Pantanal - Appendix

Collection 7

Version 2

General coordinator
Eduardo Reis Rosa

Team
Marcos Reis Rosa
Mariana Dias

1. Overview of classification method

The initial classification of the Pantanal biome within the MapBiomas project consisted of applying decision trees to generate annual maps of the predominant native vegetation (NV) types, which were distinguished in three classes: Forest, Savanna, and Grassland. The method used to generate these annual maps evolved over time, with significant improvements from the first MapBiomas Collection to the present.

Collection 1.0 covered the period of 2008 to 2015 and was published in 2016. Collections 2.0 and 2.3 covered the period of 2000 to 2016 and were published in 2018. The classification using Random Forest was implemented in Collection 2.3, and from this point onward, the empirical decision tree was used for the purpose of generating stable samples, which were classified as the same NV type over the considered period (2000-2016). These stable samples were used to train the Random Forest models for the classification of the entire time series. Collections 3.0 and 3.1 expanded the period covered to 1985–2017. Collections 4.0 and 5 used training samples collected based on the stable samples from the previous collection and reference maps. Collection 6 used stable samples from collection 5.

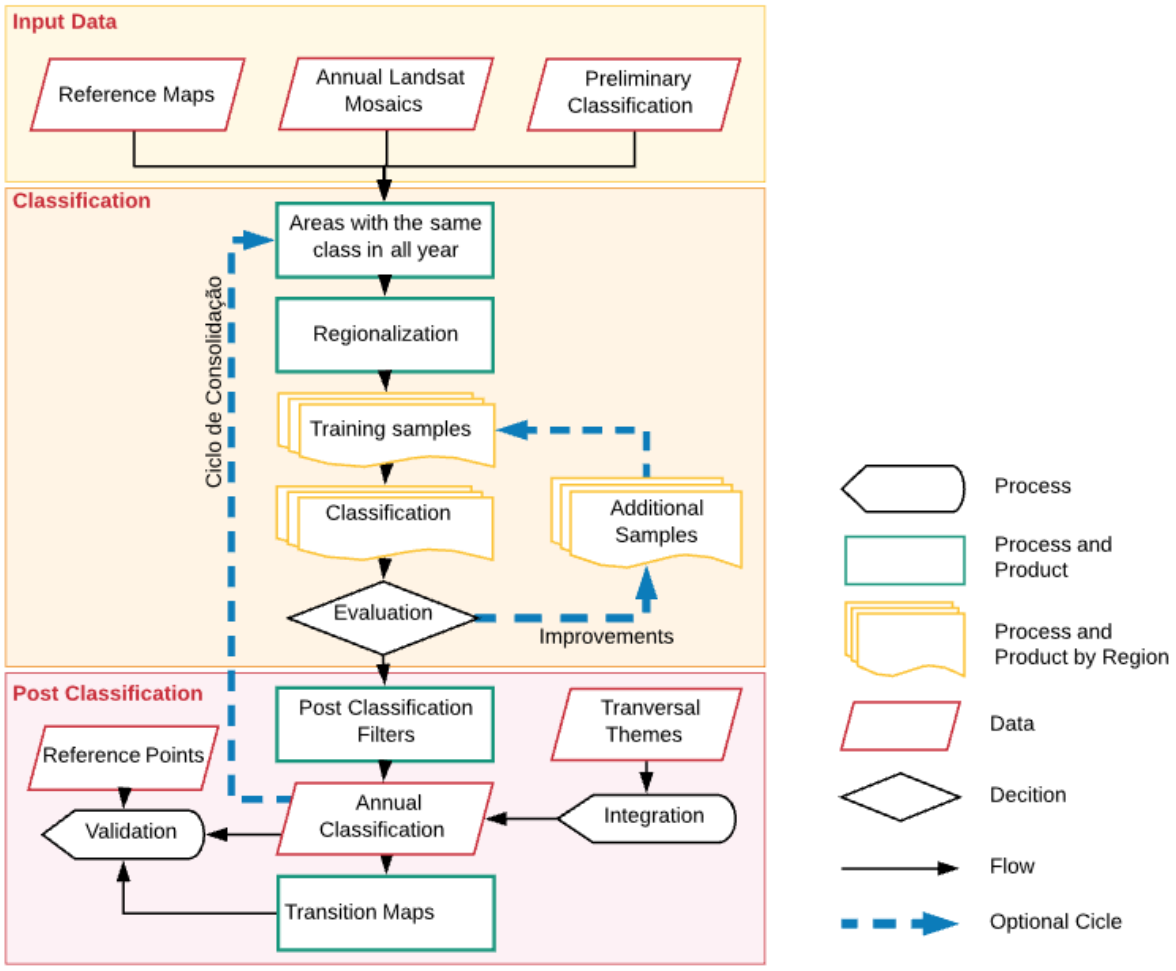
Table 1. The evolution of the Pantanal mapping collections in the MapBiomas Project, its periods, level and number of classes, brief methodological description, and global accuracy in Level 1 and 2.

