MapBiomas Land Use and Land Cover (LULC) Collections

CAATINGA

Algorithm Theoretical Basis Document (ATBD)   Appendix

Collection 11 · Product: LULC · Country: Brazil
Version 4
Date: 08/2026

General Coordinator
Washington de Jesus Sant'anna da Franca Rocha (UEFS)

Team / Authors
Diego Pereira Costa (GEODATIN/UEFS);

Rodrigo Nogueira de Vasconcelos (GEODATIN/UEFS);

Nerivaldo Afonso Santos (GEODATIN/UEFS);

Rafael Oliveira Franca Rocha (GEODATIN/UEFS);

Soltan Galano Duverger (GEODATIN/UEFS);

Deorgia Tayane Mendes de Souza (UEFS/PPGM);

Jocimara Souza Lobão (UEFS/PPGM).

Code repository
https://github.com/mapbiomas/brazil-caatinga/tree/main/lulc_30m_landsat/collection_11

How to cite
MAPBIOMAS. Caatinga: Algorithm Theoretical Basis Document (ATBD)   Appendix, Collection 11, Version 4. Brazil, 2026. Available at: https://github.com/mapbiomas/brazil-caatinga/tree/main/lulc_30m_landsat/collection_11.

Table of Contents

1. Overview

The Caatinga is a semi-arid biome unique to Brazil, dominated by seasonally dry tropical vegetation (dry forest, shrubland and savanna-like physiognomies) and marked by strong rainfall seasonality and pronounced phenological dynamics, including seasonal leaf loss. These characteristics make land-use and land-cover (LULC) mapping particularly challenging because the spectral response of the same cover type can vary sharply between wet and dry seasons. The biome covers approximately 86.2 million hectares, about 10% of Brazil, and spans portions of ten states. This appendix documents the biome-specific methodology used to generate the annual Caatinga LULC series for MapBiomas Collection 11, covering 41 years (1985–2025) across 49 hydrographic basins. It complements, and does not replace, the MapBiomas General ATBD. The document is organized into image preprocessing and mosaics, classification, post-classification, validation, practical considerations, concluding perspectives and references. The maps support environmental monitoring, territorial planning, conservation assessment and transparent analysis of long-term LULC change. Methodologies from previous collections remain available on the MapBiomas ATBD page (https://brasil.mapbiomas.org/download-dos-atbds/).

Across collections the methodology evolved from empirical decision trees to machine-learning classifiers. Random Forest (BREIMAN, 2001) replaced the initial trial-and-error parameterisation from Collection 2.3 onward. In Collection 8.0, Gradient Tree Boosting (GTB) (LAWRENCE et al., 2004) was run in parallel with Random Forest, and from Collection 9.0 onward GTB became the main classifier for all mosaics, including Collection 11. Table 1 summarises this evolution.

Main changes introduced in Collection 11.

A. The series was extended to 2025.
B. Pasture (15), Agriculture (19) and Perennial Crop (36) are now target classes trained directly by the biome classifier, rather than being resolved only through cross-theme integration.
C. Training samples are labelled from the public Collection 10 over stable ('agreement') areas.
D. Sample balancing relies exclusively on undersampling/downsampling, including a probability-band strategy that preserves hard examples;
E. The post-classification chain introduces a structure-preserving Sieve filter that protects thin features such as rivers and riparian corridors;

F. Accuracy is now assessed with the Pontius quantity/allocation/exchange/shift decomposition (PONTIUS; MILLONES, 2011) in addition to global, user's and producer's accuracies.

Reproducibility.

All processing code for Collection 11 is available in the project repository at https://github.com/mapbiomas-brazil/caatinga.

Table 1. Overview of the LULC collections of the Caatinga biome.

Collection	Time interval	Method	Classes	Main improvements	Global accuracy (Level 2)
Beta & 1	2008 – 2015	Empirical Decision Tree	Forest Formation, Non-Forest, Water Mask	Proof of concept	
2.0 & 2.3	2000 – 2016	Empirical Decision Tree / Random Forest	Forest Formation, Savanna, Grassland, Mosaic of Agriculture and Pasture, Water, Other non-vegetated Areas	LULC sample collection / spatio-temporal filters	
3.0 & 3.1	1985 – 2017	Random Forest	Same as 2.3	Samples based on current mapped classes / added Mosaic of Agriculture and Pasture / new filters	71.3%
4.0 & 4.1	1985 – 2018	Random Forest	Same as 2.3	Samples based on current mapped classes / new spatio-temporal filters	74.3%
5.0	1985 – 2019	Random Forest	+ Rocky Outcrop	Stable points (5-yr windows) / feature importance / new RF parameters / processing by watershed	75.4%
6.0	1985 – 2020	Random Forest	Same as 5.0	New mosaic collection	74.9%
7.0 & 7.1	1985 – 2021	Random Forest	+ Herbaceous Sandbank Vegetation	New class (Herbaceous Sandbank Vegetation)	76.9%
8.0	1985 – 2022	Random Forest / GTB	Same as 7.1	GTB introduced in parallel with RF	76.9%
9.0	1985 – 2023	GTB / cluster	Same as 7.1	Rocky Outcrop mapped with a cluster model	79.3%
10.0	1985 – 2024	GTB / cluster / Deep Learning	+ Photovoltaic power plant	Photovoltaic plants mapped with U-Net (RONNEBERGER et al., 2015; PATIL et al., 2022); on-the-fly annual/seasonal mosaics	79.88%
11.0	1985 – 2025	GTB (+ cluster for Rocky Outcrop)	+ Herbaceous and Shrub Formation, Wind farm (beta)	Series to 2025; direct classification of pasture / agriculture / perennial crop; structure-preserving	80.38% (Level 2; consolidated 1985–2024).

Collection	Time interval	Method	Classes	Main improvements	Global accuracy (Level 2)
				Sieve; Pontius error decomposition	

Note: from Collection 11, the Photovoltaic power plant class (ID 75) and Wind farm (beta) are documented in a dedicated cross-cutting ATBD and is only cross-referenced here.

2. Image Pre-processing / Landsat Images / Mosaics

Describe how the Landsat imagery was selected, pre-processed and organised into the mosaics used for sample collection and classification.

2.1 Sensors and image collection

Classification uses Landsat Surface Reflectance data from Landsat 5 (TM), 7 (ETM+) and 8 (OLI). Since Collection 6.0 only Surface Reflectance products are used, and from Collection 7.0 onward Landsat Collection 2, Tier 1, Level-2 products are adopted. For Collection 11 the mosaics are built on-the-fly from Google's 32-day composite asset *LANDSAT/COMPOSITES/C02/T1_L2_32DAY*, using the six reflective bands blue, green, red, NIR, SWIR1 and SWIR2.

2.2 Time period and compositing periods

The series covers 1985–2025 (41 years), processed independently for each of the 49 hydrographic basins. For every year, three seasonal composites are generated:

- **Annual** (suffix _median): 01 Jan – 31 Dec;
- **Wet season** (suffix _median_wet): 01 Jan – 31 Jul;
- **Dry season** (suffix _median_dry): 01 Aug – 31 Dec.

The wet/dry split reflects the strong rainfall seasonality of the Caatinga, defined from long-term INMET rainfall records for the Brazilian Northeast, which directly drives vegetation phenology. **Fidelity note:** although the band suffixes read '_median', the active reducer in the production code is the per-pixel maximum (*.max()*); the classical MapBiomas variant based on EVI percentiles with a median reducer exists in the code but is disabled.

2.3 Image selection and quality control

Collection 11 relies on the pre-composited 32-day product rather than an explicit per-scene cloud/shadow threshold. Quality control is achieved through (i) the sampling 'agreement' mask, (ii) the removal of null observations (*dropNulls*), and (iii) the gap-fill procedure described below.

The 32-day composites present more pixel gaps than earlier products, particularly over bare soil. To correct this, gaps are filled from the previous MapBiomas mosaic (GEE path *projects/nexgenmap/MapBiomas2/LANDSAT/BRAZIL/ mosaics-2*), rescaled band-by-band by linear regression (per-band coefficients and intercepts) so the two datasets are radiometrically aligned before filling, figure 1.

Figure 1. *Result of the band-by-band linear regression between the 32-day (GEE) mosaic and the MapBiomas mosaic used for gap-fill.*

2.4 Feature bands of the mosaic

Beyond the reflective bands, the mosaic incorporates, for each of the three periods: about 31 spectral indices, spectral-mixture-analysis (SMA) fractions (green vegetation, non-photosynthetic vegetation, soil, shade, cloud) and the NDFIa, plus terrain (slope from NASADEM) and structural-context texture metrics (neighbourhood mean and standard deviation). The full band set and the corresponding feature space used in classification are detailed in Section 3.4.

2.5 Mosaic quality assessment

Mosaic quality is evaluated from the frequency of valid pixels available per year across the biome; the selection criteria yield mosaics with reduced noise from clouds, relief and shadows, figure 2. The biome boundary follows IBGE (2019).

