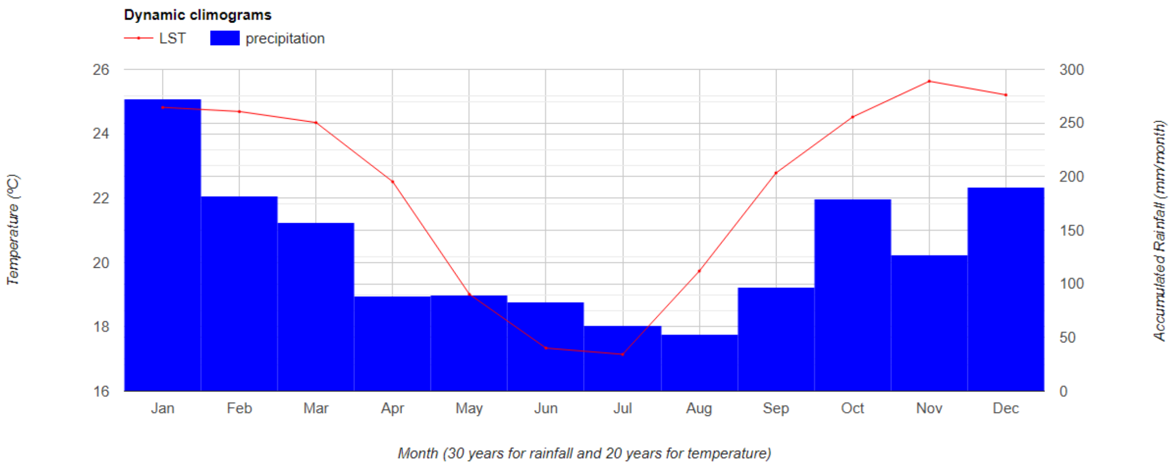


Figure 3. Climograph from 1988 to 2018 with CHIRPS (Climate Hazards Group InfraRed Precipitation with Station) precipitation data and MODIS (Moderate-Resolution Imaging Spectroradiometer) temperature data. (FUNK; PETERSON; LANDSFELD, 2015).

2.1.2 Image selection

For the selection of Landsat scenes to build the mosaics of each chart for each year, within the acceptable period, a threshold of 50% of cloud cover was applied (i.e., any available scene with up to 50% of cloud cover was accepted). This limit was established based on a visual analysis after many trials observing the results of the cloud removing/masking algorithm. 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 holes. Whenever possible, this was made by including months at the beginning of the period, in the winter season.

In most cases, the period from April 1st to August 30th was good for getting a mosaic with no or few missing information caused by clouds and shades. In some specific cases, however, it was needed to significantly extend the temporal period to include images from September and October. In the Northeast states, the period was from February 1st to 30th of October to maximize the visible areas and avoid missing areas caused by clouds (Figure 4).

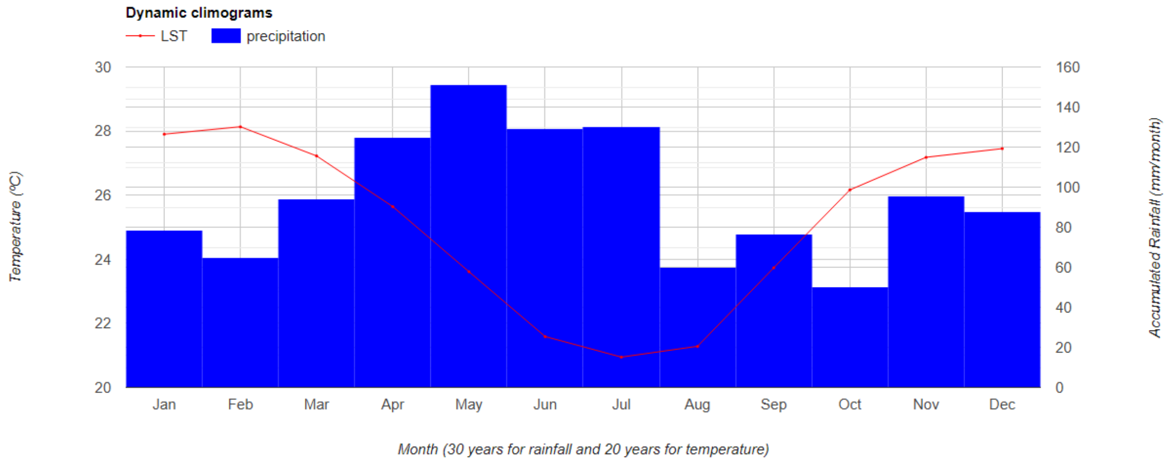


Figure 4. Climograph from 1988 to 2018 of Northeast states with CHIRPS (Climate Hazards Group InfraRed Precipitation with Station) precipitation data and MODIS (Moderate-Resolution Imaging Spectroradiometer) temperature data. (FUNK; PETERSON; LANDSFELD, 2015).

For each year, we used images from the best Landsat available:

- 1985 to 1999 – Landsat 5
- 2000 to 2002 – Landsat 7
- 2003 to 2011 – Landsat 5
- 2012 – Landsat 7
- 2013 to 2023 – Landsat 8

We made a visual analysis on the preliminary mosaics to identify and manually remove images with noises (clouds, shadow, or sensor defect) for each year (Figure 5).

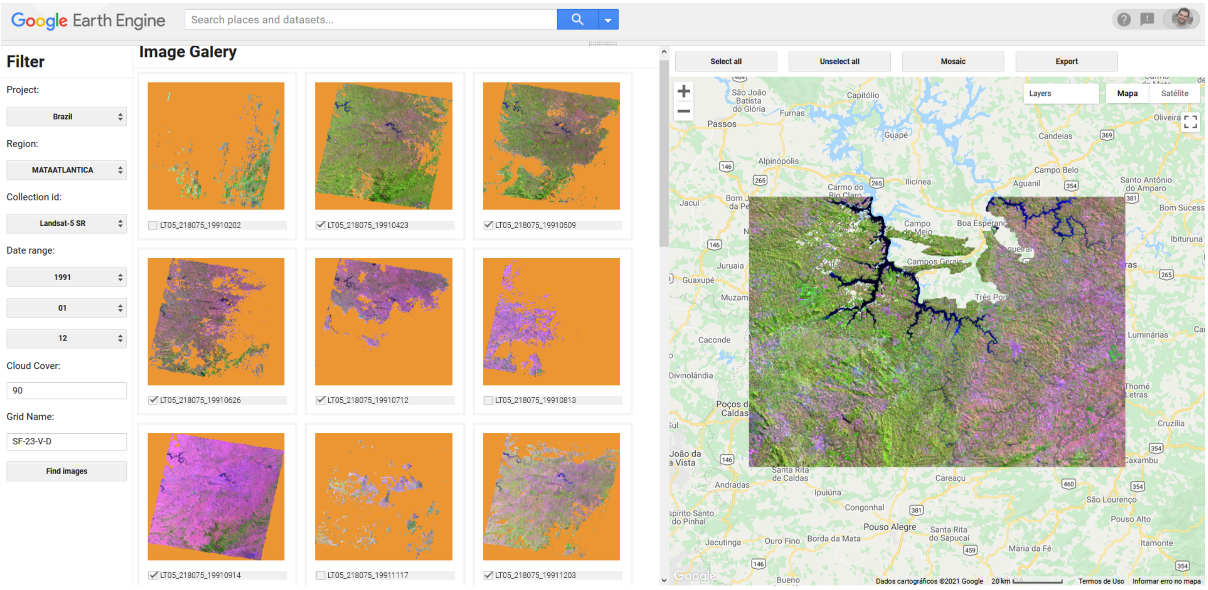


Figure 5. Google Earth Engine tool to manually identify and remove scenes with noise

2.2 Collection 10

For collection 10, the annual mosaics were generated from Google's monthly mosaics ("LANDSAT/COMPOSITES/C02/T1_L2_32DAY") which are already corrected. Due to these corrections and to the use of Deep Learning methods to recover data information, the mosaics have less gaps than the one used on previous collections. Due to less noise from cloud and shadow removal, it was possible to use all the monthly images to generate the annual mosaic.

