MapBiomas Atlantic Forest

Collection 5

Version 1

General coordinator
Ana Eljall

Technical coordinator
Pablo Baldassini

Technical support
Andrés Leszczuk
Juan Pablo Zurano
Facundo Skromeda
Gonzalo Dieguez Gaviola
Luna Schteingart

1. Introduction

1.1. Scope and content of the document

The document presents a general description of the satellite image processing, the feature inputs and the process step by step applied to obtain the annual classifications. The objective of this document is to describe the theoretical basis, justification and methods applied to produce annual maps of land use and land cover (LULC) in the Atlantic Forest of Argentina from 1985 to 2024 of the MapBiomas Atlantic Forest Collection 5.

The Atlantic Forest Collection 5 of Argentina followed a sequence of steps like those used in the Collection 4. The Collection 5 is based on the stable samples taken from the Collection 4, that covered the period of 1985 to 2023, and it was published in 2024. For this, a revalidation of all the samples was carried out. Due to the impossibility of finding a significant number of stable samples for the entire period, samples were taken considering two subperiods, 1985-2024 and 2000-2024.

1.2. Region of Interest

MapBiomas Atlantic Forest was created to produce LULC annual maps for the Atlantic Forest corresponding to Argentina territory. Other biomes located around the region were partially included to allow better regional integration between them. Thus, the northeast portion of the flooded grassland and savannas corresponding to Argentina was included. The study area was divided in 3 homogenous subregions to reduce confusion of samples and classes, as well as to allow a better balance of samples and results. A total of 43.553 km^2 was considered.

2. Remote Sensing Data

2.1. Landsat Collection

The imagery dataset used in the *MapBiomas Atlantic Forest* Collection 5 was obtained by the Landsat sensors Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+) and

the Operational Land Imager and Thermal Infrared Sensor (OLI-TIRS), on board of Landsat 5, Landsat 7 and Landsat 8, respectively. The Landsat imagery collections with 30-pixel resolution were accessible via Google Earth Engine, and source by NASA and USGS. The *MapBiomas Atlantic Forest* Collection 5 has used Collection 2 Tier 1 Level 2 surface reflectance (SR), which underwent through radiometric calibration and orthorectification correction based on ground control points and digital elevation model to account for pixel co-registration and correction of displacement errors. A total of 7 scenes were used for covering the entire region, where each of them is totally or partially within the area. 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 2024 – Landsat 8

2.2. Landsat Mosaics

Landsat cloud free composites obtained from images distributed along the whole year were considered. The cloud/shadow removal script takes advantage of the quality assessment (QA) band and the GEE median reducer. When used, QA values can improve data integrity by indicating which pixels might be affected by artefacts or subject to cloud contamination (USGS, 2017). In conjunction, GEE can be instructed to pick the median pixel value in a stack of images. By doing so, the engine rejects values that are too bright (e.g., clouds) or too dark (e.g., shadows) and picks the median pixel value in each band for a specific year.

3. Overview of methodological process

The methodological steps of Collection 5 are presented in the Figure 1 and detailed below. The first step was to generate annual Landsat image mosaics based on yearly periods. The

second step was to establish the spectral feature inputs derived from the Landsat bands to run the random forest classification. The acquisition of training samples started with the selection of temporally stable samples. Once selected each LULC classes in each subregion it has be able to adjust the training data set according to its statistical needs, including complement samples. Based on the adjusted training data set, the random forest classifier was run. Following that, spatial and temporal filters were applied to remove classification noise and stabilize the classification. The LULC maps of each subregion were integrated based on prevalence rules to generate the final map of Collection 5. The MapBiomas annual LULC maps were used to derive the transition analysis (with spatial filter application) and statistics. The statistical analysis covered different spatial categories, such as subregion, state and municipality.