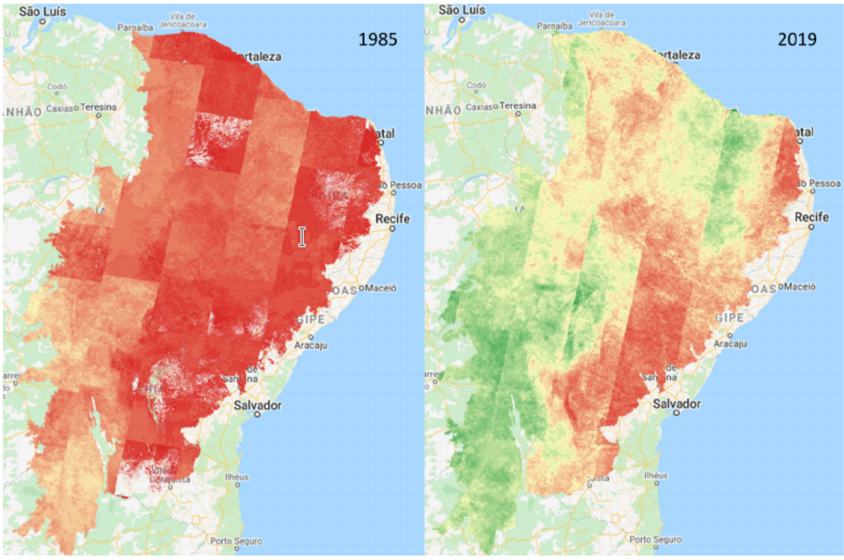

Figure 2. *Landsat valid-pixel availability across the Caatinga biome (low → high).*

2.6 Definition of regions for classification

Classifying spectrally homogeneous regions reduces within- and between-class variability and allows a consistent sample set over large areas, at a high computational cost. The Caatinga is therefore divided by hydrographic basins (Agência Nacional de Águas). Successive collections refined this division; Collection 11 uses **49 basins**. Sample collection itself is organised on a regular **30 km grid** overlaid on the basins, figure 3.

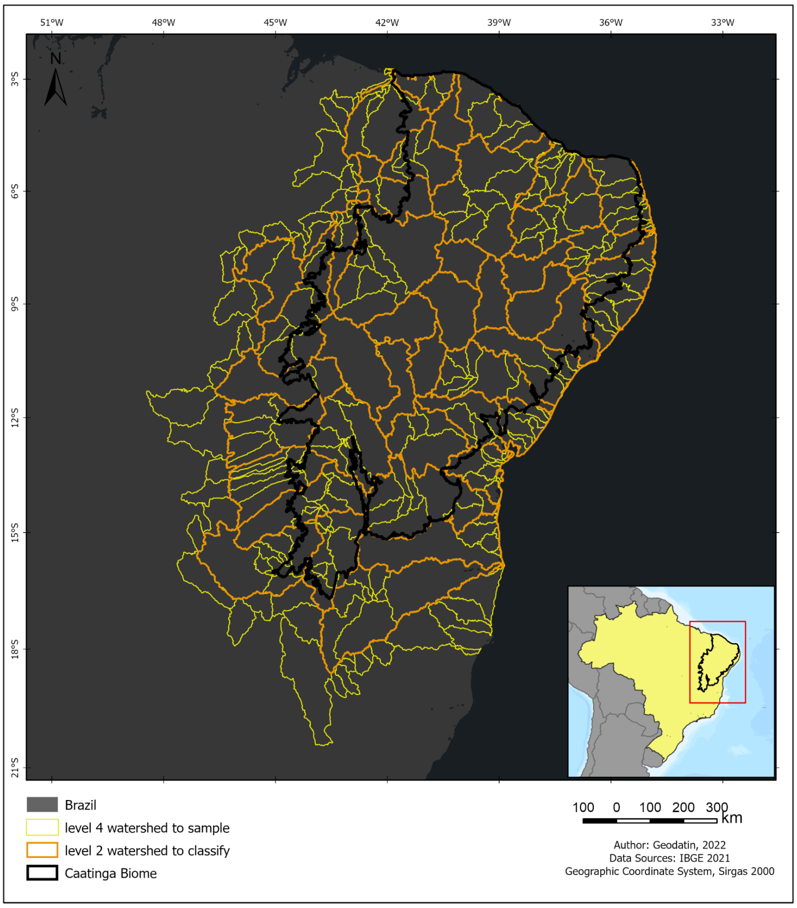

Figure 3. *Hydrographic basins (49 regions) used in the classification and the 30 km sampling grid.*

3. Classification

Describe how the classification was carried out, including the workflow, the class scheme, the feature space, the algorithm and the training samples. The classification approach relies on ensemble tree methods (BREIMAN, 2001; LAWRENCE et al., 2004).

3.1 Methodological workflow

Processing is organised per (basin, year) pair: each of the 49 basins is classified independently, year by year, and the biome map is the composition of the non-overlapping basins. The main steps are: (1) sample collection over a 30 km grid with the full spectral feature set; (2) feature selection and sample cleaning/downsampling; (3) GTB classification per basin/year; (4) a chained post-classification filter sequence; (5) integration and export; and (6) validation of accuracy and area. Figures 4 and 5 illustrate the general and detailed workflows.

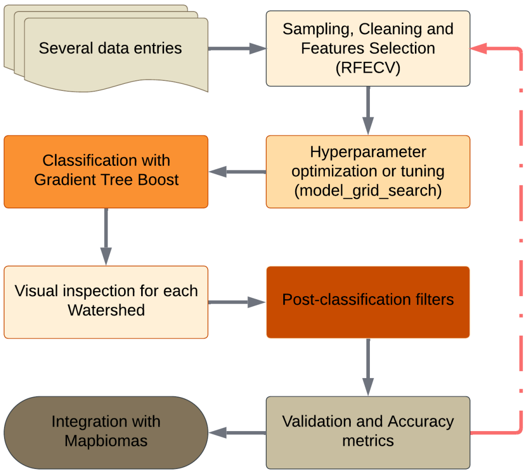

Figure 4. *Simplified general flowchart of Collection 11 for the Caatinga biome.*

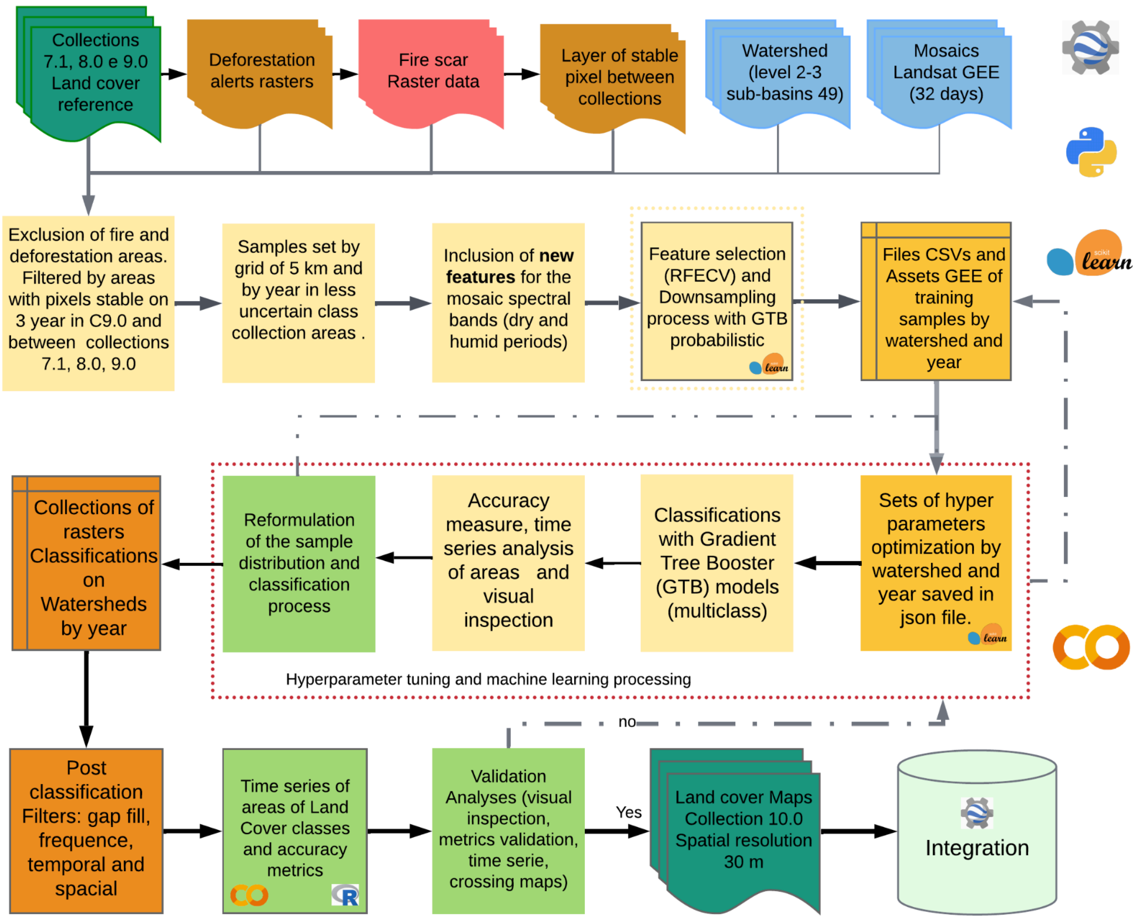

Figure 5. *Detailed classification workflow of MapBiomas Collection 11 (1985–2025) in the Caatinga biome.*

3.2 Classification scheme

The official Collection 11 legend supplied in “Legenda Coleção 11 – Descrição Detalhada.xlsx” defines the 12 Caatinga classes used in the digital product and listed in

Table 2. Non-observed (27) is not a final-product mapped class; it is retained only for validation and data-quality control. Pasture (15), Temporary Crop (19) and Perennial Crop (36) are trained directly by the biome classifier; Mosaic of Uses (21) is reserved for agro-pastoral areas where pasture and agriculture cannot be separated. Rocky Outcrop (29) is produced through a dedicated procedure that remains pending in Section 3.6.

