Amazon - Appendix

Collection 6.0

Version 1

General Coordinator
Carlos Souza Jr., *Ph.D.*

Team
Luis Oliveira Jr., *Msc*
Antônio Fonseca
João Victor Siqueira
Stefany Pinheiro
Júlia Ribeiro
Bruno Ferreira
Raissa Ferreira
Márcio Sales, *Msc*

1 Overview

The Amazon mapping in the MapBiomas Project has been evolving through the Collections launched since 2015 (Table 1). Initially, the method used decision trees for image classification. In Collection 6, the Random Forest classifier was applied to build the land use and land cover maps in the Amazon biome. The Wetlands were included as a new class using a post-classification approach in the mapping. We classified all available Landsat scenes (according to the established criteria) and then integrated the results to obtain the annual maps. In the past Collections, we ran the classification using annual Landsat mosaics. This methodological change allowed us to evaluate all spectral variations contained within a year.

Table 1. The evolution of the Amazon 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	Mapped classes	Method/ Mapping Unit	Global Accuracy
Beta & 1	8 years 2008-2015	Forest; Non-Forest; Water Mask and Cloud Mask	Empirical Decision Tree / Annual Landsat Mosaic	
2.0 & 2.3	16 years 2000-2016	Without Information; Dense Forest; Inundated Forest, Degraded Forest; Secondary Forest; Nature Non-Forest Formations; Agriculture and Pasture; Non-Vegetated Areas; Water Surface; Unobserved	Empirical Decision Tree Random Forest (2.3) / Annual Landsat Mosaic	
3.0 & 3.1	33 years 1985-2017	No Information; Forest Formation; Other Nature Non-Forest Formation; Mosaic of Agriculture and Pasture; Other Non-Vegetated Area; River, Lake and Ocean; Non Observed	Random Forest / Annual Landsat Mosaic	[3.1] Level 1: 91.8% Level 2: 90.6%
4.0 & 4.1	34 years 1985-2018	No Information; Forest Formation; Other Non-Forest Natural Formation; Pasture; Agriculture; Other Non-Vegetated Area; River, Lake and Ocean	Random Forest / All Selected Landsat Scenes	Level 1: 95.9% Level 2: 95.7%
5.0	35 years 1985-2019	No Information; Forest Formation; Savanna Formation; Grassland Formation; Pasture; Agriculture; Other Non-Vegetated Area; River, Lake and Ocean	Random Forest / All Selected Landsat Scenes	Level 1: 98% Level 2: 97.8%
6.0	36 years 1985 - 2020	No Information; Forest Formation; Savanna Formation; Wetland; Grassland Formation; Pasture; Agriculture; Other Non-Vegetated Area; Non Observed; River, Lake and Ocean	Random Forest / All Selected Landsat Scenes	Level 1: 97.5% Level 2: 97.1%

2 Landsat images

The MapBiomas Collection 6 generated annual maps of land use and land cover for 36 years (1985 to 2020). All Landsat images available for this period (Landsat 5 [L5], Landsat [L7], and Landsat 8 [L8]) were used with Cloud Cover (CC) less or equal to 50%. The mapping unit for this collection is the Landsat path-row. Figure 1 shows the distribution of Landsat path-rows in the Amazon biome. The classification results were later integrated with the mapping units used by the MapBiomas Initiative (Figure 1).

