Amazon - Appendix

Collection 4

Version 1

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

Team
Antônio Fonseca
Dalton Cardoso
Frederic Kirchhoff
Heron Martins
Jonas Paiva
Júlia Gabriela
Luis Oliveira Jr.
Marcelo Justino
Raissa Ferreira
Rodney Salomão

Consultor
João Siqueira
Marcio Sales

1 Main methodological differences (Collection 3.1 vs Collection 4)

There were main methodological differences between MapBiomas Collection 3.1 and Collection 4 for the Amazon biome, regarding satellite dataset, classification unit, sample collection, feature space and classification processing (Table 1).

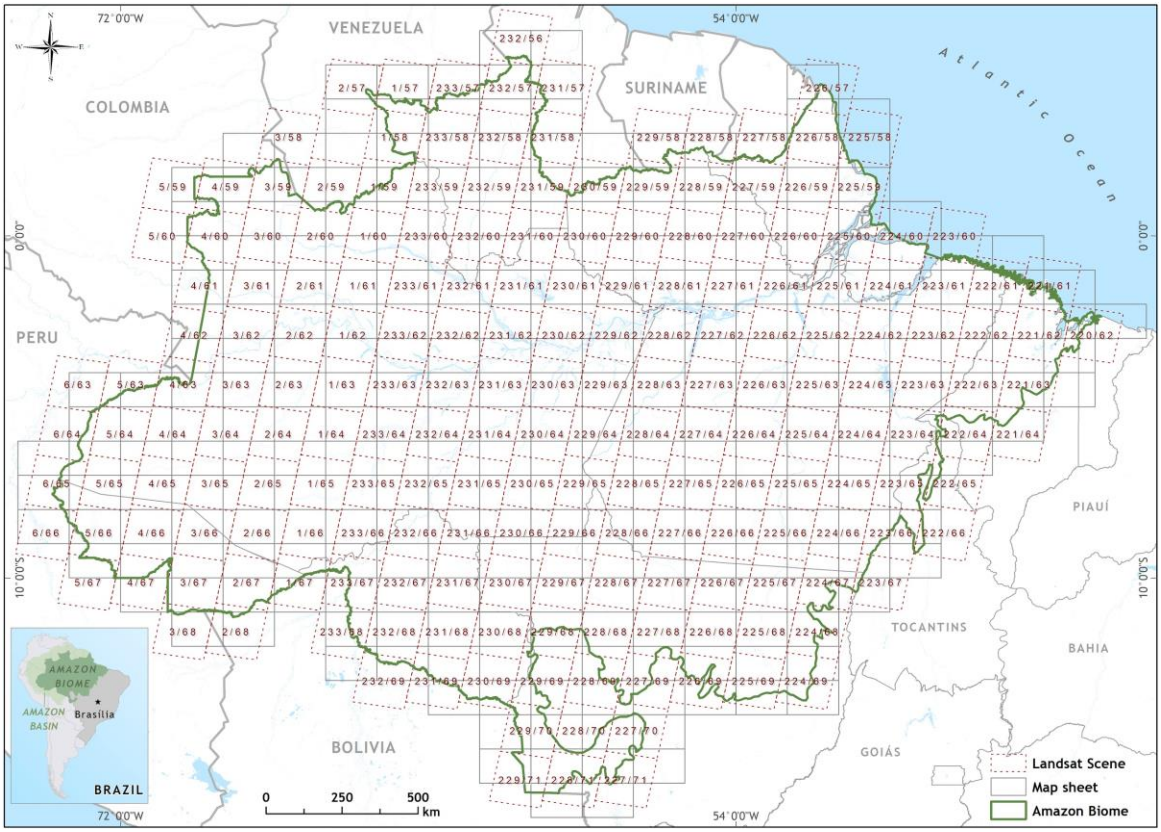
Table 1. Main differences between Collection 3.1 and Collection4 for the Amazon biome.