Collection	Period	Levels /N. Classes	Method	Global Accuracy
Beta & 1	8 years 2008-2015	1 / 7	Empirical Decision Tree	
2.0 & 2.3	16 years 2000-2016	3 / 13	Empirical Decision Tree & Random Forest (2.3)	
3.0 & 3.1	33 years 1985-2017	3 / 19	Random Forest	Level 1: 73.2% Level 3: 62.1% *
4.0 & 4.1	34 years 1985-2018	3 / 19	Random Forest	Level 1: 80.7% Level 3: 71.0% *
5.0	35 years 1985-2019	4 / 21	Random Forest	Level 1: 85.1% Level 3: 78.3% *
6.0	36 years 1985-2020	4 / 24	Random Forest	Level 1: 81.6% Level 2: 73.5%
7.0	37 years 1985-2021	4 / 24	Random Forest	Level 1: 85.9% Level 2: 79.5%

** Due to hierarchy changes in the forest classes, level 2 of collection 6 and 7 is being compared to level 3 of previous collections.*

The production of the Collection 7, with land cover and land use annual maps for the period of 1985-2021, followed a sequence of steps in the Pantanal biome, similar to those used in the previous Collection 4 ,5 and 6 (**Figure 1**). However, some improvements were added up, particularly in the mosaics, balance of samples and in the post classification filters.

