# MapBiomas Trinational Pampa

## Collection 1

## Version 1

2021

**General Coordinator**
Tasso Azevedo

**Country coordinators**
Diego de Abelleyra (Argentina)
Heinrich Hasenack (Brazil)
Santiago Baeza (Uruguay)

**Argentina team**
Santiago Verón
Santiago Banchero
Germán Baldi
Camilo Bagnato
Mariana Guerra Lara
Sofía Sarrailhe
Mariana Petek
Mayra Milkovic

**Brazil team**
Eduardo Vélez-Martin
Juliano Schirmbeck
Eliseu José Weber

**Uruguay team**
Federico Gallego
Maria Vallejos
Andrea Barbieri
Laura Bruzzone
Sebastián Ramos
Virginia Fernández Ramos
Carlos Miguel González
Bruno Guigou
Juan Hernández
Adrián Cal

# 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 of Argentina, Brazil and Uruguay 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 South American Pampa* was created to produce LULC annual maps for the Pampa Region corresponding to Argentina, Brazil and Uruguay territories. Other phytogeografic regions closed or interspersed with Pampa were partially added to allow a better regional delimitation. Thus, a neighbor area of *Espinal* around Pampa bioma as well as the Paraná river Delta located in Argentina were also included (**Figure 1**).

The total mapped area was 1,005,772 km$^2$, being 807,759 km$^2$ in the Pampa, 176,745 km$^2$ in the Espinal and 21,268 km$^2$ in the Paraná river Delta.

