



## **Trinational MapBiomias Pampa**

**Collection 2**

**Version 1**

**2022**

**General Coordinator**

Tasso Azevedo

**Country coordinators**

Diego de Abelleira (Argentina)

Heinrich Hasenack (Brazil)

Santiago Baeza (Uruguay)

**Argentina team**

Santiago Verón

Santiago Banchemo

Germán Baldi

Mariano Oyarzabal

Mariana Guerra Lara

Sofía Sarrailhe

Mariana Petek

Marisa Suklje

Ana Eljall

Daniel Blum

**Brazil team**

Eduardo Vélez-Martin

Juliano Schirmbeck

Eliseu José Weber

**Uruguay team**

Federico Gallego

Maria Vallejos

Andrea Barbieri

Laura Bruzzone

Sebastián Ramos

Ana Lucia Mary

Sara Morales

# 1 INTRODUCTION

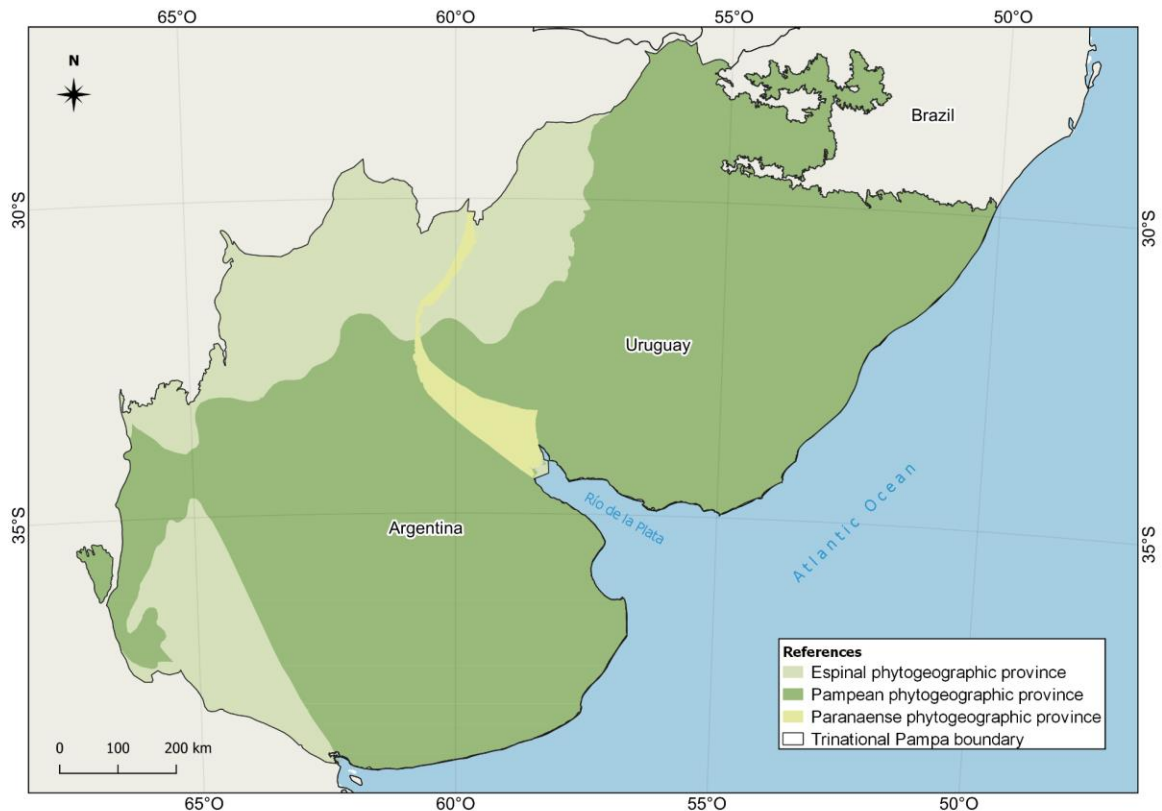
## 1.1 Scope and content of the document

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 South American Pampa biome of Argentina, Brazil and Uruguay from 1985 to 2021 (Collection 2). 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

*Trinational MapBiomias Pampa* initiative was created to produce LULC annual maps for the Pampa biome corresponding to Argentina, Brazil and Uruguay territories. Other phytogeographic regions closed or interspersed with Pampa were partially added to allow a better regional delimitation. Thus, a neighboring area of *Espinal* phytogeographic province, around the Pampa biome, as well as the *Paranaense* phytogeographic province, located both in Argentina were also included (**Figure 1**).