Table 2. Official land-cover and land-use classes used in the digital classification of the Caatinga biome, MapBiomas Collection 11; Non-observed (27) is shown only as an internal validation code.

Legend class	ID	Natural / Anthropic	Colour (hex)	General description
Forest Formation	3	Natural	#1f8d49	Vegetation with a predominantly continuous canopy, including Forested Steppe Savanna and Seasonal Semideciduous or Deciduous Forest.
Savanna Formation	4	Natural	#7dc975	Vegetation with a predominantly semi-continuous canopy, including Wooded Steppe Savanna and Wooded Savanna.
Grassland	12	Natural	#d6bc74	Predominantly herbaceous vegetation, including Park/Grass-Woody Savanna and floodable herbaceous or shrubby areas associated with drainage depressions.
Herbaceous and Shrub Formation	77	Natural	#86b074	Predominantly herbaceous and/or shrubby vegetation, including grasses, dwarf woody plants, shrubs and subshrubs.
Herbaceous Sandbank Vegetation	50	Natural	#ffaa5f	Predominantly herbaceous vegetation on sandy soils and coastal dunes under fluvial-marine influence.
Rocky Outcrop	29	Natural	#ad5100	Naturally exposed rocky surfaces with little or no soil and sparse or absent vegetation.
Mosaic of Uses	21	Anthropic	#ffefc3	Agro-pastoral areas where pasture and agriculture cannot be separated; may include peri-urban smallholdings.
Pasture	15	Anthropic	#edde8e	Areas dominated by planted pasture used for livestock; natural pasture is represented primarily by Grassland or Wetland classes in the national legend.
Temporary Crop	19	Anthropic	#c27ba0	Areas occupied by short-cycle agricultural crops, mapped as Temporary Crop (ID 19) in the official legend.
Perennial Crop	36	Anthropic	#d082de	Areas occupied by perennial agricultural crops, mapped as Perennial Crop (ID 36) in the official legend.
Other Non-vegetated Areas	25	Anthropic	#db4d4f	Impermeable or exposed surfaces, including infrastructure, urban expansion and mining, when not mapped in a separate class.
River, Lake and Ocean	33	Natural / Anthropic	#2532e4	Rivers, lakes, oceans, dams, reservoirs and other water bodies.

Legend class	ID	Natural / Anthropic	Colour (hex)	General description
Non-observed	27			Internal code for non-observed data in validation; not a mapped class of the final Collection 11 product.

3.3 Collect samples process and downsampling method

Sample collection prioritizes areas where pixels are least likely to have misclassified labels. Several aforementioned criteria serve as filters for these collection areas. These areas were subsequently divided into 761 grids, covering the 49 basins, figure 6. For Collection 10.0 and 11.0, 500 pixels were collected per grid for each class. This strategy resulted in an average of approximately 500,000 points per sample set (basin/year).

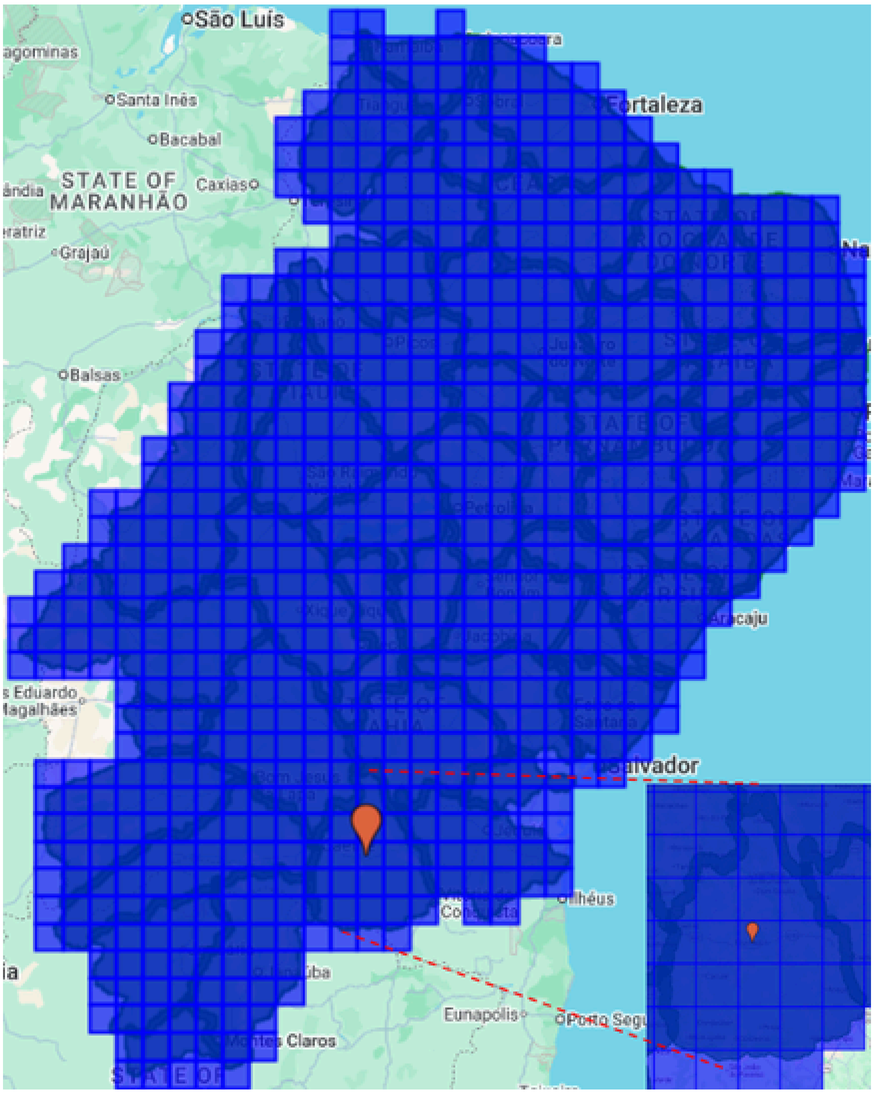

Figure 6. Collection areas by grid and their watershed will be grouped.

As mentioned earlier, the GTB classifier demands more computational power than RF, making it challenging to use very large sample volumes for training. This is precisely why we employ *downsampling*. This methodology allows us to extract a significantly smaller, yet representative, subset of the initial data, while also removing potential outlier pixels from the overall sample.

Enhanced Probabilistic Gradient Tree Boosting Methodology

Inspired by the "Isolation Forest" algorithm (LIU et al., 2008), we developed a new methodology using *probabilistic Gradient Tree Boosting (GTB)*. This approach enhances the selection of training samples, replacing the previous method that relied on the Google Earth Engine (GEE) API.

Sample Selection Process

Our process begins by selecting subsets of samples for each year and for each basin within our dataset. The core idea is to refine the training data by identifying the most confident and representative pixels.

Here's how it works:

1. Initial Separation*:* We first separate samples belonging to the three primary natural vegetation classes:
    - Forest Formation
    - Savanna Formation
    - Grassland Formation
2. Probabilistic Output*:* For every pixel classified by the GTB model, the output isn't just a single class label. Instead, the model provides a *probability vector* indicating the likelihood of that pixel belonging to each class.

    Example: A probability vector like (0.2,0.85,0.12) means:
    - 20% chance of being Forest Formation
    - 85% chance of being Savanna Formation
    - 12% chance of being Grassland Formation
3. *Identifying High-Confidence Candidates:* Even though the pixel above would be ultimately labeled as "Savanna" (since 0.85 is the highest probability), the crucial insight comes from its high confidence score for that class. In the *feature space* (the multi-dimensional representation of the pixel's characteristics), the classifier effectively "sees" this pixel as having an 85% probability of being Savanna. This makes it an ideal candidate for inclusion in the training set for the Savanna class.

Reducing Training Set Size

To significantly optimize and reduce the size of the overall training dataset, we applied a strategic selection process:

- For each of the natural vegetation classes, we specifically chose 100 pixels whose classification probabilities fell within high-confidence intervals: (0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0). This ensures that only the most confidently classified pixels are retained for training.

