MapBiomas Patagonia

Collection 1

Version 1

General coordinators
Ana Eljall

Technical coordinators
Camilo Bagnato

Technical support
Paula Torre Zaffaroni

Team

Thomas Kitzberger	Sol Rodriguez
Fernando Salvare	Micaela Abrigo
Juan Gowda	Camila Mateo
Daniela Arpigiani	Pablo Aceñolaza
Emilia Giustiniani	Virginia Piani
Micaela Gambino	Gonzalo Dieguez Gaviola
Fabio Daniel Trinco	Facundo Oddi
Romina Gonzalez Musso	Carolina Mansilla Ferro

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 Patagonia from 1998 to 2022.

1.2. Overview

The MapBiomas initiative combines people, algorithms, satellite information and large-scale processing in a methodology that has revolutionized the operational large-scale generation of LULC maps. MapBiomas provides an ideal environment to enhance and share skills and abilities by collaborators from different locations and perspectives but with the mutual core value of learning by doing. Thanks to Google Earth Engine and open source technology it was possible to access and process large-scale datasets of satellite imagery such as the one generated by the MapBiomas project.

MapBiomas collections aim to contribute to developing a fast, reliable, collaborative, and low-cost method to process large-scale datasets and generate historical time series of land use and land cover annual maps. All data, classification maps, codes, statistics, and further analyses are openly accessible through the MapBiomas Platform. This is possible thanks to i) the Google Earth Engine platform, which provides access to data, image processing, standard algorithms, and the cloud computing facilities; ii) the freely available Landsat time-series dataset; and iii) the MapBiomas collaborative network of organizations and experts that share knowledge and mapping tools. MapBiomas maps are an evolving product and subsequent collections will be available in the future.

1.3. Region of Interest

The Patagonia initiative was created to produce annual land use and land cover (LULC) maps for south of Argentina. The region of interest extends over 838.206 km², including most of the Patagonian forests or Valdivian temperate forests and Magellanic subpolar forests, Patagonian steppe, the southern part of the Monte and South Atlantic Islands ecoregions (Burkart et al., 1999; Olson et al., 2001). The study area was divided into 13 subregions to minimize confusion between samples and classes, and to ensure a more balanced distribution of samples and results (Figure 1).
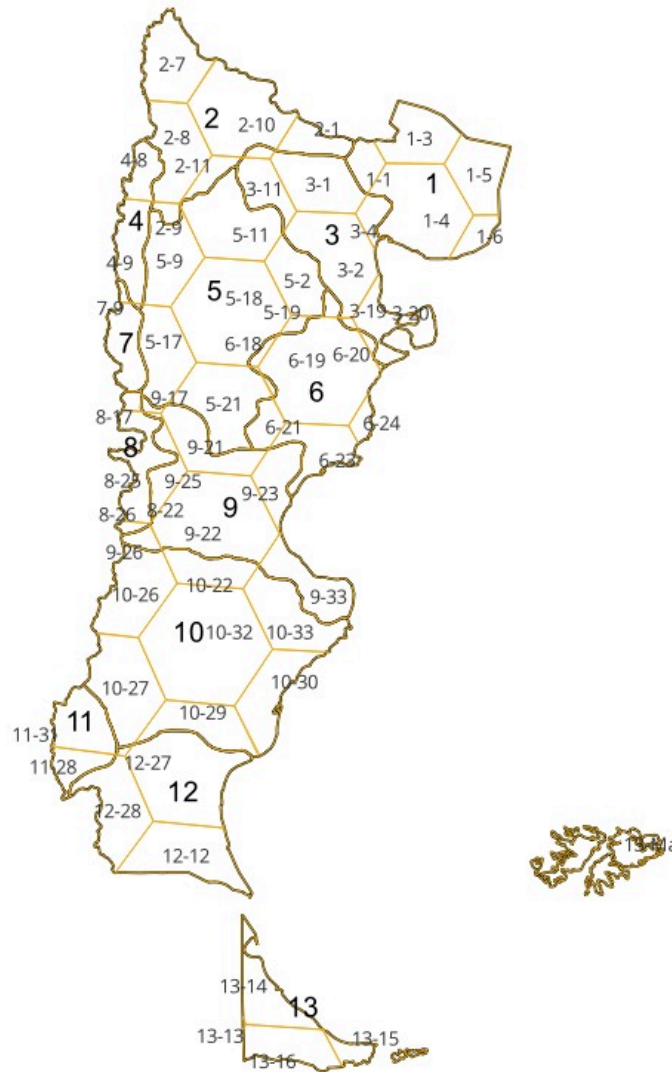
Figure 1. Region mapped by the Patagonia initiative and delimitation of subregions and hexagons cells where the sample was conducted.