Figure 1. Classification process

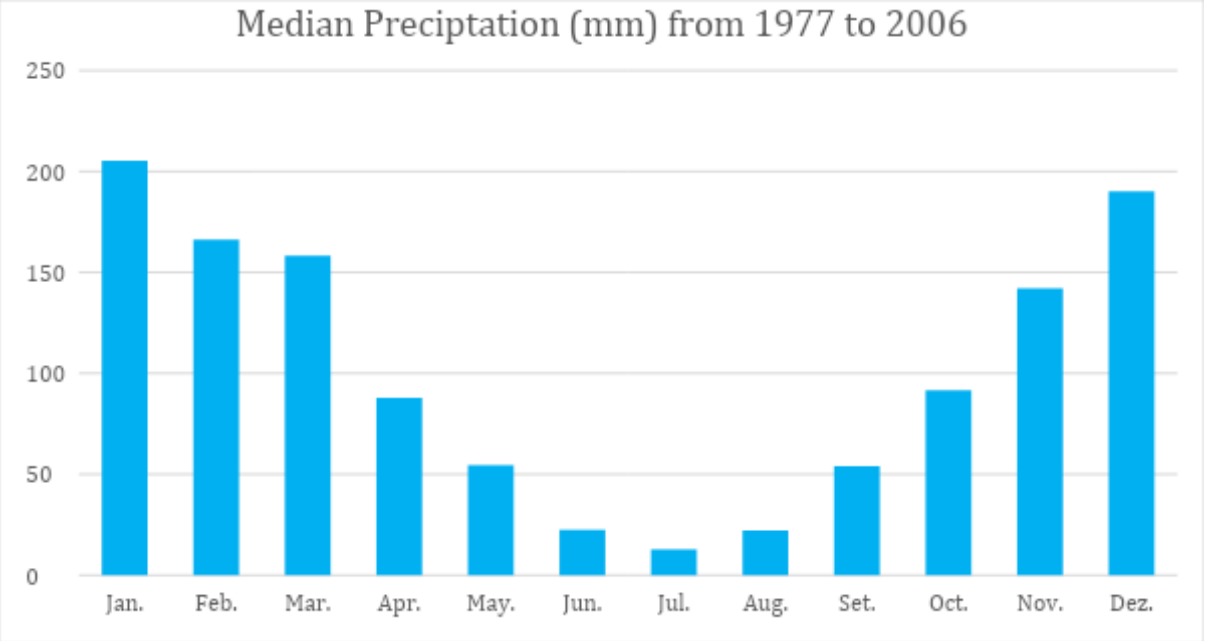


1. Landsat image mosaics

1.1. Definition of the temporal period

The image selection period for the Pantanal biome was defined aiming the selection in the dry season to reduce the wetlands. The use of images in the driest period of the Pantanal reduces the occurrence of wetlands that can reach areas of natural fields and pastures. It also helps to detect the variations in the natural fields and pastures and reduces possible confusions in the identification of the areas of Forested Savannas and Forests existing in the plain and that also is influenced by the periodic floods.

Figure 2. Median monthly precipitation values from 1977 to 2006. (MARCUZZO et al., 2010)



1.2. Image selection

Until collection 5.0 the classification was performed by using Landsat 5 (TM), 7 (ETM+) and 8 (OLI) top of atmosphere (TOA) data. In the collection 6.0, we adopted the use of surface reflectance (SR) data, being the use of TOA discontinued. In collection 7 we adopted USGS Landsat 8 Level 2, Collection 2, Tier 1.

The mosaic of images consists of a composition of the best pixels that are extracted from all the images available in a defined period within a year. Once the initial and final dates of this period were defined, the median pixel from that period was calculated, generating one median image with several bands. The aggregation of these composed pixels was conducted for each year, producing the annual Landsat mosaics, which were then submitted to classification.

For the selection of Landsat scenes to be used for building the mosaics of each chart for each year, within the acceptable period, a threshold of 50% of cloud cover was applied (i.e., any available scene with up to 50% of cloud cover was accepted). This limit was established based on a visual analysis, after many trials observing the results of the could removing/masking algorithm. When needed, due to excessive cloud cover and/or lack of data, the acceptable period was extended to encompass a larger number of scenes in order to allow the generation of a mosaic without holes. Ever when possible, this was mad by including months in the beginning of the period, in the winter season.