The mosaic is generated using six Landsat original bands: blue, red, green, nir, swir1 and swir2. After, the Spectral Mixture Analysis is applied to extract fractions of soil, vegetation and shadow, using Landsat's specific endmembers. The SMA indexes as other spectral indexes are calculated and appended as bands in the mosaic. Auxiliary layers as slope and entropy are also added as bands. Through Landsat image segmentation function (ee.Algorithms.Image.Segmentation.SNIC), using median bands (blue green, red, nir, swir1, ndfi, and green texture) clusterized bands were generated and also integrated in the mosaic.

All these bands (184) will participate as candidates to the feature space used to classify the Atlantic Forest biome.

2.3. Final quality

As a result of the selection criteria, most mosaics presented satisfactory quality based on empirical knowledge about the biome.

The Atlantic Forest biome spans vast and climatically diverse regions of Brazil, where persistent cloud cover, topographic complexity, and seasonal variability pose major challenges for generating a consistent, high-quality satellite mosaic.The northeast of Brazil and some regions in Santa Catarina and São Paulo offer more difficulties to build clean mosaics, and the information still has some noise or missing data.

3. Definition of regions for classification

The classification was done in homogenous regions to reduce confusion of samples and classes, as well as to allow a better balance of samples and results. The Atlantic Forest biome was divided into 30 regions based in native vegetation types in the Atlantic Forest biome (IBGE, 2017) (Figure 6).

Due to great diversity of phytophysiognomies in Atlantic Forest, different characteristics are considered in mapping the forest formation within the biome, primarily related to canopy cover and height.

Forest Formation includes natural forest (exclude Forest Plantation) areas of more than 0.5 hectares (ha) with trees with a minimum height of 5 meters (m) and tree canopy cover that varies for each type of original forest formation:

- Dense Ombrophiles Forest - tree crown cover of more than 80%
- Mixed Ombrophiles Forest- tree crown cover of more than 80%
- Open Ombrophiles Forest - tree crown cover of more than 60%
- Seasonal Deciduous Forest- tree crown cover of more than 60%
- Seasonal Semideciduous Forest- tree crown cover of more than 60%

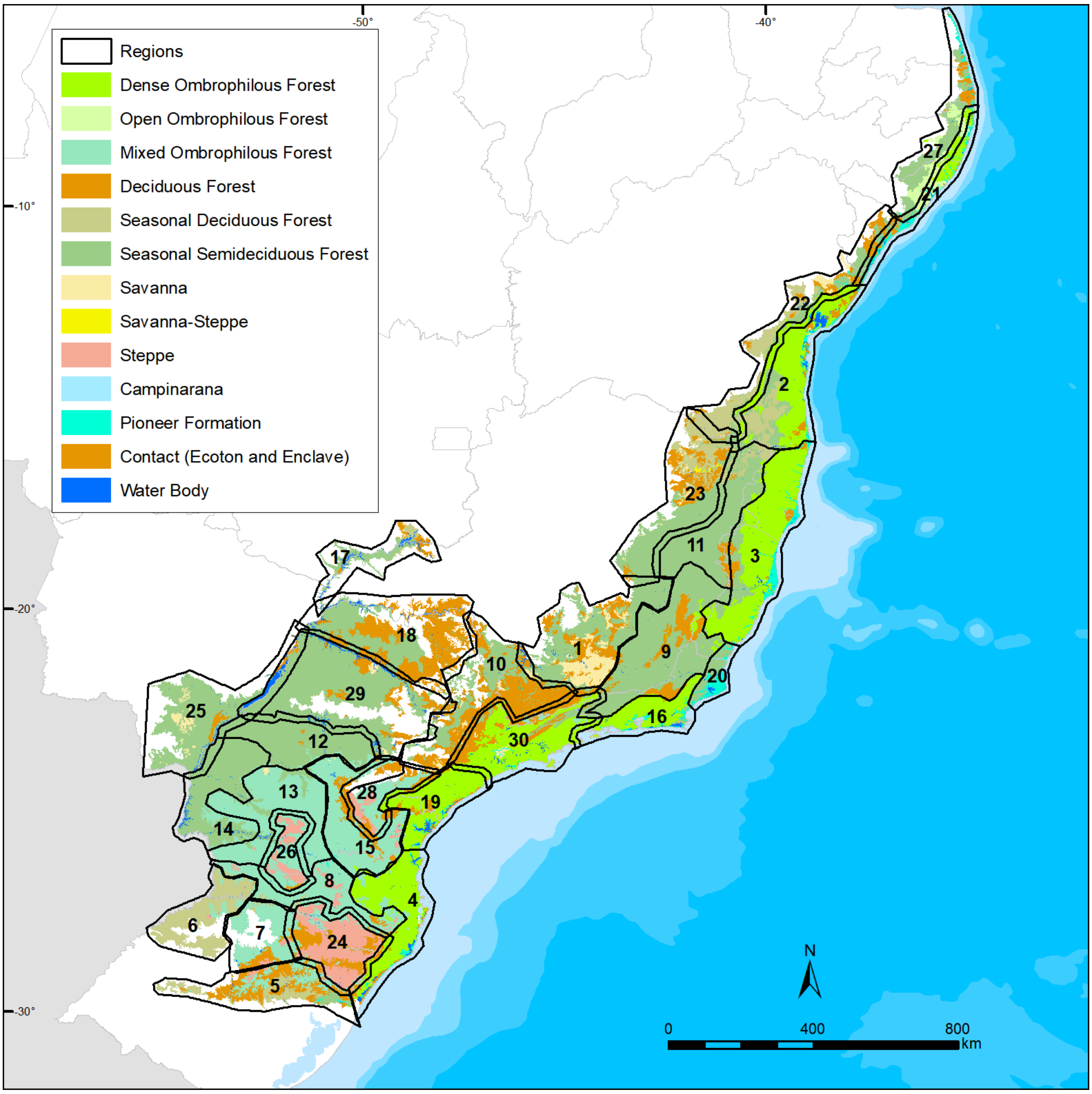


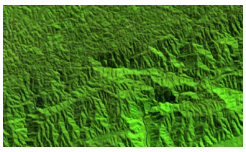
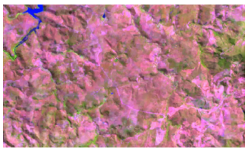
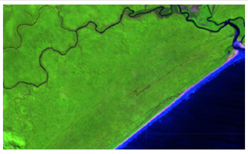
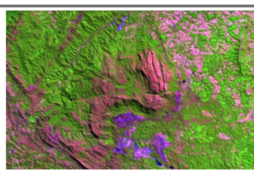
Figure 6. Regions used in the classification of Atlantic Forest biome. Each black polygon represents a classification region, while numbers within each polygon indicate the region ID.