The total mapped area was 1,088,821 km<sup>2</sup>, being 835,397 km<sup>2</sup> in the Pampa, 231,490 km<sup>2</sup> in the Espinal and 21,934 km<sup>2</sup> in the Paraná river Delta.



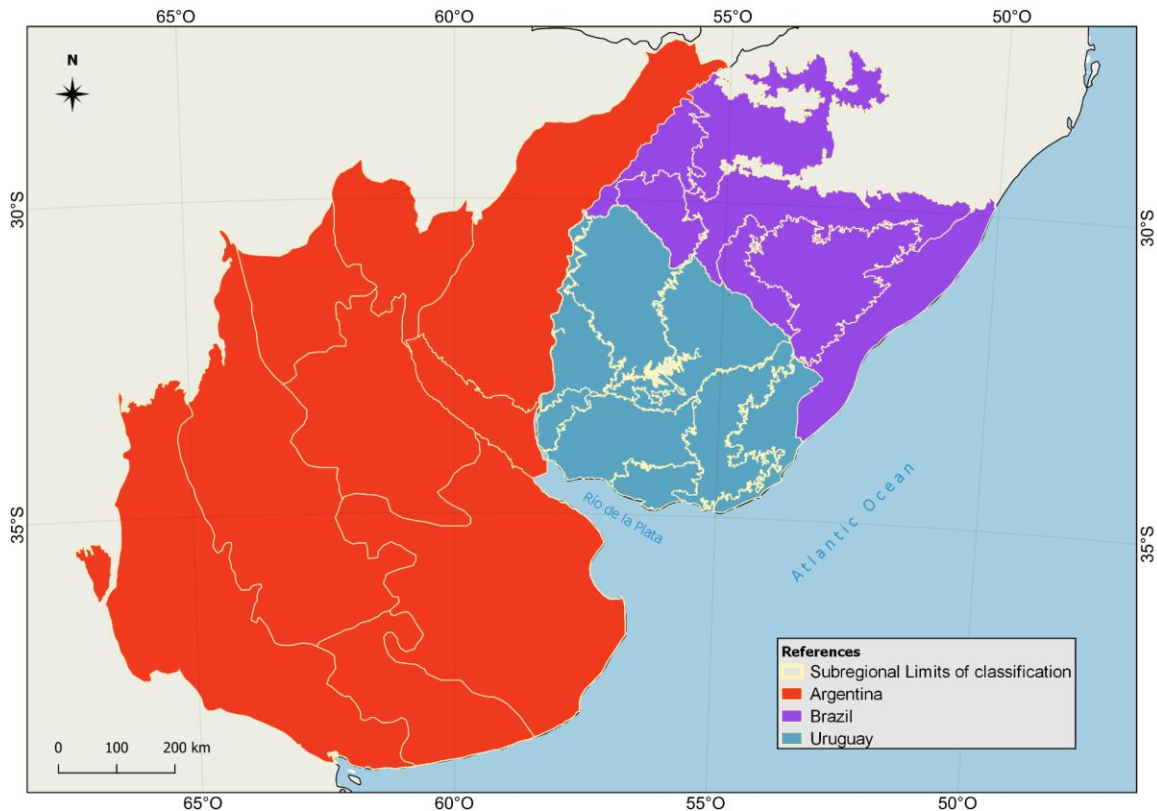
**Figure 1.** Region of interest mapped in the Trinational MapBiomias Pampa initiative (collection 2), including the typical areas of the Pampa, Espinal, and Paraná river Delta.

## 2 GEOGRAPHICAL UNITS OF CLASSIFICATION

In each country, the classification process was carried out in smaller spatial units. These units correspond to subregional homogeneous zones based on several criteria, nationally defined, including geomorphology, soils, vegetation types and land use patterns.