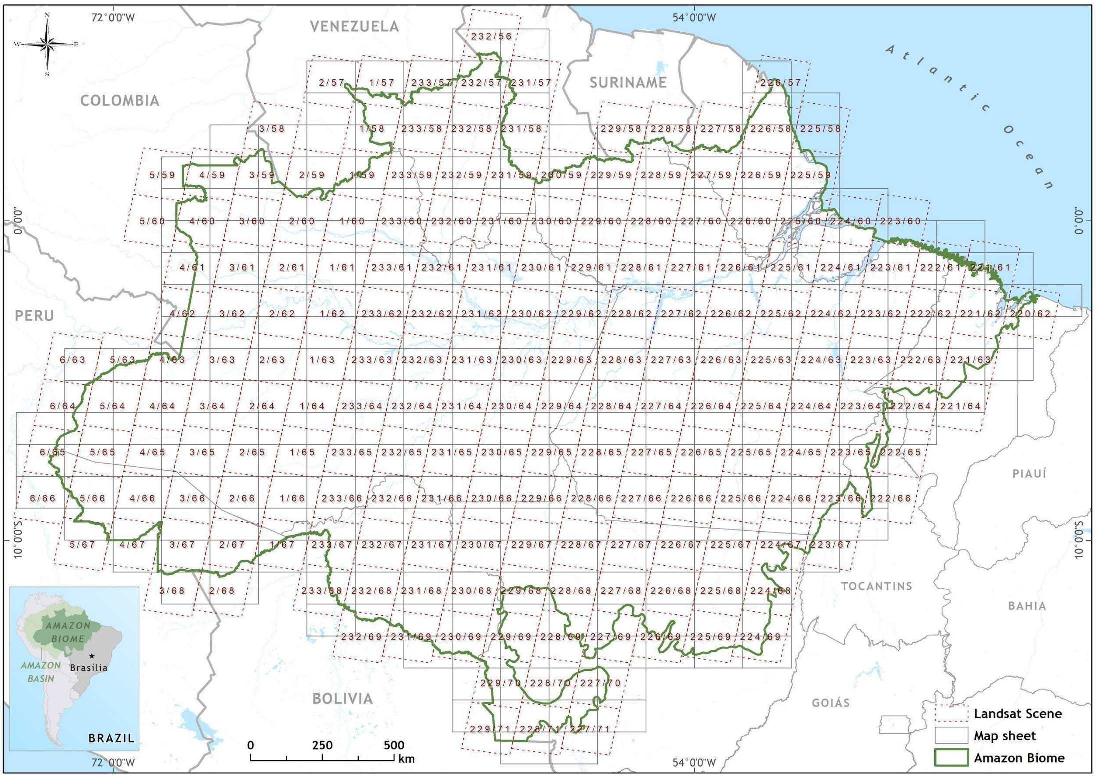

Figure 1. Distribution of Landsat path-rows for MapBiomas Amazon biome.

A total of 201 path-rows cover the entire Amazon biome, representing over 77,000 Landsat images in the time series. Figure 2 shows the number of images used each year by Landsat sensors for the Amazon biome.

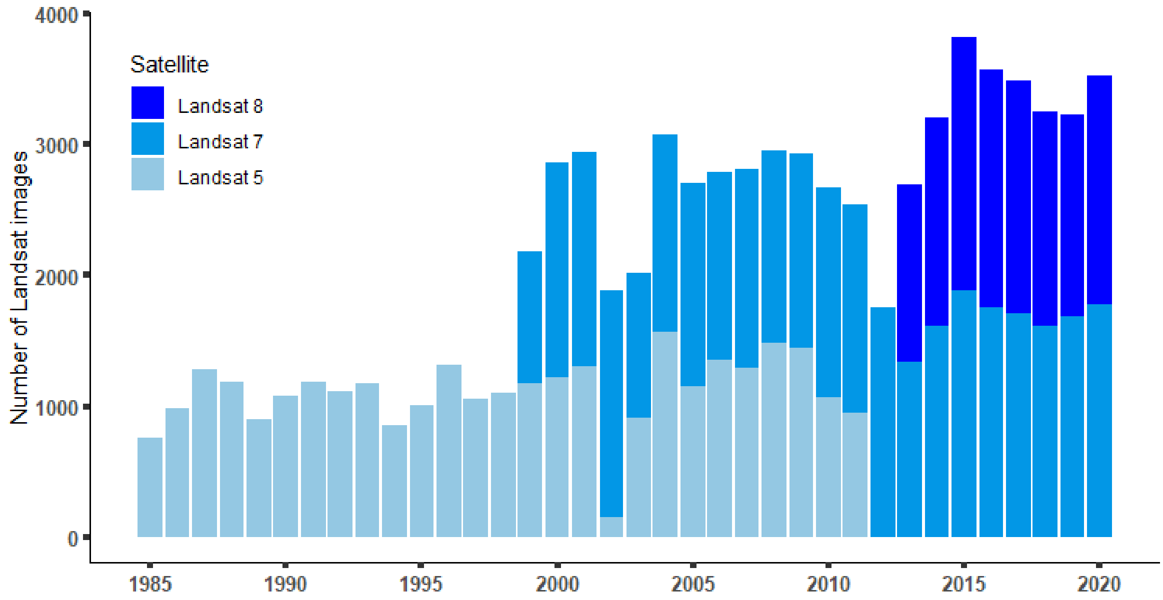

Figure 2. The number of Landsat images used per year and by Landsat sensors in the Amazon biome in Collection 6.