Parameter	Collection 3.1	Collection 4
Satellite Dataset	Landsat Median Mosaic (top-of-atmosphere)	Landsat Scene (Surface Reflectance)
Classification Unit	International Chart of the World to the Milionth (1:250,000 scale)	Landsat Path-Row
Sample Collection	Sample collection from reference maps	Aleatory samples from LAPIG (10k calibration / 25k validation)
Feature Space	median_(gvs, npv, soil, shade, ndfi) e stdDev_(gvs, npv, soil, shade, ndfi)	gv, npv, soil, cloud, gvs, ndfi, shade e csfi
Classifier	Random Forest	Optimized Random Forest (Mode + specific adjustment rules)

2 Landsat images

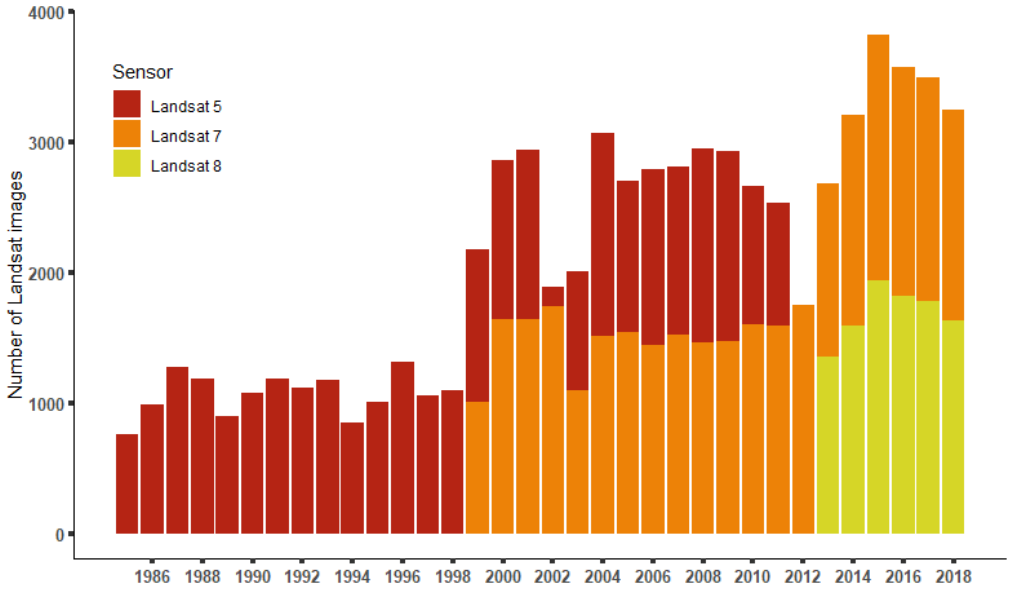
The MapBiomas Collection 4 generated annual maps of land use and land cover for 34 years (1985 to 2018). All Landsat images available for this period (Landsat 5 [L5], Landsat [L7], and Landsat 8 [L8]) were used with Cloud Cover (CC) less or equal 50%. The mapping unit for this collection is the Landsat 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 which represents over 71,000 Landsat images in the time series. Figure 2 shows the number of images used in each year by Landsat sensors for the Amazon biome.