The study area was divided in 23 homogeneous zones, nine in Argentina, seven in Brazil and seven in Uruguay (**Figure 2**).

The purpose of these geographical units of classification was to try to reduce samples and classes confusion and to allow a better balance of samples and results to improve accuracy.



**Figure 2.** Country defined homogeneous subregions used in the classification process of the South American Pampa biome.

### 3 REMOTE SENSING DATA

#### 3.1 Landsat Collection

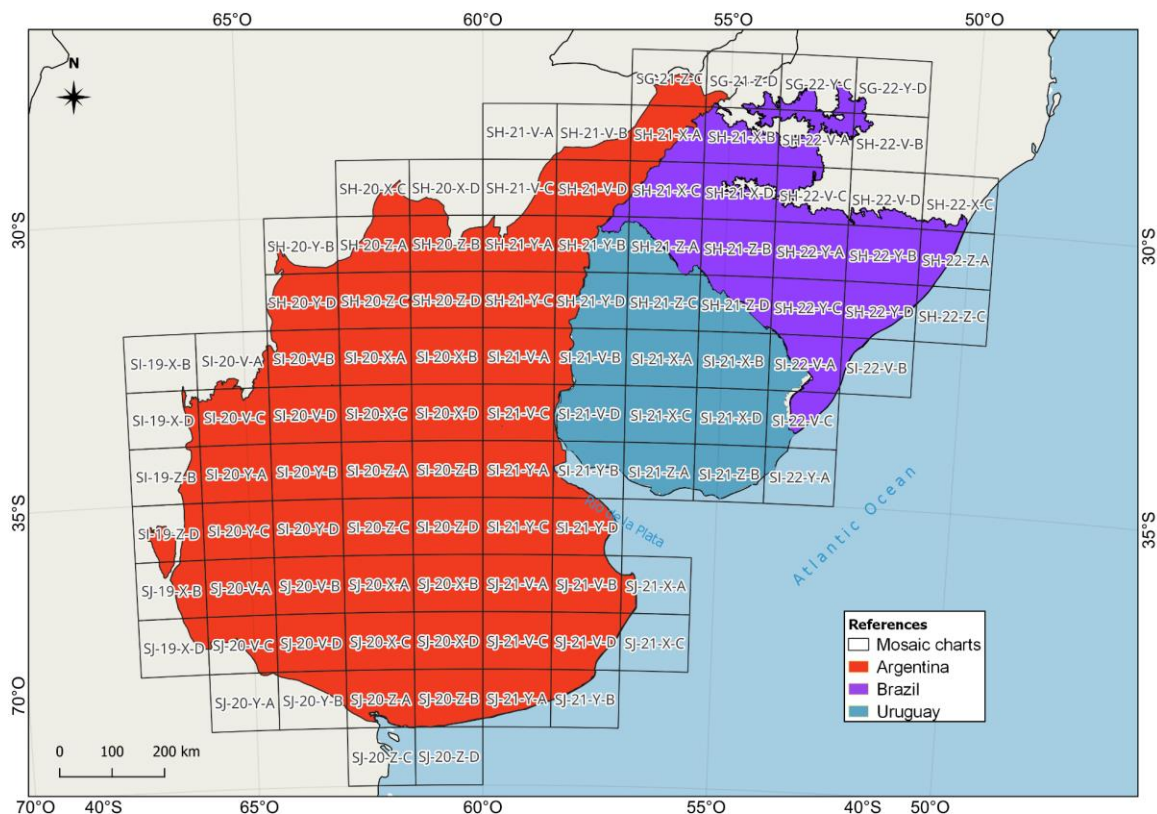
The imagery dataset used in the *Trinational MapBiomias Pampa Collection 2* was obtained from the Landsat sensors Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+) and the Operational Land Imager and Thermal Infrared Sensor (OLI-TIRS), on board of Landsat 5, Landsat 7 and Landsat 8, respectively. The Landsat imagery collections with 30-pixel resolution were accessible via Google Earth Engine, and were provided by NASA and USGS. The *Trinational MapBiomias Pampa Collection 2* 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 74 scenes 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 5 (Brazil and Uruguay) and Landsat 7 (Argentina),
- years 2001, 2002 and 2012: Landsat 7,
- from 2003 to 2011: Landsat 5,
- from 2013 to 2021: 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 99 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 formed by the composition of pixels in each set of images for a certain time period. The periods of the year in which the images are selected vary by country and result from the balance between the probability of maximizing the differences in classes spectral behavior and the availability of cloud-free images. In Uruguay and Brazil, the considered period was from September to November of each year, while in Argentina from May to July. 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 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 a visual analysis, after many trials observing the results of the cloud removing/masking algorithm. Time periods were extended for some years and portions of the study area when the availability of cloud-free images was low.

