



Atlantic Forest - Appendix

Collection 8

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

Collection 1.0 covered the period of 2008 to 2015 and was published in 2016. Collections 2.0 and 2.3 covered the period of 2000 to 2016 and were published in 2018. The classification using Random Forest was implemented in Collection 2.3, and from this point onward, the empirical decision tree was used for the purpose of generating 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 for the classification of the entire time series. Collections 3.0 and 3.1 expanded the period covered to 1985–2017 and added Rocky outcrop to the map. 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. Wooded Sandbank Vegetation was added in Collection 6 and Herbaceous Sandbank Vegetation added to the map in Collection 7. Collection 8 uses the same legend from collection 7.

The Google Earth Engine codes of atlantic forest classification are available in javascript format in the github: <https://github.com/mapbiomas-brazil/atlantic-forest>

Table 1. The evolution of the Atlantic Forest mapping collections in the MapBiomass Project, its periods, level and number of classes, brief methodological description, and global accuracy in Level 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 / 24	Random Forest	Level 1: 90.6% Level 2: 85.5%
7.0	37 years 1985-2021	4 / 24	Random Forest	Level 1: 90.1% Level 2: 84.4%
8.0	38 years 1985-2022	4 / 24	Random Forest	Level 1: 87.9% Level 2: 83.1%

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

The production of the Collection 8, with land cover and land use annual maps for the period of 1985-2022, followed a sequence of steps in the Atlantic Forest biome, similar to those used in the previous Collections 4, 5, 6 and 7 (**Figure 1**). However, some improvements were added up, particularly in the mosaics, balance of samples and in the post classification filters.

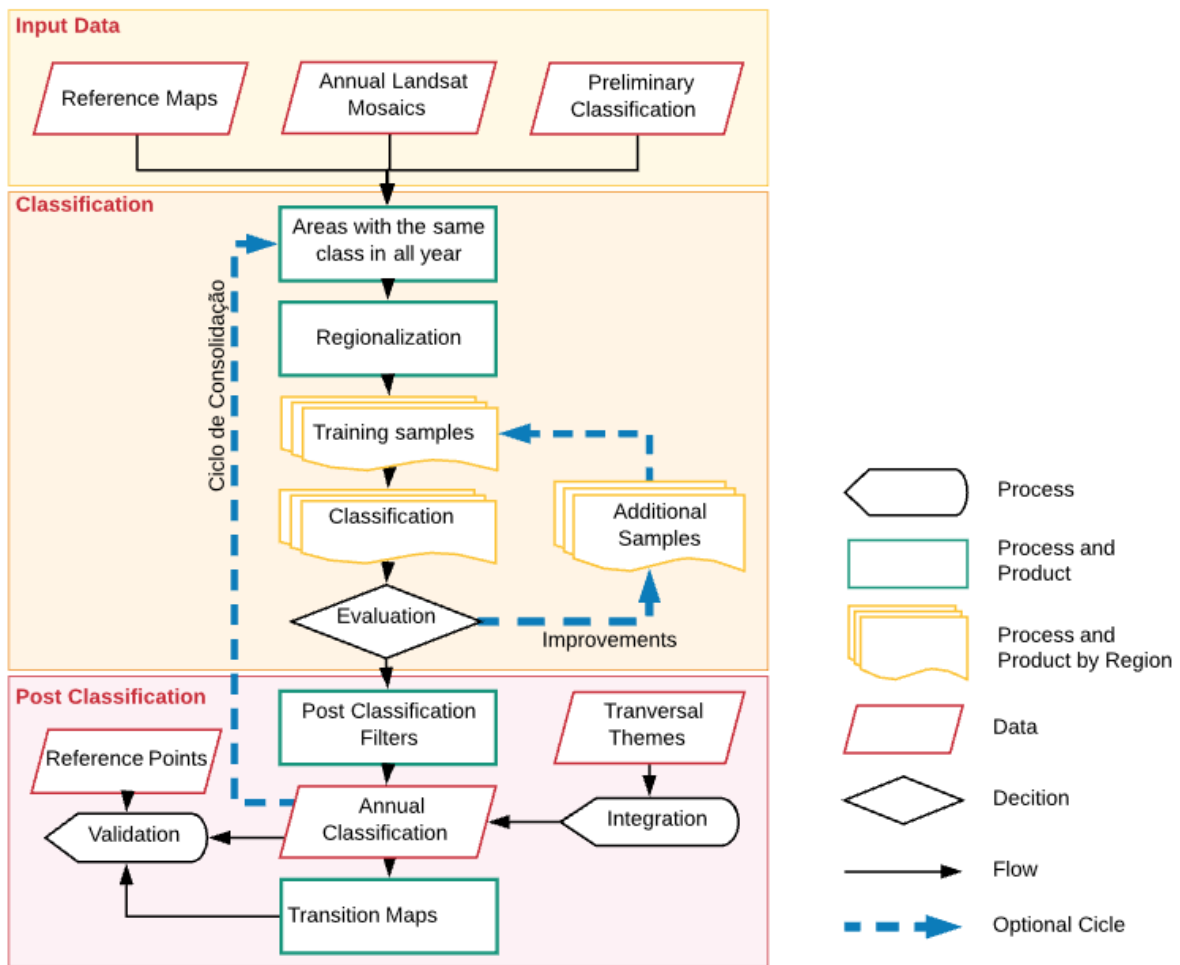


Figure 1. Classification process of Collection 8 in the Atlantic Forest biome.

