Atlantic Forest - Appendix

Collection 9.0

Version 1

General coordinator

Natalia Crusco

Luis Guedes Pinto

Team

Eduardo Reis Rosa

Fernando Frizeira Paternost

Jacqueline Freitas

Marcos Reis Rosa

Mariana Dias Ramos

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.

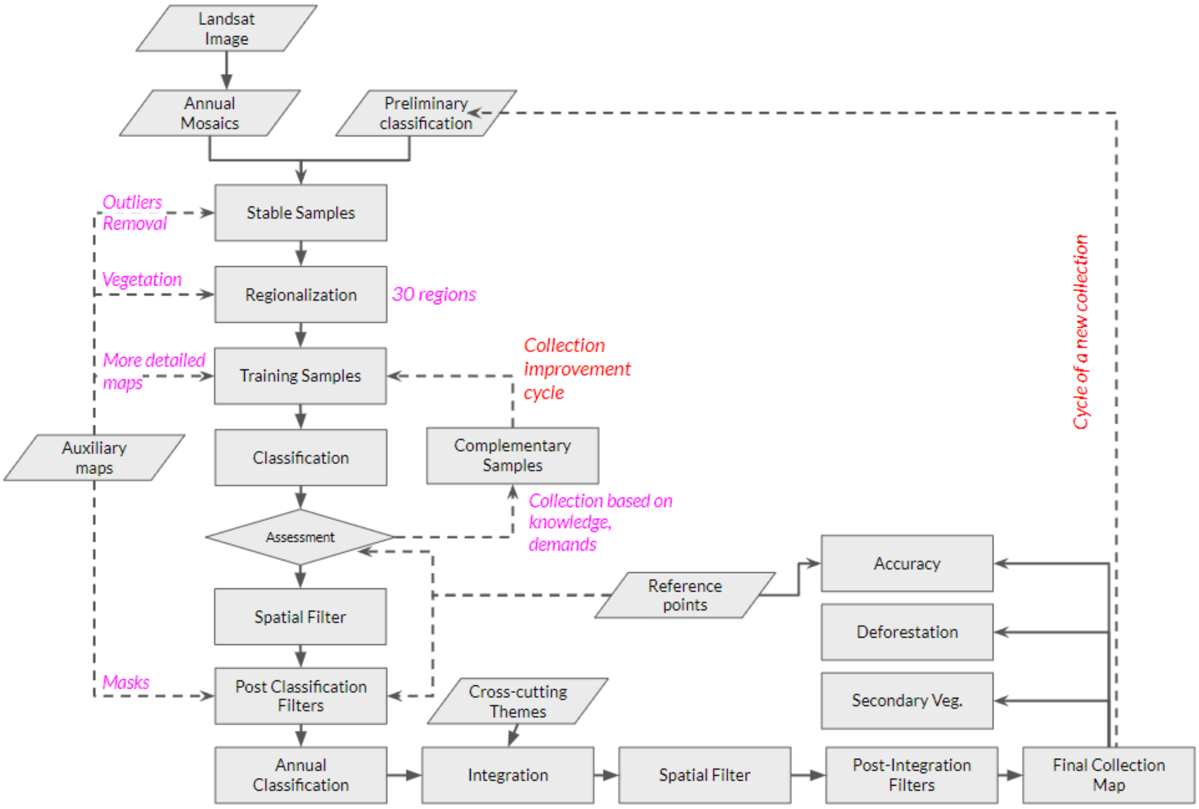
Table 1. The evolution of the Atlantic Forest mapping collections in the MapBiomas Project, its periods, level and number of classes, brief methodological description, and global accuracy in Levels 1 and 2.

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: 90.7% Level 3: 86.6% *
6.0	36 years 1985-2020	4 / 25	Random Forest	Level 1: 90.6% Level 2: 85.5%
7.0	37 years 1985-2021	4 / 27	Random Forest	Level 1: 90.1% Level 2: 84.4%
8.0	38 years 1985-2022	4 / 28	Random Forest	Level 1: 90.5% Level 2: 85,0%
9.0	39 years 1985-2023	4 / 29	Random Forest	Level 1: 90.5% Level 2: 85.4%

** 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 9 in the Atlantic Forest biome. Inclined gray rectangles represents databases, while linear rectangles point key-steps in the workflow. Solid arrows indicate main flow, and dashed arrows points assessment and evaluation cycles. Pink and red text shows additional information for specific steps.



2. Landsat image mosaics

2.1. 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. Native vegetation types in the Atlantic Forest biome (IBGE, 2017).