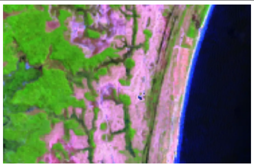
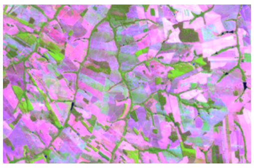
4. Classification

4.1. Classification scheme

The classification of the Landsat mosaics for the Atlantic Forest biome aimed to individualize a subset of 12 land use and land cover (Table 2), which were integrated with the cross-cutting themes in a further step.

Table 2. Land use land cover categories used for the Landsat mosaics classification for the Atlantic Forest biome in MapBiomas Collection 10.

Classes Level 1	Classes Level 2	ID	Color	RGB composite (SWIR1-NIR-Red)	Description
Forest	Forest Formation	3			Vegetation types characterized by the predominance of tree species, with high tree density, closed canopy, and vertical stratification. It includes forest typologies: Dense, Open and Mixed Ombrophilous Forest, Semi-deciduous and Deciduous Seasonal Forest, and Pioneer Formation.
	Savanna Formation	4			Vegetation type characterized by the presence of sparse tree and shrub species, with a semi-continuous canopy. It includes: Steppe, Forested and Wooded Savannah.
	Wooded Sandbank Vegetation	49			Forest formations on sandy soils in the coastal region
Herbaceous and Shrubby Vegetation	Wetland	11			Floodplain or grassland vegetation influenced by fluvial and/or lacustrine dynamics, characterized by the predominance of hygrophilous vegetation, including emergent, submerged, or floating aquatic plants.
	Grassland	12			Vegetation dominated by herbaceous species and grasses, with few scattered trees and shrubs, generally featuring an open or absent canopy. It occurs on soils ranging from deep to shallow, including rocky terrains (rupestrian grasslands). Included: Park and Grassland Steppe Savannas, Steppe and Shrub and Herbaceous Pioneers.
	Rocky Outcrop	29			Naturally exposed rocks without soil cover, often with the partial presence of rupicolous vegetation and high slope.

Classes Level 1	Classes Level 2	ID	Color	RGB composite (SWIR1-NIR-Red)	Description
	Herbaceous Sandbank Vegetation	50			Herbaceous vegetation that is established on sandy soils or on dunes in the coastal zone.
Farming	Other Temporary Crops*	19			Areas occupied with short or medium-term agricultural crops, generally with a vegetative cycle of less than one year, which after harvesting need to be planted again to produce.
	Forest Plantation*	9			Tree species planted for commercial purposes (e.g. pinus, eucalyptus, araucaria).
	Mosaic os Uses	21			Areas intended for agricultural and livestock use where it was not possible to distinguish between pastures and croplands, including fallow lands. These areas may also include peri-urban zones such as small farms, rural properties, and residential estates. Transitional areas are also included, where secondary vegetation is developing in abandoned pastures or in ecological restoration processes, prior to reaching forest size and structure.
Non Vegetated Area	Other non Vegetated Areas	25			Natural areas with exposed soil resulting from climatic events (landslides, flooding) and areas with non-permeable surfaces (infrastructure, urban expansion, or mining) not mapped within their respective classes.
Water	River, Lake and Ocean	33			Rivers, lakes, dams, reservoir and other water bodies

*Exceptionally, in regions 01, 03, 09, 10, 11, 16, 19, 20, 21, 27 and 30 we also included the class "3.2.1. Temporary Crop" (ID: 19) and in regions 01, 03, 08, 10, 13, 15, 23, 24, 28 and 30 we also included the class "3.3 Forest Plantations" (ID: 9). These two classes are shared with the agriculture team to pass through the specific filters and are then converted to 21 in the final Atlantic Forest dataset.

4.2 Feature space

The feature space used to classify the Atlantic Forest biome comprised a subset of 184 variables. They include the original Landsat reflectance bands, as well as vegetation indexes, spectral mixture modeling-derived variables and terrain morphometry (slope).

Since Collection 9, the Landsat image segmentation function (ee.Algorithms.Image.Segmentation.SNIC) was used with the following median bands: blue green, red, nir, swir1, ndfi, and green texture. This function generated a "clusters" band, "clusters_green_text" band and "clusters_ndfi_median" band for each year. For Collection 10, a “amp_ndfi_3anos” band was included as a fourth clusterized band. These bands were also included in the feature space.

A feature importance analysis was carried out for each one of the 30 Atlantic Forest regions. This analysis performs a classification with 500 trees for a specific year (set as 2024) and outputs a list showing the main bands by region. A set of the 60 most important variables were used in the classification.

To ensure that the selected set of variables was representative for each region, some regions were selected, and the importance analysis was repeated for the other years in the historical series. As a result, at least 85% of the variables remained consistent over the years, indicating that the analysis made for 2024 could be used throughout the entire series. In other words, there is a reliable set of variables per region that remains stable over the 40 years of the historical series.

Some conditions were considered while selecting the candidates variables to all regions to avoid noise, outliers, and data redundancy in the classification. These include: mandatory presence of the latitude, longitude, and slope variables in all regions and removal of all MAX and MIN variables.

The table below (Table 3) describes the 60 most commonly present variables across all regions. Among the 184 variables, only 21 were not included in any region by the feature importance analysis and 7 were included in all regions. This suggests that, besides some variables being very important for most regions, their individual characteristics and spectral behavior were responsible for assigning different feature importance to each of them.

Table 3. Feature space subset with the most frequent variables between regions considered in the classification of the Atlantic Forest biome Landsat image mosaics in the MapBiomas Collection 10 (1985-2024).