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, being the use of TOA discontinued. In Collections 7 and 8 we adopted USGS Landsat 8 Level 2, Collection 2, Tier 1.

The mosaic of images consists of a composition of the best pixels that are extracted from all the images available in a defined period within a year. Once the initial and final dates of this period were defined, the median pixel from that period was 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 to classification.

The image selection period for the Atlantic Forest biome was defined aiming to maximize the coverage of Landsat images after cloud removal/masking.

Despite the diversity of ecosystems and the great extent of the biome (**Figure 2**), both in latitudinal amplitude and in coast extension, the Atlantic Forest has a well-defined dry period between the months of April to September, as exemplified by data from Cunha-SP station (**Figure 3**).

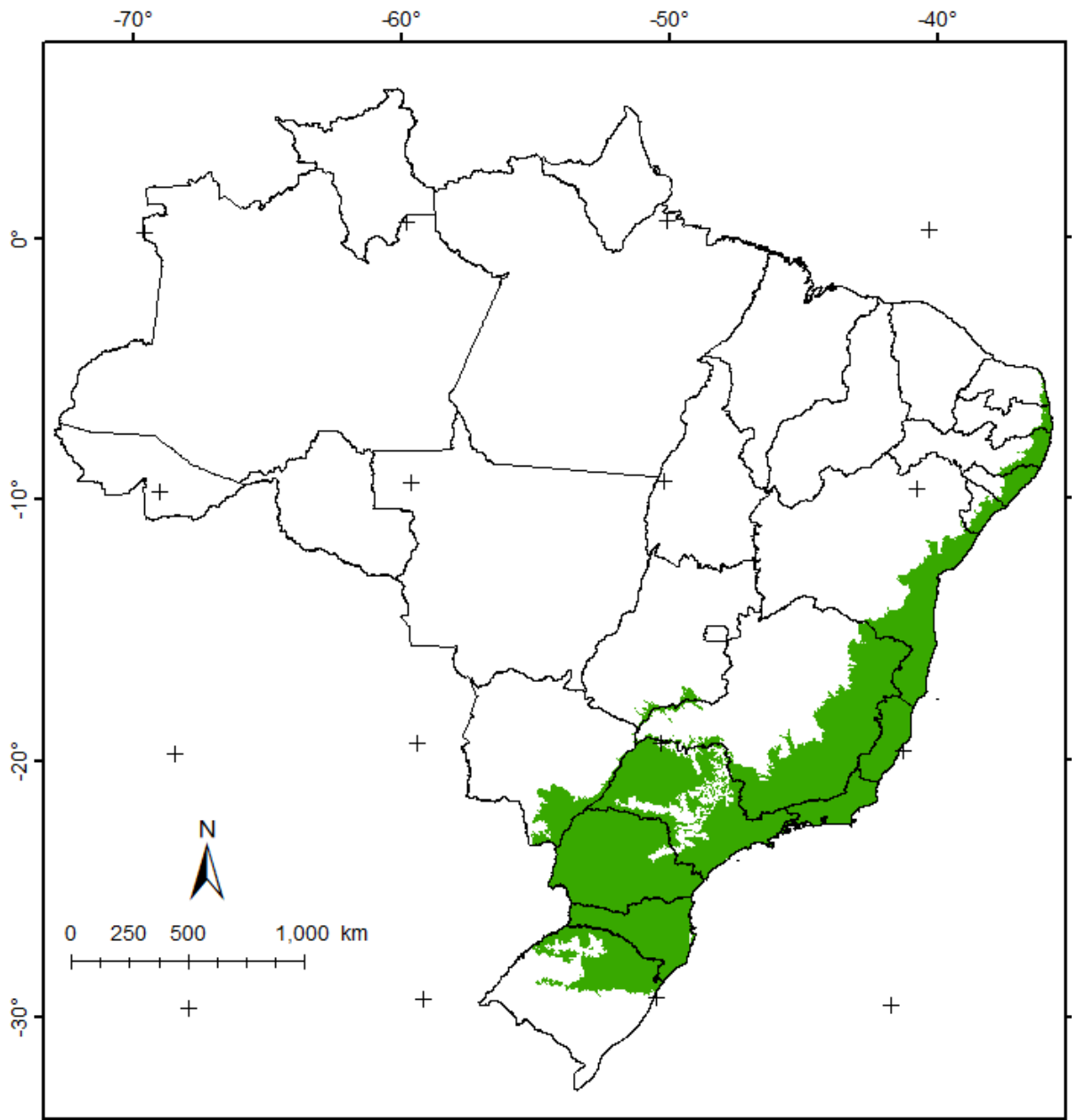


Figure 2. Atlantic Forest biome (IBGE, 2019).