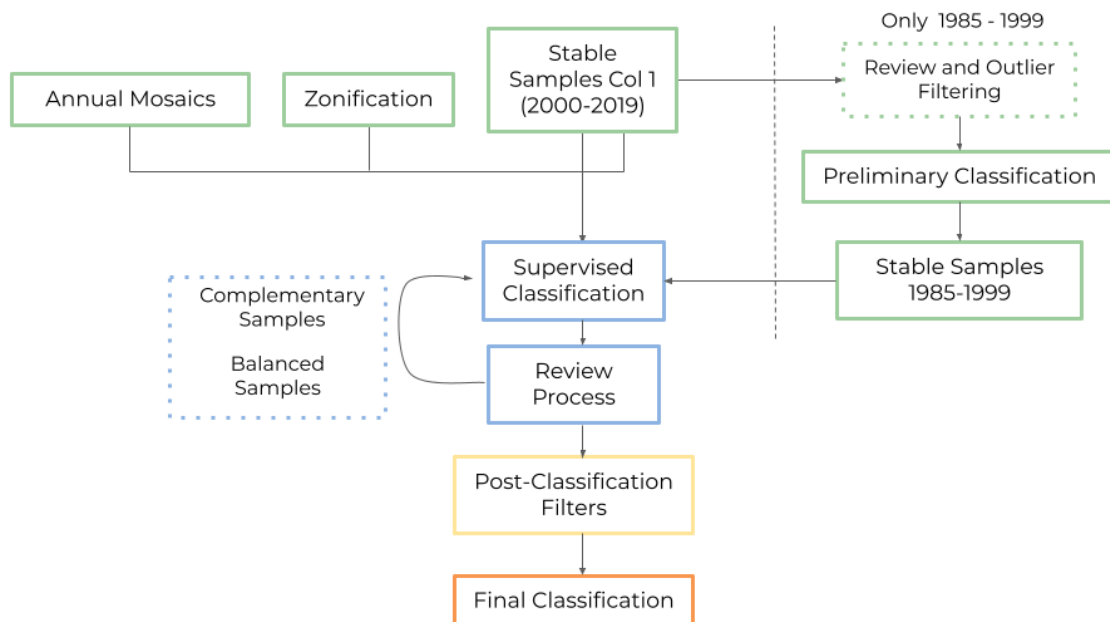
## **4 CLASSIFICATION**

### **4.1 Overview of methodological process**

The methodological procedures of Collection 2 included several steps (**Figure 4**). The first step was to generate annual Landsat image mosaics based on yearly periods. The second step was to generate a new selection of temporally stable samples derived from the stable areas of the maps of Collection 1. Then, the spectral feature inputs derived from the Landsat bands were extracted. Once the samples of 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 samples for training for each class was defined according to the proportion of the area of each class and its variation along the classification period (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. Based on the adjusted training data set, a supervised classification using the random forest algorithm was run.

To classify the period 1985-1999, samples derived from stable areas of Collection 1 maps (corresponding to the 2000-2019 period) were subjected to a filtering process to remove samples that were assigned as outliers in the feature space for the period 1985-1999. Based on the preliminary classification for the period 1985-1999 obtained with this training dataset, a new set of temporally stable samples was extracted, but now derived from the stable areas of the preliminary classification, to perform a new classification. Then, a similar sampling method used for the period 2000-2021 was implemented, and complementary samples were also added when necessary for to get a final classification for 1985-1999.

Following that, gap, spatial and temporal 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 2. The MapBiomass 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 countries, state similar and municipality similar levels of each country, water basin and phytogeographic provinces.














