Caatinga - Appendix

Collection 3

Version 1

General coordinator
Washington J. S. Franca Rocha (UEFS)

Team
Diego Pereira Costa (UEFS/GEODATIN)
Frans Pareyn (APNE)
José Luiz Vieira (APNE)
Rodrigo N. Vasconcelos (UEFS/GEODATIN)
Soltan Galano Duverger (UEFS/GEODATIN)
Taisson Monteiro (UEFS)

1 Landsat image mosaics

1.1 Definition of the temporal period

The image selection period for the Caatinga biome was defined aiming to minimize confusion between different natural vegetation and others land use and land cover (LULC) (*e.g.* cultivated areas) due to extreme phenological changes, while trying to maximize the coverage of Landsat images after cloud removing/masking. Unlike most of other Brazilian biomes, the climate of the Caatinga biome has a large seasonal variation of precipitation being the main factor determining the physiological behavior of vegetation throughout the year. Caatinga vegetation is classified as seasonal in their majority, expressing great deciduousness over the year. In fact, only a small fraction of tree species does not lose leaves during dry station, so that Caatinga Savanic formations are expected to show great variation in spectral response through the year. In order to define the periods for the mosaic construction, we used the rainfall data of the Northeast region of Brazil, considering the strong seasonal component in this region. Initially, an evaluation of the entire available time series (1961-2015) was made. This dataset was obtained from the INMET (www.inmet.gov.br). The data evaluation was performed through visual inspection of the annual graphs and historical averages for each of the climatic stations with data available for the Caatinga biome (Figure 1).

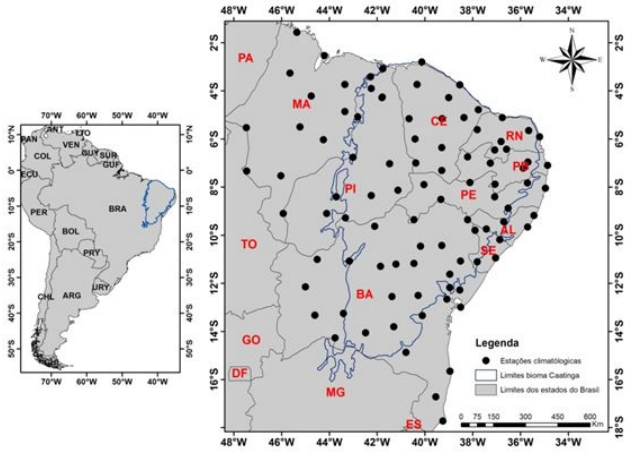

Figure 1. Location of the climatic stations used for the construction of the rainfall series for selection of the mosaic periods in the Caatinga biome.

Then, a periodic window scan was carried out for the entire Caatinga biome, indicating that the period between January to July (with higher levels of rainfall in the Caatinga biome) (Figure 2) is more likely to obtain images with spectral contrast capable of separating different classes of LULC for the biome. The choice of these sets of parameters helped to define the mosaics with better spectral quality and less amount of noise and clouds in the images for the biome.