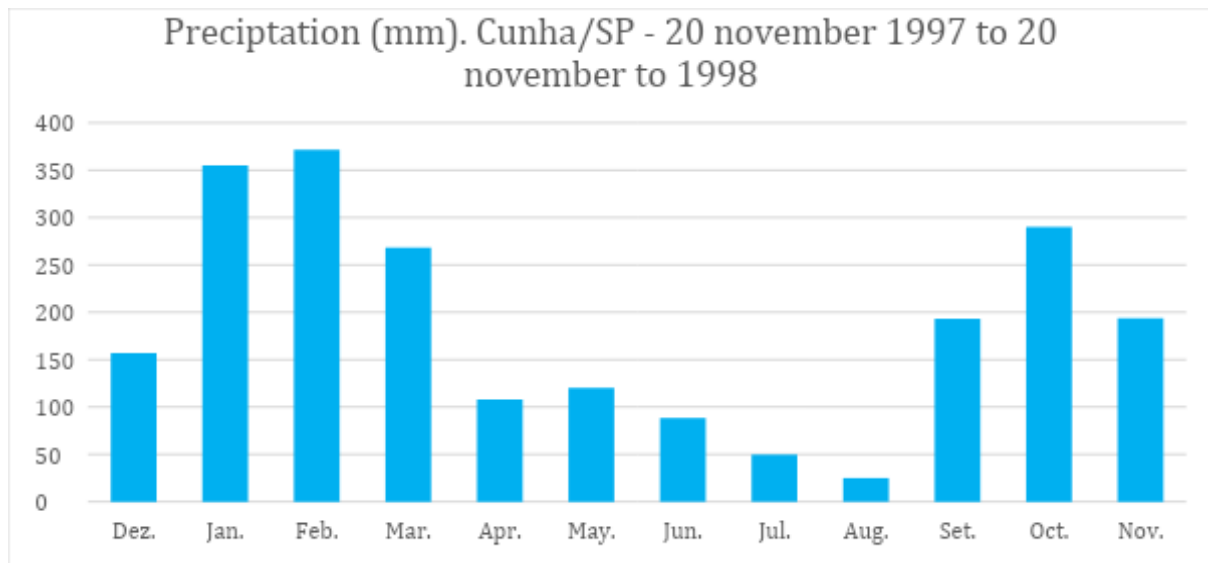


Figure 3. Monthly precipitation values of the period from November 20, 1997 to November 20, 1998, in Cunha-SP (ARCOVA et al., 2003).

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 in the beginning of the period, in the winter season.

In most cases the period from April 1st to August 30th was suitable to get a mosaic with none or few missing information caused by clouds and shades. In the Northeast states the period was February 1st to 30 of October to maximize the visible areas and avoid missing areas caused by clouds.

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 2021 – Landsat 8
- 2022 – Landsat 9

We made a visual analysis on the preliminary mosaics to identify and remove images with noises (clouds, shadow, or sensor defect) for each year (**Figures 4 and 5**).