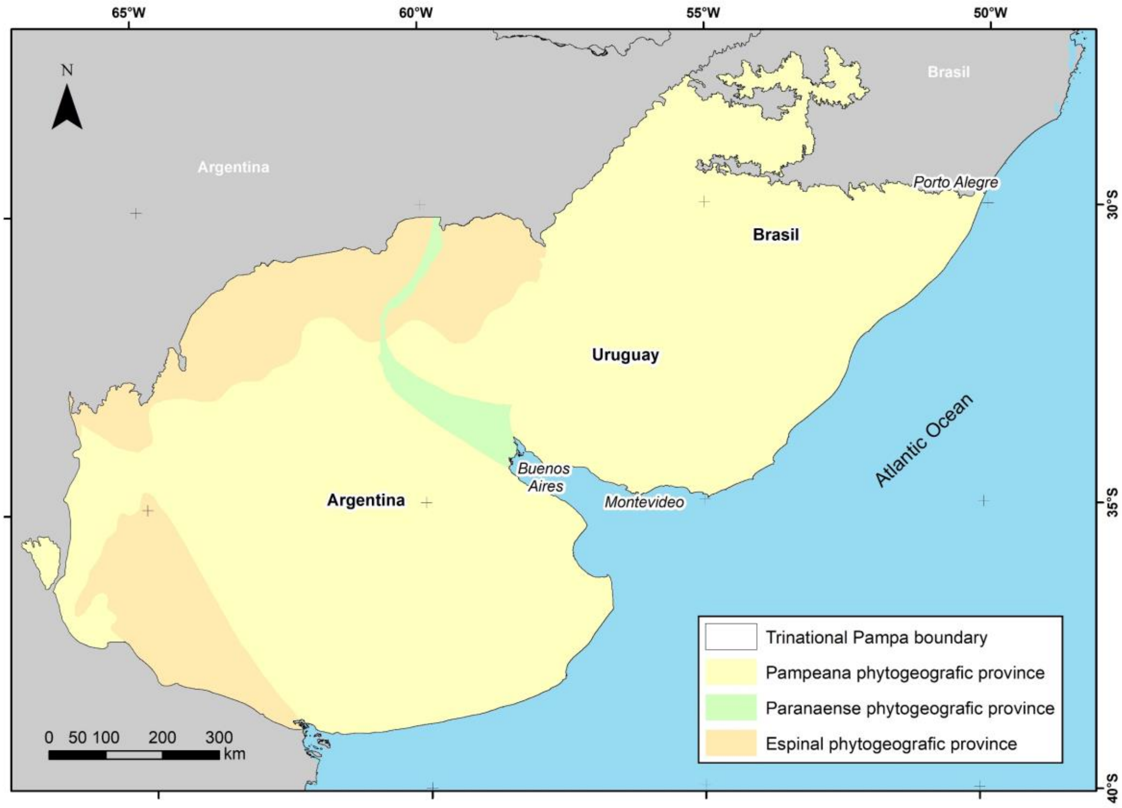

**Figure 1** Region of interest mapped in the Trinational Pampa project, including the typical areas of the Pampa, Espinal, and Paraná river Delta.

## 2 GEOGRAPHICAL UNITS OF CLASSIFICATION

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

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

The purpose of these geographical unites of classification was an attempt to reduce confusion of samples and classes, to allow a better balance of samples and results, improving accuracy.

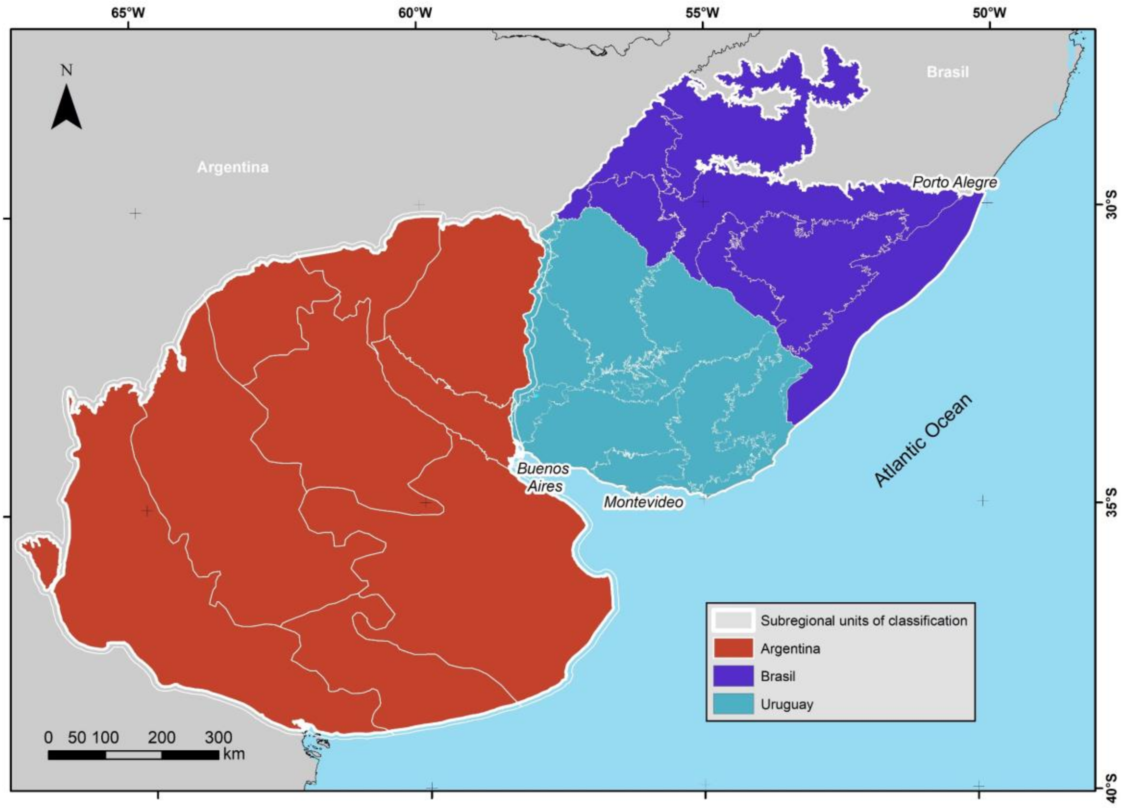

**Figure 2.** Country defined homogeneous subregions used in the classification process of the South American Pampa.

# 3 REMOTE SENSING DATA

## 3.1 Landsat Collection

The imagery dataset used in the *MapBiomas South American Pampa* Collection 1 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 source by NASA and USGS. The *MapBiomas South American Pampa* 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 71 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:

- 2000: Landsat 5 (Brazil and Uruguay) and Landsat 7 (Argentina),
- 2001, 2002 and 2012: Landsat 7,
- 2003 to 2011:  Landsat 5,
- 2013 to 2019: 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.

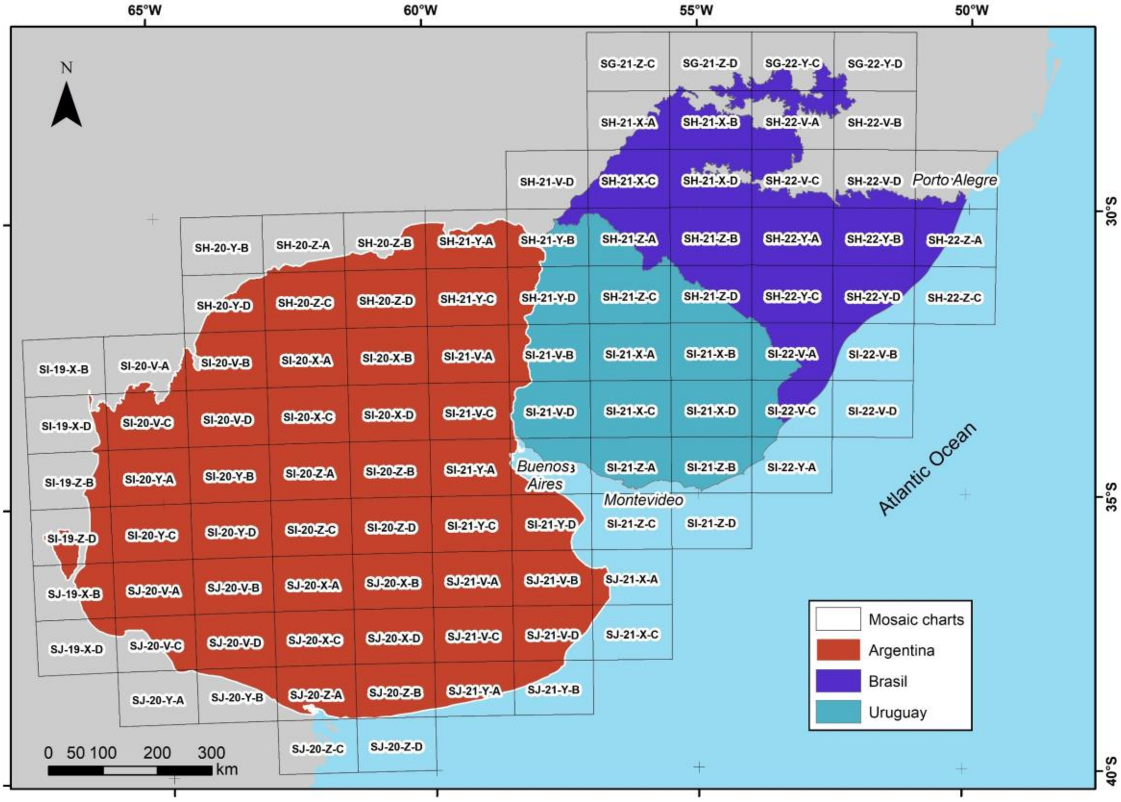

**Figure 2** Charts scheme used to build up Landsat mosaics used throughout the classification process.

## 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.

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 1 included several steps (**Figure 4**). 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 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, including complementary 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 to generate the final map of Collection 1. The MapBiomas annual LULC maps were used to derive the transition analysis (with spatial filter application) and statistics. The statistical analysis covered different spatial categories, such as subregion, state similar and municipality similar levels of each country