**Figure 4.** Classification process of Collection 2 in the *MapBiomass South American Pampa* biome for the period 2000-2021 and 1985-1999.



## 4.2 Map Legend

The classification for the *Trinational MapBiomass Pampa* initiative using Landsat mosaics included nine land use and land cover (LULC) classes (**Table 1**): Forest formation (3), Savanna formation (4), Wetland (11), Grassland (12), Agriculture or pasture (21), Silviculture (9), Non vegetated area (22), River, lake or ocean (33) and Not observed (27). In Argentina, Agriculture or pasture class (21) was also separated in Agriculture (18) and Pasture (15). A full description of the legend is described in the [document Legend Description](#).

**Table 1.** Land cover and land use classes considered for digital classification of Landsat mosaics for the South American Pampa biome - Collection 2.

Legend class of Collection 2	Numeric ID	Color
1.1. Forest formation	3	
1.2. Savanna formation	4	
2.1. Wetland	11	
2.2. Grassland	12	
3.1. Agriculture	18	
3.2. Pasture	15	
3.3. Agriculture or pasture	21	
3.4. Silviculture	9	
4. Non vegetated area	22	
5.1. River, lake or ocean	33	
6. Not observed	27	

### 4.3 Annual Mosaics

The total available bands of the Trinational MapBiomass Pampa feature space is composed of 93 input variables, including the original Landsat bands, fractional and textural information derived from these bands (**Table 2**). Reducers were used to generate temporal features such as:

- Median: median of the pixel values of the best mapping period defined by each country.
- Median\_dry: median of the quartile of pixels with the lowest NDVI values.
- Median\_wet: median of the quartile of pixels with the highest 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.

**Table 2.** List of the variables included in the feature space used in the classification of the Trinational Mapbiomas Pampa Landsat image mosaics in the Collection 2 (1985-2021).

<b>ID</b>	<b>Variable</b>	<b>Description</b>	<b>Statistics</b>	<b>Temporal range</b>	<b>Script acronym</b>	<b>Group</b>
<b>0</b>	Evi 2	Enhanced Vegetation Index 2	amplitude	mosaic months	evi2_amp	Spectral index
<b>1</b>	Gv	Green vegetation fraction	amplitude	mosaic months	gv_amp	Spectral Mixture Modeling
<b>2</b>	Ndfi	Normalized Difference Fraction Index	amplitude	mosaic months	ndfi_amp	Spectral Mixture Modeling
<b>3</b>	Ndvi	Normalized Difference Vegetation Index	amplitude	mosaic months	ndvi_amp	Spectral index
<b>4</b>	Ndwi	Normalized Difference Water Index	amplitude	mosaic months	ndwi_amp	Water Index
<b>5</b>	Npv	Non-photosynthetic vegetation fraction	amplitude	mosaic months	soil_amp	Spectral Mixture Modeling
<b>6</b>	Wefi	Woodland ecosystem fraction index	amplitude	mosaic months	wefi_amp	Fraction index
<b>7</b>	Soil	Soil fraction	amplitude	mosaic months	blue_median	Spectral Mixture Modeling
<b>8</b>	Blue dry	Landsat band	median	year -first quartile values	blue_median_dry	Landsat band
<b>9</b>	Blue wet	Landsat band	median	year – fourth quartile	blue_median_wet	Landsat band
<b>10</b>	Cai	Cellulose Absorption Index	median	mosaic months	cai_median	Spectral index
<b>11</b>	Cai	Cellulose Absorption Index	median	year -first quartile values	cai_median_dry	Spectral index
<b>12</b>	Cloud	Cloud fraction	median	mosaic months	cloud_median	Spectral Mixture Modeling
<b>13</b>	Evi 2	Enhanced Vegetation Index 2	median	mosaic months	evi2_median	Spectral index
<b>14</b>	Evi 2 dry	Enhanced Vegetation Index 2	median	year -first quartile values	evi2_median_dry	Spectral index