2.1 List of Landsat images removed from the database

We created a list of images that could contaminate the classification results, removing them from the analysis to extract the best information from the Landsat collection. The image can be removed for reasons like cloud cover, haze, no data, and Landsat 7 lines.

3 Classification

The Collection 6 method had three main steps:

1) **Image Selection**: we selected the Landsat 5, 7, and 8 scenes filtering by the sensor, date range, and cloud cover;
2) **Random Forest Calibration, Training, and Classification**: In that step, we ran an analysis to identify the best parameters to generate an optimized Random Forest Classifier (RFC). We trained the RFC using the samples produced by LAPIG/UFG plus new samples created by data augmentation analysis and classified all selected Landsat scenes. Finally, we integrated the classification results in each path-row to generate the annual Land Use and Land Cover (LULC) maps;
3) **Post-classification**: Wetlands mapping using LULC annual maps and MapBiomas Water intra-annual frequency. Temporal and Frequency filters were applied on the annual maps. The last step was to integrate them with the cross-cutting themes and run the accuracy analysis.

Figure 3 shows the process flow used to produce MapBiomas Collection 6 to Amazon biome.

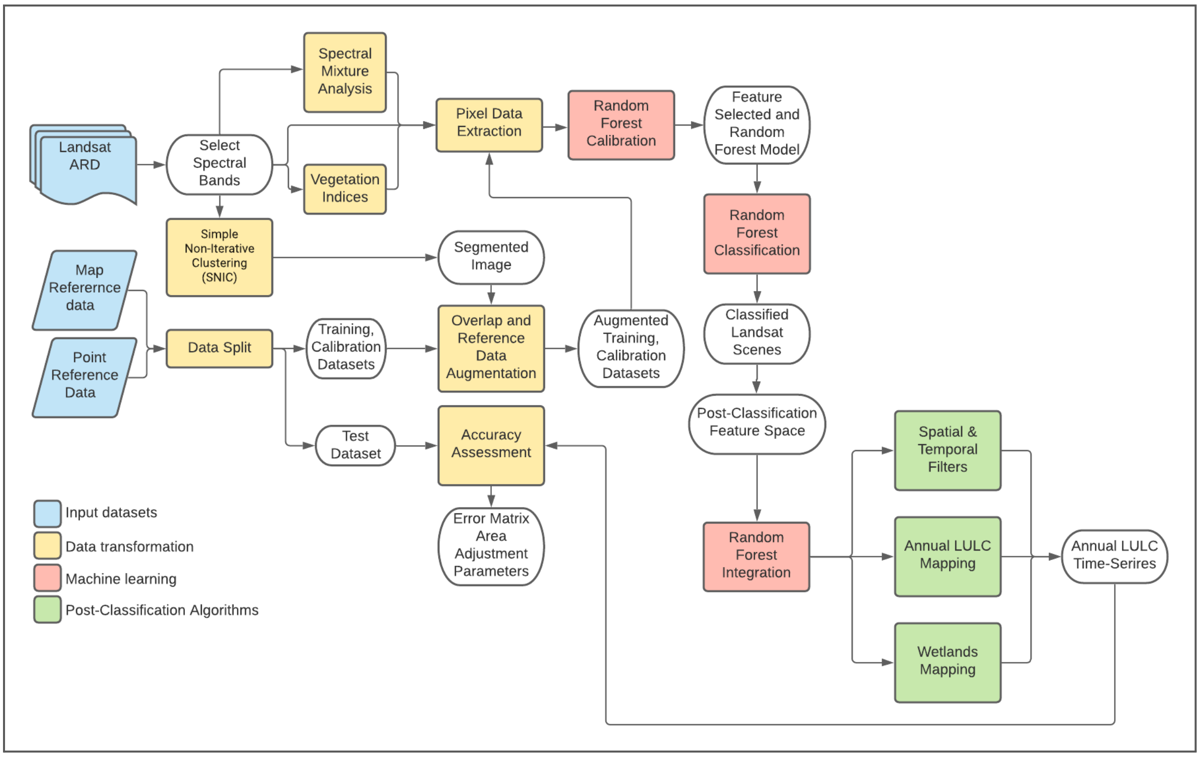

Figure 3. Classification process of Collection 6 in the Amazon biome.

3.1 Classification scheme

We mapped the same classes as the previous collection adding the Wetland as a new one. Table 2 shows all classes mapped for Collection 6 in the Amazon biome.