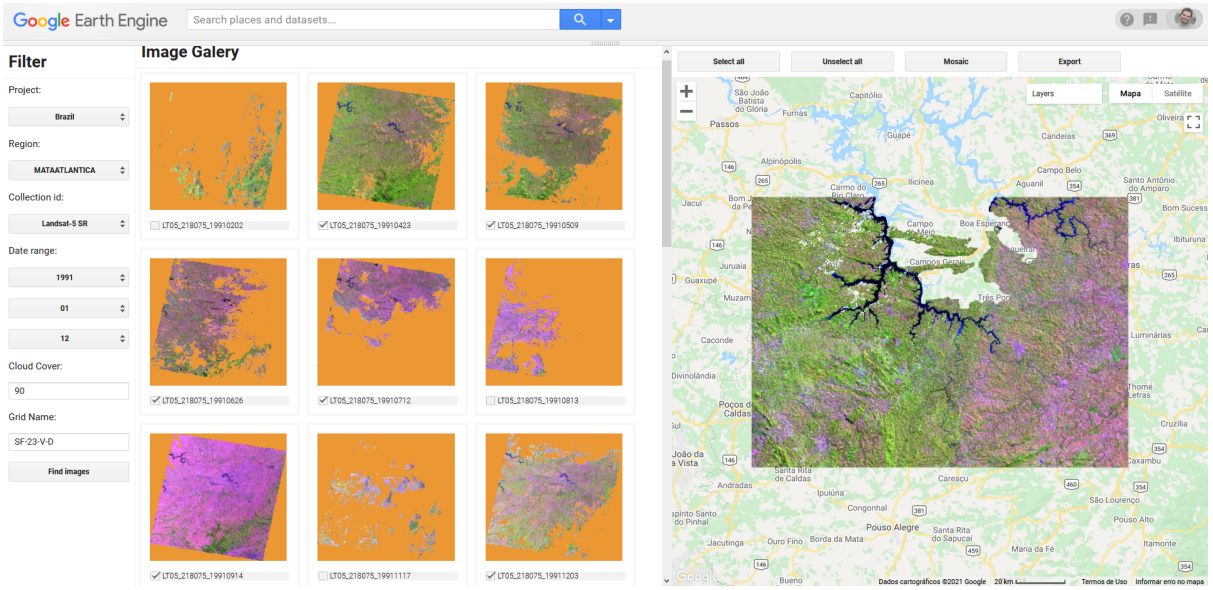
In most cases the period from May 1st to 30 of august was good to get a good mosaic with none or few missing information caused by clouds and shades.

For each year we used images from the best Landsat available:

- 1985 to 1999 – Landsat 5
- 2000 to 2002 – Landsat 7
- 2003 to 2011 – Landsat 5
- 2012 – Landsat 7
- 2013 to 2019 – Landsat 8

We made a visual analysis on the preliminary mosaics to identify and remove images with noises (clouds, shadow, or sensor defect) for each year (Figure 3).

Figure 3. Google Earth Engine tool to identify and remove scenes with noise



1.3. Final quality

As a result of the selection criteria, most of mosaics presented satisfactory quality. The first years of the collection still have some noise caused by haze.

1.4. Dry and Wet Bands

Dry and wet period bands are generated to the main spectral indices. The annual values of each pixel were divided into quartiles by the NDWI index. The median value of the quartile with the lowest NDWI value was considered as a reference for the dry period bands. The median value of the quartile with the highest NDWI value was considered a reference for the wet period bands.

2. Definition of regions to Classify

The Biome were divided in 6 regions (figure 4) based in dry and wet areas to classify more homogenous regions reduce confusion of samples and classes and allow a better balance of samples.

Figure 4. Regions of Pantanal biome

3. Classification

3.1. Classification scheme

The digital classification of the Landsat mosaics for the Pantanal biome aimed to individualize a subset of 8 classes from the complete legend of MapBiomas Collection 3 (Table 2), which were integrated with the cross-cutting themes in a further step.

Table 2. Vegetation cover/land use categories considered for digital classification of Landsat mosaics for the Pantanal biome.

Legend class of Collection 7	Numeric ID	Color
1.1.1. Forest Formation	3	
1.1.2. Savanna Formation	4	
2.1. Wetland	11	
2.2. Grassland Formation	12	
3.3. Mosaic of Agriculture or Pasture	21	
3.2.1.5. Other Temporary Crops	41	
4.5. Other non vegetated area	25	
5. Water	33	

- The Class 11 represents the maximum flood in each year without any temporal filter to better represent the Pantanal flood dynamic.

"Permanently flooded areas or flooded areas that flooded at least once a year, particularly due to the flood pulse.

The pulse of water flooding over the countryside is a natural factor. Areas with a water surface, but difficult to classify due to the amount of macrophytes, eutrophication or sediments, were also included in this category.