2. Overview of methodological process

The methodological steps of *Patagonia* initiative Collection 1 are presented in Figure 2 and detailed below. Annual Landsat mosaics based on yearly periods and the spectral feature inputs derived from the Landsat bands (i.e., feature space) to run a random forest-based classification were generated by the MapBiomas Brazil team. The *Patagonia* team defined the classes of land use or cover to be mapped (hereafter, map legend) and the subregions of the region of interest. Through visual interpretation and phenological signatures

analysis by the whole team, temporally stable training samples were acquired for each class and subregion. The training data set was derived from the spectral feature inputs over the samples, and a pre-classification was generated with a random forest classifier. The pre-classification was used to identify stable areas where randomly select stable samples through stratification based on subregion representativeness (i.e., relative class area). A new random forest classifier was run based on random, stratified, stable samples. Interpreters further collected complementary samples in areas or for classes where errors were evident, which were added to the stable samples for a new classification. Following that, spatial and temporal filters were applied to remove classification noise and stabilize the classifications. Finally, regional land use/land cover patterns were evaluated and corrected using reclassification rules designed specifically for each class and region where errors were detected.
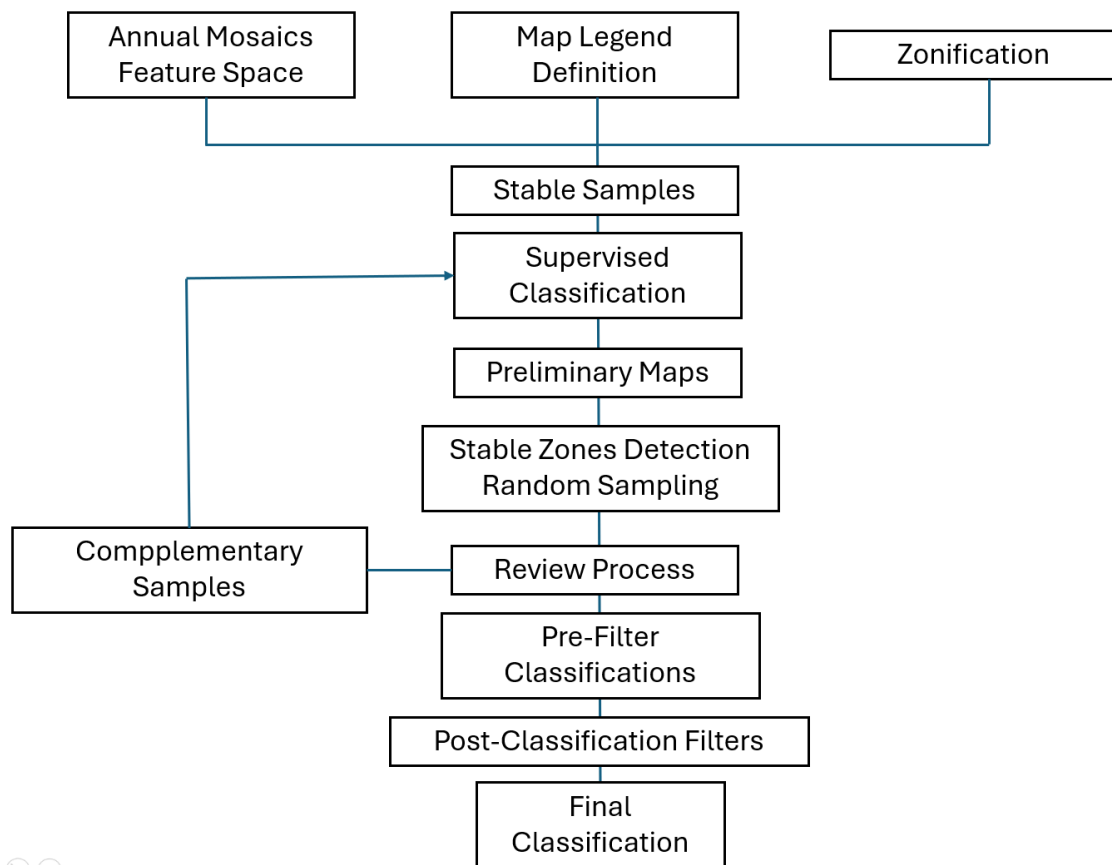