<b>ID</b>	<b>Variable</b>	<b>Description</b>	<b>Statistics</b>	<b>Temporal range</b>	<b>Script acronym</b>	<b>Group</b>
15	Evi 2 wet	Enhanced Vegetation Index 2	median	year – fourth quartile values	evi2_median_wet	Spectral index
16	Fns	$((gv + shade) - soil) / ((gv + shade) + soil)$	median	mosaic months	gcvi_median	Fraction index
17	Fns dry	$((gv + shade) - soil) / ((gv + shade) + soil)$	median	year -first quartile values	gcvi_median_dry	Fraction index
18	Fns wet	$((gv + shade) - soil) / ((gv + shade) + soil)$	median	year – fourth quartile values	gcvi_median_wet	Fraction index
19	Gcvi	$(nir / green - 1)$	median	mosaic months	green_median	Spectral index
20	Gcvi wet	$(nir / green - 1)$	median	year – fourth quartile values	green_median_dry	Spectral index
21	Green wet	Landsat band	median	year – fourth quartile values	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 values	gvs_median_dry	Spectral Mixture Modeling
25	Gvs wet	$GV / (100 - shade)$	median	year -first quartile values	gvs_median_wet	Spectral Mixture Modeling
26	Hallcover	$(-red * 0.017 - nir * 0.007 - swir2 * 0.079 + 5.22)$	median	mosaic months	hallcover_median	Spectral index
27	Ndfi	Normalized Difference Fraction Index	median	mosaic months	ndfi_median	Spectral Mixture Modeling

<b>ID</b>	<b>Variable</b>	<b>Description</b>	<b>Statistics</b>	<b>Temporal range</b>	<b>Script acronym</b>	<b>Group</b>
28	Ndfi dry	Normalized Difference Fraction Index	median	year -first quartile values	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 values	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
34	Ndwi dry	Normalized Difference Water Index	median	year -first quartile values	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) dry	Landsat band	median	mosaic months	nir_median	Landsat band
37	Near Infrared (NIR) dry	Landsat band	median	year -first quartile values	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	$(\text{blue} - \text{green}) / (\text{blue} + \text{green})$	median	mosaic months	pri_median	Spectral index
41	Pri dry	$(\text{blue} - \text{green}) / (\text{blue} + \text{green})$	median	year -first quartile values	pri_median_dry	Spectral index

<b>ID</b>	<b>Variable</b>	<b>Description</b>	<b>Statistics</b>	<b>Temporal range</b>	<b>Script acronym</b>	<b>Group</b>
42	Pri wet	(blue – green)/(blue + green)	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 values	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 values	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 values	sefi_median_dry	Fraction index
51	Sefi wet	Savanna Ecosystem Fraction Index	median	year – fourth quartile	shade_median	Fraction index
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	Landsat band	median	year -first quartile values	swir1_median_dry	Landsat band
55	Shortwave Infrared (SWIR) 2	Landsat band	median	mosaic months	swir1_median_wet	Landsat band

ID	Variable	Description	Statistics	Temporal range	Script acronym	Group
56	Shortwave Infrared (SWIR) 2 dry	Landsat band	median	mosaic months	swir2_median	Landsat band
57	Shortwave Infrared (SWIR) 2 dry	Landsat band	median	year -first quartile values	swir2_median_dry	Landsat band
58	Shortwave Infrared (SWIR) 2 wet	Landsat band	median	year – fourth quartile	swir2_median_wet	Landsat band
59	Wefi	$((gv + npv) - (soil + shade)) / ((gv + npv) + (soil + shade))$	median	mosaic months	wefi_median	Fraction index
60	Wefi wet	$((gv + npv) - (soil + shade)) / ((gv + npv) + (soil + shade))$	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	minimum	mosaic months	blue_stdDev	Landsat band

ID	Variable	Description	Statistics	Temporal range	Script acronym	Group
68	Cai	Cellulose Absorption Index	standard deviation	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	(-red * 0.017 - nir * 0.007 - swir2 * 0.079 + 5.22)	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	Pri	(blue - green)/(blue + green)	standard deviation	mosaic months	ndwi_stdDev	Spectral index