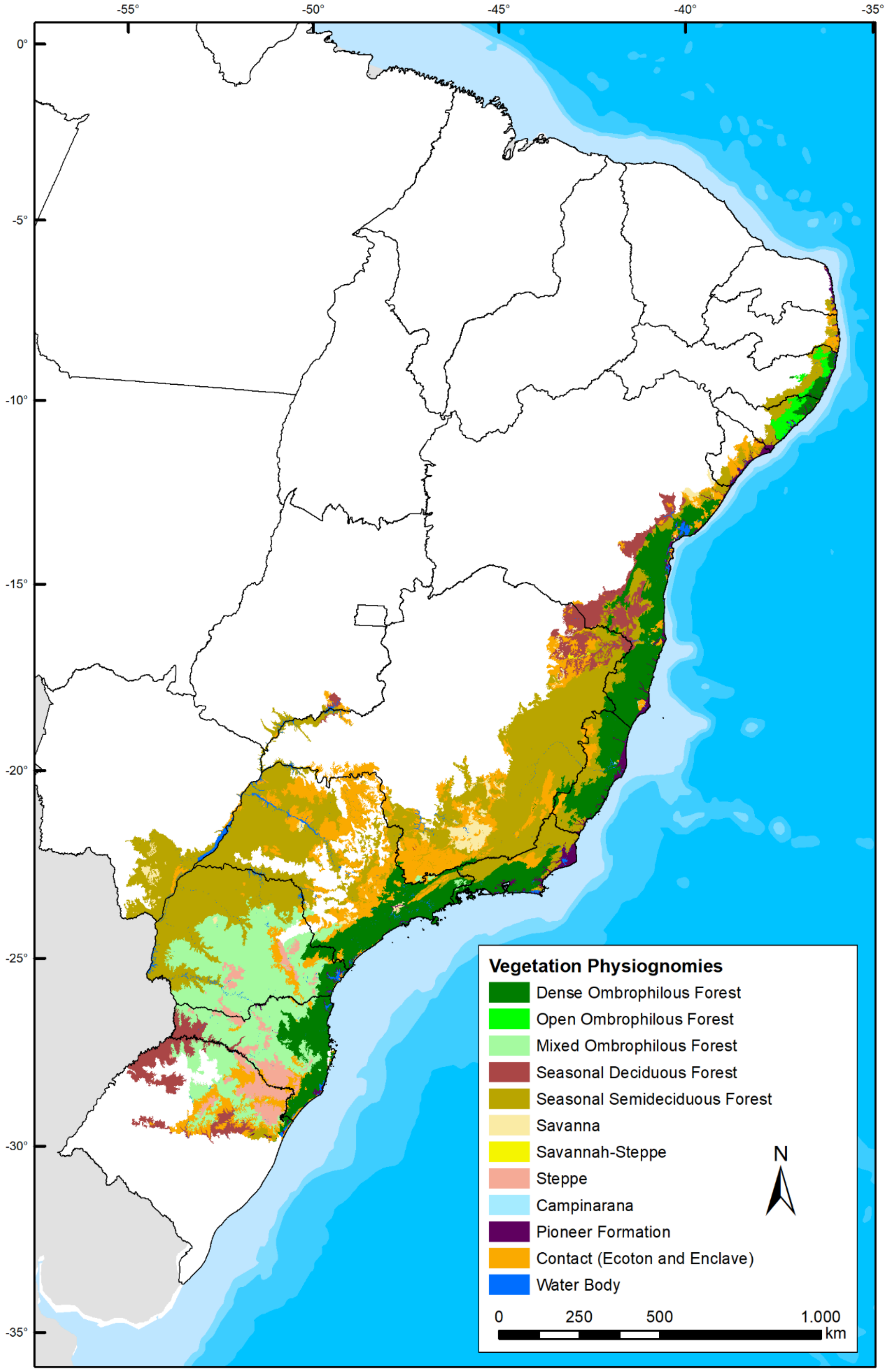
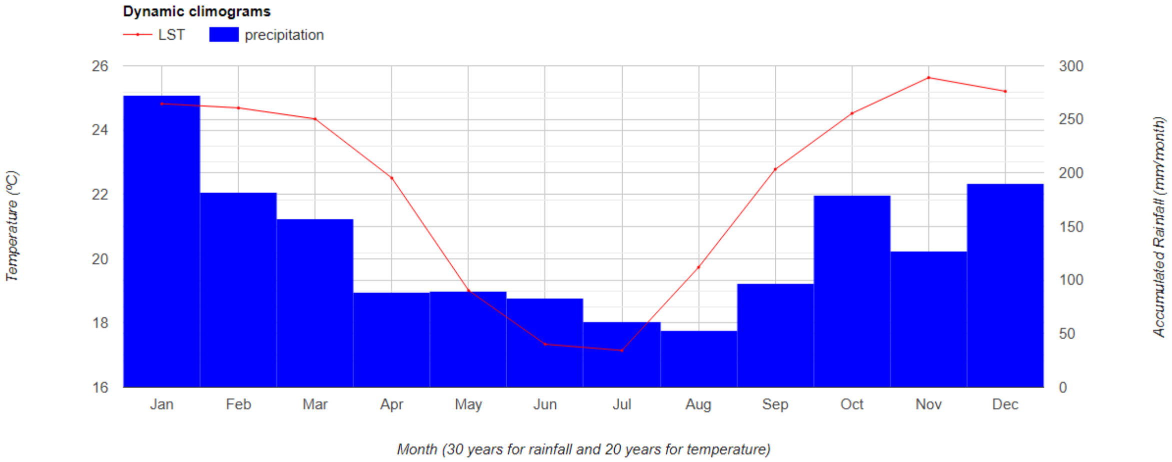