Table 2. Classification scheme of Collection 6 for the Amazon biome.

Value	Color	Color code	Class
0		#FFFFFF	No Information
3		#006400	Forest Formation
4		#32CD32	Savanna Formation
11		#45C2A5	Wetland
12		#B8AF4F	Grassland Formation
15		#FFD966	Pasture
19		#E974ED	Agriculture
25		#FF99FF	Other Non-Vegetated Area
27		#D5D5E5	Non Observed
33		#0000FF	River, Lake, and Ocean

These classes are a subset of the whole MapBiomas classification system and were the primary input for classification integration with other classes of cross-cutting themes and biomes (which is discussed in this document in the following sections).

For Collection 5, the class Other Non-Forest Formation (ONFF) was replaced by Savanna Formation (SF) and Grassland Formation (GF). We classified the Landsat images, including SF and GF samples to map these classes in the Amazon/Cerrado ecotone. In areas outside of Amazon/Cerrado ecotone, the class ONFF was replaced by GF, which is the most prevalent native vegetation class in these areas previously mapped as ONFF.

For collection 6, we revisited the ONFF samples to separate SF from GF samples, this effort enabled the mapping of SF and GF classes for the entire biome. The 2020 LULC map was built using the updated samples and added to Collection 6. The next step was to select the path-rows that had pixels classified as ONFF (replaced by GF) from 1985 to 2019 in Collection 5 for reclassification using the updated samples. Figure 4 shows the 95 path-rows reprocessed for Collection 6, in the blank areas we reused Collection 5 information.

