MapBiomas Pampa Argentina

Collection 5
(integrating MapBiomas Argentina Collection 2)

2025

Coordinator

Diego de Abelleyra (Instituto de Clima y Agua, INTA)

Team

Magdalena Bozzola (MapBiomas)
Karina Echevarria (ICBIA -CONICET-Universidad Nacional de Río Cuarto)
María del Rosario Iturralde Elortegui (INTA)
Mariano Oyarzabal (Conicet-INTA)
Jonathan Rodríguez Pérez (INTA)
Sofía Sarrailhe (MapBiomas)
Karina Zelaya (INTA)

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 Pampa, Espinal and Paraná river delta biomes of Argentina from 1985 to 2024 (Collection 5). Maps generated for this collection are integrated into the MapBiomas Argentina Collection 2 maps. 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 Pampa initiative includes the phytogeographic regions of Pampa, *Espinal* and Paranaense phytogeographic provinces (**Figure 1**). The total mapped area was 72.24 million hectares (Mha), being 46.31 Mha in the Pampa (64%), 22.67 Mha in the Espinal (31%) and 2.28 Mha in the Paraná river delta (3%).

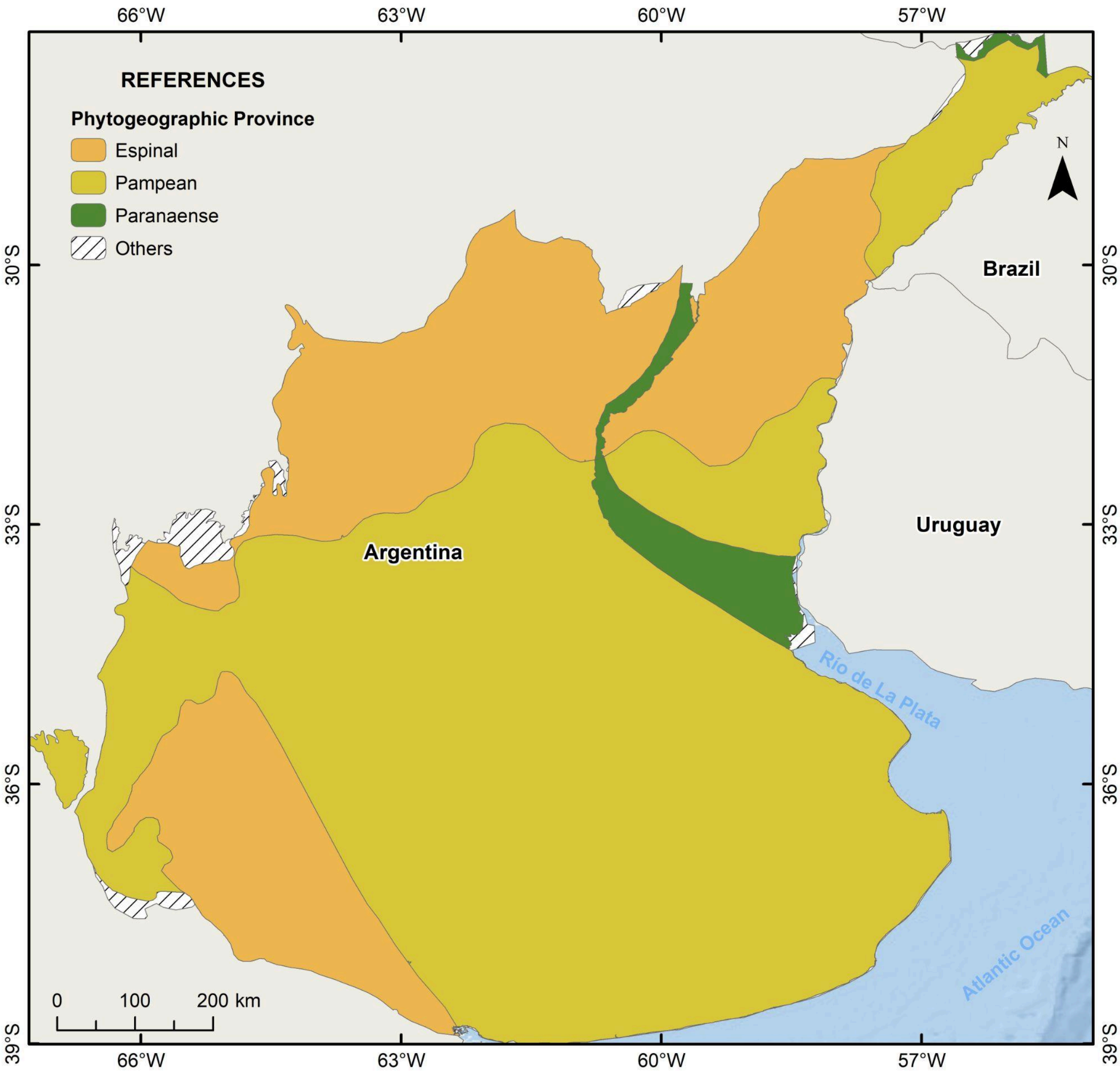

Figure 1. Region of interest mapped in the Argentine MapBiomas Pampa initiative (collection 5), including the areas of the Pampa, Espinal, and Parana river delta.

2 GEOGRAPHICAL UNITS OF CLASSIFICATION

The classification process was carried out in smaller spatial units. These units correspond to subregional homogeneous zones based on several criteria including vegetation types, land use patterns, climatology, etc. The study area was finally divided into thirteen homogeneous zones (**Figure 2**). The purpose of these homogeneous units of classification was to try to reduce samples and class confusion and to allow a better balance of samples and results to improve accuracy.