Figure 3. Climograph from 1988 to 2018 with CHIRPS (Climate Hazards Group InfraRed Precipitation with Station) precipitation data and MODIS (Moderate-Resolution Imaging Spectroradiometer) temperature data. (FUNK; PETERSON; LANDSFELD, 2015).

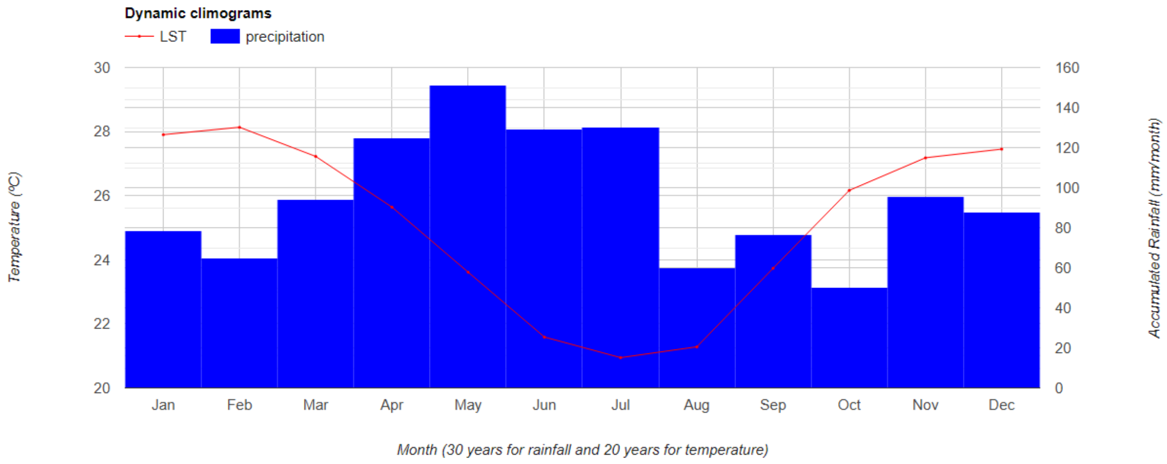


2.2. Image selection

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 and 6).