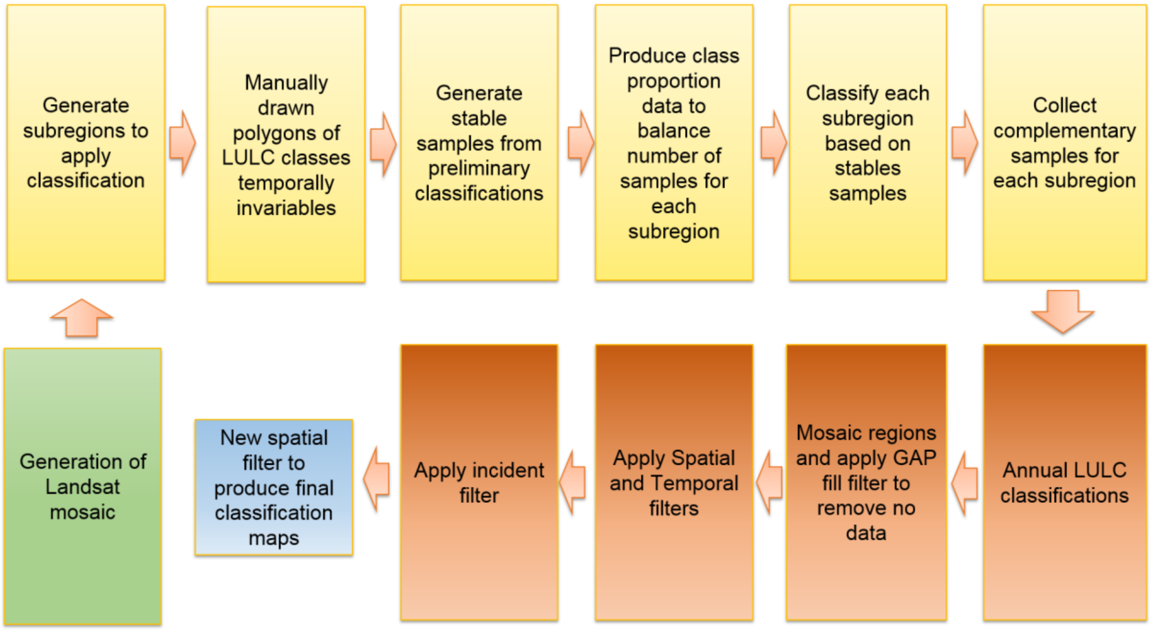
Generate subregions to apply classification
Manually drawn polygons of LULC classes temporally invariables
Generate stable samples from preliminary classifications
Produce class proportion data to balance number of samples for each subregion
Classify each subregion based on stables samples
Collect complementary samples for each subregion
Annual LULC classifications
Mosaic regions and apply GAP fill filter to remove no data
Apply Spatial and Temporal filters
Apply incident filter
New spatial filter to produce final classification maps
Generation of Landsat mosaic

**Figure 4.** Classification process of Collection 1 in the *MapBiomas South American Pampa.*

## 4.2 Classification scheme

The digital classification of the Landsat mosaics for the *MapBiomas South American Pampa* included nine land use and land cover (LULC) classes (**Table 1**): Forest Formation (3), Savanna Formation (4), Forest plantation (9), Wetland (11), Grassland (12), Farming (14), Non Vegetated Area (22), River, Lake and Ocean (33) and Non Observed (27).