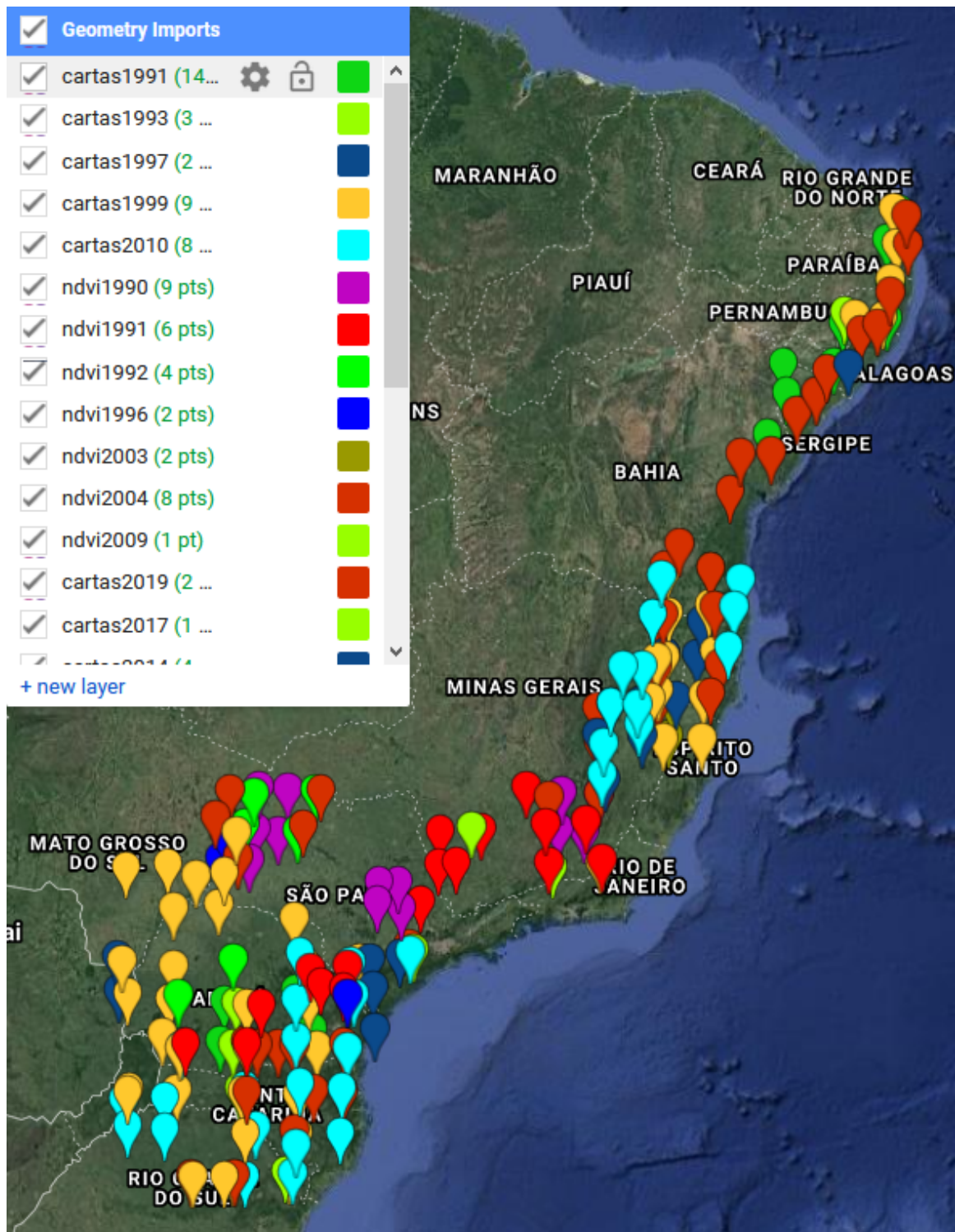


Figure 4. Example of points identifying yearly mosaics reviewed

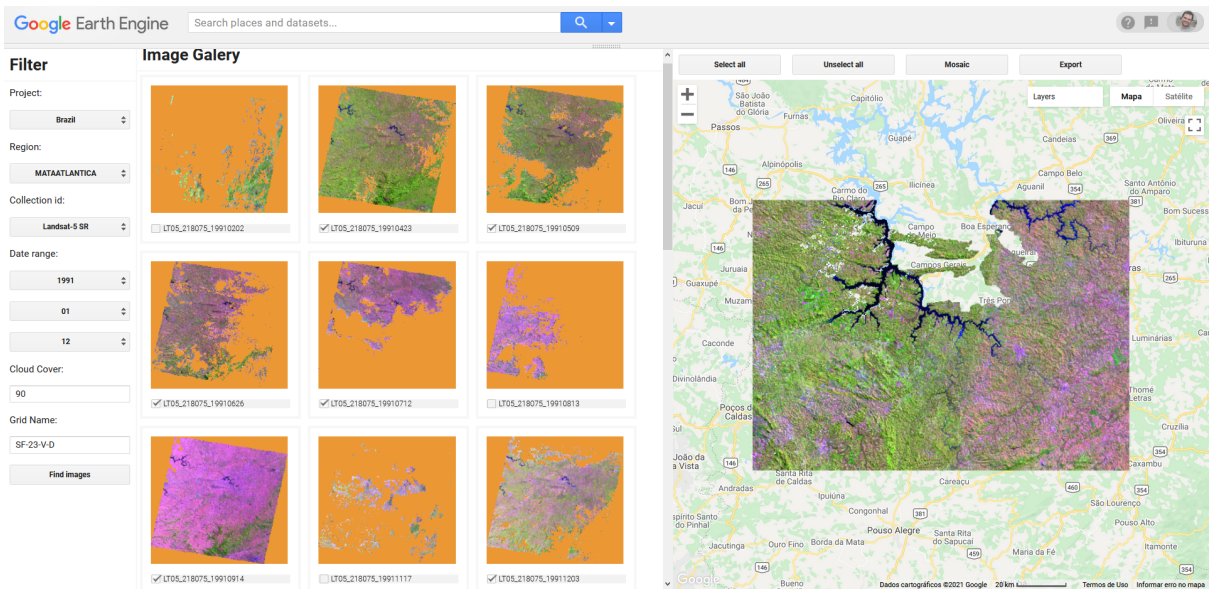


Figure 5. Google Earth Engine tool to identify and remove scenes with noise

2.3. Final quality

As a result of the selection criteria, the majority of the mosaics presented satisfactory quality. Northeast of Brazil and some regions in Santa Catarina and São Paulo offer more challenges 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 in 30 regions (**Figure 6**).