The woody element may be present on the field matrix forming a mosaic with shrub or tree plants"

- The Class "3.3 Mosaic of Agriculture or Pasture **(ID 21)**" is converted to "3.1. Pasture (ID 15)" during integration.
- The Class "3.2.1.5. Other Temporary Crops" is mapped in "Cárceres (MT)" and "Miranda (MS)" and this classes area overlayed by Other Temporary Crops **(ID 41)** during integration.

3.2. Feature space

The feature space for digital classification of the categories of interest for the Pantanal biome comprised a subset of 30 variables (Table 3), taken from the complete feature space. They include the original Landsat reflectance bands, as well as vegetation indexes, spectral mixture modeling-derived variables and wet and dry bands. The definition of the subset was made based on a feature importance analysis produced with Random Forests classification with all bands and 500 interactions.

Table 3. Feature space subset considered in the classification of the Pantanal biome Landsat image mosaics, 1985-2021.

blue_median_wet	green_min	red_min
blue_min	latitude	savi_median
cai_median	longitude	savi_median_wet
evi2_median	ndvi_median_dry	swir1_median
evi2_median_dry	ndvi_median_wet	swir1_median_dry
evi2_median_wet	nir_median	swir1_median_wet
gcvi_median	nir_median_dry	swir1_min
gcvi_median_wet	nir_median_wet	swir1_stdDev
green_median	nir_min	swir2_median_dry
green_median_wet	red_median_wet	swir2_min

3.3. Classification algorithm, training samples and parameters

Digital classification was performed chart by chart, year by year, using a *Random Forests* algorithm (Breiman, 2001) available in *Google Earth Engine*. Training samples for each chart were defined following a strategy of using pixels for which the vegetation cover/land use remained the same along the 36 years of Collection 6, so named “stable samples”. An ensemble taken from three main sources was made: extracted from Collection 5; manually drawn polygons; and complementary samples.

3.3.1. Stable samples from collection 6

The extraction of stable samples from the previous Collection 6 followed several steps aiming to ensure their confidence for use as training areas. We have identified the predominant, secondary, and rare class and in each region. The areas that did not change class from 1985 to 2020 in collection 6 were used to generate 7.000 random training points in each region. During the classification the balance was done reducing the number of stable samples adjusted for each class.

The samples from forest, savana, grassland, pasture and water were filtered using data from Global Forest Canopy Height (GFCH), 2019 (Potapov, 2019) based on GEDI data using the following rules:

- Forest sample need to be > 7
- Savana sample need to be > 3 and < 6
- Grassland and Pasture sample need to be < 2
- Water sample need to be < 1

3.3.2. Complementary samples

The need for complementary samples was evaluated by visual inspection. Complementary sample collection was done by means of drawing polygons but 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 37 years during the polygon drawing. Based in the knowledge of each regions samples from forest, savanna, grassland or wetlands where added. Samples from forests that where not well represented in the stable map where also added where need.

3.3.3. Final classification

Final classification was performed for all regions/years with stable and complementary samples. All years used the same subset of samples, and it was trained in the same mosaic of the year that was classified.

4. Pos-classification

4.1. Temporal Gap Fill filter

In this filter, no-data values ("gaps") are theoretically not allowed and are replaced by the temporally nearest valid classification. In this procedure, if no "future" valid position is available, then the no-data value is 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.

4.2. Temporal filter

The temporal filter uses the subsequent years to replace pixels that has invalid transitions.

In the first process looks in a 3-year moving window to correct any value that is changed in the middle year and return to the same class next year. This process is applied in this order: [19,3, 4, 21, 12, 33].

In the second process is similar to the first process, but it is a 4- and 5-years moving window that corrects all middle years.

In the third process the filter looks any native vegetation class [12, 3, 4] that is not this class in 85 and is equal in 86 and 87 and then corrects 85 value to avoid any regeneration in the first year.

In the last process the filter looks pixel value in 2020 that is not [12,21] and is equal 21 in 2018 and 2019. The value in 2020 is then converted to [12,21] to avoid any regeneration in the last year.