ID	Variable	Description	Statistics	Temporal range	Script acronym	Group
79	Near Infrared (NIR)	Landsat band	standard deviation	mosaic months	nir_stdDev	Landsat band
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	$((gv + npv) - (soil + shade)) / ((gv + npv) + (soil + shade))$	standard deviation	mosaic months	wefi_stdDev	Fraction index
88	Slope	Slope	-	Permanent	slope	Geomorphometric
91	Green texture	mean	median	mosaic months	green_median_texture	

<b>ID</b>	<b>Variable</b>	<b>Description</b>	<b>Statistics</b>	<b>Temporal range</b>	<b>Script acronym</b>	<b>Group</b>
89	Latitude	Geographical coordinate	-	Permanent	latitude	Geographic
90	Longitude	Geographical coordinate	-	Permanent	longitude	Geographic
92	Ndvi	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).

Training samples for each subregion were defined following a strategy of using random pixels for which the land use and land cover remained the same along the maps of Collection 1 of 20 years for the period 2000- 2019, and along the preliminary classification for 1985-1999, named as “stable samples”.

##### **4.4.1 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 for at least a minimum number of years within the period 2000-2019 or 1985-1999 to be considered as stable. The thresholds for some classes and each country and subregion were not the same. A layer of pixels with a stable classification along the 20 years was then generated by applying such thresholds. From the resulting layer of stable samples, a subset of 2,000 samples for each subregion were randomly generated for each class.

##### **4.4.2 Preliminary classification**

Considering the absence of annual maps for the period 1985-1999, it was necessary to generate the a preliminary classification using the same training samples for each subregion used for the 2000-2021 period. However, to avoid misclassification derived from samples belonging to a different class along the period 1985-1999, we removed all those samples considered as outliers in this period using two approaches. The first one was a classical procedure to detect outliers. For each variable within each class and subregion we pooled all samples data for the years 2000, 2005, 2010, 2015 and 2020 and calculated a higher ( $+1,5 \cdot \text{interquartile range}$ ) and a lower threshold ( $-1,5 \cdot \text{interquartile range}$ ). Then, each sample value for each variable within the same class and subregion was compared with both thresholds and classified as an outlier observation whenever above or beneath those values, respectively. This was repeated for each one of the 93 variables of the feature space. Then we counted, for each annual sample, the number of variables classified as an outlier. All samples with more

than six variables with outliers in each year were removed from the dataset. Then we calculated for each sample, considering the 15 years (1985-1999), the frequency of years in which they were observed after the previous filtering. All the samples with frequencies lower than 15 were removed, considering that they probably belonged to a different class within this temporal range.

In order to get more confidence in the class stability of this filtered samples dataset we also applied the Isolation Forest (IF) approach. This unsupervised algorithm can directly detect anomalies based on the concept of isolation without using any metrics (Liu, 2012). In this way, the remaining samples of each classe and subregion were ordered by score generated by IF. The threshold to define if a sample can be an outlier was then established by visual interpretation for each class and year. This process was done to reduce the presence of bad assigned visual interpretation samples and ensure that the filtered stable samples from 2000-2019 contained a high probability of membership to the same LULC class from 1985-1999.

#### **4.4.3 Sample size balance**

We generated a fixed number of samples for each class and subregion of classification. However we used in the classification process only a random subset based in 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 2000 to 2019, based on the annual class area observed along the Collection 1 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 1 maps, independent crop type maps, etc.) this proportion was adjusted changing the bias (intercept of linear regression model) accordingly.

#### **4.4.4 Complementary samples**

The need for adding complementary samples was evaluated by visual inspection of the output of the 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 1 classification. Complementary sample collection was also done manually using points in Google Earth Engine Code Editor. All the false-color images

of the 22 (2000-2021) or 15 (1985-1999) years 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.

#### **4.4.5 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. For some subregions in Uruguay, the output of the preliminary classification achieved better results than the final classification and it was chosen to proceed with post-classification filters.

#### **4.4.6 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. Temporal filters were done with the aim to generate a more stable classification pattern over time, avoiding unexpected classes variation during short times.

##### **4.4.6.1. 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.

##### **4.4.6.2. 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 6 pixels (~0,5 ha).

#### **4.4.6.3. 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 2021 that for example is not 21 (Agriculture or pasture) but is equal to 21 in 2019 and 2020. The value in 2021 is then converted to 21 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.