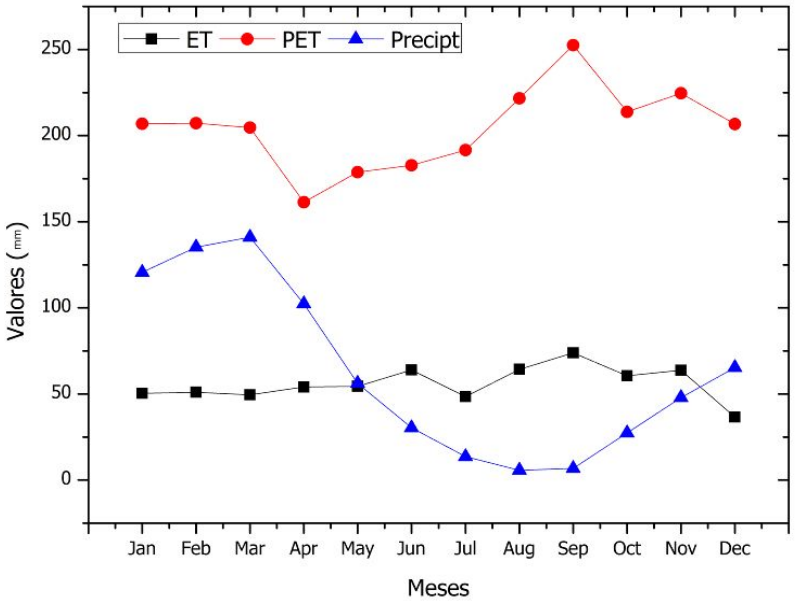

Figure 2. Temporal variation of water balance with monthly mean precipitation, evapotranspiration and potential evapotranspiration variables for Caatinga biome.

1.2 Image selection

For the selection of Landsat scenes to build the mosaics by map sheet for year, within the acceptable period, a threshold of 90% of cloud cover was applied (i.e. any available scene with up to 90% of cloud cover was accepted). 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 missing data. Whenever possible, this was made by including months in the beginning of the period, in the winter season.

For the generation of the mosaics by map sheet we used the parameters described (period and cloud cover). The selected Landsat scenes were processed to generate the temporal mosaic that covers the area of the chart.

1.3 Final quality

Considering the 68 map sheets of the Caatinga biome in a period of 33 years, a number of 2.244 mosaics were produced. The mosaic quality was evaluated using available frequency of each pixel in the Caatinga biome (Figure 3). As a result of the selection criteria, all of them presented satisfactory quality.

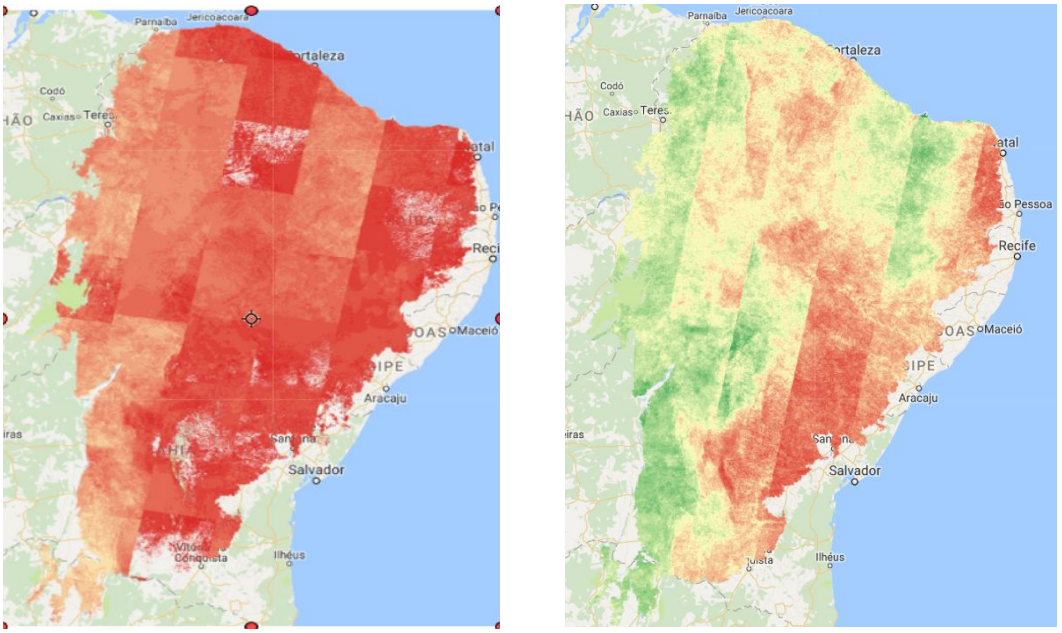

Figure 3. Landsat pixel availability in 1985 and 2017 in the Caatinga biome, where red is low, yellow is medium and green is high availability data pixel.

