MapBiomas Atlantic Forest Trinational

Collection 1

Version 1

General coordinator
Marcos Reis Rosa

Local coordinators
Mayra Milkovic
Alejandra Gill

Technical coordinators
Pablo Baldassini
José Serafini

Technical assistants
Ana Eljall
Sol Gonzalez
Patricia Insfrán
Andrea Garay

Technical support
Ariel Insaurralde
Juan Pablo Zurano
Ignacio Minoli
Andrés Leszczuk
Martín Orona
Damian Loran

1. Introduction

1.1. Scope and content of the document

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 and Paraguay from 2000 to 2019 of the MapBiomas Collection 1. 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.

1.2. Region of Interest

MapBiomas Atlantic Forest Trinational was created to produce LULC annual maps for the Atlantic Forest corresponding to Argentina and Paraguay territories. 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 and located in the south border of Atlantic Forest was included (Figure 1). The study area was divided in 11 homogenous subregions to reduce confusion of samples and classes, as well as to allow a better balance of samples and results. Six of them corresponded to Argentina and 5 to Paraguay (Figure 1). A total of 249.830 km^2 was considered.

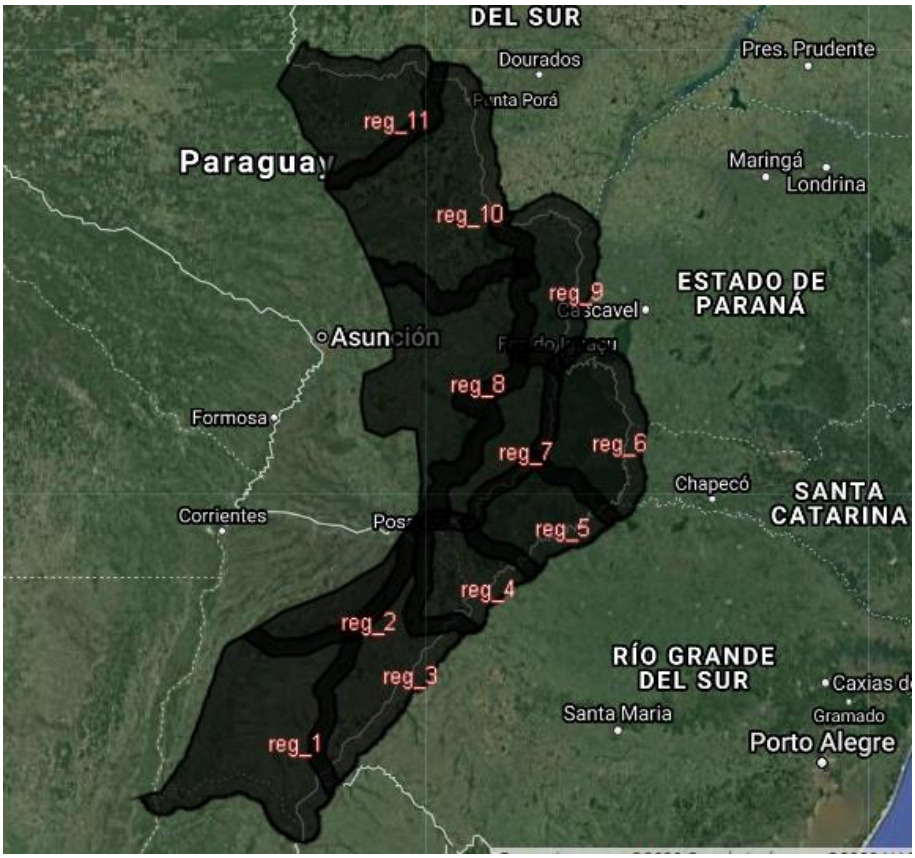

Figure 1. Area corresponded to *MapBiomas Atlantic Forest Trinational* that include Atlantic Forest of Argentina and Paraguay territory and the northeast portion of the flooded grassland and savannas corresponding to Argentina.