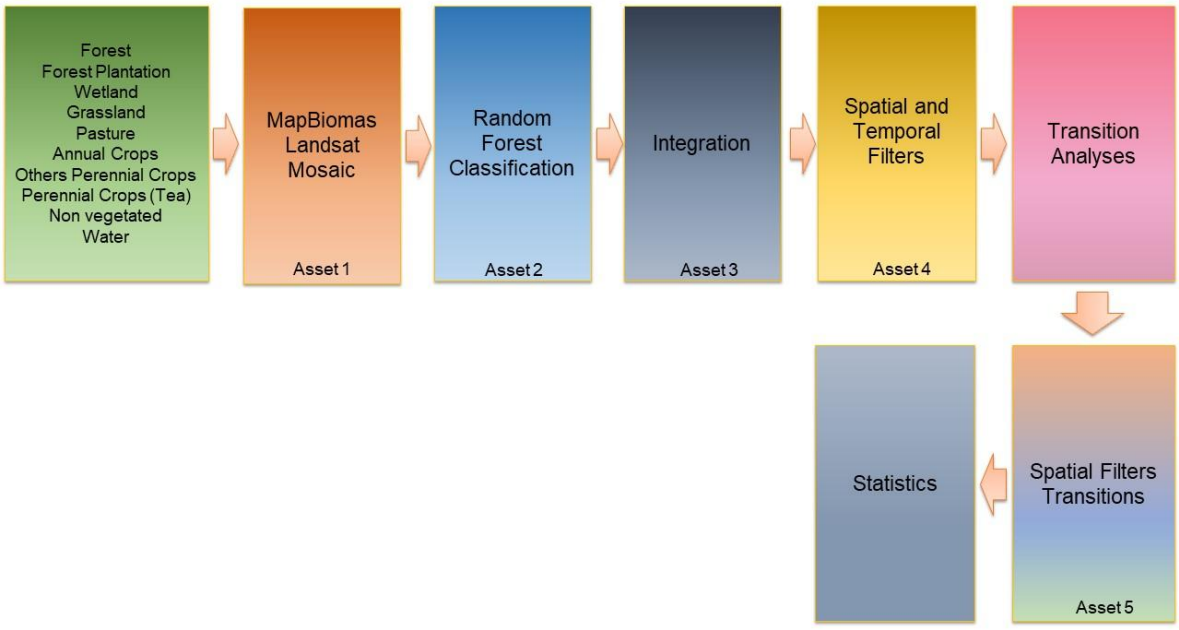


Figure 1. Methodological steps of Collection 5 to implement MapBiomas algorithms in the Google Earth Engine.

For each subregion, a temporal mosaic of Landsat images was built. All images from a specific year that presented a cloud visual pattern grouped were included (i.e, images that only presented clouds in a portion of the scene were considered). The selected Landsat data

had to allow an annual analysis and at least 4 images from different dates of the year had to be included.

4. MapBiomas feature space

The total available bands of the MapBiomas feature space are composed of 104 input variables, including the original Landsat bands, fractional and textural information derived from these bands (Table 1). Table 1 presents the equations to obtain these feature variables, as well as highlighted in green all the bands, indices and fractions available in the feature space. Reducers were used to generate temporal features such as:

- Median - Median of the pixel values of the best mapping period defined by each biome.
- Median_dry = median of the quartile of the lowest pixel NDVI values.
- Median_wet = median of the quartile of the highest pixel NDVI values.
- Amplitude = amplitude of variation of the index considering all the images of each year.
- stdDev = standard deviation of all pixel values of all images of each year.
- Min = lower annual value of the pixels of each band.

The feature space for digital classification of the categories of interest for the *MapBiomas Atlantic Forest* Collection 5 comprised a subset of 37 variables. The definition of the subset was made based on the usefulness of each variable to discriminate between LULC classes, indicated with X (Table 1). The variables selected were the same for all subregion, where the most appropriate subset of variables was chosen to later run the random forest algorithm.

			Reducer					
	band or index name	formula	median	median_dry	median_wet	amplitude	stdDev	min
bands	blue	B1 (L5 e L7); B2 (L8)	X					
	green	B2 (L5 e L7); B3 (L8)	X		X			X
	red	B3 (L5 e L7); B4 (L8)	X	X	X			X
	nir	B4 (L5 e L7); B5 (L8)	X		X			X
	swir1	B5 (L5 e L7); B6 (L8)	X	X	X			X
	swir2	B7 (L5); B8 (L7); B7 (L8)	X	X	X			X
	temp	B6 (L5 e L7); B10 (L8)						
index	ndvi	(nir - red)/(nir + red)	X		X			
	evi2	(2.5 * (nir - red)/(nir + 2.4 * red + 1)	X	X	X			
	cai	(swir2 / swir1)	X					
	ndwi	(nir - swir1)/(nir + swir1)	X		X			
	gcvi	(nir / green - 1)		X				
	hall_cover	(-red*0.017 - nir*0.007 - swir2*0.079 + 5.22)						
	pri	(blue - green)/(blue + green)						
	savi	(1 + L) * (nir - red)/(nir + red + 0,5)	X	X	X			
	textG	('median_green').entropy(ee.Kernel.square({radius: 5}))						
fraction	gv						X	
	npv							
	soil							
	cloud							
	shade	100 - (gv + npv + soil + cloud)	X					
MEM index	gvs	gv / (gv + npv + soil + cloud)			X			
	ndfi	(gvs - (npv + soil))/(gvs + (npv + soil))	X		X			
	sefi	(gv+npv_s - soil)/(gv+npv_s + soil)						
	wefi	((gv+npv) - (soil+shade)) /((gv+npv) + (soil+shade))			X			
	fns	((gv+shade) - soil) / ((gv+shade) + soil)						
slope		ALOS DSM: Global 30m						