2 Classification

2.1 Classification scheme

The digital classification of the Landsat mosaics for the Caatinga biome aimed to individualize a subset of seven LULC classes from the MapBiomas legend in the Collection 3 (Table 1), which were integrated with the cross-cutting themes in a further step. The Mosaic class of Crops and Pasture in the Caatinga was later incorporated in the category Annual and perennial Crops in Agriculture or Pasture class, remaining areas of temporary crops (very common in the Caatinga biome) or where it was not possible to distinguish between these two classes.

Table 1. Land cover and land use categories considered for digital classification of Landsat mosaics for the Caatinga biome in the MapBiomas Collection 3.

Legend class of Collection 3	Numeric ID	Color
1.1.1. Forest Formation	3	
1.1.1. Savanic Formation	4	
2.2. Grassland Formation	12	
3.3 Mosaic of Agriculture or Pasture	21	
4.4. Rocky Outcrop	29	
5. Water	26	
6. Non Observed	27	

2.2 Feature space

The feature space for digital classification of the categories of interest for the Caatinga biome comprised a subset of 29 variables (Table 2), taken from the complete feature space of MapBiomas Collection 3. These variables include the original Landsat reflectance bands, as well as vegetation indexes, spectral mixture modeling-derived variables, terrain morphometry (slope), and a spatial texture measure. Definition of the subset was made based on the expected usefulness of each variable to discriminate the targets of concern, taking into account local knowledge about their spectral, spatial and temporal dynamics.

Table 2. Feature space subset considered in the classification of the Caatinga biome Landsat image mosaics in the MapBiomas Collection 3 (1985-2017).

ID	Variable	Description	Statistics	Temporal range	Script acronym
1	Blue	Landsat band	median	mosaic months	'median_blue'
2	Green	Landsat band	median	mosaic months	'median_green'
3	Near Infrared (NIR)	Landsat band	median	mosaic months	'median_nir'
4	Red	Landsat band	median	mosaic months	'median_red'
5	Shortwave Infrared (SWIR) 1	Landsat band	median	mosaic months	'median_swir1'
6	Shortwave Infrared (SWIR) 2	Landsat band	median	mosaic months	'median_swir2'
7	Evi 2	Enhanced Vegetation Index 2	median	mosaic months	'median_evi2'
8	Ndvi	Normalized Difference Vegetation Index	median	mosaic months	'median_ndvi'
9	Ndvi Dry	Normalized Difference Vegetation Index	median	year -below first quartile	'median_ndvi_dry'
10	Ndvi Wet	Normalized Difference Vegetation Index	median	year - above third quartile	'median_ndvi_wet'
11	Savi	Soil-adjusted Vegetation Index	median	mosaic months	'median_savi'
12	Sefi	Savanna Ecosystem Fraction Index	median	mosaic months	'median_sefi'
13	Sefi	Savanna Ecosystem Fraction Index	standard deviation	complete year	'stdDev_sefi'
14	Ndwi	Normalized Difference Water Index	median	mosaic months	'median_ndwi'
15	Ndwi Dry	Normalized Difference Water Index	median	year -below first quartile	'median_ndwi_dry'
16	Ndwi Wet	Normalized Difference Water Index	median	year - above third quartile	'median_ndwi_wet'
17	Gv	green vegetation fraction	median	mosaic months	'median_gv'
18	Gvs	GV / (100 - shade)	median	mosaic months	'median_gvs'
19	Ndfi	Normalized Difference Fraction Index	median	mosaic months	'median_ndfi'
20	Ndfi	Normalized Difference Fraction Index	standard deviation	complete year	'stdDev_ndfi'
21	Ndfi Dry	Normalized Difference Fraction Index	median	year -below first quartile	'median_ndfi_dry'
22	Ndfi Wet	Normalized Difference Fraction Index	median	year - above third quartile	'median_ndfi_wet'
23	Npv	npv fraction	median	mosaic months	'median_npv'
24	Npv	npv fraction	standard deviation	complete year	'stdDev_npv'
25	Shade	shade fraction	median	mosaic months	'median_shade'
26	Soil	soil fraction	median	mosaic months	'median_soil'
27	Soil	soil fraction	standard deviation	complete year	'stdDev_soil'
28	Green spatial texture	Spatial texture	median	mosaic months	'textG'
29	Slope	Slope	-	permanent	'slope'