Type	Name	Formula	Statistics	Reference
Landsat band	Blue	Band 1 (L5 and L7) Band 2 (L8)	median, median_dry, median_wet	USGS
	Green	Band 2 (L5 and L7) Band 3 (L8)	median, median_dry, median_wet, median_texture	USGS
	Red	Band 3 (L5 and L7) Band 4 (L8)	median, median_dry,	USGS

Type	Name	Formula	Statistics	Referenc e
			median_wet	
	NIR	Band 4 (L5 and L7) Band 5 (L8)	median, median_dry, median_wet	USGS
	SWIR 1	Band 5 (L5 and L7) Band 6 (L8)	median, median_dry, median_wet	USGS
	SWIR 2	Band 7 (L5 and L7) Band 8 (L8)	median, median_dry, median_wet	USGS
Spectral Index	Cellulose Absorption Index	CAI = SWIR2 / SWIR1	median, median_dry, median_wet	Nagler et al. 2003
	Enhanced Vegetation Index 2	EVI 2 = 2.5 × (NIR - Red) / (NIR + 2.4 × Red + 1)	median, median_dry, median_wet, stdDev	Parente et al., 2018
	Green Chlorophyll Vegetation Index	GCVI = (NIR / Green - 1)	amplitude, median, median_dry, median_wet, stdDev	Burke et al., 2017
	Normalized Difference Vegetation Index	NDVI = (NIR - Red) / (NIR + Red)	amplitude, median, median_dry, median_wet	Rouse et al., 1974
	Normalized Difference Water Index	NDWI = (NIR - SWIR1) / (NIR + SWIR1)	amplitude, median, median_dry, median_wet, stdDev	Gao et a., 1996
	Normalized Difference Fraction Index	NDFI = (GV − (NPV + SOIL)) / (GV + NPV + SOIL)	stdDev	USGS
	Soil-Adjusted Vegetation Index	SAVI = 1.5 × (NIR - Red) / (NIR + Red + 0.5)	amplitude, median, median_dry, median_wet, stdDev	Huete, 1988
Surface Index	Hall's Forest Cover	HFC = −0.017 × RED − 0.007 × NIR − 0.079 × SWIR2 + 5.22	median, median_dry, median_wet	?
	Hall's Forest Height	HFH = −0.039 × RED − 0.011 × NIR − 0.026 × SWIR1 + 4.13	median, median_dry, median_wet	

Type	Name	Formula	Statistics	Reference
Spectral Mixture Analysis	Green Vegetation	GV = SMA(GV) × 100	stdDev	Fraction from SMA
	Soil	SOIL = SMA(Soil) × 100	stdDev	Fraction from SMA
	Forest/Non-Forest Index	FNS = 100 × (GVshade − SOIL) / (GVshade + SOIL) + 100 GVshade = GV + |GV + NPV + SOIL − 100|	stdDev	Adapted from NDFI
	Scaled Water-Enhanced Forest Index	WEFI = ((GV + NPV) − (SOIL + SHADE)) / ((GV + NPV) + (SOIL + SHADE)) × 100 + 100	Median, stdDev	Fraction from SMA
Terrain	Slope	ALOS DSM: Global 30 m	identity	Tadono et al., 2014
Coords	Latitude and Longitude		Latitude, Longitude	

4.3. Classification algorithm, training samples and parameters

The classification was performed region by region, year by year, using a *Random Forest* algorithm (Breiman, 2001) available in Google Earth Engine, running 100 iterations (random forest trees). Training samples for each region were defined following a strategy of using pixels for which the land cover and land use remained the same over the 39 years of Collection 9, so named “stable samples”. An ensemble was made from three main sources: extracted from Collection 9; manually drawn polygons; and complementary samples.

4.3.1. Stable samples from collection 9

The extraction of stable training samples from the previous Collection 9 followed several steps to ensure their confidence for use.

The first step defines these stable samples by selecting areas where no changes were detected between 1985 and 2024.

We have identified each region’s predominant, secondary, and rare classes. The areas that did not change class from 1985 to 2023 in collection 9 were used to generate random training points balanced with the rule:

- 3000 or 4000 training samples to predominant class
- 1000 or 2000 training samples to secondary class
- 300 or 500 training samples to rare class

After exporting these samples, reference maps from the states of São Paulo, Minas Gerais, Paraná, and Espírito Santo were used to filter stable samples located in natural areas. All reference links are available on the MapBiomas Brazil website:

https://brasil.mapbiomas.org/en/mapas-de-referencia/

The stables samples from forest and grassland were also filtered using data from Global Forest Canopy Height (GFCH), 2019 (Potapov, 2019) based on GEDI data using the following rules:

- Forest sample need to be >= 9m canopy height
- Grassland sample need to be < 7m canopy height

In Collection 10, for the first time, data from MapBiomas Alertas were also used as a filter for stable samples. In this case, all stable samples that intersected any deforestation alert published by MapBiomas Alert were removed from the set of stable samples used in the classification process.

The number of samples of each class was defined for each region based on the visual and accuracy analysis of the Collection 9 classification, and it is available in the GitHub script "02-20_Exporta_Amostras_Estratificadas_Regioes_v3".

4.3.2 Classification Process of *Herbaceous Sandbank Vegetation* Class

The *Herbaceous Sandbank Vegetation* class, since Collection 7, when it began to be mapped, was classified only during the post-classification process. At that time, only areas originally classified as "Other Non-Forest Natural Formation" (ID13) and located within predefined regions based on the IBGE Soil Map (ESPODOSOL and NEOSOL) were converted to class 50 (Herbaceous Sandbank Vegetation).

Starting with Collection 10, this class began to follow the full classification workflow applied to other natural classes in the biome — from the definition of stable samples to the application of post-classification filters. Nevertheless, the IBGE Soil Map continued to be used as a spatial constraint after the classification to prevent this class from appearing in regions where it is not naturally present.

4.3.3 Classification Process of the *Rock Outcrop* Class

Another improvement in Collection 10 relates to the classification of the *Rock Outcrop* class. The goal of this change was to better represent this class within the biome, avoiding commission errors in grassland formation areas.

An analysis of previous collections showed that *Rock Outcrop* areas were overestimated, including not only regions with exposed rock but also areas of grassland formations (Figure 7). For this collection, we chose to restrict the classification strictly to zones of rock exposure. To achieve this, all areas initially classified as *Rock Outcrop* (ID29) were converted to *Grassland Formation* (ID12) starting from the stable sample extraction phase.