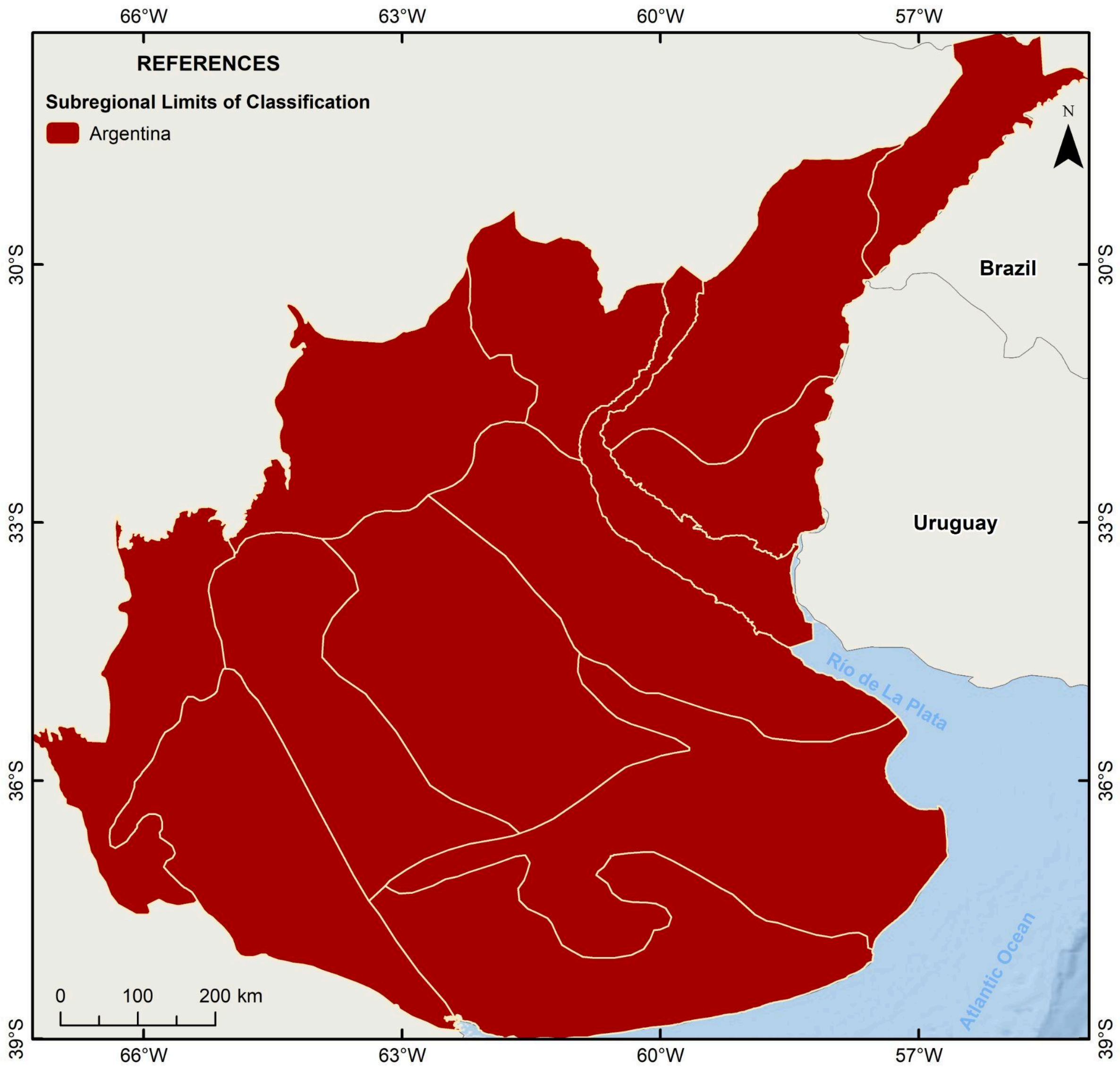

Figure 2. Homogeneous subregions used in the classification process for the Pampa Argentina initiative.

3 REMOTE SENSING DATA

3.1 Landsat Collection

The imagery dataset used in the *MapBiomas* Pampa Argentina Collection 5 was obtained from 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 m-pixel resolution were accessible via Google Earth Engine, and were provided by NASA and USGS. The MapBiomas Pampa Argentina Collection 5 used Collection 2, Tier 1 Landsat Surface Reflectance

products from USGS, 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 49 scene boundaries were used to cover the entire region, where each of them is totally or partially within the area.
According to the year and the quality of available images, a specific Landsat collection was selected:

- from 1985 to 1999: Landsat 5,
- year 2000: Landsat 7,
- years 2001, 2002 and 2012: Landsat 7,
- from 2003 to 2011: Landsat 5,
- from 2013 to 2024: Landsat 8.

3.2 Landsat Mosaics

All Landsat scenes were merged and clipped within standardized spatial units for data processing, hereafter called ‘charts’, based on the grid of the World International Chart to the Millionth, at the 1:250,000 scale level. A total of 74 charts were used to cover the biome (**Figure 3**). Each chart sets the geographical limits to build up the temporal and spatial Landsat mosaics and to proceed with digital classification procedures. Each geographical classification unit was generated by merging the correspondent mosaic charts.