2.3 Classification algorithm, training samples and parameters

Digital classification was performed chart by chart, year by year, using a Random Forest algorithm (Breiman, 2001) available in Google Earth Engine. Training samples for each chart were defined following a strategy of using pixels for which the LULC remained the same along the 33 years of Collection 3, so named “stable samples”. An ensemble taken from three main sources of samples was made extracted from: Collection 2.3, manually drawn polygons and Collection 3.

2.3.1 Stable samples from Collection 2.3

The extraction of stable samples from the previous Collection 2.3 followed several steps aiming to ensure their confidence for use as training areas. First, based on a visual analysis, a threshold was established for each class, specifying a minimum number of years in which a pixel should remained with that class to be eligible as a stable sample. A layer of pixels with a stable classification along the 17 years of Collection 2.3 was then generated by applying such thresholds. Later, a set of polygons in delineating zones with errors in some classes (*e.g.* omission or commission) was drawn and used as a mask to delete misclassified pixels. From the resulting layer of stable samples, a subset of pixels was randomly selected and used as training areas to classify all charts for each of the 33 years with the Random Forest algorithm, by running 50 iterations.

After this classification, a temporal filter was applied to each chart in order to improve the classification consistency of each pixel along the period 1985-2017. The output of the temporal filter was then submitted to the same procedures described above: definition and application of a threshold for the selection of stable pixels along the 33 years, followed by the exclusion of misclassified pixels by drawing mask polygons, and by comparison with a reference map of 2009.

2.3.2 Manually drawn polygons

Manually drawn polygons were used to add samples for classes with little occurrence, as well as to help to enrich class representation in zones which presented classification problems in the Collection 2.3. The polygons delineation was performed using WebCollect application, developed by theMapBiomas, and false-color composites of the Landsat mosaics as backdrop. Once more the concept of stable samples was applied: each of the polygons should delineate areas in which LULC remained unchanged, checking the mosaics for all the 33 years.

2.3.3 Preliminary classification

From both the sets of stable samples (stable samples from Collection 2.3 and manually drawn polygons), a subset of 5,000 pixels was randomly selected and used as training areas to classify all charts for each of the 33 years with the Random Forest algorithm, now running 100 iterations.

2.3.4 Final classification

Final classification was performed only for charts/years that had the need for complementary samples. These were previously merged with that from the manually drawn polygons in WebCollect, and then used as a source of training pixels for the Random Forest algorithm. Now 5,000 training pixels were randomly selected from this merge product, with the other parameters maintained the same used in the preliminary classification.

3 Post-classification

3.1 Temporal filter

The temporal filter rules were adapted for the classes used in the Caatinga biome and were complemented by specific rules to adjust cases where a pixel appeared two subsequent years in the class "Non Observed". A number of 79 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 (Table 3).

Table 3. Temporal filter general and specific rules for the Caatinga biome in the MapBiomas Collection 3. RG = General Rule, RP = First Year Rule, RU = Last Year Rule, FF = Forest Formation (3), AU = Savana Formation (4), FC = Grassland (12), AG = Mosaic of Agriculture and Pasture (21), AR = Rocky Outcrop (25), CD = Water Bodies (26), NO = Non Observed (27).