Figure 5. Landsat scenes that need to be reviewed in each year

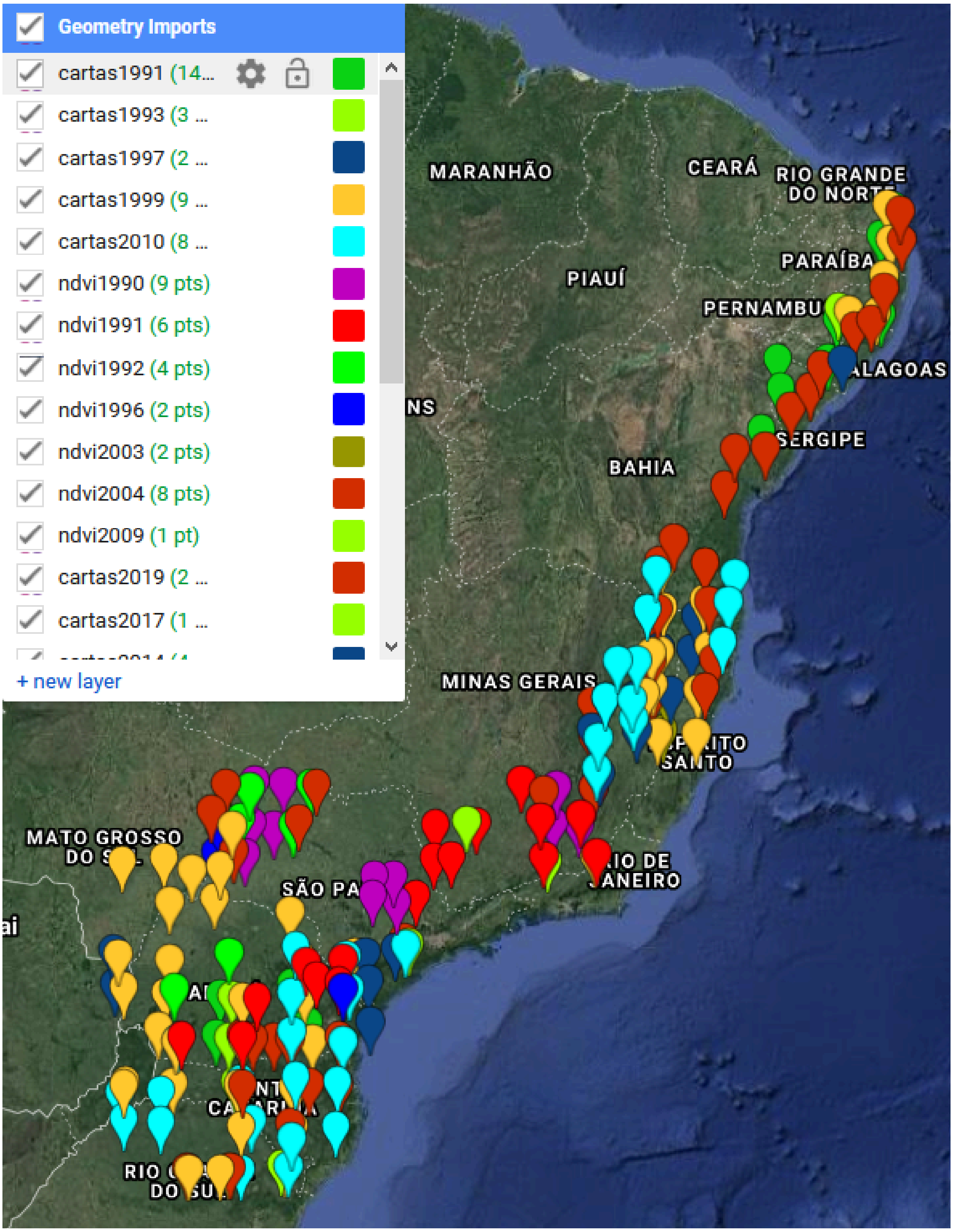
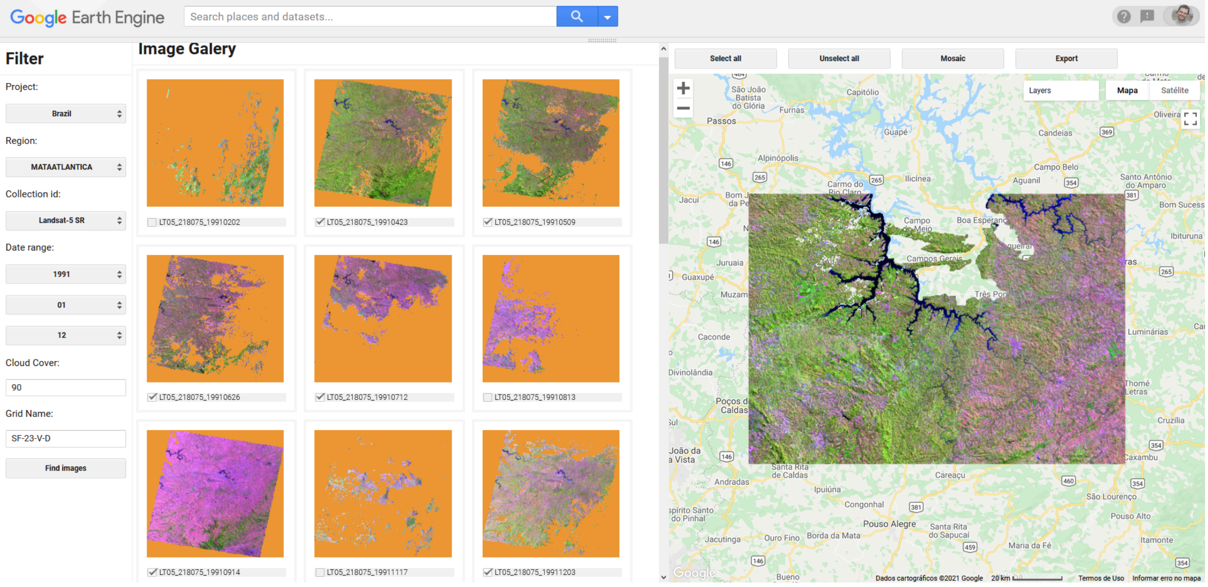


Figure 6. Google Earth Engine tool to manually identify and remove scenes with noise



Filter

Project: Brazil

Region: MATAATLANTICA

Collection id: Landsat-5 SR

Date range: 1991 01 12

Cloud Cover: 90

Grid Name: SF-23-V-D

□ LT05_218075_19910202
☑ LT05_218075_19910423
☑ LT05_218075_19910509
☑ LT05_218075_19910626
☑ LT05_218075_19910712
□ LT05_218075_19910813
☑ LT05_218075_19910914
□ LT05_218075_19911117
☑ LT05_218075_19911203