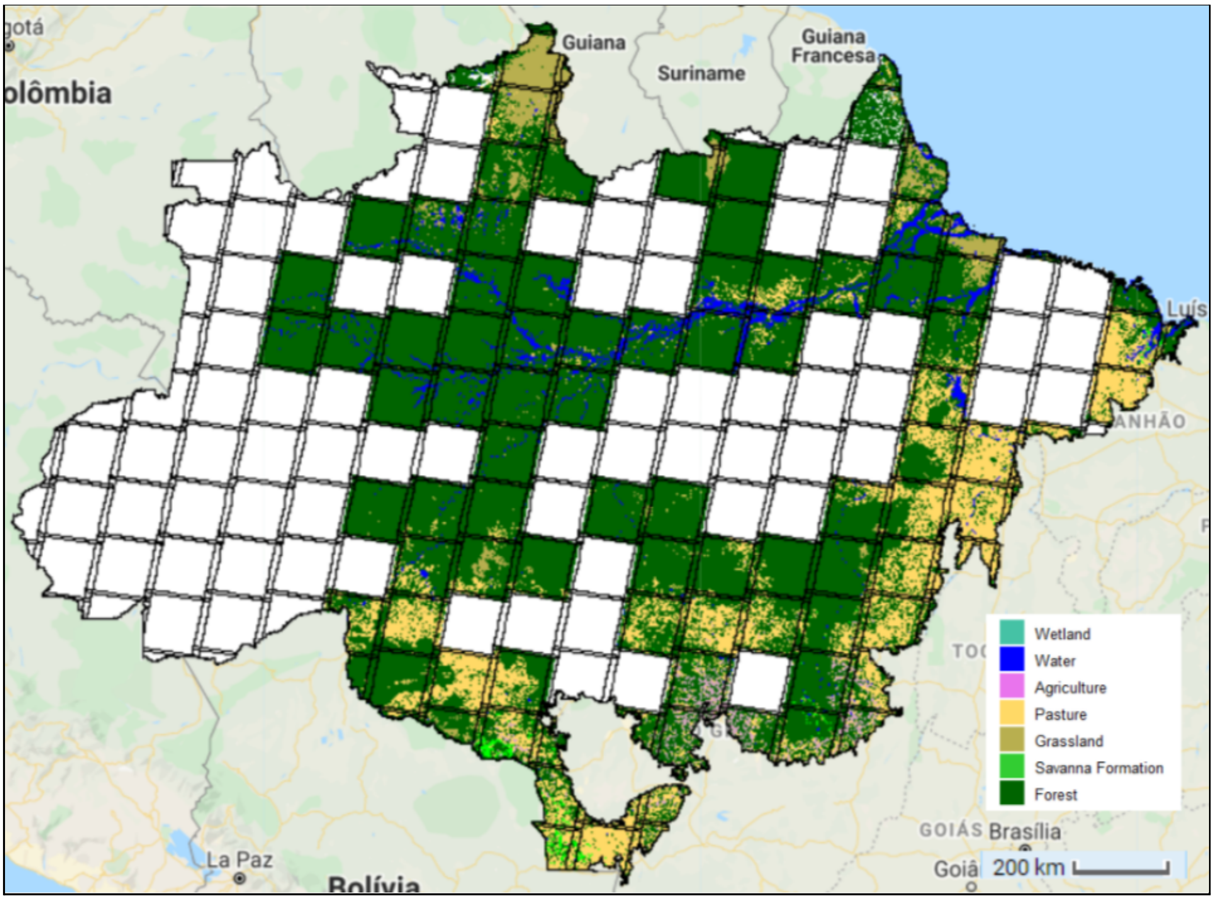

Figure 4. Path-rows reclassified for Collection 6 in the Amazon biome.

For more details about the description of classes mapped by MapBiomas Project see the document “Legend Description Collection 6” in MapBiomas website.

3.2 Feature space, classification algorithm, and training samples

The full feature space produced for the MapBiomas Collection 6 was analyzed using 35,000 random points for the Amazon biome, obtained from the reference dataset provided by LAPIG/UFG (from the one hundred thousand points collected for the whole country). Statistical analysis was done to define the minimum number of samples to estimate the accuracy assessment of all Level 2 classes in the Amazon biome. Therefore, the full reference dataset from LAPIG/UFG was split into two sets: training/calibration of the Random Forest Algorithm (RFA) classifier (10k), and accuracy assessment (~25k). The objective was to identify the most optimal features to be used in the Random Forest classifier to reduce computational cost and allow a better understanding of the response of the spectral features to map the target classes.

The feature selection process was conducted in R Language because Google Earth Engine does not have specialized statistical libraries for that. The final feature space ended up with eight variables, including Green Vegetation (GV), Non-Photosynthetic Vegetation (NPV), Soil, Cloud, Green Vegetation Shade (GVS), Normalized Difference Fraction Index (NDFI), Shade and Canopy Shade Fraction (CSFI). These features were selected using the feature importance algorithm available R Language RFA implementation (Table 3).

Table 3. Feature space subset used in the classification in the Amazon biome in the Collection 6.

ID	Variable	Description
1	GV	gv fraction
2	NPV	npv fraction
3	SOIL	soil fraction
4	CLOUD	cloud fraction
5	GVS	gv normalized fraction
6	NDFI	normalized difference fraction index
7	SHADE	shade fraction
8	CSFI	canopy shade fraction index

3.3 Additional samples for Collection 6

In addition to the 10,000 samples produced by LAPIG/UFG used as a reference dataset in the Random Forest classification, we added new samples in the classification using the following approach:

Regionalized samples for Amazon biome

To increase the number of samples in each Landsat scene and improve the RFA's training, we applied a segmentation technique in all images. As a result, we have segmented images that later were crossed with the samples from LAPIG/UFG. The segment touched by the reference dataset was used to sort new samples (new regional samples) randomly. We guarantee that 70% of RFA's training was run directly on the classified images. This approach was applied in the entire Amazon biome through the time series.

3.4 Accuracy sensitivity to inspected parameters

A sensitivity analysis was run to evaluate the effect of input parameters of the RFA on per-class user's and producer's accuracies of the classification outputs. The results indicated that these metrics had low sensitivity to input parameters. Three parameters were used for the RFA: *ntree* (number of trees to be estimated), *mtry* (number of variables in each tree), and *nodesize* (size of the tree). The user's and producer's accuracies were estimated for each of the parameters to define their values that optimize the computation time and accuracy. As a result, we defined a set of parameters that reduces the computational cost and increases the efficiency of the RFA. This analysis shows that the optimal values for the parameters were: *ntree*= 50, *mtry* = 7 and *nodesize* = 25.

3.5 Classification algorithm and training samples

The optimized version of RFA was implemented to produce Collection 6 using Google Earth Engine. The classifier's training dataset used 10,000 random samples from LAPIG/UFG plus the additional samples described in section 3.3 collected for the Amazon biome. All the selected Landsat scenes were classified based on the RFA. Each year in the time series has 201 Landsat path-rows, and each Landsat path-row can have from 0 to 56 Landsat scenes, according to Landsat sensors overlapping, and 0 to 23 when only one is in operation (Figure 2).