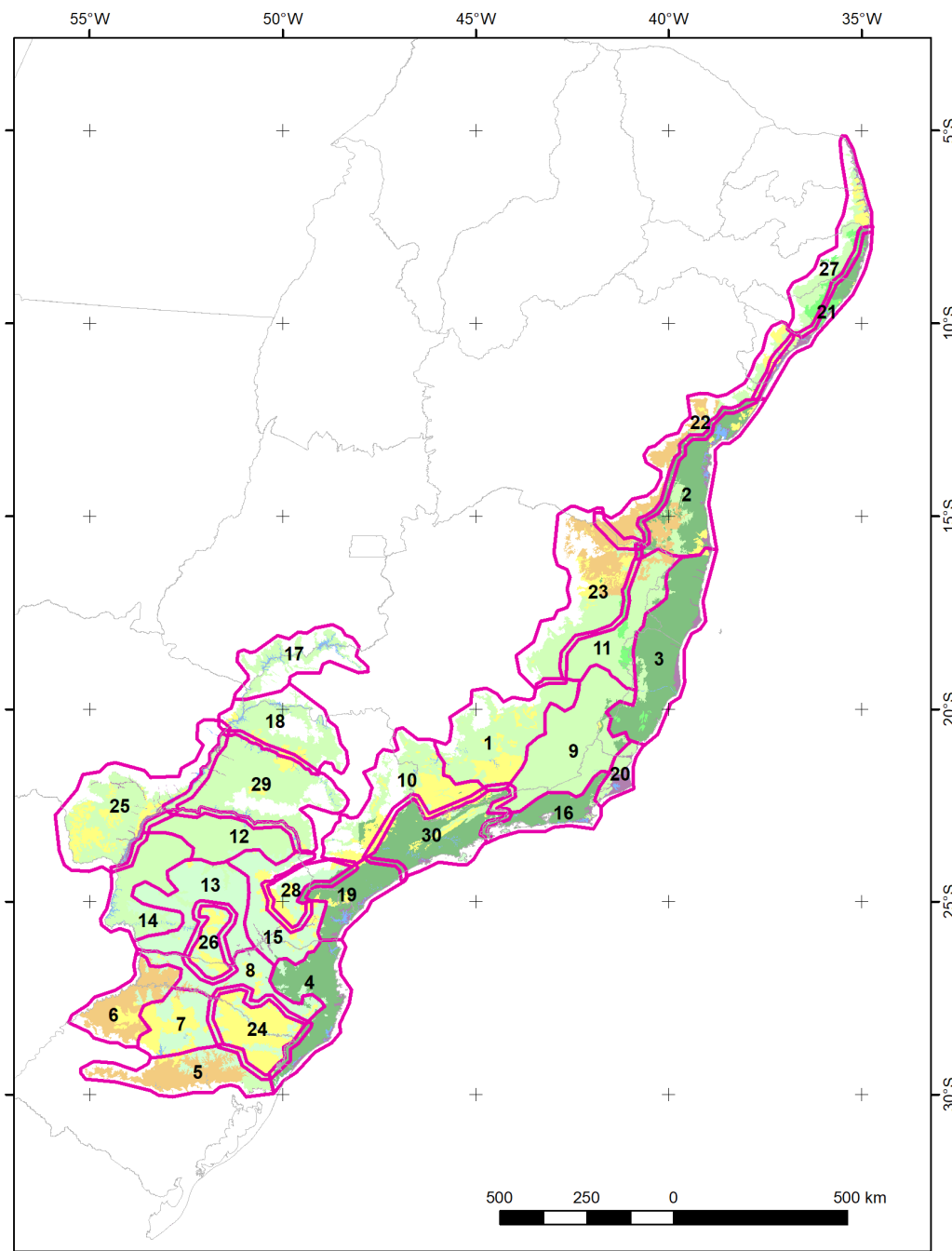













Figure 6. Regions used in the classification of Atlantic Forest biome.

4. Classification

4.1. Classification scheme

The digital classification of the Landsat mosaics for the Atlantic Forest biome aimed to individualize a subset of 10 land cover and land use classes (**Table 2**), which were integrated with the cross-cutting themes in a further step.

Table 2. Land cover and land use categories considered for digital classification of Landsat mosaics for the Atlantic Forest biome in the MapBiomias Collection 8.

Legend class of Collection 8	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	
2.6. Other non Forest Formations	13	
3.4 Mosaic of Uses	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 Plantation (id: 9).

A) Forest Formation

Forest Formation include natural forest (exclude Forest Plantation) areas of more than 0.5 hectares (ha) with trees with minimum height of 5 meters (m) and tree canopy cover that varied for each type of original forest formation (**Figure 7**):