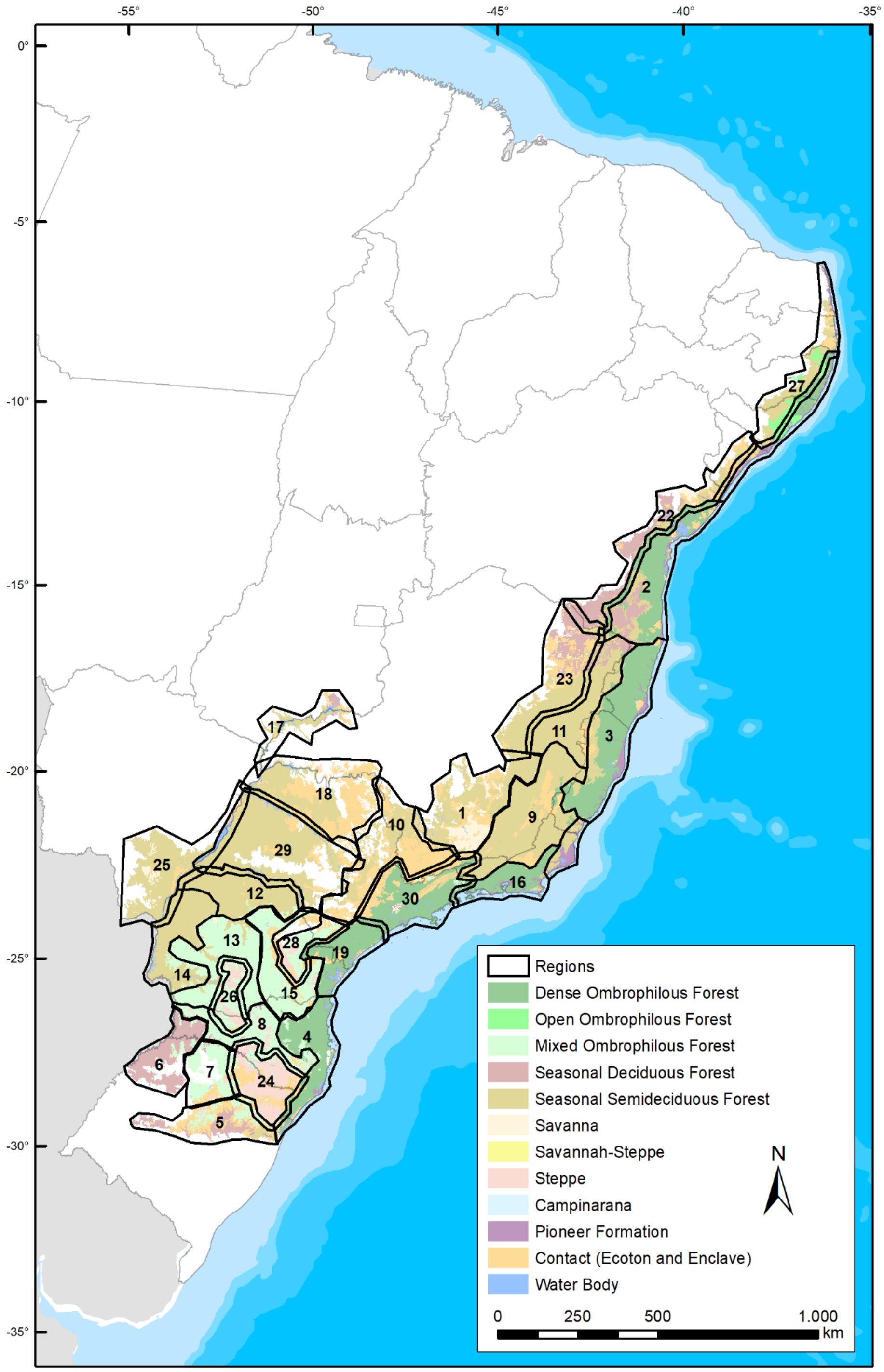
2.3. Final quality

As a result of the selection criteria, most mosaics presented satisfactory quality based on empirical knowledge about the biome. The northeast of Brazil and some regions in Santa Catarina and São Paulo offer more challenges to building 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 in 30 regions based in native vegetation types in the Atlantic Forest biome (IBGE, 2017) (Figure 7):

Figure 7. 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 cover and land use (Table 2), which were integrated with the cross-cutting themes in a further step.

Table 2. Land cover and land use classes considered for classification of Landsat mosaics for the Atlantic Forest biome in the MapBiomas Collection 9.

Legend class (level 2) of Collection 9	Numeric ID	Color
1.1. Forest Formation	3	🟩
1.2. Savanna Formation	4	🟩
1.5. Wooded Sandbank Vegetation	49	🟩
2.1. Wetland	11	⬜
2.2. Grassland	12	🟨
2.4. Rocky Outcrop	29	🟧
2.5. Herbaceous Sandbank Vegetation	50	🟫
3.2.1.5 Other Temporary Crops*	41	🟪
3.3 Forest Plantation*	9	🟫
3.4 Mosaic of Agriculture or Pasture	21	🟨
4.4 Other non Vegetated Areas	25	🟥
5. Water	33	🟦

*Exceptionally, in regions 01, 10, 19, 21, 27 and 30 we also included the class "3.2.1.5 Other Temporary Crop" (ID: 41) 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.