Table 1. List and reference of bands, fractions and indices available in the feature space (green color). The feature space subset considered by *MapBiomas Atlantic Forest* Collection 5 (1985-2024) for the LULC classification are indicated with X.

5. Classification of LULC

The production of the Collection 5, with land use and land cover annual maps for the period 1985-2024 in each subregion, included a) manually drawn polygons of LULC classes temporally invariant based on photointerpretation of annual Landsat images and temporal behavior of spectral indices, b) generation of stable samples thought LULC preliminary classifications, c) balance of samples based on proportion stats of each class, d) collect of complementary samples, e) annual LULC classifications, f) apply of temporal and spatial post classification filters (Figure 2). Due to the impossibility of finding a significant number of stable samples for the entire period for each LULC classes, samples were taken considering two subperiods, 1985-2024 and 2000-2024. Thus, preliminary classifications were made for each of the subperiods, which were then integrated for randomly sample stable samples in the entire period.

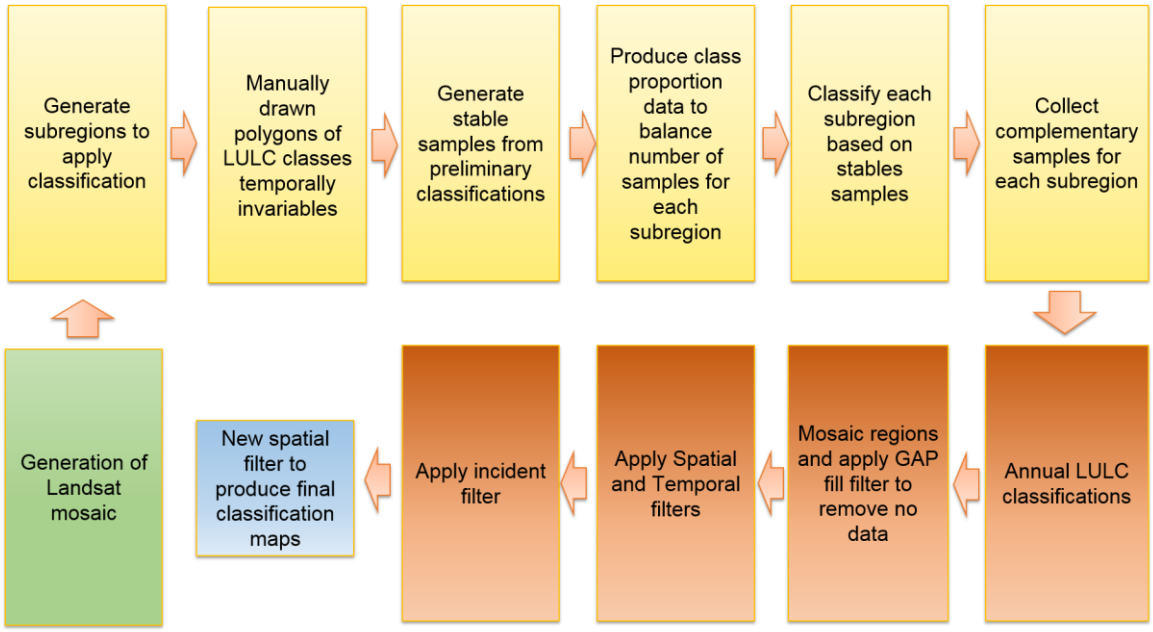


Figure 2. Classification process of Collection 5 in the *MapBiomas Atlantic Forest.*

5.1. Classification scheme

The digital classification of the Landsat mosaics for the *MapBiomas Atlantic Forest* Collection 5 aimed to individualize a subset of ten land use and land cover classes: Forest (3), Forest Plantation (9), Wetland (11), Grassland (12), Pasture (15), Annual crops (19), Non vegetated areas (22), Water (33), Others Perennial crops (Yerba and citrus) (48) and Tea (65) (Table 2). Not all classes were identified in each subregion.