2. Remote Sensing Data

2.1. Landsat Collection

The imagery dataset used in the *MapBiomas Atlantic Forest Trinational* Collection 1 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 Trinational* Collection 1 has used Tier 1 from USGS and 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 20 scenes were used for cover the entire region, where each of them is totally or partially within the area (Figure 2). For each year we used images from the best Landsat available:

- 2000 to 2002 – Landsat 7
- 2003 – Landsat 5 and 7
- 2004 to 2010 – Landsat 5
- 2011 – Landsat 5 and 7
- 2012 – Landsat 7
- 2013 to 2019 – Landsat 8

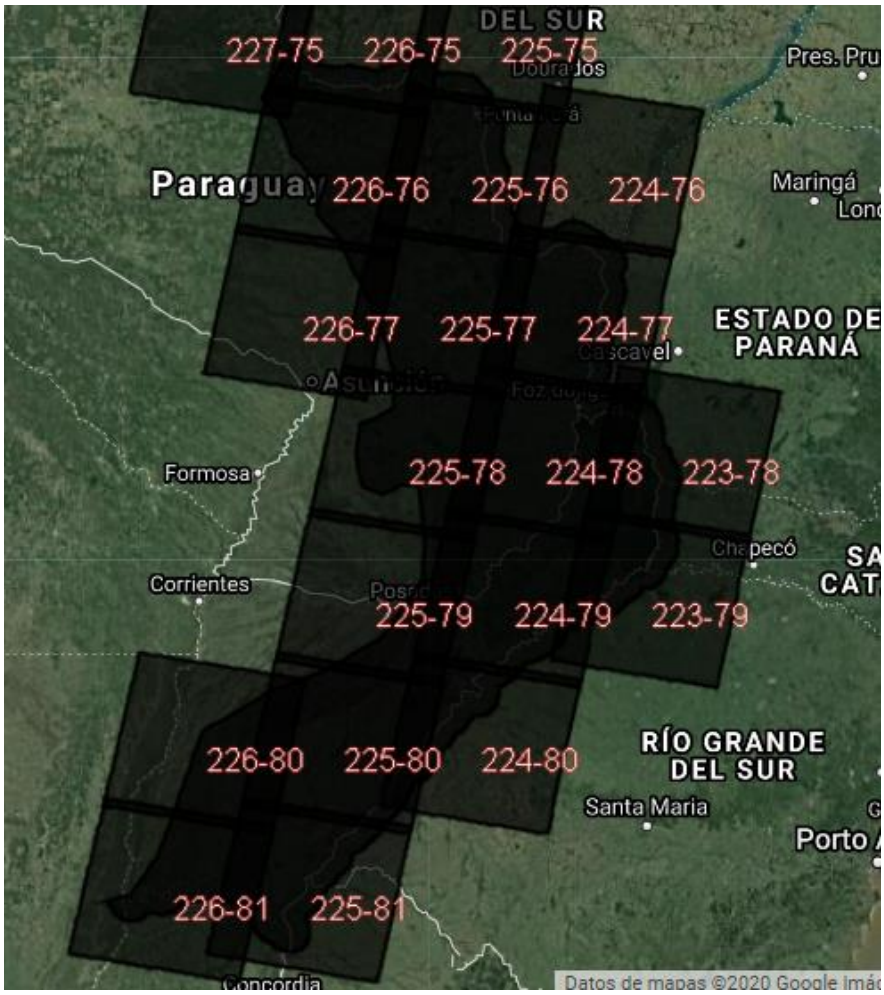

Figure 2. Path/row of Landsat scenes used in *MapBiomas Atlantic Forest Trinational* to generate annual LULC classifications for the period 2000-2019.

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 1 are presented in the Figure 3 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 1. The MapBiomas annual LULC maps were used to derived the transition analysis (with spatial filter application) and statistics. The statistical analysis covered different spatial categories, such as subregion, state and municipality.