#### **4.4.6.4. 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 cases of false positives. The general logic of the frequency filter is to search for each pixel a specific combination of classes throughout the 20 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 should be used with parsimony to solve very well delimited cases.

## **5 VALIDATION STRATEGIES**

Validation was performed for the classifications of the years 2001 and 2018 following the good practices recommendations proposed by Olofsson et al. (2014) for area and error estimation. A total of 2,330 samples were defined as necessary for the analysis. The number of samples for each class was proportional to the area of each class obtained from Collection 1 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. More details of the validation methodology are described in Baeza et al. (2022).

Validation results showed an overall accuracy of 79 % for 2001 and 80% for 2018. **Tables 3** and **4** show the contingency matrices performed to evaluate the outcome of the classifications of collection 2 for the years 2001 and 2018 from the independent data set explained above. Each table also shows the errors of omission and commission of each category. The most important errors were associated with the classification of forest plantations as other categories, mainly native forests, especially in 2021, and the classification of “Swampy area and flooded grassland“ class as grasslands. The largest number of points incorrectly classified (by inclusion or omission) was due to confusion between grassland and the agriculture/pasture class. The high physiognomic similarity and the intra class heterogeneity of the spectral response of both natural grasslands (different communities, landscape positions, location in the study area) and sown pastures (different species, sowing dates, pasture age, etc.) generate an overlapping of the spectral signatures of these coverages, which explains the confusion in some of the evaluated points.

**Table 3:** Contingency matrix and omission and commission errors resulting from comparing the 2001 map with the independent data set of the same year: Classes codes: 3 - Native woody vegetation, 9 - Forest plantation, 11 - Swampy area and flooded grassland, 12 - Grassland, 14 - Farming, 22 - Non-Vegetated Area, 33 – Water. O. Errors: Omission Errors I. Errors: Inclusion Errors

Classes		Reference							Total	O. Errors	I. Errors
		3	9	11	12	21	22	33			
Classification	3	149	12	4	26	11	0	0	202	0,20	0,26
	9	4	22	0	6	1	0	0	33	0,50	0,33
	11	3	3	91	9	9	0	7	122	0,38	0,25
	12	29	6	40	564	176	2	1	818	0,18	0,31
	21	1	1	10	65	727	0	0	804	0,21	0,10
	22	0	0	1	15	2	16	0	34	0,11	0,53
	33	0	0	1	1	0	0	73	75	0,10	0,03
	<b>Total</b>	<b>186</b>	<b>44</b>	<b>147</b>	<b>686</b>	<b>926</b>	<b>18</b>	<b>81</b>	<b>2088</b>		



**Table 4** Contingency matrix and omission and commission errors resulting from comparing the 2018 map with the independent data set of the same year: Classes codes: 3 - Native woody vegetation, 9 - Forest plantation, 11 - Swampy area and flooded grassland, 12 - Grassland, 14 - Farming, 22 - Non-Vegetated Area, 33 – Water. O. Errors: Omission Errors I. Errors: Inclusion Errors.

Classes		Reference							Total	O. Errors	I. Errors
		3	9	11	12	21	22	33			
Classification	3	154	12	4	22	7	0	0	199	0,16	0,23
	9	4	53	0	3	4	0	0	64	0,29	0,17
	11	8	3	88	7	8	0	6	120	0,38	0,27
	12	17	5	39	439	201	1	1	703	0,14	0,38
	21	1	1	7	30	840	0	0	879	0,21	0,04
	22	0	1	2	10	2	19	0	34	0,05	0,44
	33	0	0	3	0	1	0	74	78	0,09	0,05
	<b>Total</b>	184	75	143	511	1063	20	81	<b>2077</b>		

## 6 REFERENCES

- Baeza, S., Vélez-Martin, E., De Abelleira, 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 MapBiomass Pampa initiative. *Remote Sensing Applications: Society and Environment*, 28, 100834.
- Breiman, L. (2001). Random forests. *Machine learning*, v. 45, n. 1, p. 5-32.
- Liu F. T., Ting K. M., Zhou H. (2012). Isolation-based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, 6(1), 1556-4681.
- 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.