- 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%

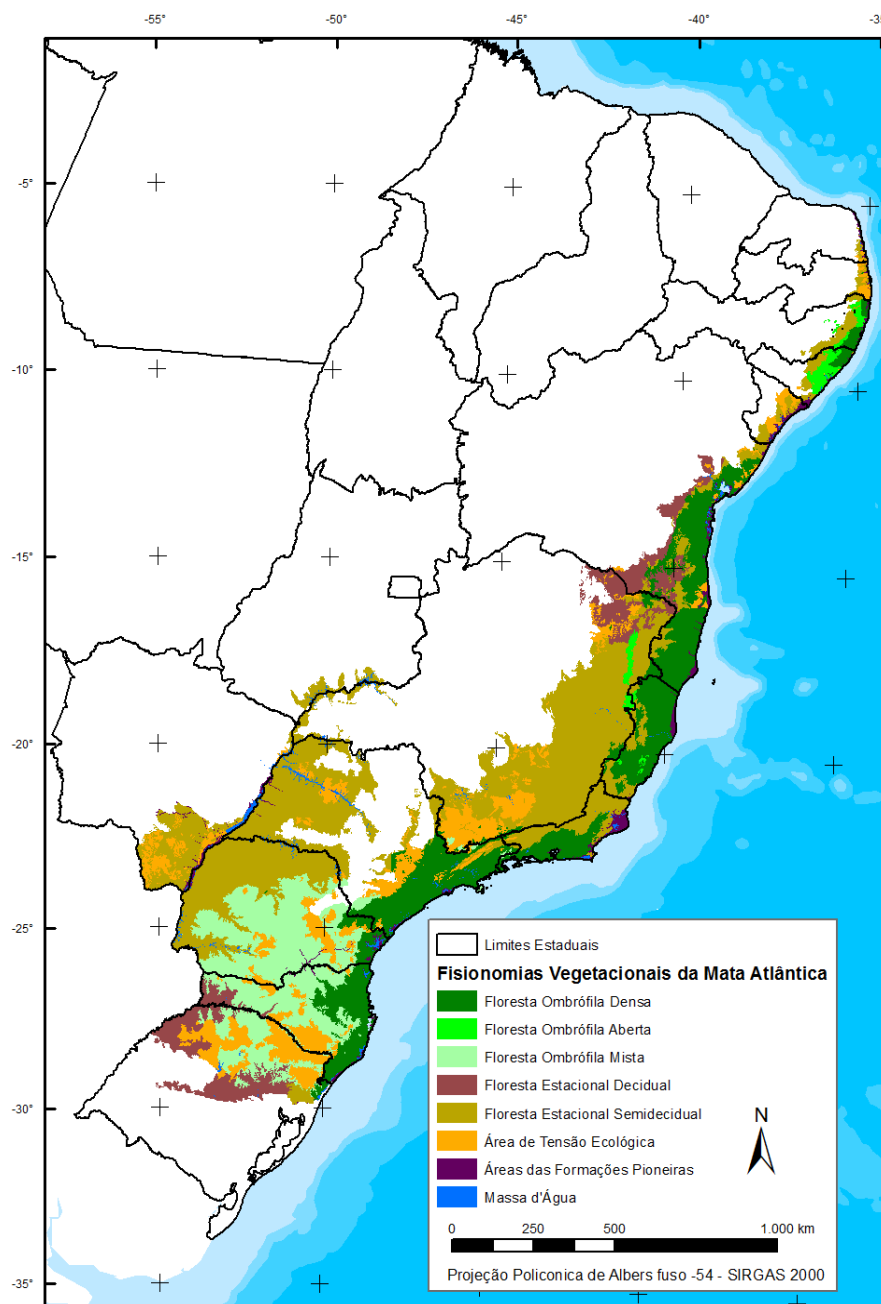


Figure 7. Native vegetation types in the Atlantic Forest biome (IBGE, 2017).

4.2 Feature space

The feature space for digital classification of the classes of interest for 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 Forests 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 MapBiomias Collection 8 (1985-2022).

amp_ndvi_3anos	latitude	red_min
cai_median	longitude	savi_median
evi2_median	ndvi_median_dry	savi_median_dry
evi2_median_dry	ndvi_median_wet	savi_median_wet
evi2_median_wet	ndwi_median	slope
gcvl_median	ndwi_median_wet	swir1_median
gcvl_median_dry	ndwi_stdDev	swir1_median_dry
gcvl_median_wet	nir_median	swir1_median_wet
gcvl_stdDev	nir_median_wet	swir2_median
green_median	red_median	swir2_median_dry
green_median_wet	red_median_dry	swir2_median_wet
green_min	red_median_wet	wefi_median_wet

4.3. Classification algorithm, training samples and parameters

Digital classification was performed region by region, year by year, using a *Random Forest* algorithm (Breiman, 2001) available in Google Earth Engine, running 70 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 along the 37 years of Collection 7, so named “stable samples”. An ensemble taken from three main sources was made: extracted from Collection 7; manually drawn polygons; and complementary samples.

4.3.1. Stable samples to Collection 8

The extraction of stable training samples from the previous Collection 7 followed several steps aiming to ensure their confidence for use as training areas. We have identified the predominant, secondary, and rare class and in each region. The areas that did not change class from 1985 to 2021 in Collection 7 were used to generate random training points balanced with the rule:

- 3.000 or 4.000 points to predominant class
- 1.000 or 2.000 points to secondary class
- 200 or 500 points to rare class

The number of samples of each class were defined for each region based on the visual and accuracy analysis of the Collection 7 classification. 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 $\geq 9\text{m}$
- Grassland sample need to be $< 7\text{m}$