This same refined approach was also extended to agricultural classes, including:

- Pasture
- Agriculture
- Mosaic of Uses (mixed land use)

Classes with low representation in the initial dataset, such as Water, Other Non-Vegetated Areas and Rocky Outcrop, were not included in this specific sampling strategy. They had too few samples to meaningfully apply this high-confidence selection method.

3.4 Feature space and feature-selection process

Collection 11 maintains a versioned catalogue of 31 raw-band formulas in FORMULAS_INDICES_ESPECTRAIS. The repository documentation explicitly identifies NDVI, NDWI, NDTI, GLI and AWEI as active spectral indices in the VC4 configuration; NDFIa is additionally derived from spectral-mixture-analysis fractions. Table 3 records these six features. The remaining formula-dictionary entries are listed below for reproducibility and must not be interpreted as proof that every entry reached each final basin-year model after feature selection. Band symbols are Blue, Green, Red, NIR, SWIR1 and SWIR2; the implementation source is the Collection 11 repository.

Table 3. Spectral indices explicitly documented as active in VC4, plus the SMA-derived NDFIa feature, for Caatinga Collection 11.

Index	Formula	Note / reference
NDVI	(NIR − Red) / (NIR + Red)	Vegetation vigour Tucker (1979), doi:10.1016/0034-4257(79)90013-0.
NDWI	(NIR − SWIR2) / (NIR + SWIR2)	Vegetation liquid-water content Gao (1996), doi:10.1016/S0034-4257(96)00067-3.
NDTI	(SWIR1 − SWIR2) / (SWIR1 + SWIR2)	Tillage/exposed-soil sensitivity Van Deventer et al. (1997); no DOI recorded in the versioned catalogue.
AWEI	4·(Green − SWIR2) − (0.25·NIR + 2.75·SWIR1)	Automated water extraction Feyisa et al. (2014), doi:10.1016/j.rse.2013.08.029.
GLI	(2·Green − Red − Blue) / (2·Green + Red + Blue)	Green leaf index Louhaichi et al. (2001), doi:10.1080/10106040108542184.
NDFIa	((GV/(1−Shade)) − Soil) / ((GV/(1−Shade)) + NPV + Soil)	SMA-derived forest index Souza et al. (2005), doi:10.1016/j.rse.2005.03.011.

Versioned formula catalogue in the Collection 11 implementation (additional entries):

RATIO = NIR / Red Jordan (1969), doi:10.2307/1936256.

RVI = Red / NIR Pearson and Miller (1972); no DOI recorded.

NDBI = (SWIR1 − NIR) / (SWIR1 + NIR) Zha et al. (2003), doi:10.1080/01431160304987.

NDMI = (NIR − SWIR1) / (NIR + SWIR1) Wilson and Sader (2002), doi:10.1016/S0034-4257(01)00318-2.

IIA = (Green − 4·NIR) / (Green + 4·NIR) implementation catalogue; no bibliographic reference recorded.

EVI = 2.4·(NIR − Red) / (1 + NIR + Red) two-band implementation in the code; Huete et al. (2002), doi:10.1016/S0034-4257(02)00096-2.

GCVI = (NIR / Green) − 1 Gitelson et al. (2003), doi:10.1078/0176-1617-00887.

GEMI = [2·(NIR² − Red²) + 1.5·NIR + 0.5·Red] / (NIR + Green + 0.5) implementation in the versioned code; Pinty and Verstraete (1992), doi:10.1007/BF00031911.

CVI = NIR·(Green / Blue²) Vincini et al. (2008), doi:10.1007/s11119-008-9075-z.

SHAPE = (2·Red − Green − Blue) / (Green − Blue) implementation catalogue; no bibliographic reference recorded.

AFVI = (NIR − 0.5·SWIR2) / (NIR + 0.5·SWIR2) Karnieli et al. (2001), doi:10.1016/S0034-4257(01)00190-0.

AVI = [NIR·(1 − Red)·(NIR − Red)]^(1/3) Ashburn (1979); no DOI recorded.

BSI = [(SWIR1 − Red) − (NIR + Blue)] / [(SWIR1 + Red) + (NIR + Blue)] Rikimaru et al. (2002); no DOI recorded.

BRBA = Red / SWIR1 He et al. (2010), doi:10.1080/01431161003638343.

DSWI5 = (NIR + Green) / (SWIR1 + Red) Apan et al. (2004), doi:10.1080/0143116031000161803 1.

LSWI = (NIR − SWIR1) / (NIR + SWIR1) Xiao et al. (2004), doi:10.1016/j.rse.2003.11.008.

MBI = [(SWIR1 − SWIR2 − NIR) / (SWIR1 + SWIR2 + NIR)] + 0.5 Rasul et al. (2018), doi:10.3390/rs10121928.

UI = (SWIR2 − NIR) / (SWIR2 + NIR)   Kawamura et al. (1996); no DOI recorded.

OSAVI = (NIR − Red) / (0.16 + NIR + Red)   Rondeaux et al. (1996), doi:10.1016/0034-4257(95)00186-7.

GNDVI = (NIR − Green) / (NIR + Green)   Gitelson and Merzlyak (1996), doi:10.1016/S0176-1617(96)80284-7.

Brightness = 0.3037·Blue + 0.2793·Green + 0.4743·Red + 0.5585·NIR + 0.5082·SWIR1 + 0.1863·SWIR2   Crist (1985), doi:10.1016/0034-4257(85)90102-6.

Wetness = 0.1509·Blue + 0.1973·Green + 0.3279·Red + 0.3406·NIR + 0.7112·SWIR1 + 0.4572·SWIR2   implementation coefficients in the versioned code; Crist (1985), doi:10.1016/0034-4257(85)90102-6.

MSI = NIR / SWIR1   implementation in the versioned code; Rock et al. (1986), doi:10.2307/1310339.

GVMI = [(NIR + 0.1) − (SWIR1 + 0.02)] / [(NIR + 0.1) + (SWIR1 + 0.02)]   Ceccato et al. (2002), doi:10.1016/S0034-4257(02)00036-6.

PRI = (Green − Blue) / (Green + Blue)   Peñuelas et al. (1994), doi:10.1016/0034-4257(94)90136-8.

NBR = (NIR − SWIR1) / (NIR + SWIR1)   Key and Benson (1999); no DOI recorded.

Implementation sources: https://github.com/mapbiomas/brazil-caatinga/blob/main/lulc_30m_landsat/collection_11/src/classification_process/arqParametros_class.py and https://github.com/mapbiomas/brazil-caatinga/blob/main/lulc_30m_landsat/collection_11/src/samples_process/README_indices_espectrais.md.

Table 4. Spectral-mixture-analysis endmembers (reflectance) used to derive the SMA fractions and NDFIa.

Endmember	Blue	Green	Red	NIR	SWIR1	SWIR2
Green vegetation (GV)	0.05	0.09	0.04	0.61	0.30	0.10
Non-photosynthetic veg. (NPV)	0.14	0.17	0.22	0.30	0.55	0.30
Soil	0.20	0.30	0.34	0.58	0.60	0.58
Cloud	0.90	0.96	0.80	0.78	0.72	0.65
Shade	0	0	0	0	0	0

The image below, figure 7, depicts an instance of the samples corresponding to sub-basin “744” which have an unbalanced distribution due to the nature of the data.

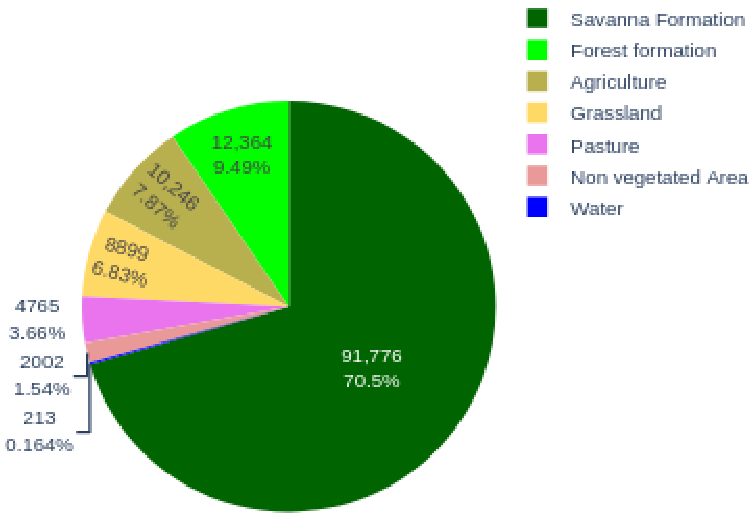

Figure 7: Distribution of samples for sub-basin 744 in the year 2000.

Achieving separability in the feature space is a prevalent challenge when performing remote sensing image classification in the Caatinga Biome. Figure 8 demonstrates that separability within a spectral band is limited for various targets in the image. Another way of visualizing this can be seen in the figure 9, which plots the "blue_median", "green_median", "red_median", "nir_median" bands of the mosaic for six coverage classes.