Rule	type	kernel	active	biome	tminus2	tminus1	t	tplus1	tplus2	result
R01	RG	3	1	CAATINGA	null	3	27	4	null	3
R02	RG	3	1	CAATINGA	null	3	27	12	null	3
R03	RG	3	1	CAATINGA	null	3	27	21	null	3
R05	RG	3	1	CAATINGA	null	3	27	26	null	3
R06	RG	3	1	CAATINGA	null	3	27	27	null	3
R08	RG	3	1	CAATINGA	null	4	27	3	null	4
R09	RG	3	1	CAATINGA	null	4	27	12	null	4
R10	RG	3	1	CAATINGA	null	4	27	21	null	4
R12	RG	3	1	CAATINGA	null	4	27	26	null	4
R15	RG	3	1	CAATINGA	null	12	27	3	null	12
R16	RG	3	1	CAATINGA	null	12	27	4	null	12
R17	RG	3	1	CAATINGA	null	12	27	21	null	12
R19	RG	3	1	CAATINGA	null	12	27	26	null	12
R21	RG	3	1	CAATINGA	null	4	27	27	null	4
R21	RG	3	1	CAATINGA	null	12	27	27	null	12
R23	RG	3	1	CAATINGA	null	21	27	4	null	21
R24	RG	3	1	CAATINGA	null	21	27	12	null	21
R25	RG	3	1	CAATINGA	null	21	27	3	null	21
R26	RG	3	1	CAATINGA	null	21	27	26	null	21
R27	RG	3	1	CAATINGA	null	26	27	3	null	26
R28	RG	3	1	CAATINGA	null	25	27	4	null	28
R33	RG	3	1	CAATINGA	null	26	27	12	null	26
R34	RG	3	1	CAATINGA	null	26	27	21	null	26
R35	RG	3	1	CAATINGA	null	26	27	25	null	26
R36	RG	3	1	CAATINGA	null	26	27	4	null	26
R38	RG	3	1	CAATINGA	null	26	27	27	null	26
R46	RG	3	1	CAATINGA	null	3	12	3	null	3
R47	RG	3	1	CAATINGA	null	3	21	3	null	3
R49	RG	3	1	CAATINGA	null	3	27	3	null	3
R50	RG	3	1	CAATINGA	null	4	27	4	null	4
R52	RG	3	1	CAATINGA	null	4	21	4	null	4
R53	RG	3	1	CAATINGA	null	12	3	12	null	12
R54	RG	3	1	CAATINGA	null	12	21	12	null	12
R56	RG	3	1	CAATINGA	null	12	27	12	null	12
R57	RG	3	1	CAATINGA	null	21	3	21	null	21
R58	RG	3	1	CAATINGA	null	21	4	21	null	21
R59	RG	3	1	CAATINGA	null	21	12	21	null	21
R61	RG	3	1	CAATINGA	null	21	27	21	null	21
R79	RG	3	1	CAATINGA	null	26	27	26	null	26
R81	RG	3	1	CAATINGA	null	21	27	27	null	21
R82	RG	3	1	CAATINGA	null	25	27	27	null	25
R83	RG	3	1	CAATINGA	null	3	26	3	null	3
R84	RG	3	1	CAATINGA	null	3	4	3	null	3
R85	RG	3	1	CAATINGA	null	3	25	3	null	3
R86	RG	3	1	CAATINGA	null	4	12	4	null	4
R87	RG	3	1	CAATINGA	null	4	3	4	null	4
R88	RG	3	1	CAATINGA	null	4	26	4	null	4
R89	RG	3	1	CAATINGA	null	12	4	12	null	12
R90	RG	3	1	CAATINGA	null	12	26	12	null	12
R91	RG	3	1	CAATINGA	null	21	26	21	null	21
R91	RG	3	1	CAATINGA	null	21	26	21	null	21
R92	RG	5	1	CAATINGA	21	21	3	3	21	21
RP07	RP	3	1	CAATINGA	null	27	3	3	null	3
RP08	RP	3	1	CAATINGA	null	27	4	4	null	4
RP09	RP	3	1	CAATINGA	null	27	13	13	null	13
RP10	RP	3	1	CAATINGA	null	27	12	12	null	12
RP11	RP	3	1	CAATINGA	null	27	21	21	null	21
RP13	RP	3	1	CAATINGA	null	27	26	26	null	26
RP14	RP	3	1	CAATINGA	null	27	27	3	null	3
RP15	RP	3	1	CAATINGA	null	27	27	4	null	4
RP16	RP	5	1	CAATINGA	27	27	27	3	3	3
RP17	RP	5	1	CAATINGA	27	27	27	27	3	3
RP18	RP	5	1	CAATINGA	27	27	27	4	4	4
RP19	RP	5	1	CAATINGA	27	27	27	27	4	4
RP20	RP	3	1	CAATINGA	null	27	27	12	null	12
RP21	RP	3	1	CAATINGA	null	27	27	26	null	26
RP22	RP	3	1	CAATINGA	null	27	21	21	null	21
RU009	RU	3	1	CAATINGA	null	3	3	27	null	3
RU010	RU	3	1	CAATINGA	null	4	4	27	null	4
RU011	RU	3	1	CAATINGA	null	11	11	27	null	11
RU012	RU	3	1	CAATINGA	null	12	12	27	null	12
RU013	RU	3	1	CAATINGA	null	21	21	27	null	21
RU015	RU	3	1	CAATINGA	null	26	26	27	null	26
RU016	RU	3	1	CAATINGA	null	21	27	27	null	21
RU017	RU	3	1	CAATINGA	null	3	27	27	null	3
RU018	RU	3	1	CAATINGA	null	4	27	27	null	4
RU019	RU	3	1	CAATINGA	null	12	27	27	null	12
RU020	RU	3	1	CAATINGA	null	21	21	25	null	21