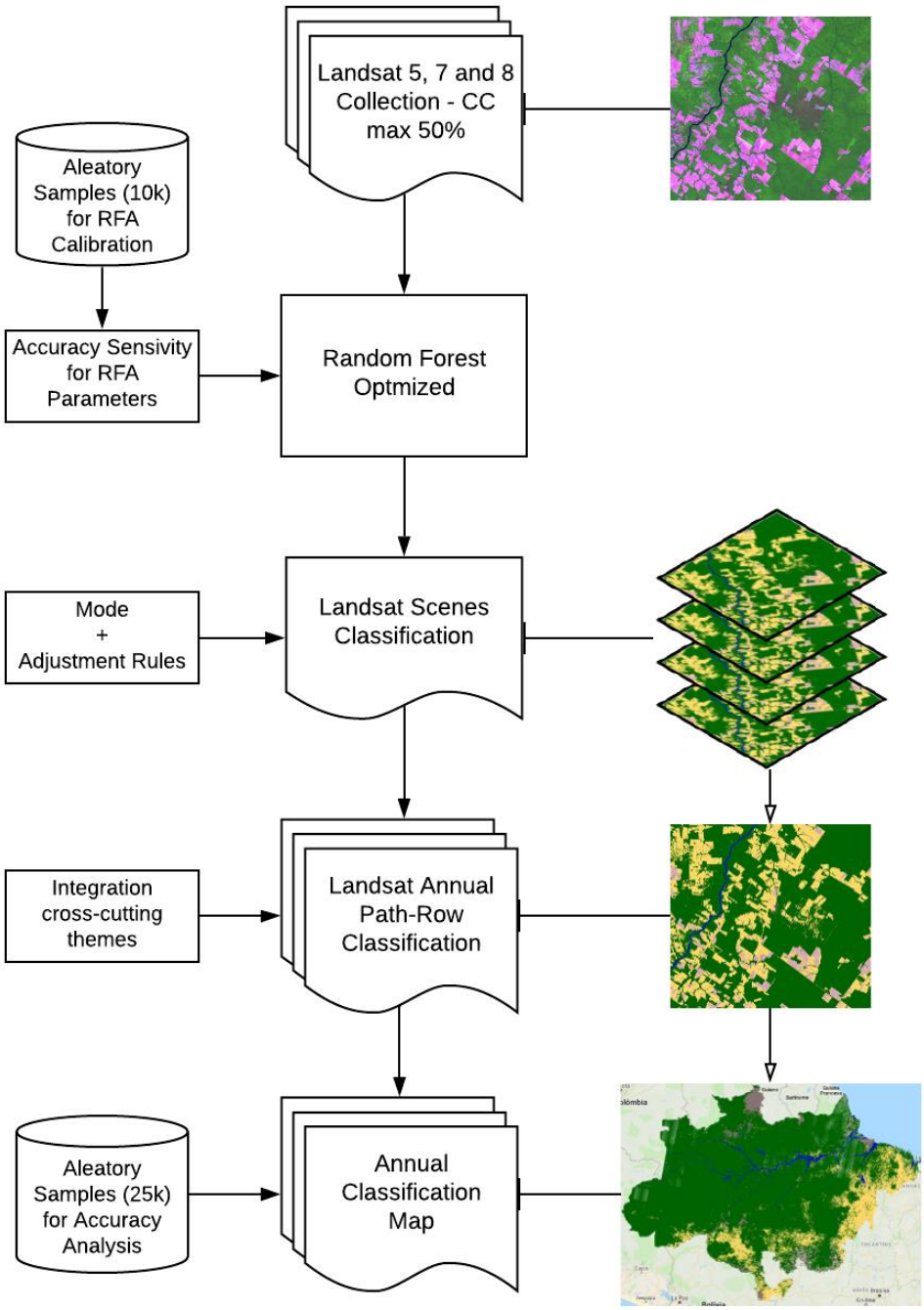
Figure 2. The number of Landsat images used per year and by Landsat sensors in the Amazon biome in the Collection 4.



3 Classification

Figure 3 shows the process flow used to produce MapBiomas Collection 4 to Amazon biome.

Figure 3. Classification process of Collection 4 in the Amazon biome.



3.1 Classification scheme

The classification scheme for Collection 4 of MapBiomas Amazon is presented in Table 2. These classes are a subset of the whole MapBiomas classification scheme and was the main input for classification integration with other classes of cross-cutting themes (which is discussed in this document in the following sections).

Table 2. Classification scheme of Collection 4 for the Amazon biome.

Value	Color	Color code	Class
0		#FFFFFF	No Information
3		#006400	Forest Formation
13		#BBFCAC	Other Non-Forest Natural Formation
15		#FFEFC3	Pasture
19		#FF99FF	Agriculture
25		#FF99FF	Other Non-Vegetated Area
33		#0000FF	River, lake and Ocean

3.2 Feature space, classification algorithm and training samples

The full feature space produced for the MapBiomas Collection 4 was analyzed using a 35,000 random points selected for the Amazon biome, obtained from the reference dataset produced by LAPIG (from the one hundred thousand points). Statistical analysis was done to define the minimal number of samples to estimate the accuracy assessment of all Level 2 classes of the Amazon biome. Therefore, the full reference dataset from LAPIG was split in 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 ending up with eight variables including: Green Vegetation (GV), Non-Photosynthetic Vegetation (NPV), Soil, Cloud, Green Vegetation Shade (GVS), NDFI, Shade and CSFI (Table 3). These features were selected using the feature importance algorithm available R Language RFA implementation.

Table 3. Feature space subset used in the classification of the Amazon biome in the Collection 4.

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 fracton
8	CSFI	canopy shade fraction index

3.3 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 above parameters to define their values that optimize the computation time and accuracy. As a result, we defined a set of parameters that reduces computational cost and increase the efficiency of the RFA. This analysis shows that the optimal values for the parameters were: *ntree*= 50, *mtry* = 7 and *nodesize* = 25.