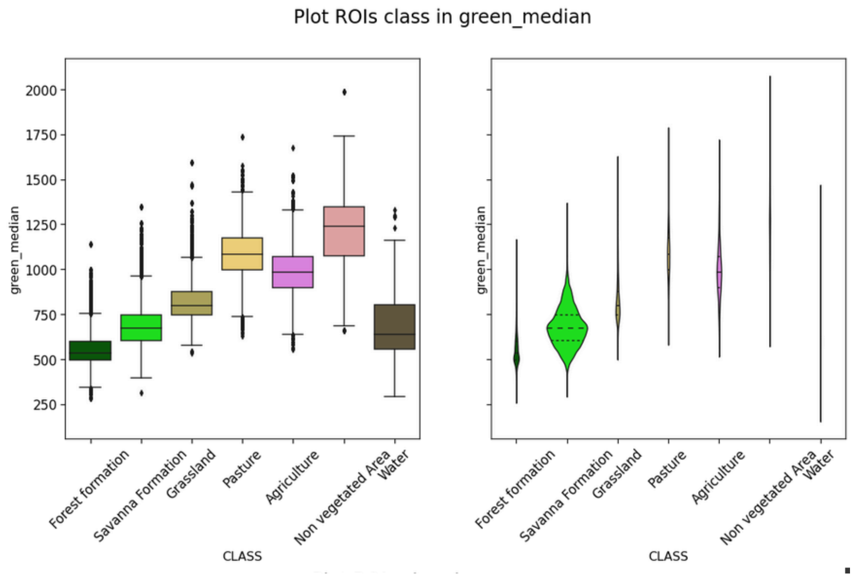

Figure 8: Box and violin plots from samples of spectral band “GREEN” in the main land cover classes mapped by the Caatinga team.

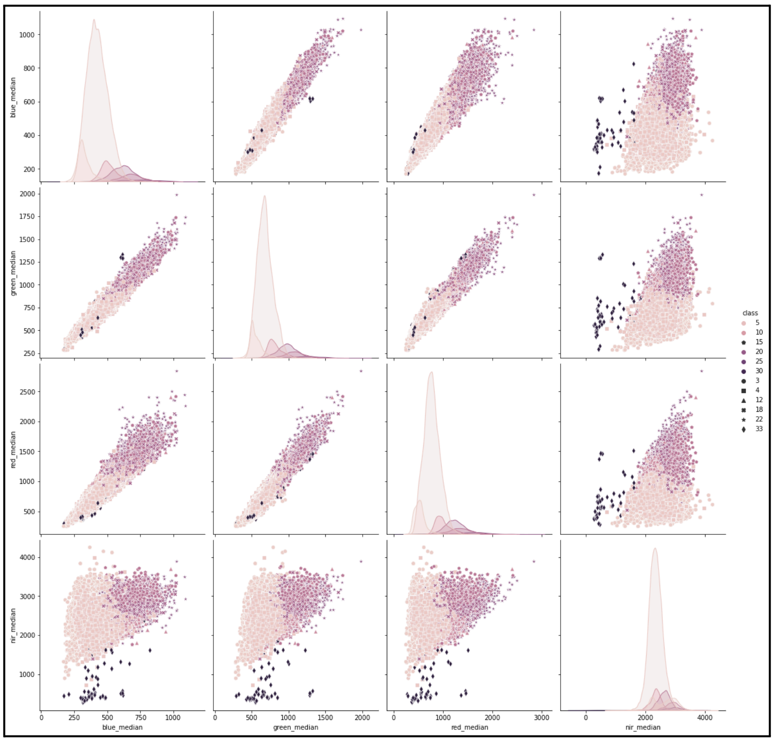

Figure 9: Spatial distribution of samples for the variables, “blue_median”, “green_median”, “red_median”, “nir_median”.

All watersheds were analyzed individually in terms of feature importance. These variables included the original Landsat reflectance bands, as well as vegetation indexes and spectral mixture modeling-derived variables. The first step was measuring the correlation between feature Collection variables (figure 10), and some variables would be eliminated from the least important criteria following the score.

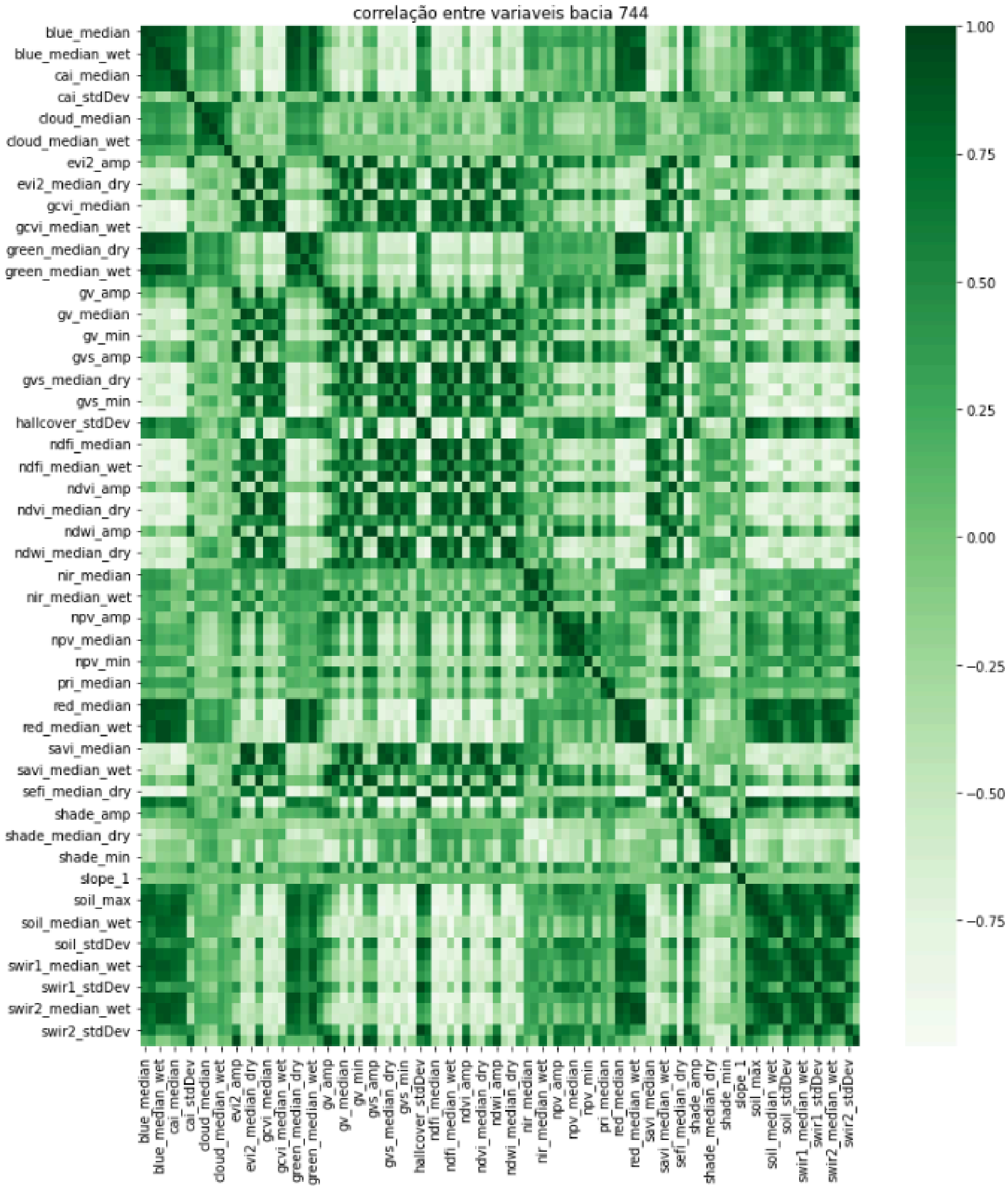

Figure 10. *Example correlation matrix of candidate features for a basin sample set (year 2020).*

Feature selection. Since Collection 8.0 the project uses Recursive Feature Elimination with Cross-Validation (RFECV; ZHANG; MA, 2009; RAMEZAN, 2022). In Collection 11 the RFECV is run locally with scikit-learn using a Random Forest estimator (*n_estimators = 800*), *StratifiedKFold(3)*, *step = 0.05* and *min_features_to_select = 15*, producing one selected-feature list per basin/year. A basic example may be found at the link below:

https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html

The *RFECV()* function can be accessed by using the python Sklearn library. There are two methods in which the class can be used to filter the selected variables: the "support_()" method and the "ranking_" method. With the former we can choose the surviving variables from a list of "TRUE" or "FALSE", and with the latter we can extract the ranking of the "TRUE" variables.

If the number of variables in "TRUE" is less than 10, then banking consecutive to 1 is taken as a condition (e.g. 2,3,4,5 etc.), figure 11.