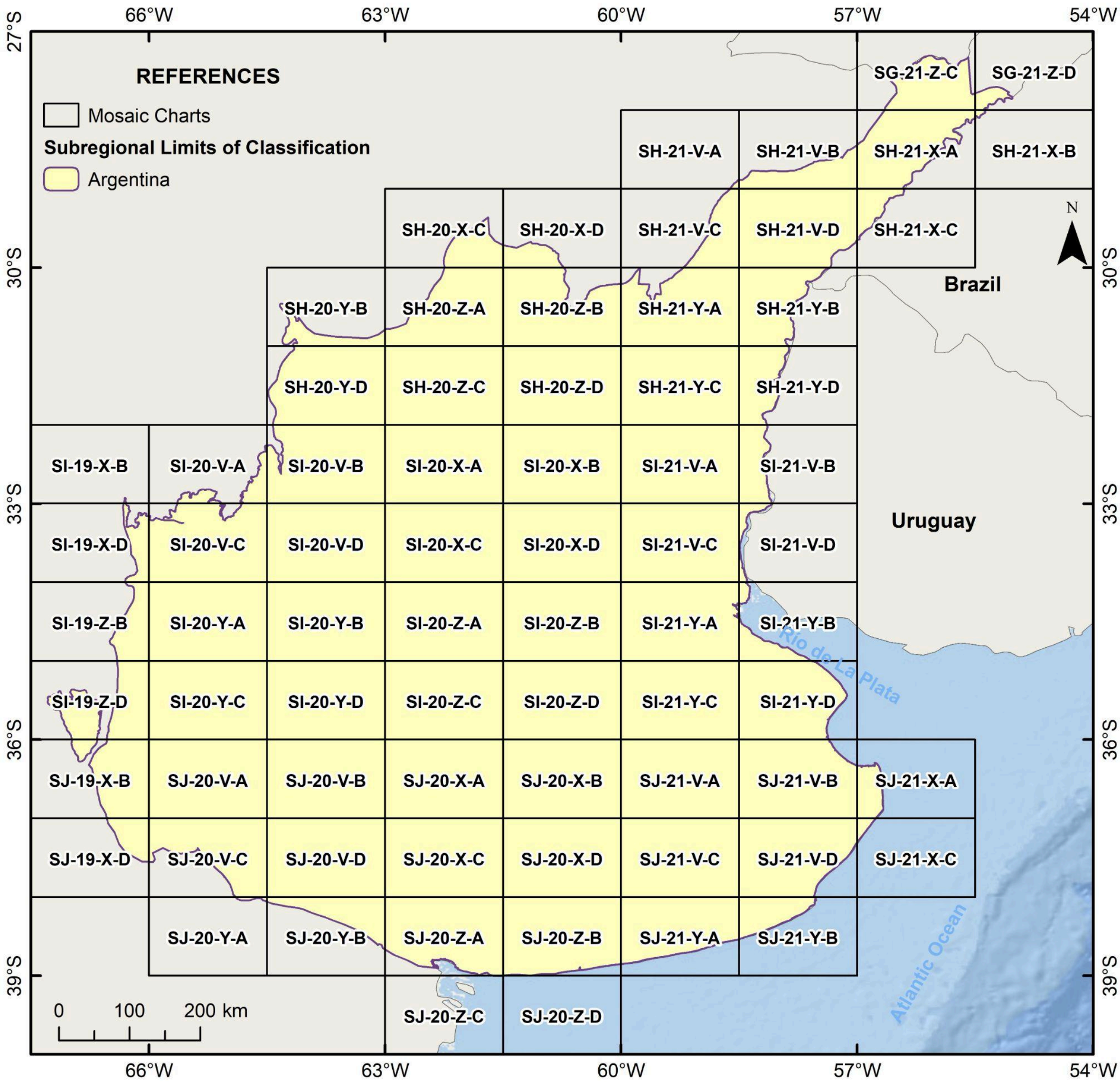

Figure 3. Charts scheme used to build up Landsat mosaics used throughout the classification process.

3.3 Definition of the temporal period

The mosaics were generated by the composition of pixels in each set of images for a certain time period. Two periods were used: 1) yearly based, considering all available images from January to December of each year, and 2) trimestral based, considering a short period considering the balance between the probability of maximizing the differences in classes spectral behavior and the availability of cloud-free images. This period was determined from May to July of each year. Nevertheless, for some years this period was adapted (extended one to three months) for each chart according to the availability of cloud-free images. For example, if during the three-months period a cloud free mosaic could not be generated, the trimestral period was extended to four, five or six months to get a complete or almost complete mosaic.

For the selection of Landsat scenes a threshold of 90% of cloud cover was applied (i.e., any available scene with up to 90% of cloud cover was accepted). This limit was established based on visual analysis, after many trials observing the results of the cloud removing/masking algorithm.

4 CLASSIFICATION

4.1 Overview of methodological process

The methodological procedures of Collection 5 included several steps (**Figure 4**).

The first step was to generate annual Landsat image mosaics based on yearly and trimestral metrics. The second step was to generate a new selection of temporally stable samples derived from the stable areas of the maps of Collection 4. Stable areas were defined in sub-periods of 10 years-length (1985-1994, 1995-2004, 2005-2014 and 2015-2024). Then, the spectral feature inputs derived from the Landsat bands were extracted and associated to each sample point. Once the samples for each LULC class were selected for each of the subregions, it was possible to adjust the training data set according to its statistical needs. The number of training samples for each class was defined initially according to the proportion of the area of each class taken from Collection 4 and its variation over time (sample size balance). Additionally, to improve the classification results, complementary samples were generated, defining georeferenced points of different classes by visual interpretation of historical satellite images (high and very high resolution images) and time series of vegetation indices. In addition, complementary samples generated for Collections 3 and 4 were included when improvements in the classification were observed. Based on the adjusted training data set, a supervised classification using the random forest algorithm was run.

Following that, gap, spatial, temporal and frequency filters were applied to remove classification noise and stabilize the classification. The LULC maps of each subregion were integrated to generate the final map of Collection 5.

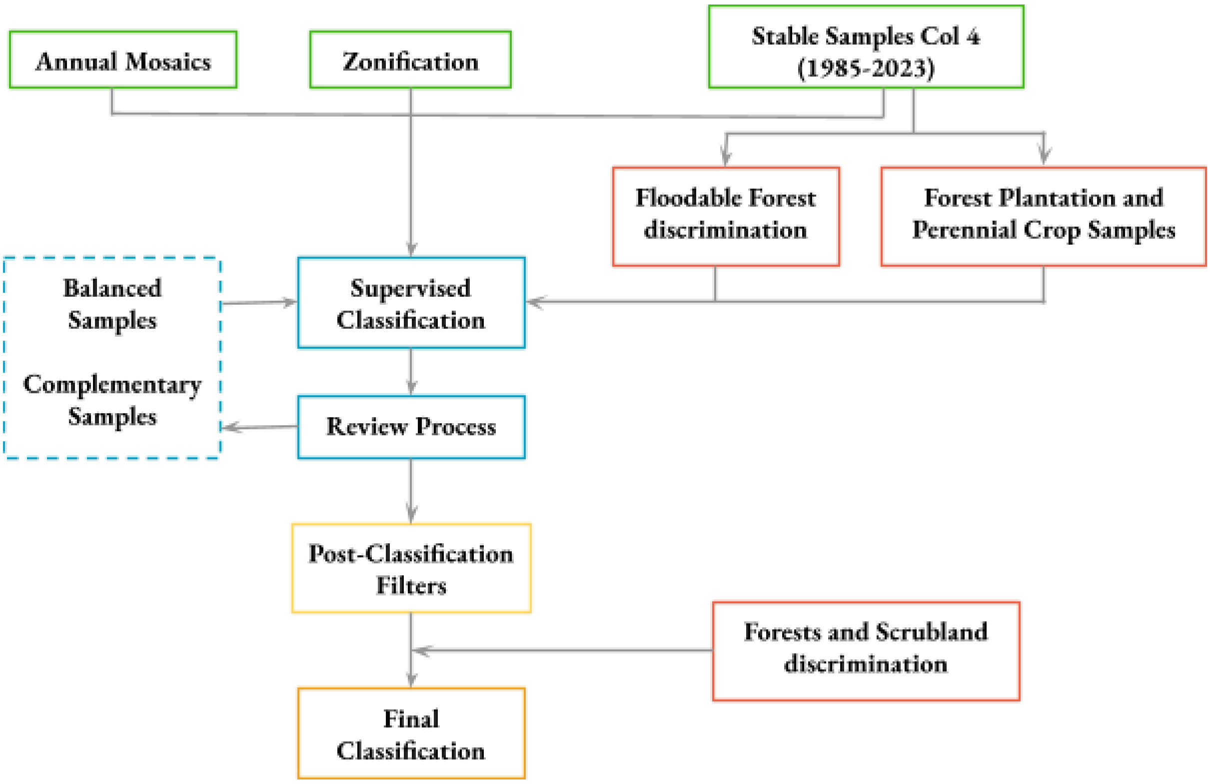

Figure 4. Classification process of Collection 5 in the MapBiomas Pampa Argentina initiative for the period 1985-2024.