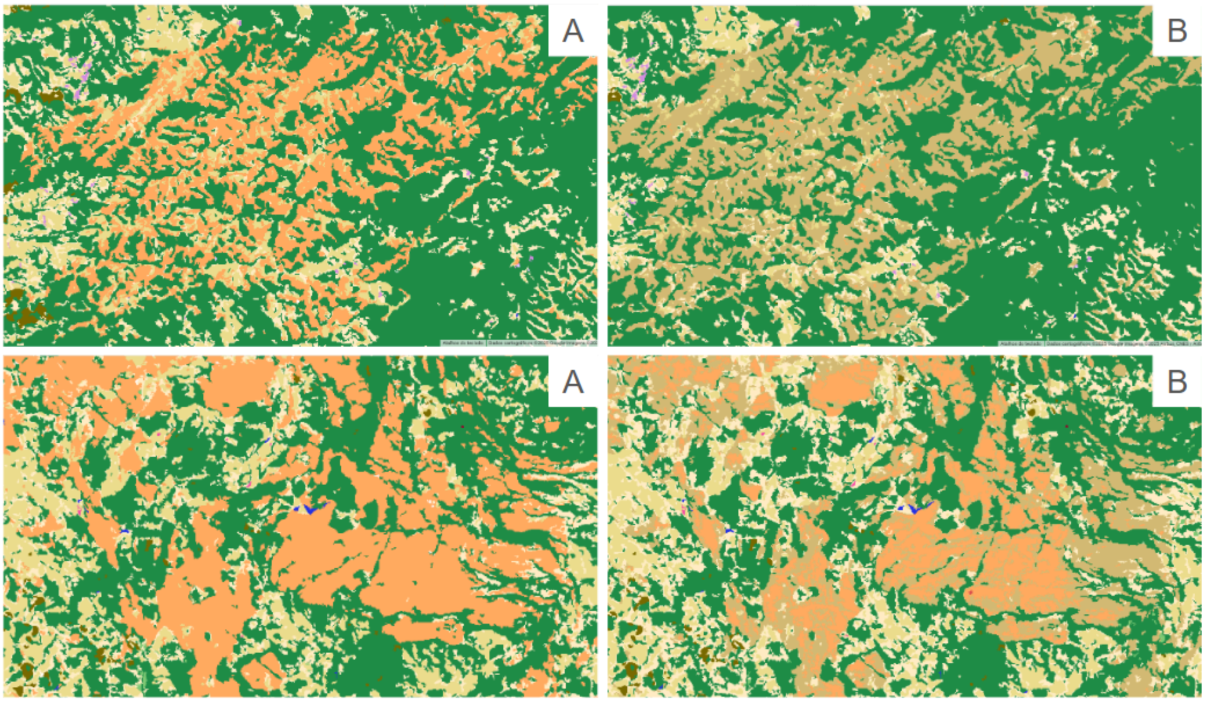
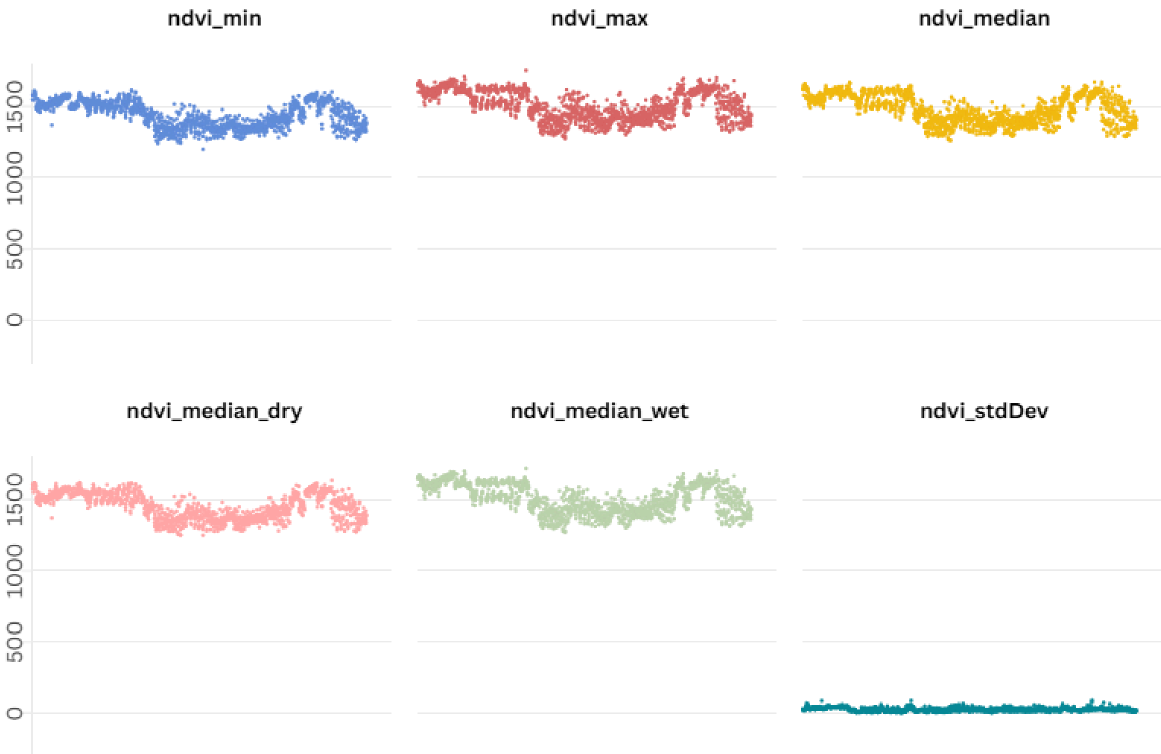


Figure 7. Differences between Rocky outcrop classification process in collection 9 (A) and collection 10 (B).

The *Rock Outcrop* class began to be identified during the post-classification stage, using band intervals related to NDVI that differentiate class 29 (*Rock Outcrop*) from class 12 (*Grassland Formation*). These intervals were established based on spectral analysis of *Rock Outcrop* sample areas (Figure 8).



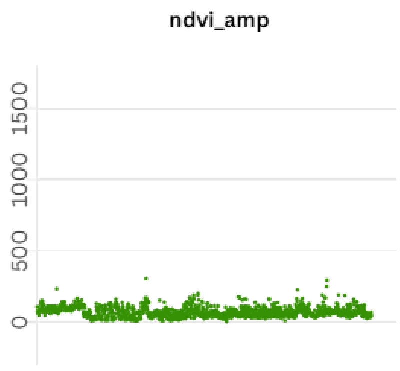


Figure 8. Spectral analysis of 7 bands of *Rock Outcrop* sample areas.

The NDVI derived bands values are varying between 0 and 2000, the commonly used range of NDVI (-1 ; 1) was rescaled to work with integer numbers. The most representative bands, along with their corresponding value ranges used in the classification, are listed in Table 4:

Table 4. NDVI-based variables and value ranges used to differentiate Rock Outcrop (ID29) from Grassland Formation (ID12).

Description	**Variable**	**Range**
Amplitude	'ndvi_amp'	20-200
Standard Deviation	'ndvi_stdDev'	0-50
Median	'ndvi_median'	1200-1650
Median during dry season	'ndvi_median_dry'	1200-1650
Median during wet season	'ndvi_median_wet'	1200-1650
Minimum	'ndvi_min'	1200-1400
Maximum	'ndvi_max'	1300-1500

A mask was created using the pixels that met the NDVI value ranges described in Table 4. This mask was applied exclusively to areas previously classified as *Grassland Formation*. In other words, all *Grassland Formation* areas falling within the NDVI-based mask were reclassified as *Rock Outcrop*.

4.3.4. Multi Probability

Collection 8 was produced with 10 different classifications for each region and each year. Each classification used a different seed to create training samples, which affects the location of the pixel within the stable classes. The value of the seed, with positive or negative values, was also used to change the balance of the main and secondary classes in each region, according to the code below:

```
var lista_seeds = [1, 5, 10, 25, -10, -25, -35, -50, -75, -100]
var n_pr2 = 4000 + (seed * 5)
var n_pr1 = 3000 + (seed * 4)
var n_se1 = 2000 + (seed * 3)
var n_se2 = 1000 + (seed * 2)
```

The final class of each pixel in each year was defined by the MODE value. The number of times the pixel was classified in the final class will be analyzed to estimate the degree of reliability.

This approach was abandoned and replaced, since collection 9, by the use of Multi Probability. This function of RF evaluates the probability that a pixel belongs to each of the possible classes and assigns the class with the highest probability in the final classification, by region for each year in the series. Additionally, the method records the highest probability value for the pixel among all the classes present in the classification.