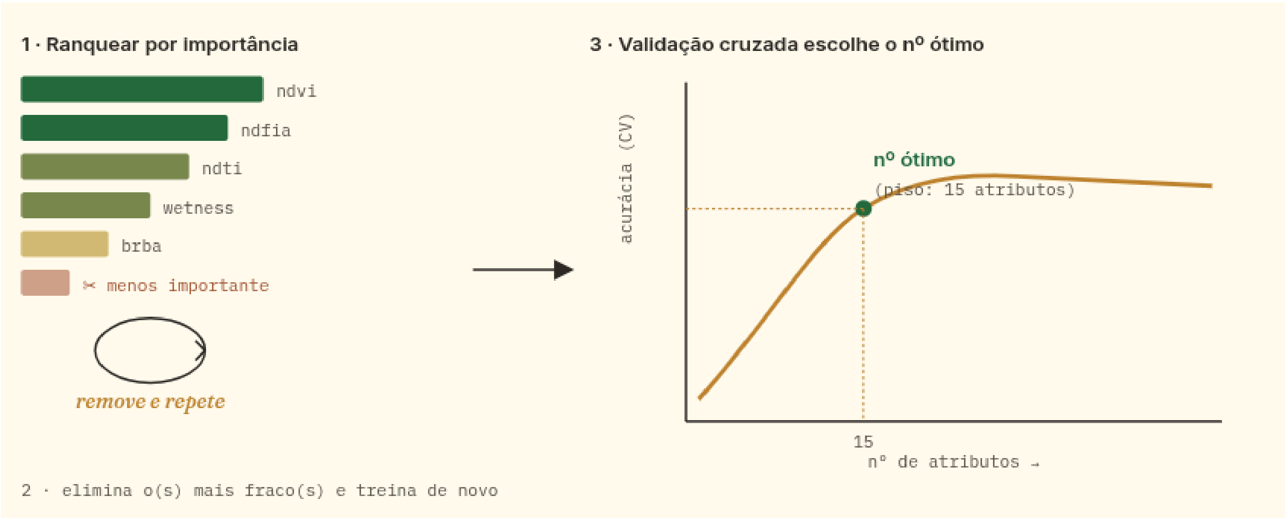

Figure 11: The curve rises while useful attributes are retained and stabilizes when only redundant ones remain. The peak indicates how many attributes are worth keeping.

HYPERPARAMETER TUNING PROCESS

A script was implemented for the **Hyperparameter Tuning** process after selecting the variable sets by drainage basin and year. The GridSearchCV() function, along with the Pipeline() function, is capable of testing various parameter combinations for the model. It is then possible to establish which combination of parameters represents the best score or accuracy. The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid. An example from data of collection 10.0, the "learning rate" parameters and "n estimators" is shown in figure 12, where the optimal pair of parameters would be (40, 0.175).

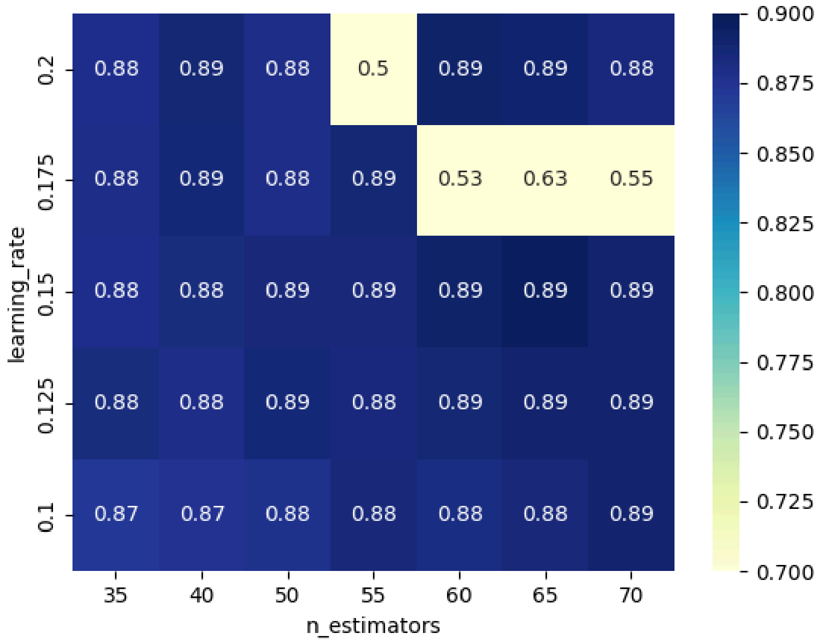

Figure 12. Example of the plot of combination of "learning rate" parameters and "n estimators".

The GridSearchCV() (Grid Search Cross-Validation) is an exhaustive, or "brute-force," hyperparameter optimization technique, which means it can be computationally expensive. It operates as follows:

1. Defining a Hyperparameter "Grid": You define a dictionary where keys are the hyperparameter names for your model, and values are lists of all possible values you want to test for each hyperparameter. This creates a "grid" of every possible combination.

2. Exhaustive Training and Cross-Validation: For each hyperparameter combination within your defined grid, `GridSearchCV()` performs the following steps:
   - It trains the model using cross-validation. This involves splitting the training dataset into k "folds" (subsets). The model is then trained k times; each time, it uses k-1 folds for training and one fold for validation.
   - The model's performance is evaluated on each validation fold using a pre-defined scoring metric (e.g., accuracy, F1-score, Mean Squared Error (MSE), etc.).
   - The final score for that specific hyperparameter combination is the average of the scores obtained across all k folds.

3. Selecting the Best Combination: After evaluating all possible combinations in the grid via cross-validation, `GridSearchCV()` identifies the combination of hyperparameters that yielded the best average validation score.
4. Final Model Refitting: Once the optimal hyperparameter combination is found, `GridSearchCV()` (by default, if `refit=True`) retrains the model using the entire original training dataset with these winning hyperparameters. This ensures you have a robust final model trained on all available data.

3.5 Classification algorithm

The classifier is Gradient Tree Boosting via *ee.Classifier.smileGradientTreeBoost*, trained per basin/year on the class label and the selected bands. The production configuration is: *numberOfTrees = 30*, *shrinkage = 0.1*, *samplingRate = 0.65*, *loss = LeastSquares*, *seed = 0*. Per-basin hyperparameters (learning rate and number of estimators) are loaded from a tuning dictionary. Two overrides apply for robustness: basins with a history of memory overflow are forced to a reduced number of trees (≈18), and specific difficult basin/years use a 'simple' setting (6 trees, shrinkage 0.05, first 15 features). Some basins with too few samples reuse training samples donated from a neighbouring basin.

3.5.1 Random Forest: historical and supporting role

Random Forest was the production classifier through Collection 7.1 and was evaluated in parallel with GTB in Collection 8.0. It is not used to generate the final Collection 11 maps. In Collection 11, Random Forest is retained only as the estimator within RFECV because it provides stable variable-importance rankings for the high-dimensional feature space; its role is methodological support rather than production classification.

3.5.2 Gradient Tree Boosting: Collection 11 production classifier

GTB is the production classifier used for all basin-year mosaics in Collection 11. It was selected because the tuned sequential ensemble improves discrimination under strong seasonality and class imbalance. The classifier configuration, basin-specific overrides and training-sample rules are reported in Section 3.5. Its principal operational limitation is sequential training, which increases computation time and the risk of Google

Earth Engine memory or timeout errors; reduced-tree and simplified-model overrides are safeguards for affected basin-years.

4. Post-classification

The temporal filter rules were specifically adapted to account for the phenological and spectral dynamics of the land cover classes characteristic of the Caatinga biome. In addition to the standard temporal consistency checks, custom rules were incorporated to handle anomalous transitions particularly cases in which a class appears abruptly in a pixel's time series. These adjustments aim to improve classification stability and reduce spurious temporal fluctuations in areas with high seasonal variability and low vegetation cover.

Purpose. Explain how the raw per-basin/year classification is refined into the final product. The chain runs, in order, Gap-fill → Sieve → Temporal → Frequency → Spatial All, each operating per basin over the 41 annual bands at 30 m resolution.

4.1 Filter sequence

Gap-fill. Fills unobserved pixels using the public Collection 10 as reference (remapped to seven classes). Class-specific rules constrain the fill (e.g., Water 33 is kept only where Collection 10 is 33 or 3; Agriculture 19 only where Collection 10 ∈ {19, 21, 15}; Perennial Crop 36 only where ∈ {36, 21, 15}). Remaining gaps are filled backward (2025→1986) using the first non-null value of subsequent years, with a Collection-10 fallback at the edges and a zero→Water correction, figure 16.

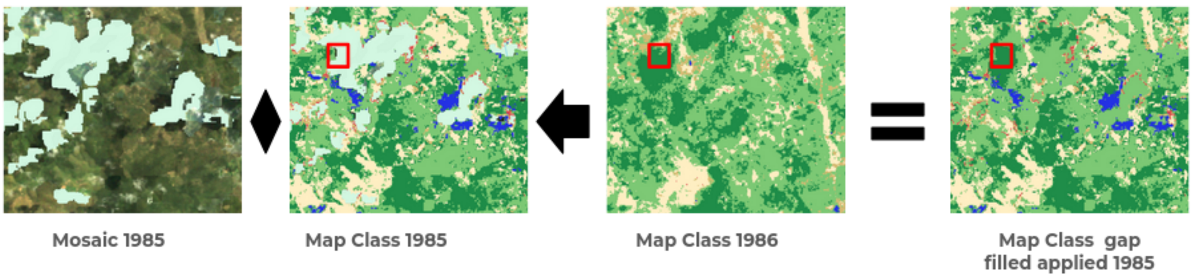

Figure 16: Example of the process gap fill.