4.2 Map Legend

The classification for the MapBiomas Pampa Argentina initiative using Landsat mosaics included fourteen land use and land cover (LULC) classes (**Table 1**): Forest formation (3), Savanna formation (4), Flooded forests (6), Closed shrubland (66), Open shrubland (76), Wetland (11), Grassland (12), Pasture (15), Silviculture (9), Temporary crop (19) , Perennial crop (36), Industrial crops (78), Non vegetated area (22) and River, lake or ocean (33). A full description of the legend is described in the document Legend Description of MapBiomas Argentina Collection 2.

Table 1. Land cover and land use classes considered for digital classification of Landsat mosaics for the MapBiomas Pampa Argentina - Collection 5.

Legend class of Collection 5	Numeric ID	Color
1.1. Forest formation	3	
1.2. Savanna formation	4	
1.3 Flooded Forest	6	
1.4 Closed shrubland	66	
1.5 Open shrubland	77	
2.1. Wetland	11	
2.2. Grassland	12	
3.1. Pasture	15	
3.2. Temporary crop	19	
3.3. Forest plantation	9	
3.4 Perennial Crop	36	
3.5. Industrial crops	78	
4. Non vegetated area	22	
5.1. River, lake or ocean	33	

4.3 Annual Mosaics

The total available bands of the MapBiomas Pampa Argentina feature space is composed of 93 input variables, including the original Landsat bands, fractional and textural information derived from these bands (**Table 2**). As mentioned above, some bands were generated with annual mosaics and others with trimestral mosaics (mosaic months). Reducers were used to generate temporal features such as:

● Median: median of the pixel values of the best mapping trimestral period defined.

● Median_dry: median of the quartile of pixels with the lowest NDVI values of each year.

● Median_wet: median of the quartile of pixels with the highest NDVI values of each year.

● 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.

Table 2. Variables included in the feature space used in the classification of the Mapbiomas Pampa Argentina Landsat image mosaics. Collection 5 (1985-2024).

ID	Variable	Description	Statistics	Temporal range	Script acronym	Group
0	Evi 2	Enhanced Vegetation Index 2	amplitude	mosaic months	evi2_amp	Spectral index
1	Gv	Green vegetation fraction	amplitude	mosaic months	gv_amp	Spectral Mixture Modeling
2	Ndfi	Normalized Difference Fraction Index	amplitude	mosaic months	ndfi_amp	Spectral Mixture Modeling
3	Ndvi	Normalized Difference Vegetation Index	amplitude	mosaic months	ndvi_amp	Spectral index
4	Ndwi	Normalized Difference Water Index	amplitude	mosaic months	ndwi_amp	Water Index
5	Soil	Soil fraction	amplitude	mosaic months	soil_amp	Spectral Mixture Modeling
6	Wefi	Woodland ecosystem fraction index	amplitude	mosaic months	wefi_amp	Fraction index
7	Blue	Landsat band	median	mosaic months	blue_median	Landsat band
8	Blue dry	Landsat band	median	year -first quartile	blue_median_dry	Landsat band
9	Blue wet	Landsat band	median	year – fourth quartile	blue_median_wet	Landsat band
10	Cai	Cellulose Absorption Index	median	mosaic months	cai_median	Spectral index
11	Cai dry	Cellulose Absorption Index	median	year -first quartile	cai_median_dry	Spectral index
12	Cloud	Cloud fraction	median	mosaic months	cloud_median	Spectral Mixture Modeling
13	Evi 2	Enhanced Vegetation Index 2	median	mosaic months	evi2_median	Spectral index
14	Evi 2 dry	Enhanced Vegetation Index 2	median	year -first quartile	evi2_median_dry	Spectral index
15	Evi 2 wet	Enhanced Vegetation Index 2	median	year – fourth quartile	evi2_median_wet	Spectral index
16	Gcvi	(nir/green – 1)	median	mosaic months	gcvi_median	Spectral index
17	Gcvi dry	(nir/green – 1)	median	year -first quartile	gcvi_median_dry	Spectral index
18	Gcvi wet	(nir/green – 1)	median	year – fourth quartile	gcvi_median_wet	Spectral index
19	Green	Landsat band	median	mosaic months	green_median	Landsat band
20	Green dry	Landsat band	median	year -first quartile	green_median_dry	Landsat band
21	Green wet	Landsat band	median	year – fourth quartile	green_median_wet	Landsat band
22	Gv	Green vegetation fraction	median	mosaic months	gv_median	Spectral Mixture Modeling
23	Gvs	GV / (100 - shade)	median	mosaic months	gvs_median	Spectral Mixture Modeling
24	Gvs dry	GV / (100 - shade)	median	year -first quartile	gvs_median_dry	Spectral Mixture Modeling
25	Gvs wet	GV / (100 - shade)	median	year – fourth quartile	gvs_median_wet	Spectral Mixture Modeling
26	Hallcover	Hall cover vegetation index	median	mosaic months	hallcover_median	Spectral index
27	Ndfi	Normalized Difference Fraction Index	median	mosaic months	ndfi_median	Spectral Mixture Modeling
28	Ndfi dry	Normalized Difference Fraction Index	median	year -first quartile	ndfi_median_dry	Spectral Mixture Modeling
29	Ndfi wet	Normalized Difference Fraction Index	median	year – fourth quartile	ndfi_median_wet	Spectral Mixture Modeling
30	Ndvi	Normalized Difference Vegetation Index	median	mosaic months	ndvi_median	Spectral index
31	Ndvi dry	Normalized Difference Vegetation Index	median	year -first quartile	ndvi_median_dry	Spectral index
32	Ndvi wet	Normalized Difference Vegetation Index	median	year – fourth quartile	ndvi_median_wet	Spectral index
33	Ndwi	Normalized Difference Water Index	median	mosaic months	ndwi_median	Water Index