4.3.5. Complementary samples

The need for complementary samples was evaluated by visual inspection and by comparing the output of the preliminary accuracy of each region. Complementary sample collection was also done drawing polygons using Google Earth Engine Code Editor. The same concept of stable samples was applied, checking the false-color composites of the Landsat mosaics for all the 40 years during the polygon drawing. Based on the expert knowledge about each region, polygon samples from each class were collected, and the number of random points in these polygons was defined to balance the samples.

4.3.6. Final classification

Final classification was performed for all regions and years with stable and complementary samples. All years used the same subset of samples, which was trained in the same mosaic as the year that was classified.

At this stage, the classifications of the 30 regions within the biome are merged. The boundaries between regions are overlapped to prevent gaps in classification along their edges. The priority order for merging the regions is defined visually, based on the classifications from previous collections combined with the classification generated for the current collection.

5. Post-classification

A list of post-classification spatial and temporal filters was applied, since the pixel-based classification method and the extended temporal series implies some unwanted effects on the mosaics, like gaps in the data and spectral changes among the years coming either from weather variations or from changes in intra-annual seasonality. The post-classification process includes the application of gap-fill, temporal, spatial, and frequency filters to refine the results. The temporal filter rules were adapted for the land cover and land use classes used in the Atlantic Forest biome and were complemented by specific rules to adjust for cases where a pixel appeared.

5.1. Temporal Gap Fill filter

In this filter, no-data values (“gaps”) are replaced by the temporally nearest valid classification. In this procedure, if no “future” valid position is available, then the no-data value is replaced by its previous valid class. Therefore, gaps should only exist if a given pixel has been permanently classified as no-data throughout the entire temporal domain.

5.2. Spatial filter

The spatial filter avoids unwanted modifications to the edges of the pixel groups (blobs), a spatial filter was built based on the "connectedPixelCount" function. Native to the GEE platform, this function locates connected components (neighbors) that share the same pixel value. Thus, only pixels that do not share connections to a predefined number of identical neighbors are considered isolated. In this filter, at least six connected pixels are needed to reach the minimum connection value. Consequently, the minimum mapping unit is directly affected by the spatial filter applied, and it was defined as 6 pixels (~0,5 ha).

To avoid the exclusion of small forested areas along rivers, a mask was created using HAND data so this filter does not affect riparian forest regions.

5.3. Temporal filter

The temporal filter uses the subsequent years to replace pixels that have invalid transitions.

The first process looks into a 3-year moving window to correct any value that is changed in the middle year and return to the same class as in next year. This process is applied in this order: [11, 12, 21, 4, 3, 29, 50, 22].

The second process is similar to the first process, but it is a 4- and 5-years moving window that corrects all middle years. And follows the sequence [4, 11, 12, 3, 29, 50, 21, 22].

In the third process the filter looks at any native vegetation class (3, 4, 12, 29, 50) that is not this class in 85 and is equal in 86 and 87 and then corrects 85 values to avoid any regeneration in the first year.

In the last process the filter looks for pixels value in 2024 that is not 21 (Mosaic of Agriculture and Pasture) and is equal to 21 in 2022 and 2023. The value in 2024 is then converted to 21 to avoid any regeneration in the last year.

5.4. Frequency filter

Frequency filters were applied only to natural classes (3, 4, 11, 12, 29, 50). It identifies the most common natural class assigned to that pixel throughout the years and applies that class to all the years where that pixel was classified as another natural class .

The result of these frequency filters is a classification with more stable classification between native classes (e.g. Forest and Savanna). Another important result is the removal of noises in the first and last year in the classification.

5.4. Wetland filter

We used the 'Height Above Nearest Drainage' product (HAND) as a proxy to represent the 'groundwater depth' and assumed the premise that if a pixel classified as wetland (ID=11) had a HAND value greater than 15 meters, this pixel was converted to Mosaic of Agriculture or Pasture (ID=21).

5.5. Herbaceous Sandbank Vegetation filter

Considering the geographic restriction of this class, a filter was applied using the IBGE Soil Map as a reference. This process aims to avoid commission errors in areas where it does not naturally occur, restricting its classification only to regions with the presence of Spodosols and Neosols.

5.6. Incident filter

An incident filter was applied in collection 6 and 7 and was abandoned in collection 8. This filter was used to remove pixels that change too many times in the 36 and 37 years. All pixels that change more than 6 times are replaced with Savanna (ID=4) or Mosaic of Agriculture or Pasture (ID+21) according to the mode value. This avoids changes in the border of the classes. It has not been used since collection 8.

5.7. Transition filter

The transition filter operates with both spatial and temporal corrections. Yearly deforestation or natural classes recovery in small-areas (< 0.5 ha) are treated as likely misclassification while abrupt changes at the beginning and end of the time series or isolated rare classes in intermediate years are removed. This means that some small and temporary changes in natural classes classification will not be wrongly considered as regrowth of secondary vegetation nor as deforestation.

5.8. Classification of Wooded Sandbank Vegetation

Wooded sandbank vegetation was mapped resulting from the post-classification. The ALOS DSM: Global 30 m was used to identify coastal forest with less than 25m altitude and it was converted to this class using a spatial mask to exclude some regions in northeast of Brazil.

6. Integration with cross-cutting themes

The final classification result, including the application of post-classification filters generated by the Atlantic Forest team, is integrated with the data produced by the cross-cutting themes for all years in the historical series (1985–2024). This step is carried out based on predefined prevalence rules (Table 5). The final integrated map for the Atlantic Forest biome consists of 27 classes at level 4 of the legend (Figure 9).

A new class was added in Collection 10: Photovoltaic power plant. This class represents large-scale areas with interconnected photovoltaic panels, designed to generate electricity from sunlight and integrated into the main power grid.

Table 5. General prevalence rules - MapBiomas Collection 10

Class	ID	Prevalence order	Color
Photovoltaic power plant	75	1	
Mining	30	2	
Beach, Dune and Sand Spot	23	3	
Mangrove	5	4	
Aquaculture	31	5	
Hypersaline Tidal Flat	32	6	
Urban Infrastructure	24	7	
Forest Plantation	9	8	
Rocky Outcrop	29	9	
Sugar Cane	20	10	
Soybean	39	11	
Rice	40	12	
Cotton	62	13	
Other Temporary Crops	41	14	
Coffee	46	15	
Citrus	47	16	
Other Perennial Crops	48	17	
Herbaceous Sandbank Vegetation	50	18	
Other Non Vegetated Areas	25	19	
River, Lake and Ocean	33	20	
Forest Formation	3	21	
Savanna Formation	4	22	
Wooded Sandbank Vegetation	49	23	
Wetland	11	24	
Grassland Formation	12	25	