**Table 1** Land cover and land use classes considered for digital classification of Landsat mosaics for the South American Pampa - Collection 15.

| Legend class of Collection 5 | 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  Farming | 14 | 🟨 |
| 4. Non-Vegetated Area | 22 | 🟥 |
| 5. River, Lake and Ocean | 33 | 🟦 |
| 6. Non Observed | 27 | ⬜ |

## 4.3 Feature space

The total available bands of the MapBiomas feature space is composed of 107 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** Feature space considered in the classification of the South American Pampa Landsat image mosaics in the MapBiomas Collection 1 (2000-2019).

| ID | Variable | Description | Statistics | Temporal range | Script acronym | Group |
|---|---|---|---|---|---|---|
| 0 | Evi 2 | Enhanced Vegetation Index 2 | amplitude | mosaic months | 'amp_evi2' | Spectral index |
| 1 | Gv | Green vegetation fraction | amplitude | mosaic months | 'amp_gv' | Spectral Mixture Modeling |
| 2 | Ndfi | Normalized Difference Fraction Index | amplitude | mosaic months | 'amp_ndfi' | Spectral Mixture Modeling |
| 3 | Ndvi | Normalized Difference Vegetation Index | amplitude | mosaic months | 'amp_ndvi' | Spectral index |
| 4 | Ndwi | Normalized Difference Water Index | amplitude | mosaic months | 'amp_ndwi' | Water Index |
| 5 | Npv | Non-photosynthetic vegetation fraction | amplitude | mosaic months | 'amp_npv' | Spectral Mixture Modeling |
| 6 | Sefi | Savanna Ecosystem Fraction Index | amplitude | mosaic months | 'amp_sefi' | Fraction index |
| 7 | Soil | soil fraction | amplitude | mosaic months | 'amp_soil' | Spectral Mixture Modeling |
| 10 | Blue dry | Landsat band | median | year -first quartile values | 'median_blue_dry' | Landsat band |
| 11 | Blue wet | Landsat band | median | year – fourth quartile | 'median_blue_wet' | Landsat band |
| 15 | Cloud | Cloud fraction | median | mosaic months | 'median_cloud' | Spectral Mixture Modeling |
| 16 | Evi 2 | Enhanced Vegetation Index 2 | median | mosaic months | 'median_evi2' | Spectral index |
| 17 | Evi 2 dry | Enhanced Vegetation Index 2 | median | year -first quartile values | 'median_evi2_dry' | Spectral index |
| 18 | Evi 2 wet | Enhanced Vegetation Index 2 | median | year – fourth quartile values | 'median_evi2_wet' | Spectral index |
| 19 | Fns | ((gv + shade) − soil)/((gv + shade) + soil) | median | mosaic months | 'median_fns' | Fraction index |
| 20 | Fns dry | ((gv + shade) − soil)/((gv + shade) + soil) | median | year -first quartile values | 'median_fns_dry' | Fraction index |
| 21 | Fns wet | ((gv + shade) − soil)/((gv + shade) + soil) | median | year – fourth quartile values | 'median_fns_wet' | Fraction index |
| 22 | Gcvi | (nir/green − 1) | median | mosaic months | 'median_gcvi' | Spectral index |
| 24 | Gcvi wet | (nir/green − 1) | median | year -first quartile values | 'median_gcvi_wet' | Spectral index |
| 27 | Green wet | Landsat band | median | year -first quartile values | 'median_green_wet' | Landsat band |
| 31 | Gvs wet | GV / (100 - shade) | median | year -first quartile values | 'median_gvs_wet' | Spectral Mixture Modeling |
| 32 | Hallcover | (−red ∗ 0.017 − nir ∗ 0.007 − swir2 ∗ 0.079 + 5.22) | median | mosaic months | 'median_hallcover' | Spectral index |
| 34 | Ndfi wet | Normalized Difference Fraction Index | median | year – fourth quartile | 'median_ndfi_wet' | Spectral Mixture Modeling |
| 36 | Ndvi | Normalized Difference Vegetation | median | mosaic months | 'median_ndvi' | Spectral index |