Figure 2. Methodological scheme of *Patagonia* initiative Collection 1

3. Remote Sensing Data

3.1. Landsat Collection

The imagery dataset used in the *Patagonia* 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 *Patagonia* 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. For each year we used images from the best Landsat available:

- 1998 to 1999 – Landsat 5

- 2000 to 2002 – Landsat 7

- 2003 to 2011 – Landsat 5

- 2012 – Landsat 7

- 2013 to 2022 – Landsat 8

In the case of the Patagonian region, the time frame is from 1998 to 2022 due to the absence of quality landsat images before 1998.

3.2. 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.

3.3.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 Patagonia* Collection 1 comprised a subset of 31 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.

	band or index name	formula	Reducer					
			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					
	red	B3 (L5 e L7); B4 (L8)	X					
	nir	B4 (L5 e L7); B5 (L8)	X					
	swir1	B5 (L5 e L7); B6 (L8)	X					
	swir2	B7 (L5); B8 (L7); B7 (L8)	X					
	temp	B6 (L5 e L7); B10 (L8)						
index	ndvi	(nir - red)/(nir + red)	X	X	X			
	evi2	(2.5 * (nir - red)/(nir + 2.4 * red + 1)	X					
	cai	(swir2 / swir1)						
	ndwi	(nir - swir1)/(nir + swir1)	X		X	X		
	gcvi	(nir / green - 1)	X	X	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					
	textG	('median_green').entropy(ee.Kernel.square({radius: 5}))	X					
fraction	gv		X					
	npv		X				X	
	soil		X				X	
	cloud							
	shade	100 - (gv + npv + soil + cloud)	X					
MEM index	gvs	gv / (gv + npv + soil + cloud)	X		X			
	ndfi	(gvs - (npv + soil))/(gvs + (npv + soil))	X				X	
	sefi	(gv+npv_s - soil)/(gv+npv_s + soil)	X				X	
	wefi	((gv+npv) - (soil+shade)) /((gv+npv) + (soil+shade))						
	fns	((gv+shade) - soil) / ((gv+shade) + soil)						
slope		ALOS DSM: Global 30m	X					

Table 1. List and reference of bands, fractions and indices available in the feature space (green color). The feature space subset considered by *MapBiomas Patagonia* Collection 1 (1998-2022) for the LULC classification are indicated with X.

4. Classification of LULC

The production of the Collection 1, with land use and land cover annual maps for the period 1998-2022 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).
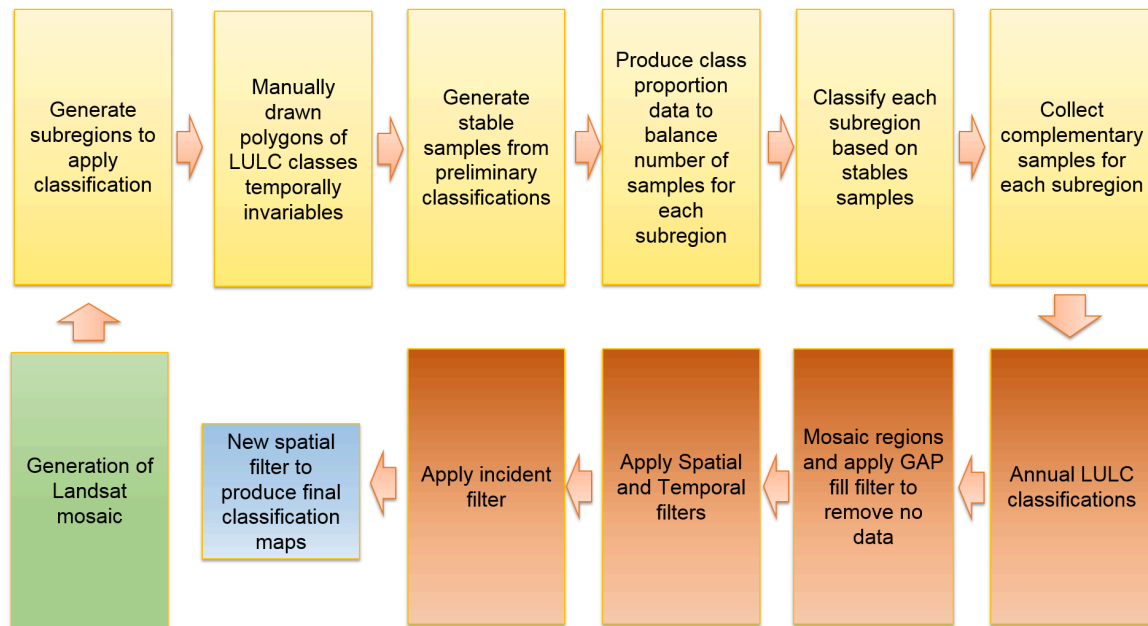
Figure 4. Classification process of Collection 1 in the *MapBiomas Patagonia*