3.3 Classification algorithm and training samples

The optimized version of RFA was implemented to produce Collection 4 using Google Earth Engine. The training dataset for the classifier used 10,000 random samples collected for the Amazon biome. All the selected Landsat scenes were classified based in the RFA. The classification of maps for all-time series makes possible to calculate the annual frequency of classes to define the final annual classification of each year.

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). In addition to that, we estimated the number of ground observation, which are pixels that had no cloud contamination.

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 defined a set of post-classification rules to deal with years and path-rows that had less than seven observed pixels observation per year and other peculiarities in the time series (Table 4).

Table 4. Post-classification rules used to integrate the Landsat path-row annual classification maps obtained with the RFA of the Amazon biome in the Collection 4.

Rule	Target/ID Class	Condition	Description
1	Pasture (15)	Path-row up to 7 observations	Pixel mapped in at least 1 observation as Pasture it will be classified as Pasture (15)
2	Pasture (15)	Path-row over 20 observations	Pixel mapped in at least 10 observation as Pasture, it will be classified as Pasture (15)
3	Pasture (15)	Forest Frequency + Pasture Frequency greater than 50% and mode = Non-Forest Natural Formation (NFNF)	When the pixel has in all time series Forest Frequency added Pasture Frequency greater than 50% and the mode is NFNF (13), it will be classified as Pasture (15)
4	Agriculture (19)	At least 3 observation as Agriculture and mode = Forest	Pixel mapped in at least 3 observation as Agriculture and the mode is Forest will be classified as Agriculture (19)
5	Agriculture (19)	At least 3 observation as Agriculture and mode = Pasture	Pixel mapped in at least 3 observation as Agriculture and the mode is Pasture will be classified as Agriculture (19)
6	Non Forest Natural Formation (13)	Non-Forest Natural Formation Frequency greater than 50%	When the pixel has in all time series Non-Forest Natural Formation Frequency greater than 50%, it will be classified as NFNF (13)

To obtain the mapsheets used in MapBiomas, we integrated all Landsat path-row in each year. This process allowed to generate the annual classification maps for MapBiomas Collection 4 of the Amazon biome.

4 Post-classification

4.1 Temporal filter

The temporal filter is a set of rules for non-allowed transitions that are applied to each image classified in a given year. That way, it is possible to remove clouds and correct non-allowed transitions. A number of 34 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 cases of implausible transitions or not observed for intermediate years (RG) (Table 5).

Table 5. Temporal filter rules applied to Amazon Collection 4 land use and land cover classes. RG = General Rule, RP = First-Year Rule, RU = Last Year Rule, FF = Forest Formation, FNNF = Other Non-Forest Natural Formation, P = Pasture, AG = Agriculture A = Water, NO = Non-Observed.