4.3.2. Multi-classification

In Collection 8 an innovation was performed to produce 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, where the variables “n_pri” and “n_se” define the number of samples in primary and secondary class for each region.

```
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.

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 37 years during the polygon drawing. Based on experts’ knowledge of each region, polygon samples from each class were collected and the number of random points in these polygons were 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, and it was trained in the same mosaic of the year that was classified.

5. Post-classification

Due to the pixel-based classification method and the long 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. 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). It 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). Pixels that do not reach the six connected pixels with the same class are dissolved by MODE value in a 3x3 kernel.

5.3. Temporal filter

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

In the first process, filter 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 of classes: 21, 9, 33, 13, 4, 29, 12, 11, 3.

The second process is similar to the first process, but it is a 4- and 5-years moving window that corrects all middle years.

In the third process, the filter looks for any native vegetation pixel (3, 4, 12, 13) that has the same value in 1986 and 1987 and a different value in 1985. Then the 1985 value is corrected to the same value as 1986 and 1987, in order to avoid any regeneration in the first year.

In the last process, the filter looks for a pixel value in 2022 that is not class 21 (Mosaic of Uses) and it is equal to class 21 in 2020 and 2021. The value in 2022 is then converted to class 21 to avoid any regeneration in the last year.

5.4. Wetland filter

We used the 'Height Above Nearest Drainage'-HAND (Rennó et al. 2008) product as a proxy to represent the 'groundwater depth'. If a pixel classified as wetland (ID=11) had a HAND value greater than 15 meters, this pixel was converted to Mosaic of Uses (ID=21).

5.5. Incident filter

An incident filter was applied in Collections 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 by Savana (ID=4) or Mosaic of Agriculture or Pasture (ID=21) according to the mode value. This avoids changes in the border of the classes.

5.6. Transition filter