3.2 Integration with cross-cutting themes

After the application of the temporal filter, for each of the 33 years in the period 1985-2017, the products of digital classification were then integrated with the cross-cutting themes, by applying a set of specific hierarchical prevalence rules (Table 4). As output of this step, a final vegetation LULC map for each chart of the Caatinga biome for each year was obtained.

Table 4. Prevalence rules for combining the output of digital classification with the cross-cutting themes in the Caatinga biome in the MapBiomas Collection 3.

Order	Class	Font
1	4.1. Beach and Dune	Crosscutting Theme
2	1.1.3. Mangrove	Crosscutting Theme
3	5.2. Aquiculture	Crosscutting Theme
4	5. Water	Biome
4	5.1. River, Lake and Ocean	Biome
5	2.3. Salt Flat	Crosscutting Theme
6	1.2. Forest Plantation	Crosscutting Theme
7	3.2. Agriculture	Crosscutting Theme
7	3.2.1. Annual and Perennial Crop	Crosscutting Theme
7	3.2.1. Semi-Perennial Crop	Crosscutting Theme
8	1.1.1. Forest Formation	Biome
8	1.1.4. Secundary Forest	Biome
9	4.2. Urban Infrastructure	Crosscutting Theme
10	4.4. Mining	Crosscutting 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	Crosscutting Theme
14	2.2. Grassland Formation	Biome
14	2.4. Other non forest natural formation	Biome
14	3.1. Pasture	Crosscutting Theme
15	4.3. Other non vegetated Area	Biome
16	3.3 Mosaic of Agriculture or Pasture	Biome
17	6. Non Observed	Biome

4 Validation strategies

4.1 Use of reference maps

Protocol validation was done based in 1,526 random points selected over the grid of the Brazilian National Forest Inventory performed by SFB-MMA (Figure 4).

4.2. Validation with independent points

WebCollect is a tool implemented to evaluate each point based on visual interpretation of the same Landsat mosaic used in the classification (Figure 5). Each point was evaluated by three different interpreters with experience in Landsat image interpretation and Caatinga mapping. The evaluation considers the exact pixel that is viewed in the image for each year. The interpreter was instructed to consider the rules of temporal filter applied in the classification. If the pixel is not available in one specific year, the interpreter should repeat the last visible class until a new image is available.