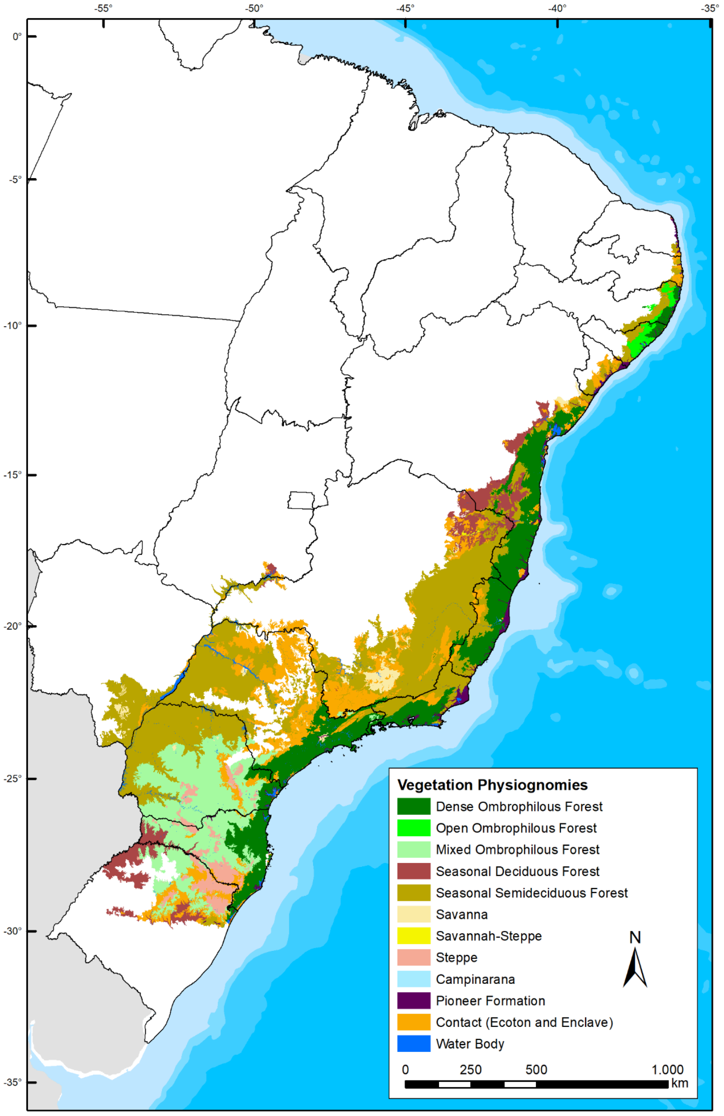
4.1.1. Forest Formation

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 (Figure 8):

- 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 8. Native vegetation types in the Atlantic Forest biome (IBGE, 2017).



4.2 Feature space

The feature space used to classify the Atlantic Forest biome comprised a subset of 36 variables (Table 3). They include the original Landsat reflectance bands, as well as vegetation indexes, spectral mixture modeling-derived variables and terrain morphometry (slope). The definition of the subset was made based on a feature importance analysis produced with Random Forest classification with all bands and 500 interactions.

Table 3. Feature space subset considered in the classification of the Atlantic Forest biome Landsat image mosaics in the MapBiomas Collection 9 (1985-2023).

Type	Name	Formula	Statistics	Reference
Landsat band	Green	Band 2 (L5 and L7) Band 3 (L8)	median	USGS
	Red	Band 3 (L5 and L7) Band 4 (L8)	median, minimum	USGS
	NIR	Band 4 (L5 and L7) Band 5 (L8)	median, 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	USGS
Spectral Index	Cellulose Absorption Index	CAI = SWIR2 / SWIR1	median	Nagler et al. 2003
	Enhanced Vegetation Index 2	EVI 2 = 2.5 × (NIR - Red) / (NIR + 2.4 × Red + 1)	median, median_dry, median_wet	Parente et al., 2018
	Green Chlorophyll Vegetation Index	GCVI = (NIR / Green - 1)	median, median_dry, median_wet, stdDev	Burke et al., 2017
	Normalized Difference Vegetation Index	NDVI = (NIR - Red) / (NIR + Red)	amplitude, median_dry, median_wet	Rouse et al., 1974
	Normalized Difference Water Index	NDWI = (NIR - SWIR1) / (NIR + SWIR1)	median, median_wet, stdDev	Gao et a., 1996
	Soil-Adjusted Vegetation Index	SAVI = 1.5 × (NIR - Red) / (NIR + Red + 0.5)	median, median_dry, median_wet	Huete, 1988
Terrain	Slope	ALOS DSM: Global 30 m	identity	Tadono et al., 2014
Coords	Latitude and Longitude		Latitude, Longitude	

In Collection 9, for the first time, the Landsat image segmentation function (ee.Algorithms.Image.Segmentation.SNIC) was used with the original median bands. This function generated a "clusters" band, "clusters_green_text" band, and "clusters_ndfi_median" band for each year. These bands were also included in the feature space.