Class	ID	Prevalence order	Color
Pasture	15	26	
Mosaic of Uses	21	27	

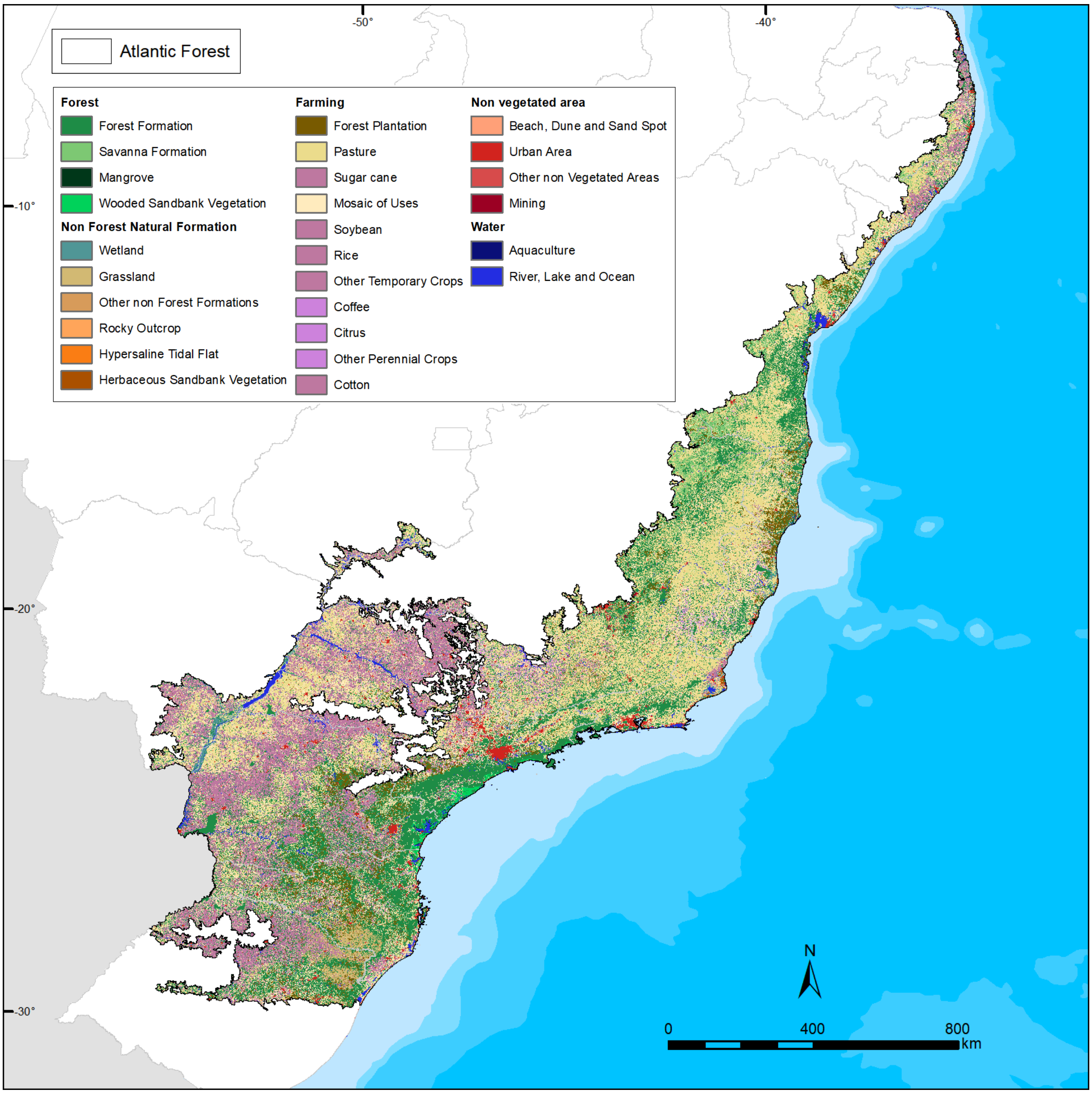


Figure 9. Final land use and land cover map of the Atlantic Forest biome - 2024

One of the important post-integration filters for the Atlantic Forest concerns the application of a mask using deforestation alerts published on the MapBiomas Alerta platform. In this filter, corrections are made only in areas classified as natural classes in the land use and cover data that have been identified as deforested by MapBiomas Alerta. As a result, the natural class is replaced by class 21 (Mosaic of Uses).

As an exception, in cases of alerts with the pressure vectors “extreme weather events” and “mining,” the land use and cover data are replaced by class 25 (Other Non-Vegetated Areas), which is more representative of these situations. In these cases, not only natural classes but also class 21, as mapped in the land use and cover data, are changed to class 25.

6. Validation strategies

The set of 10.840 independent validation points provided by Lapig (Laboratório de Processamento de Imagens e Geoprocessamento - UFG) was used to perform accuracy analysis (Figure 10). For collection 10, some of the validation points were revised. This revision, along with the improvements in the classification of the Collection 10, has altered the global accuracy for all the previous collections. The values were updated, in level 1 and 2.