Legend class of Collection 3	ID	Hexadecimal Code	Color
1.1.1 Forest Formation	3	#1f8d49	
1.2 Forest Plantation	9	#7a5900	
2.1 Wetland	11	#519799	
2.2 Grassland	12	#d6bc74	
3.1 Pasture	15	#edde8e	
3.2.1 Annual Crops	19	#E974ED	
3.2.2 Perennial Crops (Others)	48	#e6ccff	

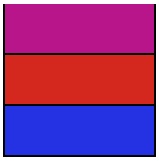
3.2.3 Perennial Crops (Tea)	65	#b9158a	
4. Non vegetated area	22	#d4271e	
5. Water	33	#2532e4	

Table 2. Land use and land cover (LULC) categories considered for digital classification of Landsat mosaics for the *MapBiomas Atlantic Forest* Argentina.

5.2. 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 250 iterations (random forest trees). Training samples for each region were defined following a strategy of using random pixels for which the land use and land cover remained the same at least 33 of the 40 years of Collection 5, so named “stable samples”. The stable areas were identified thought annual preliminary classification made using random pixels selected from manually drawn polygons. For this, false-color composites of the Landsat mosaics for all the 40 years as backdrop and graphs with the temporal behavior of spectral indices per pixel were used to stablish the LULC class. The classification process was carried out in two stages. In a first stage, a class called "perennial crops" was identified. In a second instance, only that class was reclassified into tea, other perennial crops and forest.

5.3. Preliminary classification

From manually draw polygons, a subset between 200 and 700 pixels per class was randomly selected and they were used as training areas to classify each of the 40 years with the Random Forest algorithm, running 250 iterations. A total of 40 yearly preliminary classification were obtained and the frequency with which a pixel was classified with the same LULC class was calculated to define the stables areas.

5.4. Stable samples

The identification of stable areas to extract random pixels or “stable samples” was based on a criterion of minimum frequency aiming to ensure their confidence for use as training areas. Each pixel should be classified with the same LULC class at least 33 times in the period 1985-2024 to be considered as stable, i.e. a pixel should remain with that class a minimum of 33 years to be eligible as a stable sample. A layer of pixels with a stable classification along the 40 years was then generated by applying such threshold. Additionally, reference maps were considered for the identification of different land uses and covers, which were applied as masks on the map of stable areas. In the case of forests, a consensus map was generated by combining 4 reference maps, which included Copernicus Global Land Service Land Cover 100m collection 3 (Buchhorn et al. 2020), the treecover2000 layer obtained from the Global Forest Change 2000– 2022 (Hansen et al. 2013), the ESRI 10m Annual Land Use Land Cover (2017-2022) (Karra et al. 2021) and the Global Forest Canopy Height from GEDI & Landsat (Potapov et al. 2021). The variance was calculated in a 5x5 pixels moving window, and a mask was applied to the consensus map in areas with variance greater than zero to eliminate the edges. In the case of perennial crops and forest plantations, shapefile information provided by Dirección de Desarrollo Foresto Industrial de la Secretaría de Agricultura, Ganadería y Pesca del Ministerio de Economía and Subsecretaría de Desarrollo Forestal del Ministerio del Agro y la Producción (Argentina) and by INFONA (Paraguay) were used. From the resulting layer that combined the stable samples map, and the references maps of forest, forest plantation and perennial crops, a subset of 2.000 samples for each subregion were randomly generated and balanced for each class based on the class cover percentage. A minimum of 200 samples used to rare classes that does not cover at least 10% of the region area.

5.5. Complementary samples

The need for complementary samples was evaluated by visual inspection and by comparing the output of the preliminary classification with both Landsat and high-resolution images available in GEE. 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 in the 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.

5.6. Final classification

Final classification was performed for all subregions 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.

6. 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.