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 38 years of Collection 8, so named “stable samples”. An ensemble was made from three main sources: extracted from Collection 8; manually drawn polygons; and complementary samples.

4.3.1. Stable samples from collection 8

The extraction of stable training samples from the previous Collection 8 followed several steps to ensure their confidence for use. We have identified each region’s predominant, secondary, and rare classes . The areas that did not change class from 1985 to 2022 in collection 8 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
- 200 or 500 training samples to rare class

Reference maps of the states of São Paulo, Minas Gerais, Paraná, and Espírito Santo were used as a source to filter stable samples in natural areas. All links are available on the MapBiomas Brazil website.
https://brasil.mapbiomas.org/en/mapas-de-referencia/

The number of samples of each class was defined for each region based on the visual and accuracy analysis of the Collection 8 classification, and it is available in the GitHub script “step2b_exports_samples”.

The samples from forest and grassland were 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

4.3.2. 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 in 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.

4.3.3. 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 39 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.

4.3.3. 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.

5. Post-classification

Due to the pixel-based classification method and the extended temporal series, a list of post-classification spatial and temporal filters was applied. 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 theoretically not allowed and 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).

5.3. Temporal filter

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

The first process looks in a 3-year moving window to correct any value that is changed in the middle year and return to the same class 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 2022 that is not 21 (Mosaic of Agriculture and Pasture) and is equal to 21 in 2021 and 2022. The value in 2023 is then converted to 21 to avoid any regeneration in the last year.

5.4. Frequency filter

Frequency filters were applied only in pixels that were considered “stable native vegetation” (at least 35 years as [3, 4, 11, 12,, 29]). If a “stable native vegetation” pixel is at least 80% of years of the same class, all years are changed to this 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. 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 is replaced to 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.

5.6. Transition filter

Yearly deforestation or forest recovery with less than 4 connected pixels that do not persist until 2021 were removed from the LULC annual map. This means that some small and temporary changes in forest classification will not be considered as regrowth of secondary vegetation wrongly .

5.7. 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.

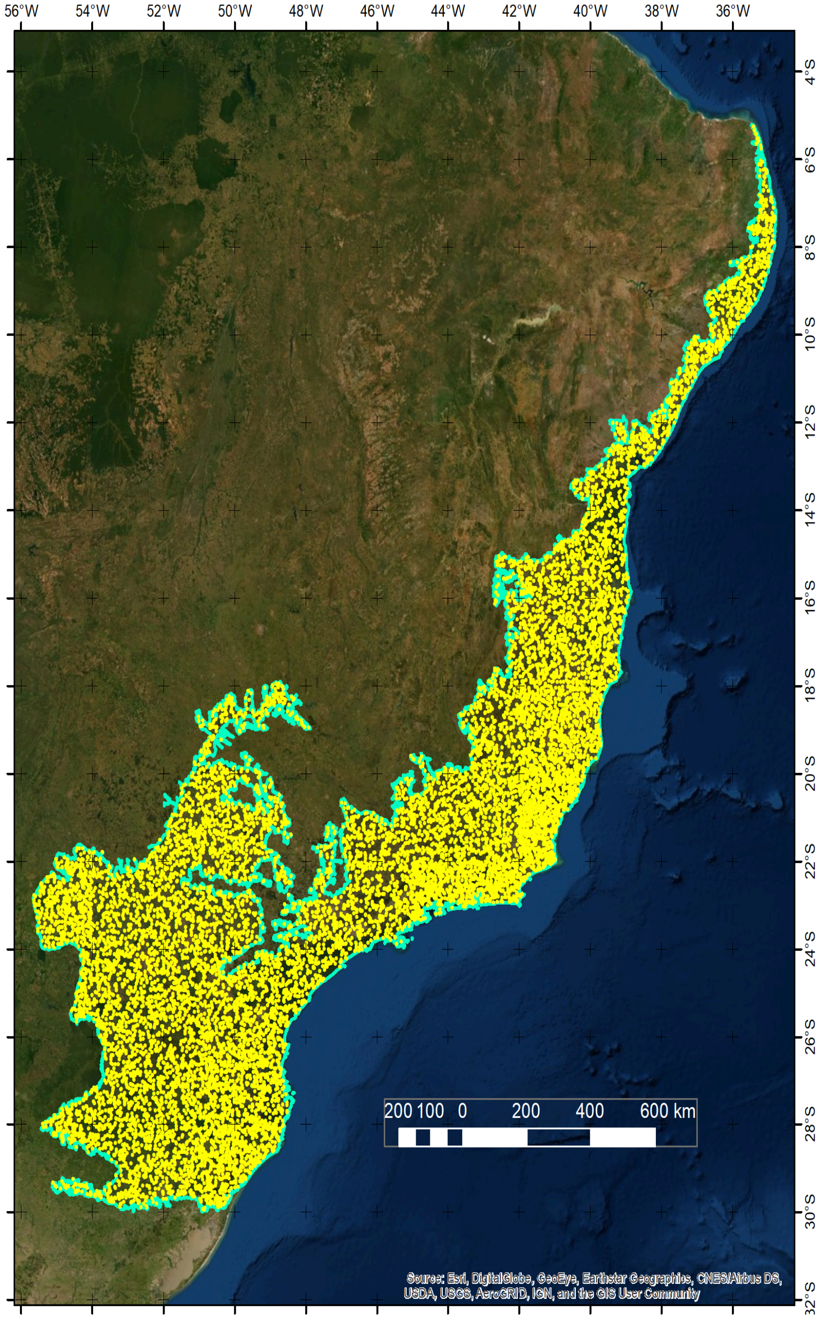
5.8. Classification of Herbaceous Sandbank Vegetation