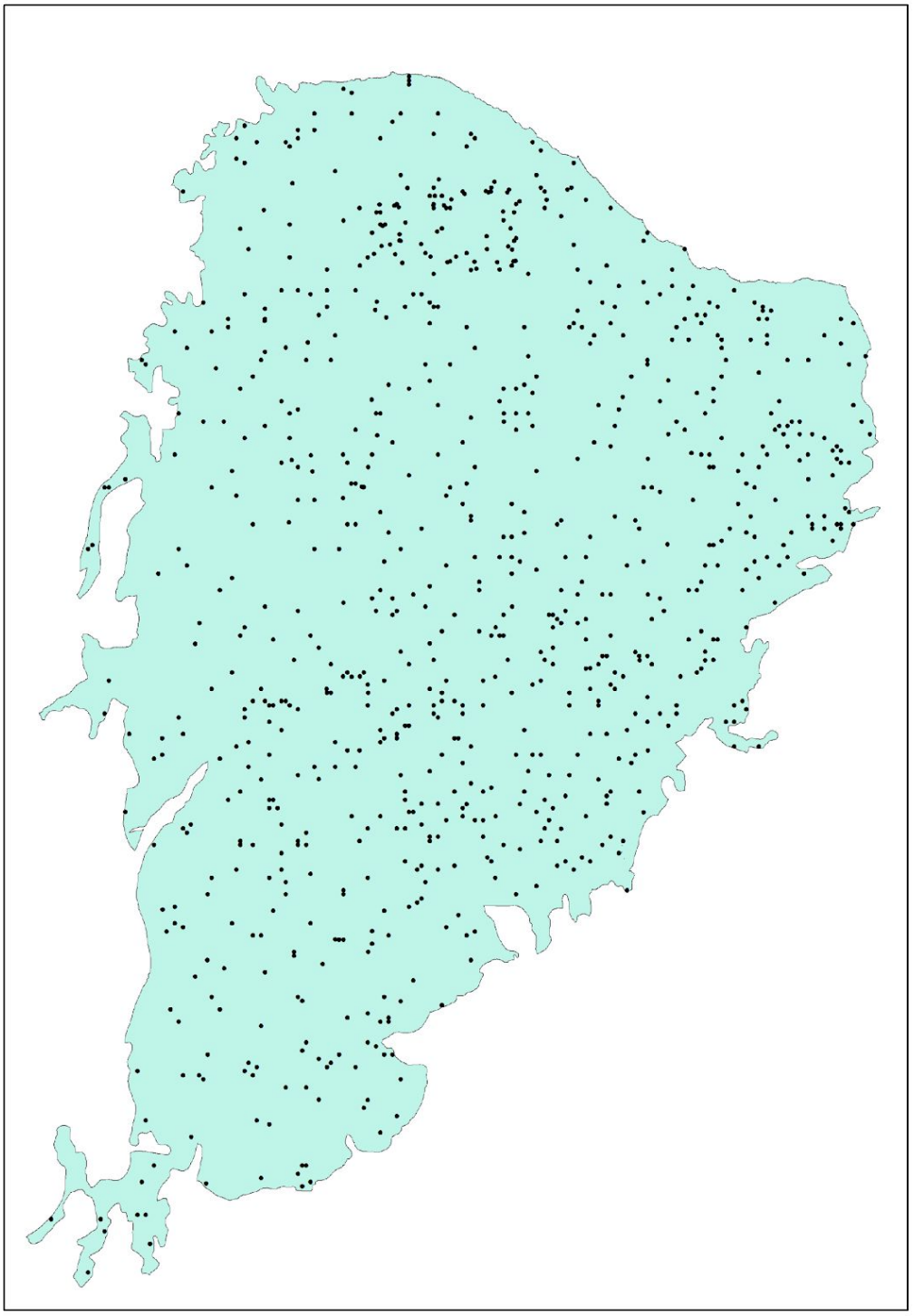

Figure 4. Spatial distribution of the 1,526 validation points in Caatinga biome in the MapBiomas Collection 3.