| ID | Variable | Description | Statistics | Temporal range | Script acronym | Group |
|---|---|---|---|---|---|---|
|  |  | Index |  |  |  |  |
| 37 | Ndvi dry | Normalized Difference Vegetation Index | median | year -first quartile values | 'median_ndvi_dry' | Spectral index |
| 39 | Ndwi | Normalized Difference Water Index | median | mosaic months | 'median_ndwi' | Water Index |
| 40 | Ndwi dry | Normalized Difference Water Index | median | year -first quartile values | 'median_ndwi_dry' | Water Index |
| 41 | Ndwi wet | Normalized Difference Water Index | median | year – fourth quartile | 'median_ndwi_wet' | Water Index |
| 43 | Near Infrared (NIR) dry | Landsat band | median | year -first quartile values | 'median_nir_dry' | Landsat band |
| 45 | Npv | Non-photosynthetic vegetation fraction | median | mosaic months | 'median_npv' | Spectral Mixture Modeling |
| 47 | Pri dry | (blue − green)/(blue + green) | median | year -first quartile values | 'median_pri_dry' | Spectral index |
| 48 | Pri wet | (blue − green)/(blue + green) | median | year – fourth quartile | 'median_pri_wet' | Spectral index |
| 49 | Red | Landsat band | median | mosaic months | 'median_red' | Landsat band |
| 50 | Red dry | Landsat band | median | year -first quartile values | 'median_red_dry' | Landsat band |
| 51 | Red wet | Landsat band | median | year – fourth quartile | 'median_red_wet' | Landsat band |
| 52 | Savi | Soil-adjusted Vegetation Index | median | mosaic months | 'median_savi' | Spectral index |
| 53 | Savi dry | Soil-adjusted Vegetation Index | median | year -first quartile values | 'median_savi_dry' | Spectral index |
| 55 | Sefi | Savanna Ecosystem Fraction Index | median | mosaic months | 'median_sefi' | Fraction index |
| 56 | Sefi dry | Savanna Ecosystem Fraction Index | median | year -first quartile values | 'median_sefi dry' | Fraction index |
| 57 | Sefi wet | Savanna Ecosystem Fraction Index | median | year – fourth quartile | 'median_sefi wet' | Fraction index |
| 63 | Shortwave Infrared (SWIR) 2 | Landsat band | median | mosaic months | 'median_swir2' | Landsat band |
| 64 | Shortwave Infrared (SWIR) 2 dry | Landsat band | median | year -first quartile values | 'median_swir2_dry' | Landsat band |
| 65 | Shortwave Infrared (SWIR) 2 wet | Landsat band | median | year – fourth quartile | 'median_swir2_wet' | Landsat band |
| 68 | Wefi dry | ((gv + npv) − (soil + shade))/ ((gv + npv) + (soil + shade)) | median | year -first quartile values | 'median_wefi_dry' | Fraction index |
| 70 | Blue min | Landsat band | minimum | mosaic months | 'min_blue' | Landsat band |
| 71 | Green min | Landsat band | minimum | mosaic months | 'min_green' | Landsat band |
| 73 | Red min | Landsat band | minimum | mosaic months | 'min_red' | Landsat band |
| 74 | Shortwave | Landsat band | minimum | mosaic months | 'min_swir1' | Landsat band |