Sieve (structure-preserving spatial filter): Removes small patches while explicitly protecting thin features (rivers, riparian corridors). Related groups are merged ([15,21] and [4,12]); connected-pixel counts (8- and 4-connectivity) and a morphological 'bridge' detection plus erosion identify compact interiors versus thin structures. Thin, non-compact structures are preserved; non-bridge patches of 25 pixels or fewer are replaced by the local mode (9-pixel kernel). Classes [33, 29, 25] are never substituted.

Temporal (unified): Corrects short-lived, implausible transitions using a single pass. A binary natural/anthropic flag is used only to detect the pattern; the correction copies the real class of the anchor year. Patterns handled are J3 (T-!T-T, one anomalous year), J4 (T-!T-!T-T, two years) and J5 (T-!T-!T-!T-T, three years), applied first to natural then to anthropic classes.

Frequency: Applies a light internal sieve (≤5-pixel patches, kernel 3; exceptions [33, 29]) and then stabilises permanently natural pixels across the full 41-year series: Forest (3) where frequency > 70%, Savanna (4) ≥ 80%, Grassland (12) > 70%, and Rocky Outcrop (29) forced where frequency ≥ 75%. This reduces spurious inter-class fluctuation, notably at the series extremes.

Spatial All (final product). A neighbourhood-mode filter (9×9 window) is applied to weakly connected pixels (connectedPixelCount < 12), producing the final post-classified series.

4.2 Integration

The post-classified biome result is integrated with the MapBiomas cross-cutting themes over the full series (1985–2025). In the pre-integration step, per year: unclassified pixels (0) are set to Mosaic of Uses (21); the rocky-outcrop layer is blended; and ; and a biome mask with a 5 km buffer is applied before export. The biome-specific prevalence rules and the final integrated legend follow the MapBiomas integration protocol.

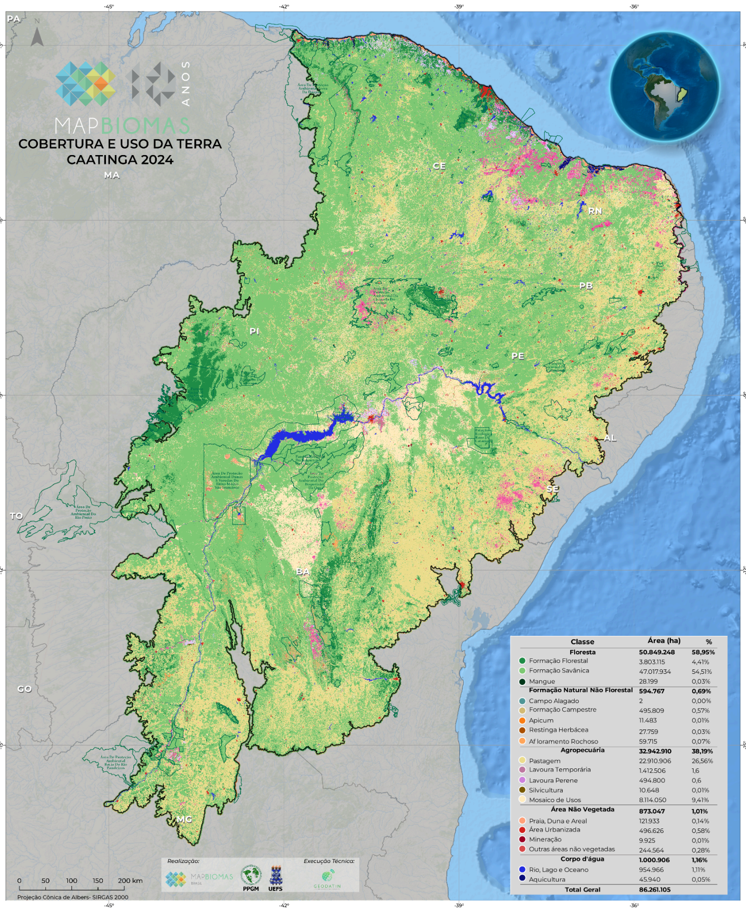

Figure 17. Final land use and land cover map of the Caatinga biome (2025).

5. Validation Strategies

Purpose. Allow the reader to assess the quality and reliability of the product through accuracy assessment and comparison with reference maps.

5.1 Accuracy assessment

The independent MapBiomas/LAPIG reference panel derives from the probability-sampling design documented in the supplied Accuracy Assessment Appendix for Collection 10. The national panel contains 85,152 spatially independent locations repeated annually; 9,738 locations are allocated to the Caatinga. The national annual accuracy calculation uses 75,152 samples because 10,000 Amazon locations were assigned to training; this does not reduce the documented Caatinga allocation. The Caatinga Collection 11 workflow samples the reference layer at 30 m, excludes Non-observed (27) from the mapped-class metrics and reports overall and balanced accuracy, precision, recall, F1, Jaccard, user's and producer's accuracy, commission and omission, confusion matrices and the Pontius quantity/allocation/exchange/shift decomposition.

The national panel is a proportionally stratified random sample (Cochran, 1977). Strata combine 127 grouped topographic-chart units each formed from four IBGE charts with six EMBRAPA slope classes. Each grouped-chart unit receives at least 500 points, with additions according to variability and number of classes; the design targets 95% confidence, a maximum 5% margin of error for grouped units and 0.5% expected national error, with Bonferroni adjustment. A supplementary disproportionate sample covers rare classes: Beach/Dune, Mining, Mangrove, Aquaculture, Hypersaline Tidal Flat and Rocky Outcrop. For Rocky Outcrop, the appendix records 450 points in accumulated class areas and 150 in a surrounding buffer.

Each location is interpreted independently by three analysts in the LAPIG/UFG TVI platform. A senior interpreter adjudicates disagreements. Evidence includes Landsat dry- and rainy-season imagery in NIR/SWIR1/Red false colour, MODIS NDVI time series and Google Earth high-resolution imagery when available. Estimation weights account for selection probability, missing observations and degree of interpreter agreement; accuracy statistics are derived from the weighted population error matrix. This description follows the supplied Collection 10 appendix and is the documented methodological basis for the Collection 11 panel; any Collection 11 departures must be versioned explicitly.

Collection 11 accuracy results were obtained from the MapBiomas Accuracy Dashboard for the Caatinga biome, Collection 11, using Levels 1 and 2. The dashboard reports the consolidated 1985–2024 series; 2024 is the latest year available in the accuracy panel, whereas the mapped Collection 11 product extends through 2025.

At Level 1, consolidated overall accuracy was 88.56%, with 9.99% allocation disagreement and 1.45% quantity disagreement; annual overall accuracy in 2024 was 88.02%. For 2024, user's/producer's accuracies were: Forest (1), 92.53%/89.58%; Shrub and Herbaceous Vegetation (10), 9.23%/11.69%; Farming (14), 80.93%/85.53%; Non-vegetated Area (22), 72.51%/93.87%; and Water Body (26), 90.88%/92.93%.

At Level 2, consolidated overall accuracy was 80.38%, with 13.20% allocation disagreement and 6.42% quantity disagreement; annual overall accuracy in 2024 was 81.92%. For 2024, user's/producer's accuracies were: Mosaic of Uses (21), 49.03%/100.00%; Grassland Formation (12), 9.23%/11.69%; Pasture (15), 82.30%/80.83%; Agriculture (18), 29.51%/52.25%; Urbanized Area (24), 72.51%/93.87%; Forest Formation (3), 60.36%/37.19%; River, Lake and Ocean (33), 90.88%/92.93%; and Savanna Formation (4), 86.81%/86.84%.

The lower Level 2 result reflects the greater difficulty of separating finer thematic classes. The most pronounced limitations in 2024 were observed for Grassland Formation, Agriculture and Forest Formation. The requested dashboard filters did not provide a Level 3 result; therefore, no Level 3 value was imputed. Source: MapBiomas Accuracy Dashboard.

Table 5. Evolution of the global accuracy of the Caatinga collections (Levels 1, 2 and 3).