ID	Variable	Description	Statistics	Temporal range	Script acronym	Group
34	Ndwi dry	Normalized Difference Water Index	median	year -first quartile	ndwi_median_dry	Water Index
35	Ndwi wet	Normalized Difference Water Index	median	year – fourth quartile	ndwi_median_wet	Water Index
36	Near Infrared (NIR)	Landsat band	median	mosaic months	nir_median	Landsat band
37	Near Infrared (NIR) dry	Landsat band	median	year -first quartile	nir_median_dry	Landsat band
38	Near Infrared (NIR) wet	Landsat band	median	year – fourth quartile	nir_median_wet	Landsat band
39	Npv	Non-photosynthetic vegetation fraction	median	mosaic months	npv_median	Spectral Mixture Modeling
40	Pri	Photochemical reflectance index	median	mosaic months	pri_median	Spectral index
41	Pri dry	Photochemical reflectance index	median	year -first quartile	pri_median_dry	Spectral index
42	Pri wet	Photochemical reflectance index	median	year – fourth quartile	pri_median_wet	Spectral index
43	Red	Landsat band	median	mosaic months	red_median	Landsat band
44	Red dry	Landsat band	median	year -first quartile	red_median_dry	Landsat band
45	Red wet	Landsat band	median	year – fourth quartile	red_median_wet	Landsat band
46	Savi	Soil-adjusted Vegetation Index	median	mosaic months	savi_median	Spectral index
47	Savi dry	Soil-adjusted Vegetation Index	median	year -first quartile	savi_median_dry	Spectral index
48	Savi wet	Soil-adjusted Vegetation Index	median	year – fourth quartile	savi_median_wet	Spectral index
49	Sefi	Savanna Ecosystem Fraction Index	median	mosaic months	sefi_median	Fraction index
50	Sefi dry	Savanna Ecosystem Fraction Index	median	year -first quartile	sefi_median_dry	Fraction index
51	Shade	Shade fraction	median	mosaic months	shade_median	Spectral Mixture Modeling
52	Soil	Soil fraction	median	mosaic months	soil_median	Spectral Mixture Modeling
53	Shortwave Infrared (SWIR) 1	Landsat band	median	mosaic months	swir1_median	Landsat band
54	Shortwave Infrared (SWIR) 1 dry	Landsat band	median	year -first quartile	swir1_median_dry	Landsat band
55	Shortwave Infrared (SWIR) 1 wet	Landsat band	median	year – fourth quartile	swir1_median_wet	Landsat band
56	Shortwave Infrared (SWIR) 2	Landsat band	median	mosaic months	swir2_median	Landsat band
57	Shortwave Infrared (SWIR) 2 dry	Landsat band	median	year -first quartile	swir2_median_dry	Landsat band
58	Shortwave Infrared (SWIR) 2	Landsat band	median	year – fourth quartile	swir2_median_wet	Landsat band

ID	Variable	Description	Statistics	Temporal range	Script acronym	Group
	wet					
59	Wefi	Woodland ecosystem fraction index	median	mosaic months	wefi_median	Fraction index
60	Wefi wet	Woodland ecosystem fraction index	median	year – fourth quartile	wefi_median_wet	Fraction index
61	Blue min	Landsat band	minimum	mosaic months	blue_min	Landsat band
62	Green min	Landsat band	minimum	mosaic months	green_min	Landsat band
63	Near Infrared (NIR) min	Landsat band	minimum	mosaic months	nir_min	Landsat band
64	Red min	Landsat band	minimum	mosaic months	red_min	Landsat band
65	Shortwave Infrared (SWIR) 1	Landsat band	minimum	mosaic months	swir1_min	Landsat band
66	Shortwave Infrared (SWIR) 2	Landsat band	minimum	mosaic months	swir2_min	Landsat band
67	Blue	Landsat band	standard deviation	mosaic months	blue_stdDev	Landsat band
68	Cai	Cellulose Absorption Index	median	mosaic months	cai_stdDev	Spectral index
69	Cloud	Cloud fraction	standard deviation	mosaic months	cloud_stdDev	Spectral Mixture Modeling
70	Evi 2	Enhanced Vegetation Index 2	standard deviation	mosaic months	evi2_stdDev	Spectral index
71	Gcvi	(nir/green – 1)	standard deviation	mosaic months	gcvi_stdDev	Spectral index
72	Green	Landsat band	standard deviation	mosaic months	green_stdDev	Landsat band
73	Gv	Green vegetation fraction	standard deviation	mosaic months	gv_stdDev	Spectral Mixture Modeling
74	Gvs	GV / (100 - shade)	standard deviation	mosaic months	gvs_stdDev	Spectral Mixture Modeling
75	Hallcover	Hall cover vegetation index)	standard deviation	mosaic months	hallcover_stdDev	Spectral index
76	Ndfi	Normalized Difference Fraction Index	standard deviation	mosaic months	ndfi_stdDev	Spectral Mixture Modeling
77	Ndvi	Normalized Difference Vegetation Index	standard deviation	mosaic months	ndvi_stdDev	Spectral index
78	Ndwi	Normalized Difference Water Index	standard deviation	mosaic months	ndwi_stdDev	Water Index
79	Near Infrared (NIR)	Landsat band	standard deviation	mosaic months	nir_stdDev	Landsat band

ID	Variable	Description	Statistics	Temporal range	Script acronym	Group
80	Red	Landsat band	standard deviation	mosaic months	red_stdDev	Landsat band
81	Savi	Soil-adjusted Vegetation Index	standard deviation	mosaic months	savi_stdDev	Spectral index
82	Sefi	Savanna Ecosystem Fraction Index	standard deviation	mosaic months	sefi_stdDev	Fraction index
83	Shade	Shade fraction	standard deviation	mosaic months	shade_stdDev	Spectral Mixture Modeling
84	Soil	soil fraction	standard deviation	mosaic months	soil_stdDev	Spectral Mixture Modeling
85	Shortwave Infrared (SWIR) 1	Landsat band	standard deviation	mosaic months	swir1_stdDev	Landsat band
86	Shortwave Infrared (SWIR) 2	Landsat band	standard deviation	mosaic months	swir2_stdDev	Landsat band
87	Wefi	Woodland ecosystem fraction index	standard deviation	mosaic months	wefi_stdDev	Fraction index
88	Slope	Terrain slope	identity	Permanent	slope	Geomorphometric
89	Green Texture	Texture from Landsat band	mean	mosaic months	green_median_texture	
90	Latitude	Geographical coordinate	-	Permanent	Latitude	Geographic
91	Longitude	Geographical coordinate	-	Permanent	Longitude	Geographic
92	Ndvi_3years	Normalized Difference Vegetation Index	amplitude	Last 3 years mosaic months	ndvi_amp_3y	Spectral index

4.4 Classification algorithm, training samples and parameters

Classification was performed subregion by subregion, year by year, using the Random Forest algorithm (Breiman, 2001) available in Google Earth Engine, running 100 iterations (random forest trees).