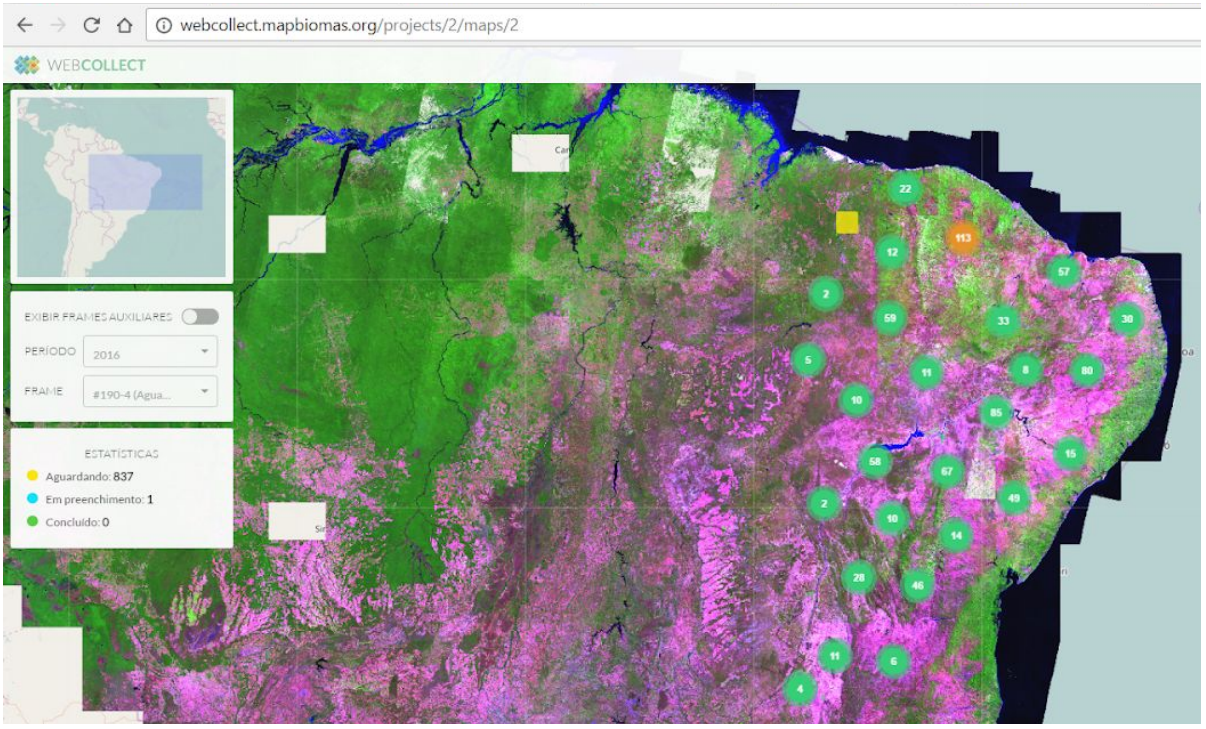

Figure 5. Data collection in WebCollect environment for validation of Collection 3 in the Caatinga biome.

The final class of each point was the class identified by at least 2 interpreters. This reference class of each year was compared with the map resulted from temporal filter to build the confusion matrix and evaluate omission and commission for each year.

In the first step of the accuracy analysis a random sampling was collected to estimate the overall accuracy of the mapping. In the second step, a random sample stratified by LULC class was collected. Mapping accuracy was inferred from the error matrix, to estimates global accuracy. These quantities was accompanied by their respective calculation of sample error and 95% confidence intervals.

5. References

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