rule	type	kernel	active	biome	notes	tminus2	tminus1	t	tplus1	tplus2	result
RP01	RP	3	1	AMAZONIA	NO-FF-FF	null	27	3	3	null	3
RP05	RP	3	1	AMAZONIA	NO-A-A	null	27	33	33	null	33
RU01	RU	3	1	AMAZONIA	FF-FF-NO	null	3	3	27	null	3
RG09	RG	5	1	AMAZONIA	FF-FF-A-FF-FF	3	3	33	3	3	3
RG01	RG	3	1	AMAZONIA	FF-NO-FF	null	3	27	3	null	3
RG05	RG	3	1	AMAZONIA	P-NO-P	null	15	27	15	null	15
RG03	RG	3	1	AMAZONIA	FNNF-NO-FNNF	null	13	27	13	null	13
RG07	RG	3	1	AMAZONIA	FF-P-FF	null	3	15	3	null	3
RG08	RG	3	1	AMAZONIA	FF-FNNF-FF	null	3	13	3	null	3
RG10	RG	3	1	AMAZONIA	P-FF-P	null	15	3	15	null	15
RG12	RG	3	1	AMAZONIA	P-FNNF-P	null	15	13	15	null	15
RG04	RG	3	1	AMAZONIA	AG-NO-AG	null	19	27	19	null	19
RG06	RG	3	1	AMAZONIA	FF-AG-FF	null	3	19	3	null	3
RG11	RG	3	1	AMAZONIA	AG-FF-AG	null	19	3	19	null	19
RG13	RG	3	1	AMAZONIA	AG-FNNF-AG	null	19	13	19	null	19
RG14	RG	5	1	AMAZONIA	AG-AG-A-AG-AG	19	19	33	19	19	19
RG15	RG	5	1	AMAZONIA	P-P-A-P-P	15	15	33	15	15	15
RG16	RG	3	1	AMAZONIA	FNNF-FF-FNNF	null	13	3	13	null	13
RG17	RG	3	1	AMAZONIA	FNNF-AG-FNNF	null	13	19	13	null	13
RG18	RG	3	1	AMAZONIA	FNNF-P-FNNF	null	13	15	13	null	13
RG19	RG	5	1	AMAZONIA	FNNF-FNNF-A-FNNF-FNNF	13	13	33	13	13	13
RG20	RG	3	1	AMAZONIA	A-P-A	null	33	15	33	null	33
RG21	RG	3	1	AMAZONIA	A-AG-A	null	33	19	33	null	33
RG22	RG	3	1	AMAZONIA	A-FNNF-A	null	33	13	33	null	33
RP02	RP	3	1	AMAZONIA	NO-FNNF-FNNF	null	27	13	13	null	13
RP03	RP	3	1	AMAZONIA	NO-P-P	null	27	15	15	null	15
RP04	RP	3	1	AMAZONIA	NO-AG-AG	null	27	19	19	null	19
RU02	RU	3	1	AMAZONIA	P-P-NO	null	15	15	27	null	15
RU05	RU	3	1	AMAZONIA	A-A-NO	null	33	33	27	null	33
RU04	RU	3	1	AMAZONIA	AG-AG-NO	null	19	19	27	null	19
RU03	RU	3	1	AMAZONIA	FNNF-FNNF-NO	null	13	13	27	null	13
RG23	RG	3	1	AMAZONIA	AG-P-AG	null	19	15	19	null	19
RG24	RG	3	1	AMAZONIA	P-AG-P	null	15	19	15	null	15
RG02	RG	3	1	AMAZONIA	A-NO-A	null	33	27	33	null	33

4.2 Integration with cross-cutting themes

The products of digital classification after the application of the temporal filter, for each of the 34 years in the period 1985-2018, 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 vegetation cover/land use map for each chart of the Amazon biome for each year was obtained.

For the Amazon biome, there was only one exception in the prevalence rule in the integration in relation to forest formation class and crosscutting theme of pasture. In the case where the pasture overlaps with the forest formation, the pasture class prevails to generate the integrated map.

Table 6. Prevalence rules for combining the output of digital classification with the crosscutting themes in the Amazon biome in the Collection 4.

Order	Class	Font
1	4.1. Beach and Dune	Cross-cutting Theme
2	1.1.3. Mangrove	Cross-cutting Theme
3	5.2. Aquaculture	Cross-cutting Theme
4	5. Water	Biome
4	5.1. River, Lake and Ocean	Biome
5	2.3. Salt Flat	Cross-cutting Theme
6	1.2. Forest Plantation	Cross-cutting Theme
7	3.2. Agriculture	Cross-cutting Theme
7	3.2.1. Annual and Perennial Crop	Cross-cutting Theme
7	3.2.1. Semi-Perennial Crop	Cross-cutting Theme
8	1.1.1. Forest Formation	Biome
8	1.1.4. Secondary Forest	Biome
9	4.2. Urban Infrastructure	Cross-cutting Theme
10	4.4. Mining	Cross-cutting Theme
11	1.1.2. Savanna Formation	Biome
12	4.5. Rocky Outcrop	Biome
13	2.1. Wetland	Biome
13	3.2.3. Mosaic of Crops	Cross-cutting Theme
14	2.2. Grassland Formation	Biome
14	2.4. Other non-forest natural formation	Biome
14	3.1. Pasture	Cross-cutting Theme
15	4.3. Other non-vegetated Area	Biome
16	3.3 Mosaic of Agriculture or Pasture	Biome
17	6. Non Observed	Biome

5 Validation strategies

5.1 Accuracy Analysis

A second dataset of ~25,000 reference samples, collected by LAPIG using the same sampling design and independently from the first dataset (10,000 samples used for RFA calibration), 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.