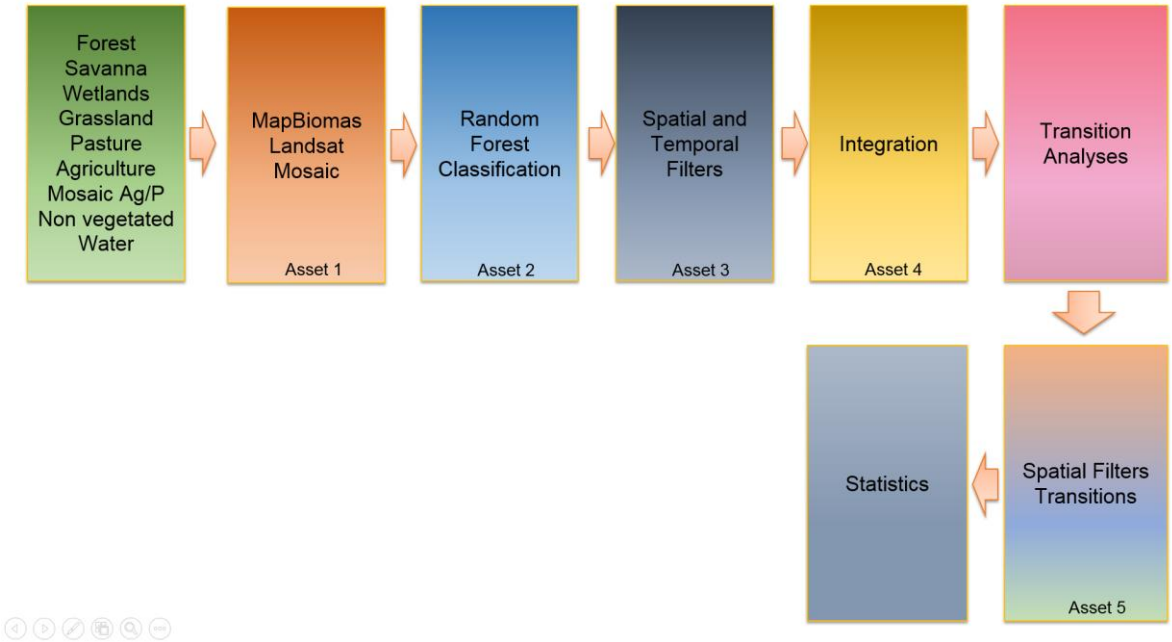

Figure 3. Methodological steps of Collection 1 to implement MapBiomas algorithms in the Google Earth Engine.

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 is 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 Trinational* Collection 1 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 Trinational* Collection 1 (2000-2019) for the LULC classification are indicated with X.

5. Classification of LULC

The production of the Collection 1, with land use and land cover annual maps for the period 2000-2019 in each subregion, included a) manually drawn polygons of LULC classes temporally invariables 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 4).