4.3. Frequency filter

A frequency filter was applied only in pixels that are considered “stable natural vegetation” (at least 35 years as [3, 11, 12, 4]). If a “stable natural vegetation” pixel is at least 60% of years of the same class, all years are changed to this class. The result of this frequency filter is a classification with more stable classification between natural classes (ex: forest and savanna). Another important result is the removal of noises in the first and last year in classification.

4.4. Regeneration filter

A new filter was applied to prevent Forest or Savanna Formations converted to anthropic to regenerate as Natural Grassland Formation. Thus, it’s eliminated part of the confusion of pasture areas and Grassland Formation. This filter is used to make sure any area converted to pasture stay as pasture for, at least, 6 consecutive years to avoid some confusion in the classification.

4.5. Wetland

The Wetland were produced applying a slice in NDDI index, produced by the normalized difference with ‘ndvi_median_wet’ and ‘ndwi_median_wet’. All pixels with NDDI less than 1.140 were classified as Wetland and integrated in land use and land cover map after all post classification filters except the spatial filter.

4.6. Spatial filter

The spatial filter avoids unwanted modifications to the edges of the pixel groups (blobs), 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 do not share connections to a predefined number of identical neighbors are considered isolated. In this filter, at least six connected pixels are 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. Validation strategies

Global accuracy (considering all years) was 85.1% and 78.3% in levels 1 and 2 of the collection 5 and collection 6 have 81.6% and 73.5% in levels 1 and 2, respectively.

In collection 7 the Global accuracy is 85,8%, 78,7% in levels 1 and 2, respectively. Wetland and Water was added in the map by the maximum in each year and this criterion is different from the interpretation methodology. To fix the sample point of grassland or wetland in water was considered correct. The detailed information about inclusion and omission error are presented in Figure 9. and Figure 10.