3.6 Path-row integration and annual maps

For Collection 4 and 5 the annual classification for each path-row was defined using a statistical measure of central tendency named *mode* (most frequent value in the observations) for each pixel. We also identified a set of post-classification rules (see Amazon ATBD Collection 5) to deal with some transitions not captured by mode in the time series. The union of all Landsat path-rows (mode product + post-classification rules) in the same year represents the LULC annual map.

For Collection 6 we calculate some metrics:

- Mode;
- Alternative Mode: mode of wet season;
- Total Transitions : number of all class changes in the time series;
- Transitions per Year: number of class changes in each year;
- Total Distinct: number of differents class changes in the time series;
- Distinct per Year: number of differents class changes in each year;
- Grassland, Savanna, Agriculture, and Water Total Occurrence: occurrence of these classes in the time series;

- Forest, Grassland, Savanna, Pasture, Agriculture and Water Occurrence per Year: occurrence of these classes in each year.

Initially, the metrics were calculated to improve the post-classification rules, but at some point, these rules got so complex that new adjustments brought new challenges to the mapping. Therefore we opted to use these metrics for training another round of RFA to integrate the results of classifications and let the algorithm decide based on these metrics which class will prevail in the final map. This approach allows us to automate this step in the Amazon mapping classification process, avoiding subjectivity brought by post-classification rules in the results integration. Finally, Collection 6 shows the changes in the Amazon landscape over the past 36 years.

4 Post-classification

4.1 Wetland Mapping

We added the Wetland class for collection 6 using a post-classification approach. The main goal was to separate areas with intense dynamics, flooded part of the year. To generate the annual wetland mapping, we crossed the LULC mapping with MapBiomas Water (monthly) product to observe how many months the dynamic areas were covered by water each year. We consider wetlands, pixels of Forest Formation, Savanna Formation, Grassland Formation, and Pasture that transitioned (changed class) more than 40% in some year and was covered up to 40% of the year by water.

4.2 Temporal filter

The temporal filter is a set of rules for non-allowed transitions applied to each image classified in a given year. That way, it was possible to remove clouds and correct non-allowed transitions. A number of 50 rules, distributed in three groups, were used: a) rules for cases not observed in the first year (RP); (b) rules for cases not observed in the final year (RU); (c) rules for examples of implausible transitions or not observed for intermediate years (RG) (Table 5).

Table 5. Temporal filter rules applied to Amazon Collection 6 land use and land cover classes. RG = General Rule, RP = First-Year Rule, RU = Last Year Rule, FF = Forest Formation, SF = Savanna Formation, GF = Grassland Formation, P = Pasture, AG = Agriculture, NO = Non-Observed, W = Water.

rule	type	kernel	active	tminus2	tminus1	t	tplus1	tplus2	result
RG01	RP	3	1	null	NO	FF	FF	null	FF
RG02	RP	3	1	null	NO	SF	SF	null	SF
RG03	RP	3	1	null	NO	GF	GF	null	GF
RG04	RP	3	1	null	NO	P	P	null	P
RG05	RP	3	1	null	NO	AG	AG	null	AG
RG06	RP	3	1	null	NO	W	W	null	W
RG07	RU	3	1	null	FF	FF	NO	null	FF
RG08	RU	3	1	null	SF	SF	NO	null	SF
RG09	RU	3	1	null	GF	GF	NO	null	GF
RG10	RU	3	1	null	P	P	NO	null	P
RG11	RU	3	1	null	AG	AG	NO	null	AG
RG12	RU	3	1	null	W	W	NO	null	W
RG13	RG	3	1	null	FF	NO	FF	null	FF
RG14	RG	3	1	null	SF	NO	SF	null	SF
RG15	RG	3	1	null	GF	NO	GF	null	GF
RG16	RG	3	1	null	P	NO	P	null	P
RG17	RG	3	1	null	AG	NO	AG	null	AG
RG18	RG	3	1	null	W	NO	W	null	W
RG19	RG	3	1	null	FF	SF	FF	null	FF
RG20	RG	3	1	null	FF	GF	FF	null	FF
RG21	RG	3	1	null	FF	P	FF	null	FF
RG22	RG	3	1	null	FF	AG	FF	null	FF
RG23	RG	3	1	null	FF	W	FF	null	FF
RG24	RG	3	1	null	SF	FF	SF	null	SF
RG25	RG	3	1	null	SF	GF	SF	null	SF
RG26	RG	3	1	null	SF	P	SF	null	SF
RG27	RG	3	1	null	SF	AG	SF	null	SF
RG28	RG	3	1	null	SF	W	SF	null	SF
RG29	RG	3	1	null	GF	FF	GF	null	GF
RG30	RG	3	1	null	GF	SF	GF	null	GF
RG31	RG	3	1	null	GF	P	GF	null	GF
RG32	RG	3	1	null	GF	AG	GF	null	GF
RG33	RG	3	1	null	GF	W	GF	null	GF
RG34	RG	3	1	null	P	FF	P	null	P
RG35	RG	3	1	null	P	SF	P	null	P
RG36	RG	3	1	null	P	GF	P	null	P
RG37	RG	3	1	null	P	AG	P	null	P
RG38	RG	3	1	null	P	W	P	null	P
RG39	RG	3	1	null	AG	FF	AG	null	AG
RG40	RG	3	1	null	AG	SF	AG	null	AG
RG41	RG	3	1	null	AG	GF	AG	null	AG
RG42	RG	3	1	null	AG	P	AG	null	AG
RG43	RG	3	1	null	AG	W	AG	null	AG
RG44	RG	3	1	null	W	FF	W	null	W
RG45	RG	3	1	null	W	SF	W	null	W
RG46	RG	3	1	null	W	GF	W	null	W
RG47	RG	3	1	null	W	P	W	null	W
RG48	RG	3	1	null	W	AG	W	null	W
RG49	RG	5	1	FF	FF	SF	P	P	P
RG50	RG	5	1	FF	FF	GF	P	P	P