6.1. Gap fill filter

First, a spatial integration between subregions was made, where the subregion classifications were merged in a unique map. A hierarchical overlap of each mapped class was considered according to specific prevalence rules. The integration process was made on a pixel-by-pixel basis, where the classes identified with a less category number (ID) prevalence over other highest. Second, a no-data values (“gaps”) filter was applied. Because theoretically the no-data values are not allowed, it was replaced by the temporally nearest valid classification. In this procedure, if no “future” valid position was available, then the no-data value was 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.

6.2. Spatial filters

The spatial filter avoids unwanted modifications to the edges of the pixel groups, so two spatial filters were applied. First, a mode filter was applied, by calculating the majority class in a moving window of 3x3 pixels around the focal pixel. This filter was applied to all classes, except for forest (ID=3), non-vegetated areas (ID=22) and water (ID=33). Second, we built a spatial filter 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 did not share connections to a predefined number of identical neighbors were considered isolated. In this filter, at least six connected pixels were 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).

On the other hand, we observed that edge pixels of patches classified as forest (ID=3) were classified as Te (ID=65). Likewise, areas of open forest or with less density of trees were included within that class. Due to this, a variance filter was applied, by calculating the variance in a moving window of 5x5 pixels around the focal pixel, only considering the classes forest (ID=3) and Te (ID=65). When a variance greater than 0 was detected, the pixel class was replaced by forest class (ID=3).

6.3. Frequency filters

We applied two frequency filters depending on the class considered. First, we quantified the frequency with which a pixel was classified as wetland and grassland over the 40 years and replaced the pixel class with the most frequent one. Secondly, a frequency filter was applied only in pixels that were considered “stable native vegetation” (at least 33 years as [3, 11, 12]). If a “stable native vegetation” pixel is at least 80% of years of the same class, all years are changed to this class when the land use and cover has not been classified as anthropogenic (i.e., pasture, annual crops, perennial crop or forest plantation). The result of these frequency filters is a classification with more stable classification between native classes. Another important result is the removal of noises in the first and last year in the

classification. This filter was applied to all classes, except for non-vegetated areas (ID=22) and water (ID=33).

6.4. Grassland filter

We observed that certain areas originally with forest formation (ID=3) were replaced by pastures (ID=15) but in some cases it was classified as grassland (ID=12). Due to this, we considered that in those areas where after a forest there were forage use, they were reclassified as pasture.

6.5. Temporal filters

The temporal filter uses the subsequent years to replace pixels that has invalid transitions. In the first process the filter looks covers (3, 11, 12, 15, 19, 22, 33 or 36) that is not this class in 1985 and is equal in 1986 and 1987 and then corrects 1985 value to avoid any regeneration in the first year. In the second process, the filter looks pixel value in 2024 that is not those classes, and they were in 2023 and 2022. In these cases, the class value in 2024 is then converted to any of those classes, as appropriate, to avoid any regeneration in the last year. The third 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 was applied in this order: [33, 3, 11, 12, 15, 19, 22, 36]. The last process is similar to the third process, but it is a 4- and 5-years moving window that corrects all middle years. Specifically for the Forest Plantation class, a moving window of more years was considered (up to 12 years) because during the first years the forest plantations were classified as non-vegetated areas, perennial crops, annual crops, or pastures. Analogously, a moving window of up to 7 years was considered for perennial crops, where during the first years use to confuse with non-vegetated areas, annual crops, or pastures.

6.6. Incident filter

An incident filter was applied to remove pixels that change too many times in the 40 years. All pixels that change more than six times and is connected to less than 33 pixels that also changes more than six times is replaced by the mode value. This avoids changes in the border of the classes.

7. References

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

- Buchhorn, M., Lesiv, M., Tsendbazar, N. E., Herold, M., Bertels, L., & Smets, B. (2020). Copernicus global land cover layers—collection 2. Remote Sensing, 12(6), 1044.

- Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science 342 (15 November): 850-53. 10.1126/science.1244693 Data available on-line at: https://glad.earthengine.app/view/global-forest-change.

- Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., & Brumby, S. P. (2021, July). Global land use/land cover with Sentinel 2 and deep learning. In 2021 IEEE international geoscience and remote sensing symposium IGARSS (pp. 4704-4707). IEEE.

- Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., ... & Hofton, M. (2021). Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sensing of Environment, 253, 112165.