A new filter was applied on Forest Class in Collection 8 to reduce noise in transitions. Yearly deforestation or forest recovery with less than 4 connected pixels that do not persist until 2022 was not changed to forest in the annual map. This means that some small and temporary changes in forest classification were removed from annual maps.

5.7. Classification of Wooded Sandbank Vegetation

Wooded Sandbank Vegetation was mapped as a result of the post-classification. The ALOS DSM: Global 30m was used to identify coastal forest with less than 25m altitude. This was then assumed to be Wooded Sandbank Vegetation. A spatial mask was applied to exclude some regions in northeast of Brazil where wooded sandbank vegetation does not exist.

5.8. Classification of Herbaceous Sandbank Vegetation

Herbaceous Sandbank Vegetation was mapped as BETA version in the collection, as a result of the post-classification. The IBGE Soil map was used as reference. The class 13 (Other non Forest Formations) in ESPODOSSOLO and NEOSSOLO was converted to Herbaceous Sandbank Vegetation.

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 in the Atlantic Forest mapping (**Figure 8**).

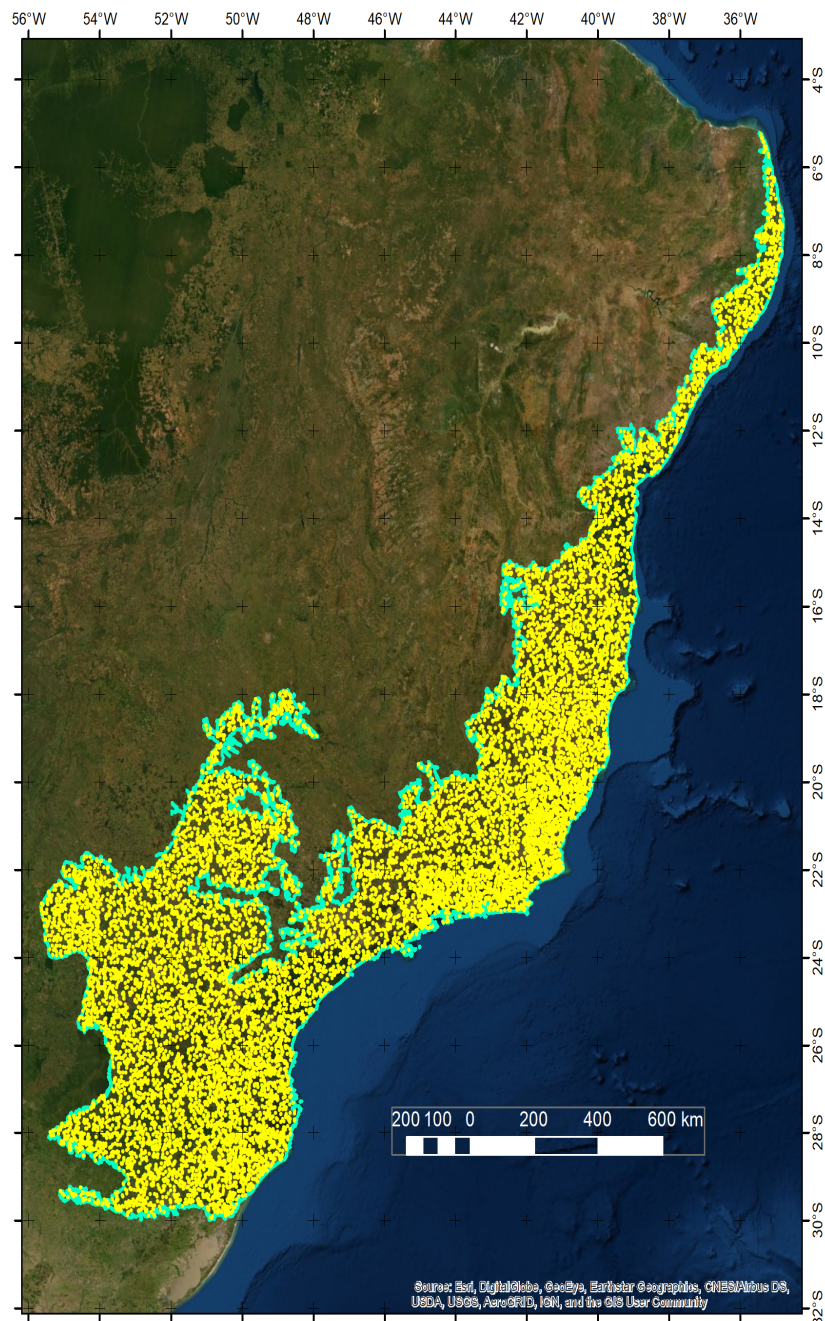


Figure 8. Accuracy points in Atlantic Forest.

The result of accuracy is presented in MapBiomass Website (https://mapbiomas.org/en/estadistica-de-accuracia?cama_set_language=en).

Global accuracy (considering all years) was 90.6%, 85.5% and 85.5% in levels 1, 2 and 3 of the Collection 6 and Collection 7 presented a similar accuracy, 90.1%, 84.4% and 84.3% in levels 1, 2 and 3, respectively.

In Collection 8 the Global accuracy was 87.9%, 83.1% and 83.1% in levels 1, 2 and 3, respectively. The detailed information about commission and omission error are presented in Figure 9 and Figure 10.

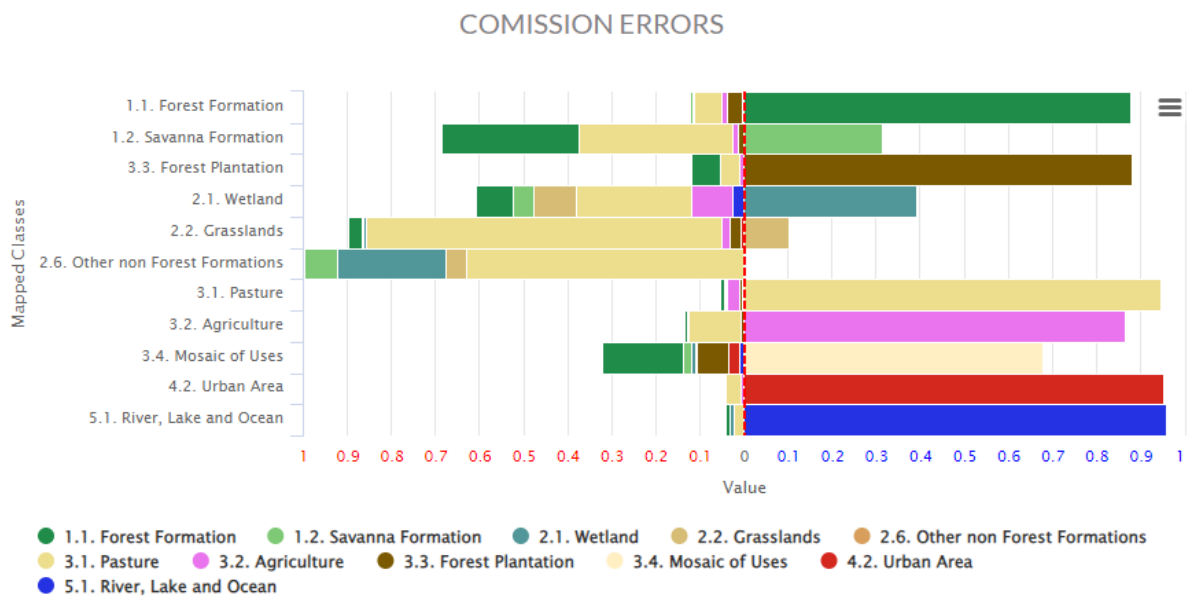


Figure 9. Commission error in 2022 in Atlantic Forest for each level 2 class.

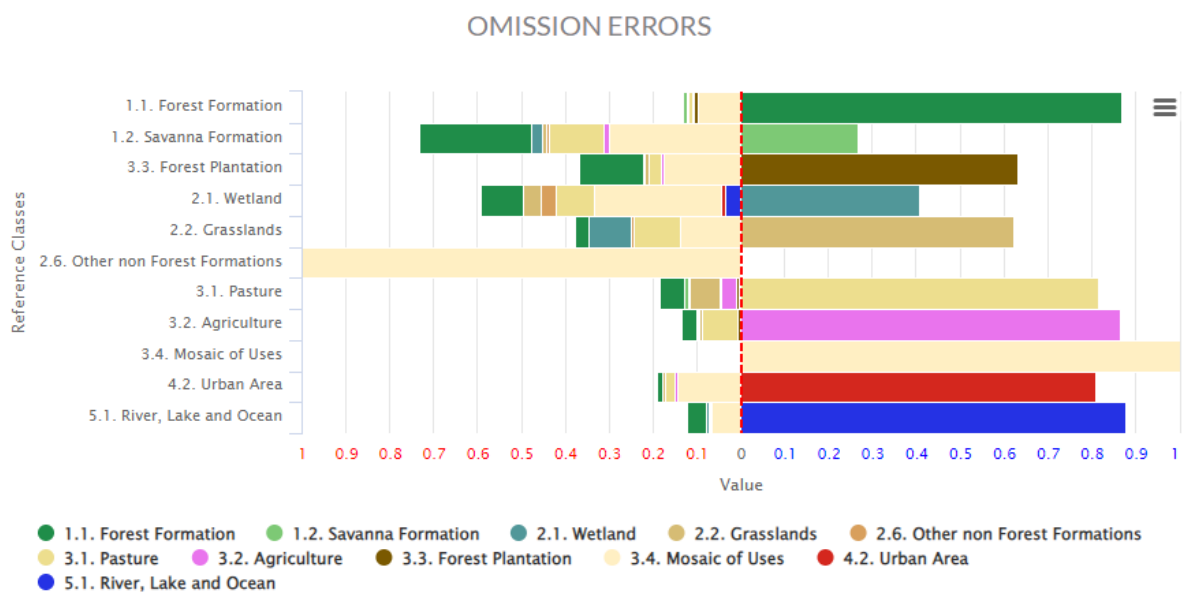


Figure 10. Omission error in 2022 in Atlantic Forest for each level 2 class.

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

IBGE. **Vegetação RADAM**. Disponível em:

<ftp://geofp.ibge.gov.br/informacoes_ambientais/acervo_radambrasil/vetores/>. Accessed in: may, 30 2018.

POTAPOV, P.; LI, X.; HERNANDEZ-SERNA, A; A. ; HANSEN, M.C.; KOMMAREDDY, A.; PICKENS, A.; TURUBANOVA, S.; TANG, H.; SILVA, C.E.; ARMSTON J.; DUBAYAH, R.; BLAIR, J. B.; HOFTON, M. (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>

Rennó, Camilo & Nobre, Antonio & Cuartas, Luz & Soares, João & Hodnett, Martin & Tomasella, Javier & Waterloo, M.J.. (2008). HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. *Remote Sensing of Environment*. 112. 3469-3481. 10.1016/j.rse.2008.03.018.