4.1. Legend

The digital classification of the Landsat mosaics for the *Patagonia initiative* Collection 1 aimed to individualize a subset of 11 land use and land cover classes: Forest (3), Shrubland (67), Wetland(11), Grassland (12),Steppe (63), Mosaic of agriculture and pasture(21), Forest Plantation(9), Non vegetated areas (22),Rivers, Lakes and Oceans (33), Glacier / Snow-Ice (34), Non Observed(27)(Table 2). Not all classes were identified in each subregion.
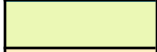
Legend class of Collection 1	ID	Hexadecimal Code	Color
1.1 Forest	3	#1f8d49	
1.2 Shrubland	66	#a89358	
2.1 Grassland	12	#d6bc74	
2.2 Wetland	11	#519799	
2.3 Steppe	63	#ebf8b5	
3.1 Mosaic of agriculture and pasture	21	#ffefc3	
3.2 Forest Plantation	9	#7a5900	
4. Non vegetated area	22	#d4271e	
5.1 Rivers, Lakes and Oceans	33	#2532e4	
5.2 Glacier / Snow-Ice	34	#93dfe6	
6 Non Observed	27	#ffffff	

Table 2. Land use and land cover (LULC) categories considered for digital classification of Landsat mosaics for Patagonia.

4.2. 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 100 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 22 of the 25 years of Collection 1, so named "stable samples". The stable areas were identified through annual preliminary classification made using random pixels selected from manually drawn polygons. For this, false-color composites of the Landsat mosaics for all the 25 years as backdrop and graphs with the temporal behavior of spectral indices per pixel were used to establish the LULC class.

4.3. Preliminary classification

From manually drawn polygons, a subset between 200 and 2000 pixels per class was randomly selected and they were used as training areas to classify each of the 25 years with the Random Forest algorithm, running 100 iterations. A total of 25 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.

4.4. 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 22 times in the period 1998-2022 to be considered as stable, i.e. a pixel should remain with that class a minimum of 22 years to be eligible as a stable sample. A layer of pixels with a stable classification along the 25 years was then generated by applying such threshold. A subset of samples for each subregion were randomly generated and balanced for each class based on the class cover percentage. A minimum of 200 samples used for rare classes that does not cover at least 10% of the region área and a maximum of 2000 samples for the most abundant classes.

4.5. 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 25 years during the polygon drawing. Based on 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.

4.6. Final classification

Final classification was performed for all subregions and years with stable and complementary samples. All years used the same subset of samples to train the classifier with the mosaic of the specific year that was classified.

5. Post-processing

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.

5.1. Gap fill filter

First, a spatial integration between subregions was made, where the subregion classifications were merged in a unique map. 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.

5.2. 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. 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 11 pixels (~1 ha).

5.3. Temporal filters

The temporal filter uses the subsequent years to replace pixels that has invalid transitions. In the first process the filter looks covers that is not this class in 1998 and is equal in 1999 and 2000 and then corrects 1998 value to avoid any regeneration in the first year. In the second process, the filter looks pixel value in 2022 that is not those classes, and they were in 2020 and 2021. In these cases, the class value in 2022 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: [3, 63, 67, 12, 21, 9, 11, 22, 34, 27]. The last process is similar to the third process, but it is a 4- and 5-years moving window that corrects all middle years.

5.4. Last spatial filter

After applying the filters detailed above, a spatial filter with the same parameters as the first one was applied to remove potential spatial artifacts produced by the temporal filters.

6. References

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

Burkart, R., N. Bárbaro, R. O. Sánchez, and D. A. Gómez. 1999. Ecorregiones de la Argentina, APN, PRODIA. Pp. 43.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., ... & Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. BioScience, 51(11), 933-938.