Herbaceous sandbank vegetation was mapped resulting from the post-classification. The IBGE Soil map was used as reference. The class id 13 (Other non Forest Formations) in ESPODOSSOLO and NEOSSOLO was converted to this class.

6. Validation strategies

The set of 14.487 independent validation points provided by Lapig (Laboratório de Processamento de Imagens e Geoprocessamento - UFG) was used to perform accuracy analysis (Figure 8). For collection 9, some of the validation points were revised. This revision, along with the improvements in the classification of the Collection 9, has altered the global accuracy for all the previous collections. The values were updated, in level 1, 2 and 3.

Figure 8. Accuracy points in Atlantic Forest biome.



Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

The result of accuracy is presented in MapBiomas Website.

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

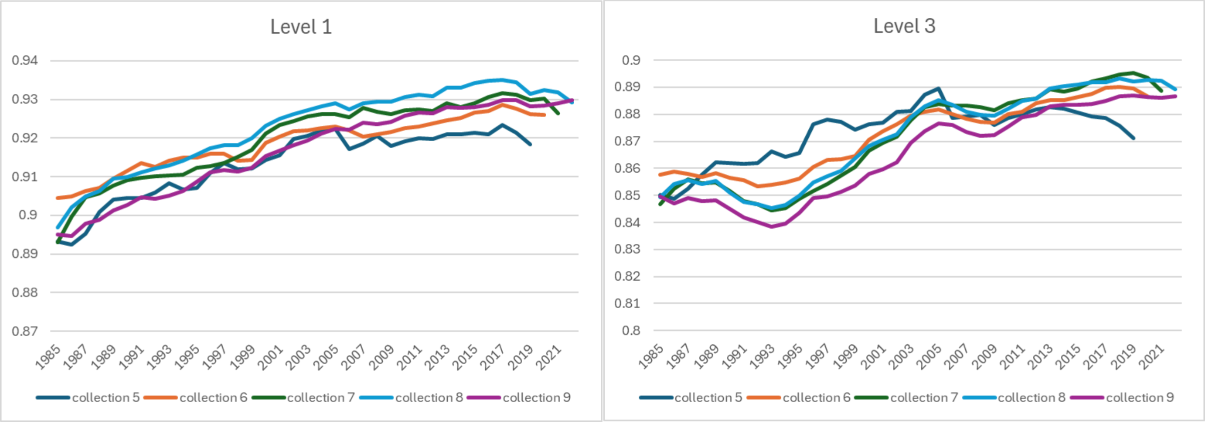
Global accuracy (considering all years) was 90.4%, 86.4%and 86.4% in levels 1, 2 and 3 for collection 5 and collection 6 have about the same values, 91.0%, 86.3%and 86.3% in levels 1, 2 and 3, respectively. The difference is explained by the reclassification of “Forest Plantation” from “1. Forest > 1.2 Forest Plantation” to “3. Farming > 3.3 Forest Plantation”, also affecting “Savanna Formation” that moved from level 3 to level 2 in collection 6.

In collection 7 the Global accuracy was 91.0%, 86.2%and 86.1% in levels 1, 2 and 3, respectively. In collection 8 the Global accuracy is 91.2%, 86,1%and 86.1% in levels 1, 2 and 3, respectively.

Collection 9 shows a Global accuracy of 90,5% in level 1, 85,4% in level 2 and 85.4% in level 3.

The detailed information about Global accuracy presented in Figure 9 for level 1 and level 3.

Figure 9. Global accuracy for Atlantic Forest biome at legend level 1 and level 3



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