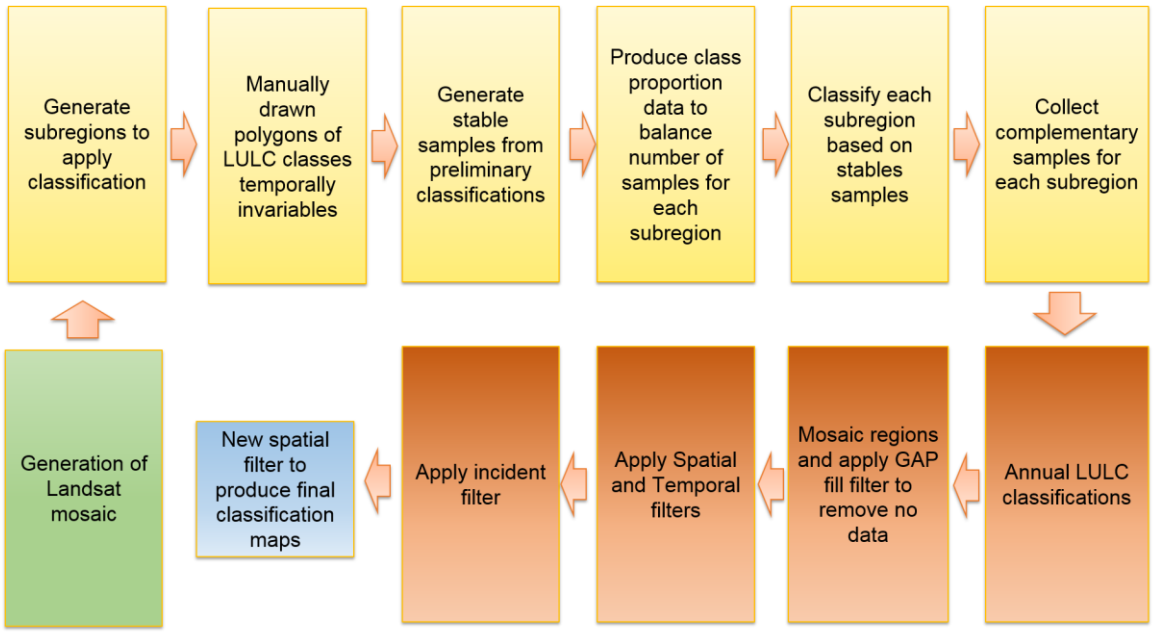

Figure 4. Classification process of Collection 1 in the *MapBiomas Atlantic Forest Trinational.*

5.1. Classification scheme

The digital classification of the Landsat mosaics for the *MapBiomas Atlantic Forest Trinational* Collection 1 aimed to individualize a subset of ten land use and land cover classes: Forest (3), Savanna (4), Wetland (11), Grassland (12), Pasture (15), Annual crops (19), Mosaic of agriculture or pasture (21), Non vegetated areas (22), Water (33) and Perennial crops (36) (Table 2).

Legend class of Collection 1	Numeric ID	Color
1.1.1 Forest Formation	3	
1.1.2 Savanna Formation	4	
1.2 Forest Plantation	9	
2.1 Wetland	11	
2.2 Grassland	12	
3.1 Pasture	15	
3.2.1 Annual Crops	19	
3.2.2 Perennial Crop	36	
3.3 Mosaic of Agriculture and Pasture	21	
4. Non vegetated area	22	
5. Water	33	

Table 2. Land use and land cover (LULC) categories considered for digital classification of Landsat mosaics for the *MapBiomas Atlantic Forest Trinational* Collection 1.

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 40 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 along the 20 years of Collection 1, 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 20 years as backdrop and graphs with the temporal behavior of spectral indices per pixel were used to stablish the LULC class.

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 20 years with the Random Forest algorithm, running 40 iterations. A total of 20 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 17 times in the period 2000-2019 to be considered as stable, i.e. a pixel should remain with that class a minimum of 17 years to be eligible as a stable sample. A layer of pixels with a stable classification along the 20 years was then generated by applying such threshold. From the resulting layer of stable samples, a subset 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 20 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 were 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) prevalenced over other highest. Second, a no-data values (“gaps”) filter was apply. Because theoretically the no-data values are not allowed, it were 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 filter

The spatial filter avoids unwanted modifications to the edges of the pixel groups, 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 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 6pixels (~0,5 ha).

6.3. Temporal filter

The temporal filter uses the subsequent years to replace pixels that has invalid transitions. In the first process the filter looks any natural cover (3, 4, 11, 12, 33) that is not this class in 2000 and is equal in 2001 and 2002 and then corrects 2000 value to avoid any regeneration in the first year. In the second process the filter looks pixel value in 2019 that is not 19, 15 or 21 (Annual crops, Pasture or Mosaic of Agriculture or Pasture) and is equal to 19, 15 or 21 in 2017 and 2018. The value in 2019 is then converted to 19, 15 or 21 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, 4, 19, 15, 21, 3, 12, 11]. The last process is similar to the third process, but it is a 4- and 5-years moving window that corrects all middle years.

6.4. Frequency filter

Frequency filters were applied only in pixels that were considered “stable native vegetation” (at least 16 years as [3, 4, 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. 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.

6.5. Incident filter

An incident filter was applied to remove pixels that change too much times in the 20 years. All pixels that changes 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.