4.3 Frequency filter for native classes

A frequency filter was applied for the Amazon/Cerrado ecotone region exclusively for the native vegetation classes: Forest Formation (FF), Savanna Formation (SF), and Grassland Formation (GF). If a pixel varied between these classes during the time series, the most frequent class would prevail, changing the classification in the years when that pixel was not classified as the most frequent class. The objective of the filter was a classification with more stable behavior between native classes. Other classes that may appear during the time series were not changed.

4.4 Integration with cross-cutting themes

After applying the temporal filter, the products of digital classification for each of the 36 years in the period 1985-2020 were then integrated with the cross-cutting themes by applying a set of specific hierarchical prevalence rules (Table 6). As the output of this step, a final land cover and land use map was obtained for each chart of the Amazon biome for each year.

There was only one exception in the prevalence rule in the integration with Forest Formation class and cross-cutting theme of Pasture for the Amazon biome. When the Pasture overlaps with the Forest Formation, the Pasture class prevailed to generate the integrated map.

Table 6. Prevalence rules for combining the output of digital classification with the cross-cutting themes in the Amazon biome in Collection 6.

Order	Class	Class ID	Source
1	Mining	30	Cross-cutting Theme
2	Beach and Dune	23	Cross-cutting Theme
3	Mangrove	5	Cross-cutting Theme
4	Aquaculture	31	Cross-cutting Theme
5	Salt Flat	32	Cross-cutting Theme
6	Urban Infrastructure	24	Cross-cutting Theme
7	Forest Plantation	9	Cross-cutting Theme
8	Sugar Cane	20	Cross-cutting Theme
9	Soybean	39	Cross-cutting Theme
10	Other Temporary Crops	41	Cross-cutting Theme
11	Other Perennial Crops	48	Cross-cutting Theme
12	Rocky Outcrop	29	Biome
13	Other non Vegetated Area	25	Biome
14	River, Lakes and Ocean	33	Biome
15	Forest Formation	3	Biome
16	Savanna Formation	4	Biome

17	Wetland	11	Biome
18	Grassland Formation	12	Biome
19	Pasture	15	Cross-cutting Theme

5 Validation strategies

5.1 Accuracy Analysis

The second dataset of ~25,000 reference samples, collected by LAPIG/UFG, was used for the validation dataset. For validation, we calculated and reported confusion matrices, user's, producer's, and overall accuracies, as well as the post-stratification class area estimates, along with 95% confidence intervals for each statistic.

The global accuracy analysis has increased over MapBiomas collections in the Amazon biome. Figure 5 shows the behavior of accuracy analysis since Collection 3.1 for the Amazon biome integrated maps.