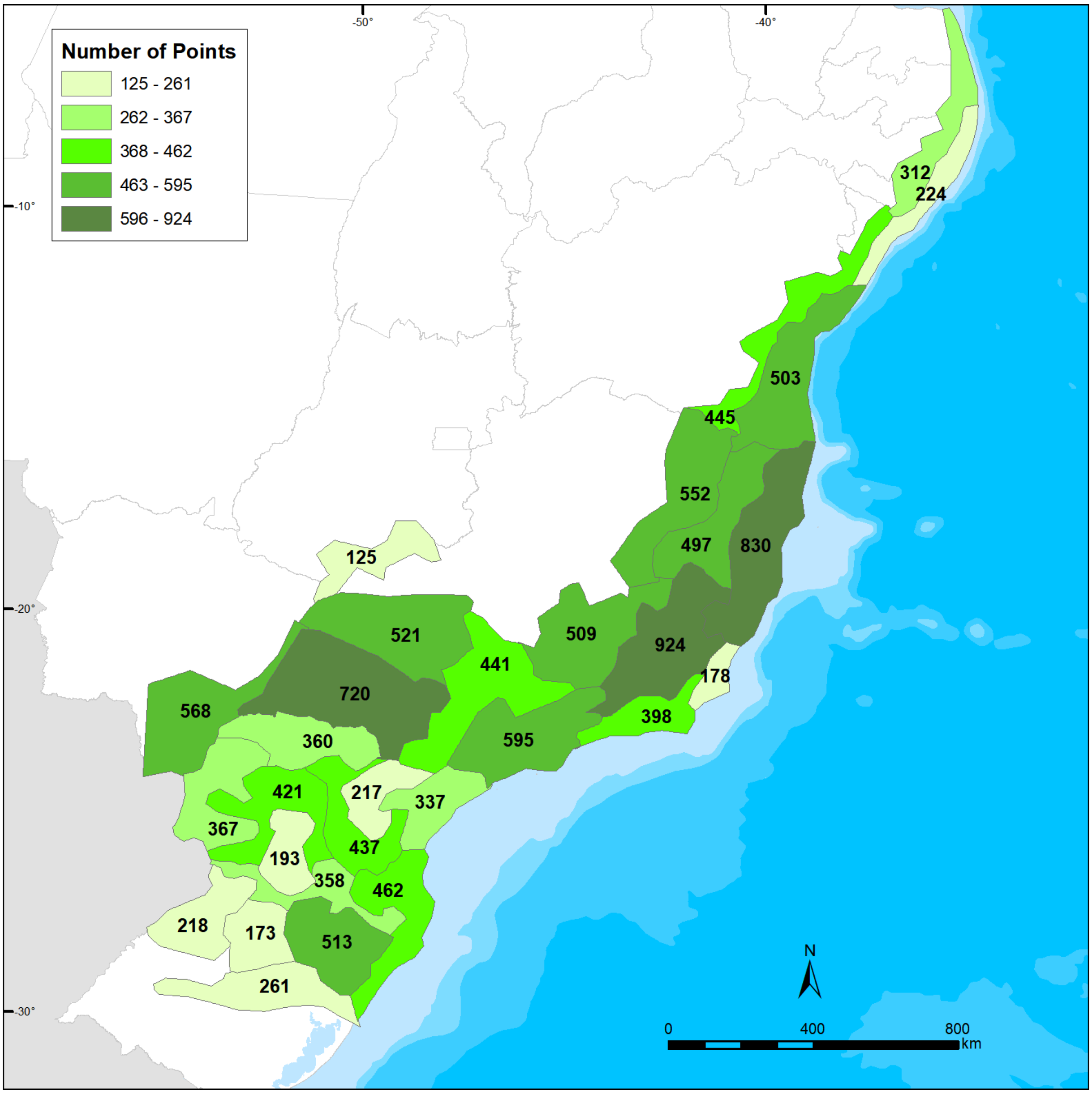


Figure 10. Accuracy points per region in Atlantic Forest biome.

The result of accuracy is presented in MapBiomas Website.

https://brasil.mapbiomas.org/en/estatistica-de-acuracia/colecao-9/

Global accuracy (considering all years) was 91.4% and 87.4% in levels 1 and 2 for collection 5 and collection 6 have about the same values, 91.9% and 87.3% in levels 1 and 2, respectively. The difference is explained by the reclassification of “Forest Plantation” from “1. Forest > 1.2 Forest Plantation” to “3. Farming > 3.3 Forest Plantation”.

In collection 7 the Global accuracy was 91.9% and 87.1% in levels 1 and 2, respectively. In collection 8 the Global accuracy is 92.3% and 87.2% in levels 1 and 2, respectively. Collection 9 shows a Global accuracy of 91.7,% in level 1 and 86.6% in level 2 .

The accuracy values for Collection 10 are similar to those observed for Collection 9. For the most recent collection, Global accuracy was 91.5% at level 1 and 86.1% at level 2. This small difference can be explained by adjustments made in the classification regarding the savanna formation and grassland formation classes, which are the classes with the highest omission and commission errors.

The detailed information about Global accuracy presented in Figure 11 for level 1.

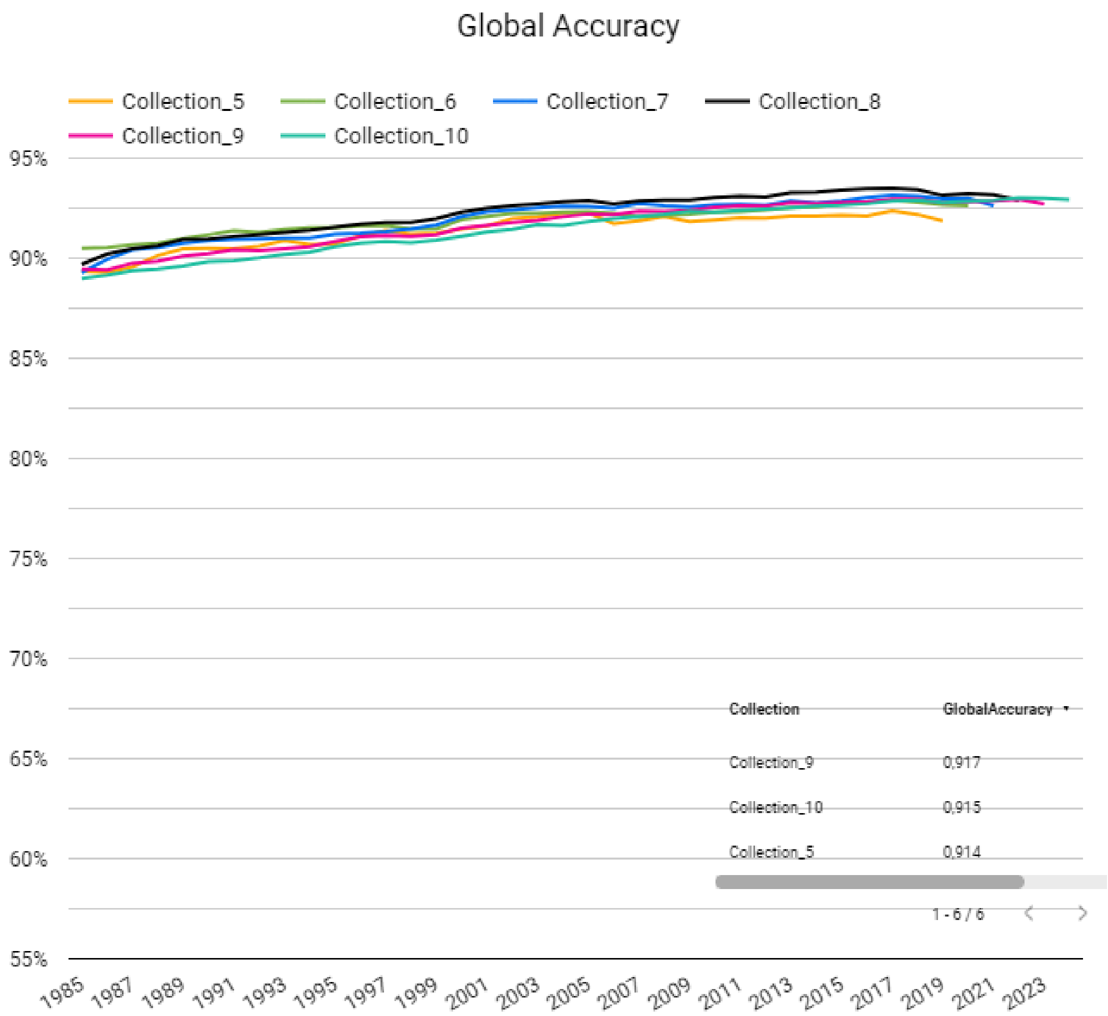


Figure 11. Global accuracy for Atlantic Forest biome at legend level 1

8. References

Breiman, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

IBGE. Vegetação RADAM. Disponível em: <ftp://geoftp.ibge.gov.br/informacoes_ambientais/acervo_radambrasil/vetores/>. Accessed in: may, 30 2018.

Funk, C.; Peterson, P.; Landsfeld, M. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. **Scientific Data 2**, [S. l.], v. 150066, 2015. DOI: 10.1038/sdata.2015.66.

P. Potapov, X. Li, A. Hernandez-Serna, A. Tyukavina, M.C. Hansen, A. Kommareddy, A. Pickens, S. Turubanova, H. Tang, C.E. Silva, J. Armston, R. Dubayah, J. B. Blair, M. Hofton (2020) Mapping and monitoring global forest canopy height through integration of GEDI and

Rosa, M. R. (2020). Metodologia de classificação de uso e cobertura da terra para análise de três décadas de ganho e perda anual da cobertura florestal nativa na Mata Atlântica (Doctoral Dissertation, Universidade de São Paulo).

Landsat data. Remote Sensing of Environment, 112165. https://doi.org/10.1016/j.rse.2020.112165