As mentioned, training samples for each subregion were defined following a strategy of using random pixels where land use and land cover remained the same (stable samples) along the maps of Collection 4 over different subperiods: 1985-1994, 1995-2004, 2005-2014 and 2015-2024, named as “stable samples”.

The identification of stable areas to extract random pixels or “stable samples” was based on a criterion of minimum temporal frequency aiming to ensure confidence to use them as training areas. Each pixel should be classified with the same LULC class throughout each sampling subperiod (1985-1994, 1995-2004, 2005-2014 and 2015-2024). A layer of pixels with a stable classification for each subperiod was then generated. From the resulting layer of stable samples, a subset of 2,000 samples for each subregion was randomly generated for each class for each subperiod. It is important to clarify that not all of these samples were necessarily used in the classification process for each year.

MapBiomas Pampa Argentina Collection 5, represents an improvement in class definition. For example, Fruit plantations were separated from Silviculture. For this purpose, stable samples of Fruit Crops and Silviculture derived from Collection 4 (where they were not differentiated) were manually resampled to perennial crops (Fruit crops) and Silviculture (Forest plantations). Visual interpretation of Landsat and Very High resolution images was implemented to generate stable samples of these classes over each 10-years sampling subperiod. In addition, woody vegetation was separated in Flooded and Non Flooded. For this purpose, stable samples were separated using a probability of wetland factor (Navarro Rau et al., 2025). These samples were then included in the Random Forest Classification for all classes.

In addition, a classical procedure to detect outliers was implemented. For each year, and within each training class, we searched for outliers in all variables. Three outliers method were considered:

1) Isolation forests. It is an anomaly detection method (outliers) that employs binary trees and the concept of isolation, without using any metric. Each tree

recursively splits the data from the root to the leaves, randomly selecting an attribute and a split threshold at each node, until each instance is isolated in a leaf node. (Banchero et al., 2021). The path length of an instance, that is, the number of splits required from the root to a leaf node, allows distinguishing between normal and anomalous instances: anomalous instances typically reach a leaf node in fewer splits, while normal instances require more partitions. This property serves as the basis for calculating a score, which reflects the probability that an instance is an outlier. The score can be modified during each classification to define a threshold for outlier detection.

2) Interquartiles. In this case, an outlier is defined as any value of a specific variable lower or higher than 1.5 times the interquartile range (the first quartile value subtracted from the third quartile value) considering all values of this variable within a specific class of a particular year. The number of variables of the feature space with values considered as outliers for each sample were registered. This value can be changed during each classification as a threshold to detect outliers.
3) Residuals. In this method, the residuals of a simple linear function between the annual mean of Red and Infrared reflectance were estimated. Then, the interquartiles of the residuals were generated. Values outside the interquartiles were identified as outliers.

Decision to consider each or none of these outlier methods and its thresholds were determined by region and subperiod according to preliminary results observed.

4.4.1 Sample size balance

We generated a fixed number of samples for each class, subregion and subperiod for classification. However we used in the classification process only a random subset based on the class area proportion within each subregion, considering each year to be classified. To do this we previously adjusted linear simple functions to estimate the area of each class for each year from 1985 to 2024, based on the annual class area observed along the Collection 4 dataset. These functions were used to estimate, for each year, the proportion of each class to train the classifier. Then, these annual proportions for each class were set to extract a subset of the available samples for the correspondent classification in each year. Whenever the classification resulted in overestimation or underestimation of the class after comparing with supplemental information (e.g.: Collection 4 maps, Landsat mosaics, independent crop type maps, etc.) this proportion was adjusted changing the bias (intercept of linear regression model) accordingly. Notwithstanding the above, a minimum number of 50 to 100 samples per class was set for each region and year, to ensure the correct detection of the less frequent categories.

4.4.2 Complementary samples

The need for adding complementary samples was evaluated by visual inspection of the output of a preliminary classification, with both Landsat and high-resolution images available in GEE and time series of vegetation indices, and also by comparing with the Collection 4 classification. Complementary sample collection was also done manually using points in Google Earth Engine Code Editor. All the false-color images of the 40 years (1985-2024) Landsat mosaics and the vegetation index time series were checked at the selected point. Based on the knowledge of each subregion, the samples for different classes were collected. As mentioned, complementary samples previously generated for Collections 3 and 4 were also added in some regions to improve the classification when necessary.

4.4.3 Final classification

The final classification was performed for all subregions and years combining stable and complementary samples. For some years the classification output resulted in anomalous results for some classes. Then, it was necessary to improve the classification through a new sample size balance and a specific set of complementary samples.

4.4.4 Post-classification

The results of the final classification were improved through a sequence of filters, to correct missing data, “salt-and-pepper” classification errors and, specially, cases of misclassification or to avoid unexpected results. Temporal filters were done with the aim to generate a more stable classification pattern over time, avoiding unexpected class variation during short times.

4.4.4.1 Gap fill filter

A filter to fill no-data pixels ("gaps") was applied. Because theoretically the no-data values are not allowed, they are 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.

4.4.4.2 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 6 pixels (~0,5 ha).

4.4.4.3 Temporal filters

The temporal filters use the information from the year before and after to identify and correct a pixel misclassification, considered as cases of invalid transitions. In a first step, the filter looks for specific cover classes (3, 4, 11, 12, 33) that are not this class in 1985 and were kept unchanged in 1986 and 1987 and then corrects the 1985's value to avoid any regeneration in the first year. In a second step, the filter looks at a pixel value in 2023 that for example is not 11 (wetland) but is equal to 11 in 2021 and 2022. The value in 2023 is then converted to 11 to avoid any regeneration in the last year. The third process looks in a 3-year moving window to correct any value that changed in the middle year and returns to the same class next year.

A temporal filter with a slightly different approach was applied to solve problems in forestry classification. To correct the problems related to the years with forestation cutting, interrupting a continuous series of years classified as forestry we used a special six-year spatial filter. The rule of application checks whether two years before and two years after the class was forestation, if this is true it shifts the classification of the two middle years to silviculture.

4.4.4.4 Frequency filter

To correct classification problems associated with some classes in specific regions, frequency filters were applied to use the temporal information available for each pixel to correct false positives cases. The general logic of the frequency filter is to search for each pixel a specific combination of classes throughout the 40 years producing a subset of pixels considered eligible for correction. Then the filter detects and overwrites only those years where cases of false positives are present using a fixed class value, that usually is the mode of classifications detected along the temporal range. This type of filter was used with parsimony to solve very well delimited cases.