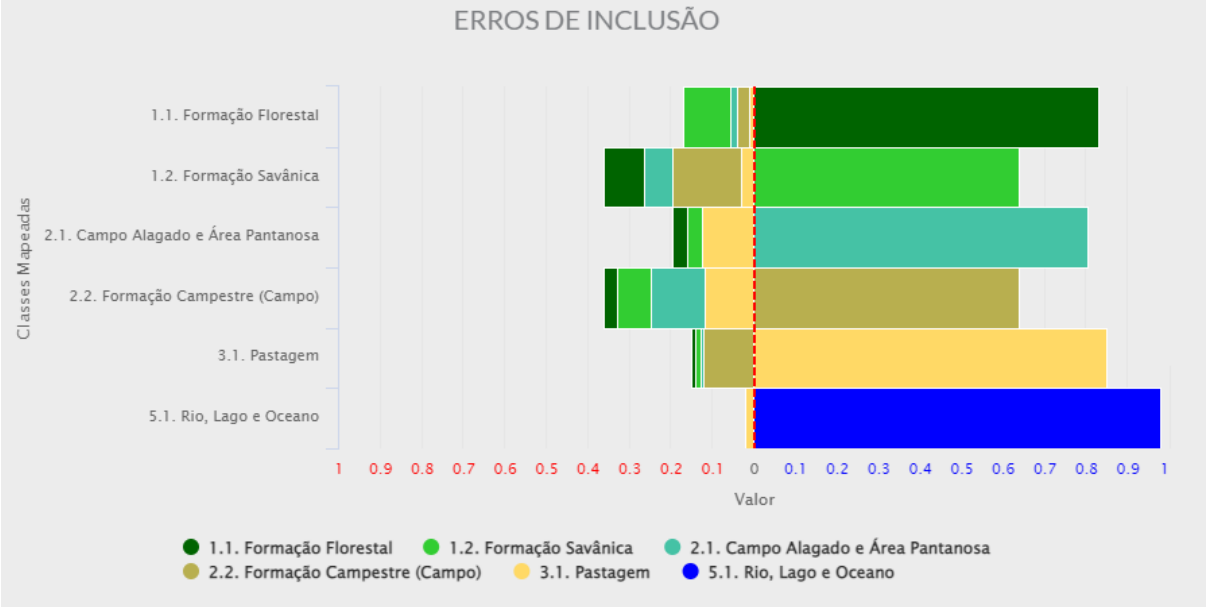


Figure 9. Inclusion error in 2018 in Pantanal for each level 2 class

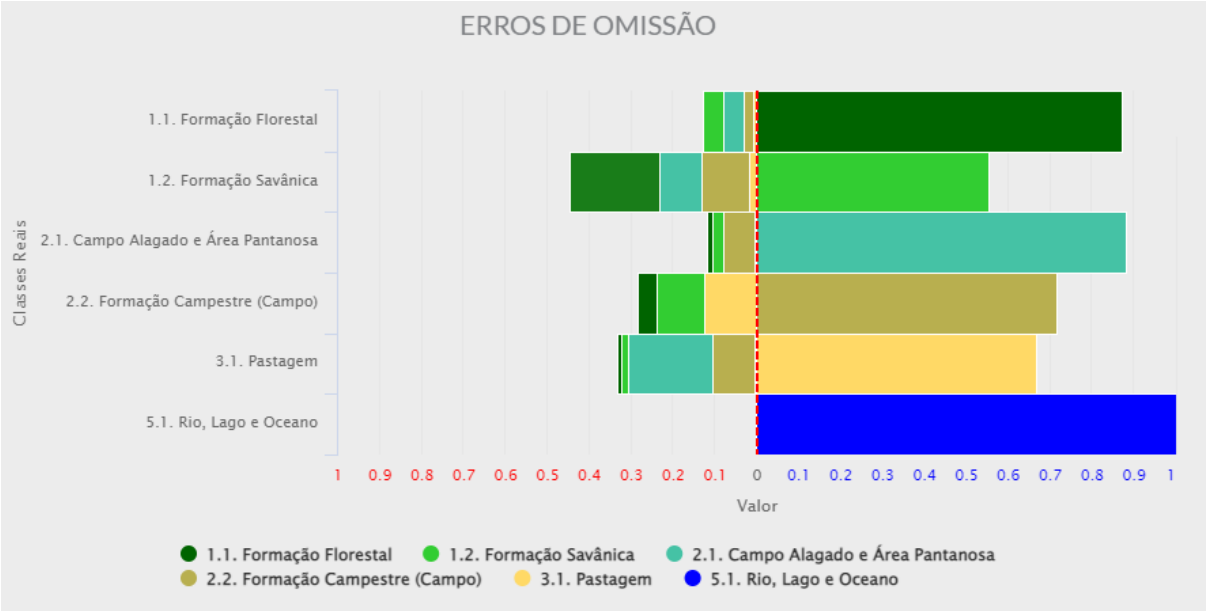


Figure 10. Omission error in 2018 in Pantanal for each level 2 class.

6. References

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

P. Potapov, X. Li, A. Hernandez-Serna, A. Tyukavina, M.C. Hansen, A. Kommareddy, A. Pickens, S. Turubanova, H. Tang, C.E. Silva, J. Armston, R. Dubayah, J. B. Blair, M. Hofton (2020) Mapping and monitoring global forest canopy height through integration of GEDI and Landsat data. *Remote Sensing of Environment,* 112165. (https://doi.org/10.1016/j.rse.2020.112165)

Monitoramento das alterações da cobertura vegetal e uso do solo na Bacia do Alto Paraguai – Porção Brasileira – Período de análise: 2002 a 2008. Relatório técnico metodológico. Brasília: CI – Conservação Internacional/ECOA - Ecologia e Ação/Fundación AVINA/Instituto SOS Pantanal/WWF-Brasil. 2009. 58 p.; Il.; 23 cm. ISBN 978-85-86440-25-0