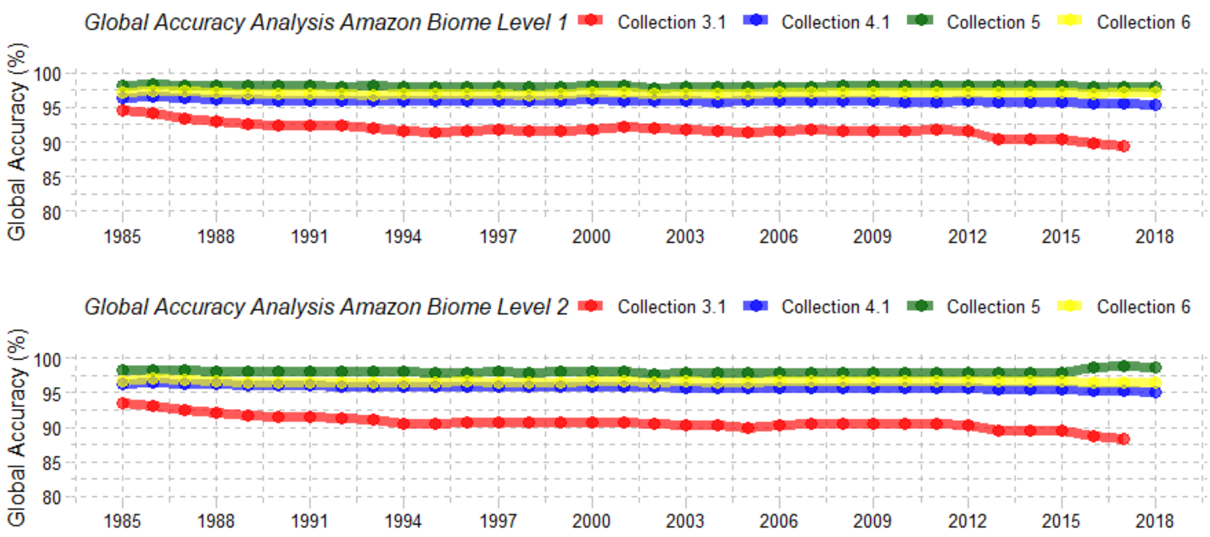

Figure 5. Accuracy analysis since Collection 3.1 for Amazon biome (Level 1 and 2).

We ran an Accuracy Analysis in the Amazon Biome Collection 6 maps before and after the integration with cross-cutting themes. Figure 6 shows the results:

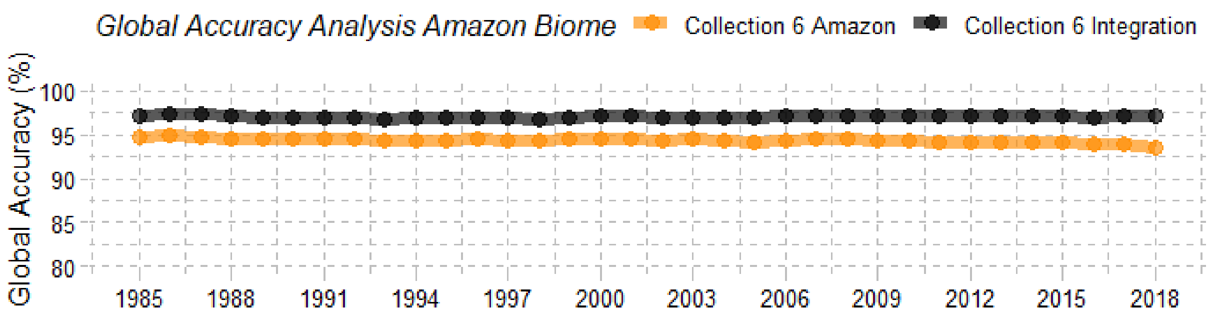

Figure 6. Comparison between Amazon's accuracy before and after integration with cross-cutting themes in Collection 6.

6 References

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R. 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone Remote Sens. Environ., Big Remotely Sensed Data: Tools, Appl. Experiences 202 18–27.

Souza, C.M., Jr.; Siqueira, J.V.; Sales, M.H.; Fonseca, A.V.; Ribeiro, J.G.; Numata, I.; Cochrane, M.A.; Barber, C.P.;Roberts, D.A.; Barlow, J. Ten-year landsat classification of deforestation and forest degradation in the Brazilian amazon. Remote Sens. 2013, 5, 5493–5513.

TerraClass 2014 Projeto TerraClass 2014 [WWW Document] (http://inpe.br/cra/projetos_pesquisas/terraclass2014. php) (Accessed: 23 July 2019).