4.4.4.5 Specific filters

Additional specific filters were generated to remove unexpected classification changes that remained after applying previous standard filters. In general, these filters operate based on frequency and incidence. Frequency is the number of years a class occurs in a pixel. The incidence is the number of times that a pixel classification changes along the entire series of years. The application of these filters was limited to fix problems of false transitions between specific classes.

We also used a filter that eliminates problems related to the shadows of the mountains. These filters use characteristics of the relief, in addition to the frequency to be applied. It corrects false positives of water and wetland in shaded slopes in regions with wavy relief. The filter selects all pixels classified as water at least in one year but in less than 38 years (<95%), or as wetland at least in one year but in less than 36 years (<90%), whenever occurring in areas of cliffs and slopes, established by a combination of slope data (SRTM derived) with HAND (Height Above the Nearest Drainage) database, to define places where it is not expected the presence of water or wetland. In such cases, both classes were replaced by the class corresponding to the pixel mode.

A filter to smooth abrupt transitions between the first and the second year (1985-1986) and the last and penultimate years (2023-2024) was applied. It has been observed in previous collections, that the last year of the series registered an unexpected increase in the area of anthropic classes and a decrease of natural classes, most likely corresponding to an artifact resulting from the set of applied filters. To alleviate the problem, a filter was developed to smooth this abrupt transition, avoiding all transitions from natural areas to anthropic areas, and vice versa, in patches equal to or smaller than 2 hectares. In these cases, the corresponding pixels from the last year receive the same classification as the penultimate year as well as pixels from the first year receive the same classification as the second year.

Exceptionally, the spatial effect of some filters was limited to a set of polygons, in such a way as not to modify the entire zone classification. Similarly, in some cases, filters were applied only for specific years. Examples of these filters include: a grassland filter that unifies wet and dry years, taking into account the coverage of that place and not the rainfall of a particular year. Or a rice filter that corrects sites classified as wet grasslands, only for certain years, as long as it has been previously classified as agriculture.

4.4.4.6 Separation between arboreous and shrub woody vegetation

This separation allowed the classification of forest and shrubland, being a new definition of classes compared to previous versions of MapBiomas Pampas, where all open and closed woody vegetation was presented together.

For this purpose GEDI measurements of height (Dubayah et al., 2021) were extracted for available Open woody vegetation (named Forest formation in Collection 4) and Closed woody vegetation (named Savanna formation in Collection 4) samples. Not all samples had GEDI height. Also GEDI values with low quality were excluded. The value of RH100 of 3 m was selected as threshold to separate samples between shrubs (below 3 m height) and arboreous (higher or equal to 3 m). Then, Random Forest classification were performed separating only the classes arboreous and shrubs, and were applied only over the classes Open woody vegetation and Closed woody vegetation generated previously.

5 VALIDATION STRATEGIES

Validation was performed for the classifications of the years 1991, 1996, 2006, 2012, and 2022 following the good practices recommendations proposed by Olofsson et al. (2014) for area and error estimation. A total of 1,416 samples were used for the analysis. The sampled area and the number of samples for each class, balanced in proportion to the area of each class, were both obtained from MapBiomas Pampa

Trinacional Collection 1 map for the year 2010. Independent samples were raffled and class classified by visual interpretation of Landsat images, very high resolution images from Google Earth and time series of vegetation indices. Two interpreters evaluated each of the sample points generated from the stratified random design. In those sample points where discordance in class classification was detected among interpreters, a third interpreter defined the final class assignment. When a final class could not be defined by the three interpreters (e.g. three different class assignments), a final class was agreed by a team of interpreters. More details of the validation methodology are described in Baeza et al. (2022).

Some Collection 5 classes were grouped to generate Collection 4 classes because validation data has the class definition of Collection 4: Closed shrubland and Forest formation were grouped in a unique class called Forest formation; Open shrublands and Savanna formation were grouped in a class named Savanna formation; Forest plantations, Perennial and Industrial crops were grouped in the class Forest plantations.

Table 3 shows the overall accuracy and Kappa coefficient for years 1991, 1996, 2006, 2012 and 2022. **Figure 7** shows the User and Producer accuracy of each class. **Figure 8** shows in a Sankey Diagram the confusion matrix of the year 2022.

Table 3. Overall accuracy and Kappa coefficient for MapBiomas Pampa Argentina Collection 5.

Year	Overall Accuracy	Kappa
1991	0.62	0.51
1996	0.64	0.53
2006	0.70	0.61
2012	0.74	0.64
2022	0.75	0.67