| ID | Variable | Description | Statistics | Temporal range | Script acronym | Group |
|---|---|---|---|---|---|---|
|  | Infrared (SWIR) 1 |  |  |  |  |  |
| 75 | Shortwave Infrared (SWIR) 2 | Landsat band | minimum | mosaic months | 'min_swir2' | Landsat band |
| 76 | Temperature | Landsat band | minimum | mosaic months | 'min_temp' | Landsat band |
| 77 | Blue | Landsat band | standard deviation | mosaic months | 'stdDev_blue' | Landsat band |
| 79 | Cloud | Cloud fraction | standard deviation | mosaic months | 'stdDev_cloud’ | Spectral Mixture Modeling |
| 82 | Gcvi | (nir/green − 1) | standard deviation | mosaic months | 'stdDev_gcvi' | Spectral index |
| 86 | Hallcover | (−red ∗ 0.017 − nir ∗ 0.007 − swir2 ∗ 0.079 + 5.22) | standard deviation | mosaic months | ' stdDev_hallcover' | Spectral index |
| 88 | Ndvi | Normalized Difference Vegetation Index | standard deviation | mosaic months | 'stdDev_ndvi' | Spectral index |
| 92 | Pri | (blue − green)/(blue + green) | standard deviation | mosaic months | 'stdDev_pri' | Spectral index |
| 93 | Red | Landsat band | standard deviation | mosaic months | 'stdDev_red' | Landsat band |
| 94 | Savi | Soil-adjusted Vegetation Index | standard deviation | mosaic months | 'stdDev_savi' | Spectral index |
| 95 | Sefi | Savanna Ecosystem Fraction Index | standard deviation | mosaic months | 'stdDev_sefi' | Fraction index |
| 97 | Soil | soil fraction | standard deviation | mosaic months | 'stdDev_soil' | Spectral Mixture Modeling |
| 98 | Shortwave Infrared (SWIR) 1 | Landsat band | standard deviation | mosaic months | 'stdDev_swir1' | Landsat band |
| 100 | Temperature | Landsat band | standard deviation | mosaic months | 'stdDev_temp' | Landsat band |
| 102 | Slope | Slope | - | Permanent | 'slope' | Geomorphometric |
| 105 | Latitude | Geographical coordinate | - | Permanent | 'latitude' | Geographic |
| 106 | Ndvi_3anos | Normalized Difference Vegetation Index | amplitude | mosaic months | 'amp_ndvi_3anos' | Spectral index |

## 4.4 Classification algorithm, training samples and parameters

Digital classification was performed region by region, year by year, using the 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, named as "stable samples". The stable areas were identified through an annual preliminary classification made using random pixels selected from on-screen-digitized polygons. For this, backdrops of false-color Landsat mosaics for all the 20 years as well as graphs showing the temporal behavior of spectral indices per pixel were used to create a stable LULC class.

### 4.4.1 Preliminary Classification

From on-screen-digitized polygons, which totalized 4,189 for Argentina and 1,703 for Uruguay, a subset between 200 and 700 pixels per class and per zone were randomly selected from the pixels of the on-screen-digitized polygons (randomly selected too) and used as training areas to classify each of the 20 years with the Random Forest algorithm. A total of 20 yearly preliminary classification were obtained and the frequency with which a pixel was classified to the same LULC class was calculated to define the temporal stable areas. In Brazil, the results of MapBiomas Brazil collection 4.1 were used to define the temporal stable areas.

### 4.4.2 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 confidence for use them as training areas. Each pixel should be classified with the same LULC class at least a a minimum number of years in the period 2000-2019 to be considered as stable. The thresholds for some classes and each country 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 based on the class cover percentage. A minimum of 200 samples was used for rare classes that did not reached a land cover at least 10% of the region area.

### 4.4.3 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 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.

### 4.4.4 Final classification

The final classification was performed for all subregions and years with stable and complementary samples. All years used the same subset of samples, but trained using the specific mosaic of the year being classified.

# 5 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.

## 5.1 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.

## 5.2 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).

### 5.3 Temporal filter

The temporal filter uses the information from the previous year and the later year to identify and correct a pixel misclassification, considered as cases of invalid transitions. In the first step, the filter looks at any natural cover (3, 4, 11, 12, 33) that is not this class in 2000 and was kept unchanged in 2001 and 2002 and then corrects the 2000's value to avoid any regeneration in the first year. In the second step, the filter looks at a pixel value in 2019 that is not 14 (Farming) but is equal to 14 in 2017 and 2018. The value in 2019 is then converted to 14 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.

### 5.4 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.

## 6 VALIDATION STRATEGIES

Collection 1 does not include an accuracy analysis. It is planned for the next collections.

## 7 REFERENCES

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