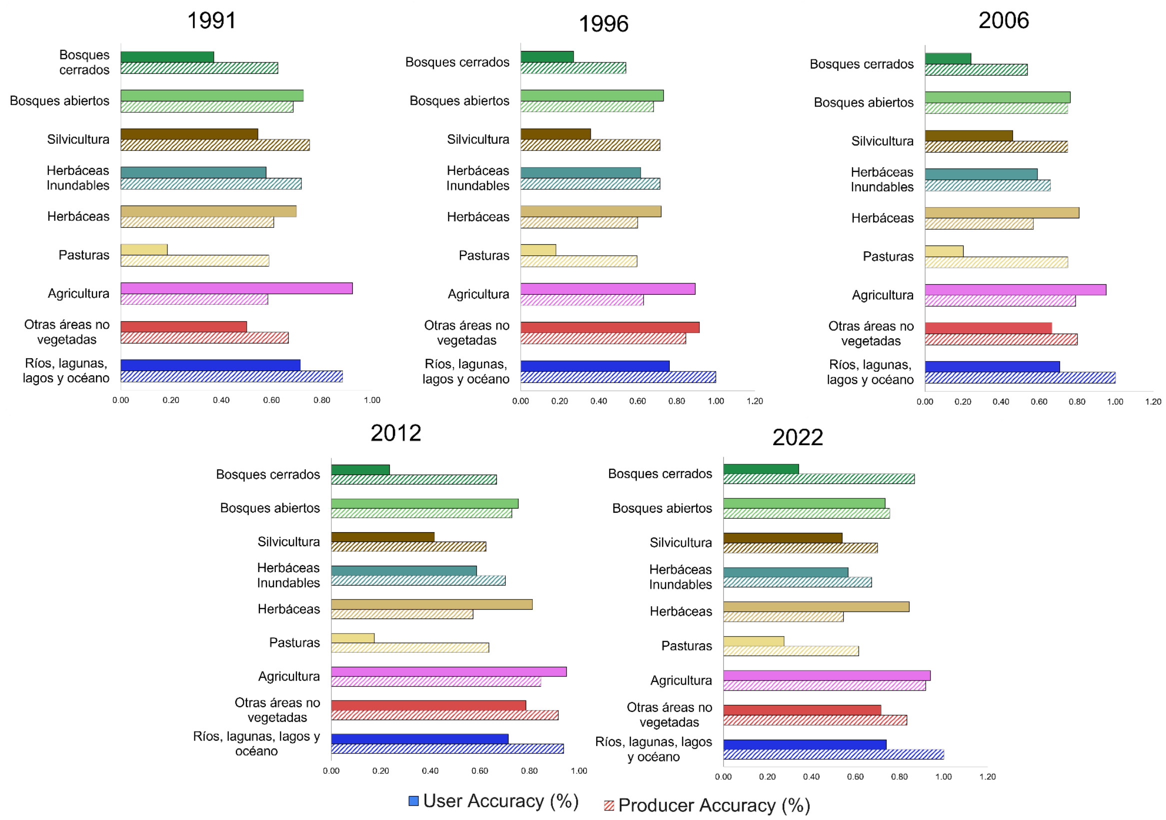

Figure 7. User and producer accuracies for each of mapped class in each evaluated year (collection 4).

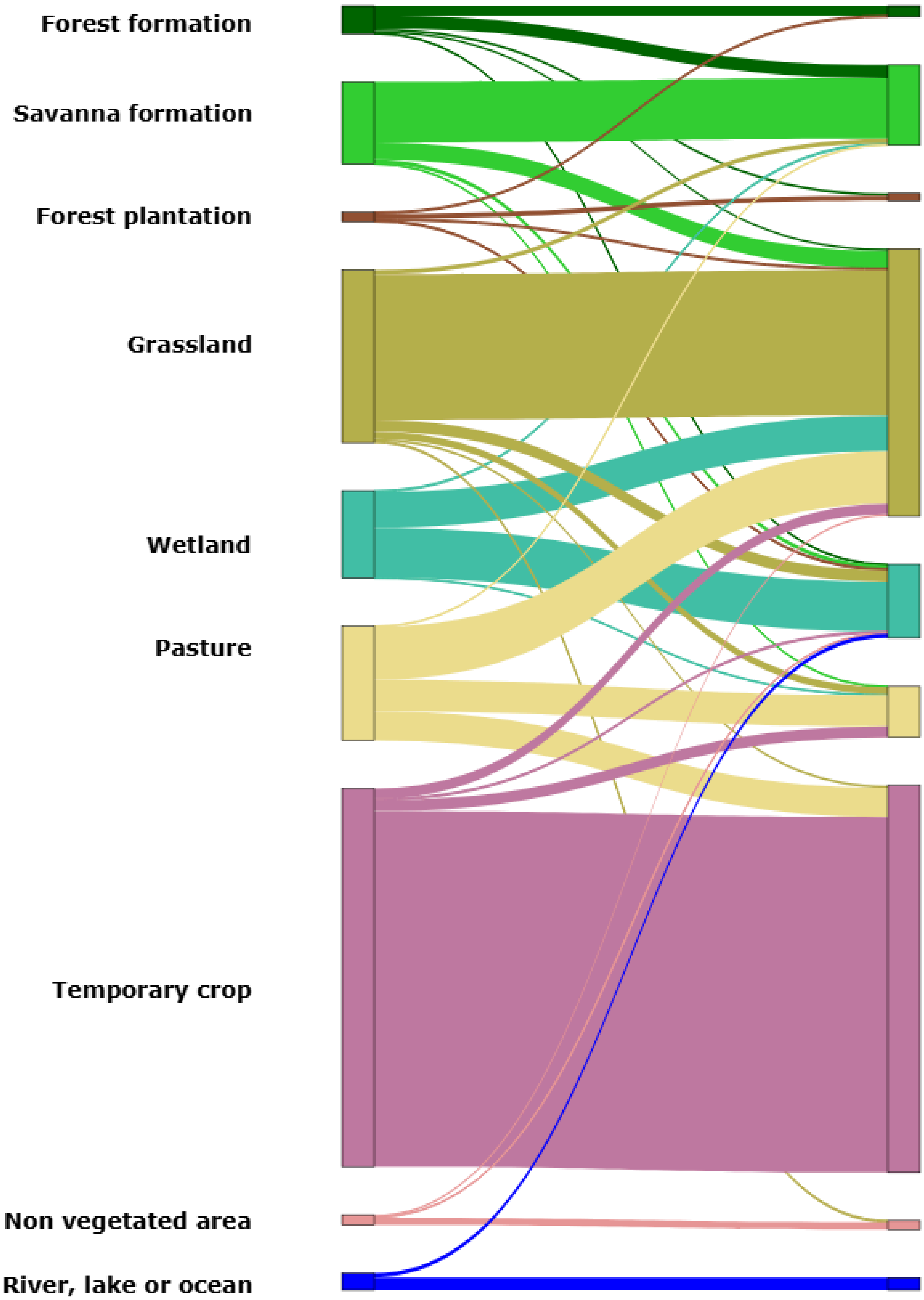

Figure 8. Confusion matrix of Argentine MapBiomas Pampa Collection 5 for year 2022, shown as a Sankey diagram.

6 REFERENCES

Baeza, S., Vélez-Martin, E., De Abelleyra, D., Banchero, S., Gallego, F., Schirmbeck, J.& Hasenack, H. (2022). Two decades of land cover mapping in the Río de la Plata grassland region: The MapBiomas Pampa initiative. Remote Sensing Applications: Society and Environment, 28, 100834.

Banchero, S., Verón, S., Petek, M., Sarrailhe, S., & De Abelleyra, D. (2021). Detección de outliers en muestras de entrenamiento generadas mediante interpretación visual. In *XIII Congreso de AgroInformática (CAI 2021)-JAIIO 50.*

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

Dubayah, R., Hofton, M., J. B. Blair, Armston, J., Tang, H., Luthcke, S. (2021). GEDI L2A Elevation and Height Metrics Data Global Footprint Level V002 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed YYYY-MM DD from https://doi.org/10.5067/GEDI/GEDI02_A.002.

Liu F. T., Ting K. M., Zhou H. (2012). Isolation-based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, 6(1), 1556-4681.

Navarro Rau, M F., Calamari, N. C., Navarro, C. S., Enriquez, A., Mosciaro, M. J., Saucedo, G., ... & Kurtz, D. B.(2025). Advancing wetland mapping in Argentina: A probabilistic approach integrating remote sensing, machine learning, and cloud computing towards sustainable ecosystem monitoring. Watershed Ecology and the Environment, 2025, vol. 7, p. 144-158.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. Remote sensing of Environment, 148, 42-57.

Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing, 32(15), 4407-4429.