Collection	Method	Global accuracy
3.1	Random Forest	L1 80.0% · L2 78.2% · L3 71.3%
4.1	Random Forest	L1 81.9% · L2 79.9% · L3 74.3%
5.0	Random Forest	L1 81.8% · L2 80.0% · L3 75.4%
6.0	Random Forest	L1 82.8% · L2 76.6% · L3 74.9%
7.1	Random Forest	L1 83.7% · L2 78.8% · L3 76.9%
8.0	Random Forest / GTB	L1 83.6% · L2 78.2% · L3 76.9%
9.0	GTB	L1 84.6% · L2 79.4% · L3 79.3%
10.0	GTB	L1 84.6% · L2 79.4% · L3 79.3%
11.0	GTB	**L1 88.56% · L2 80.38% · L3 (not reported by panel).**

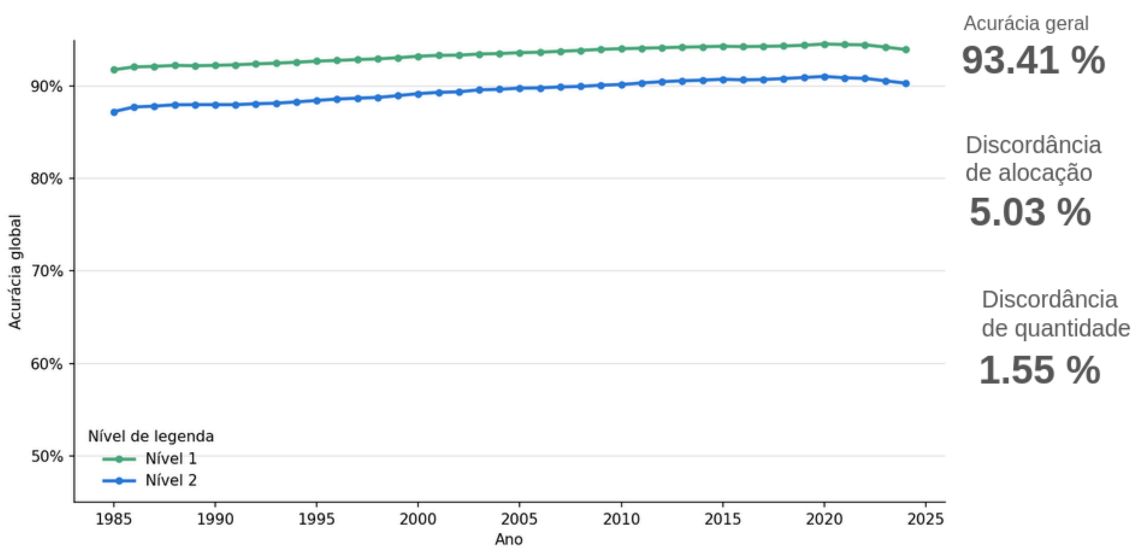

Figure 18. *Accuracy of the Caatinga LULC series across collections (Level 2, 1985–2025).*

Figures 19 and 20 are retained only as historical Collection 10 visualizations. They are not used to derive the Collection 11 indicators reported above. Updated Collection 11 class-error graphics should be generated from the same filtered panel values before publication.

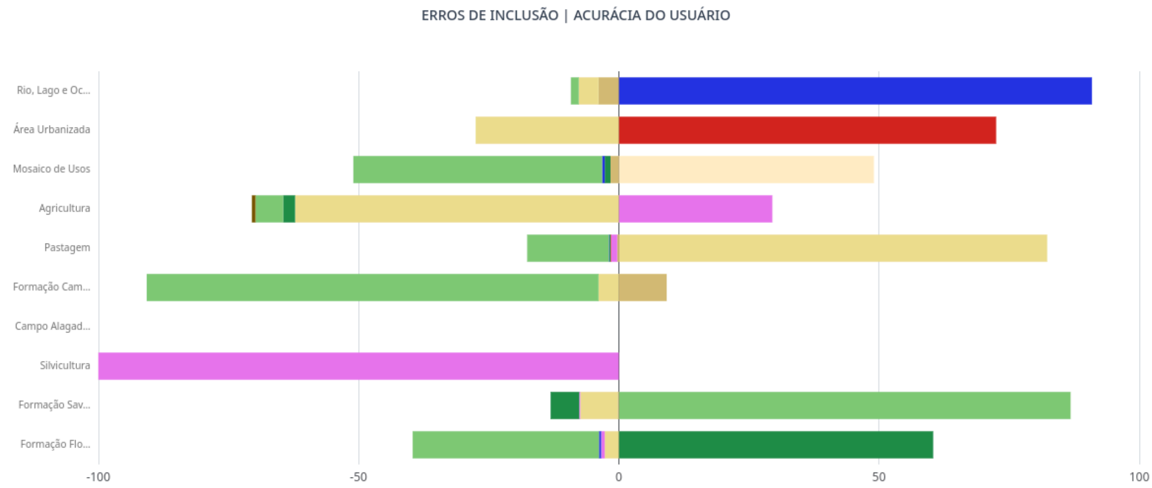

Figure 19. *Commission errors of the Caatinga LULC mapping, Collection 11 year 2024.*

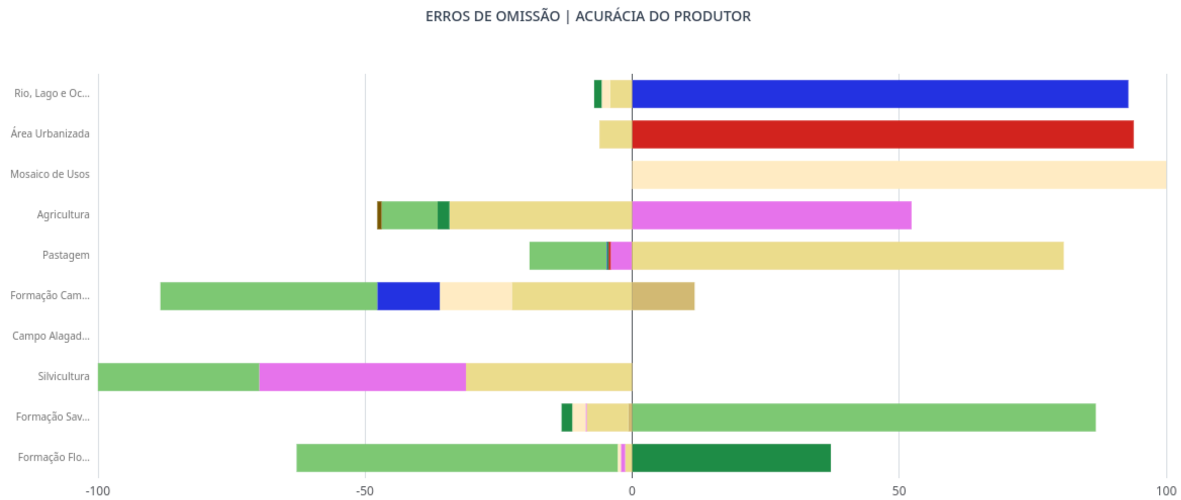

Figure 20. *Omission errors of the Caatinga LULC mapping, Collection 11 year 2024.*

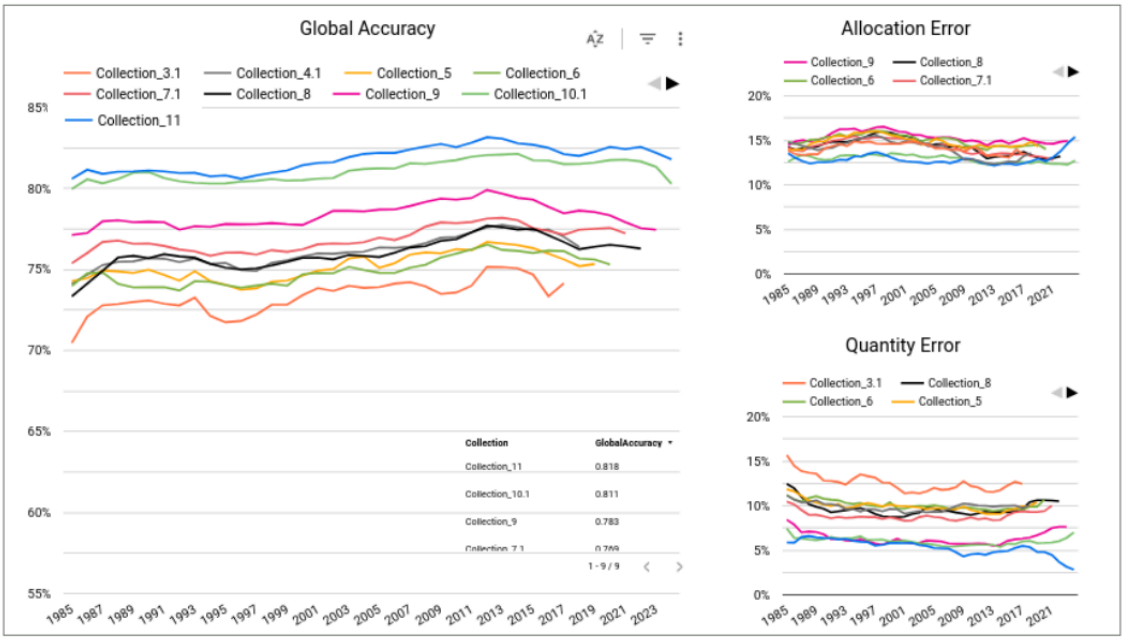

Figure 21. Plot of Accuracy of level 3 of MapBiomas Collections 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 11.0 in the Caatinga biome (1985-2024 years).

If we plot all values in the accuracy series, we can better compare and observe all of the data from the different collections (Figure 21), the same analysis was done in the collection 11.0 . Another technique to assess the quality of a map series is to examine the area's behavior in relation to each class of land cover across time. Figure 22 depicts time series plots of area by cover class. In addition, the areas of the two collections prior to the 10.0 collection are plotted. This type of analysis makes it possible to compare the areas of the classes throughout the series with other previously published collections. At this point, it is possible to understand where and why there have been changes from one collection to another.

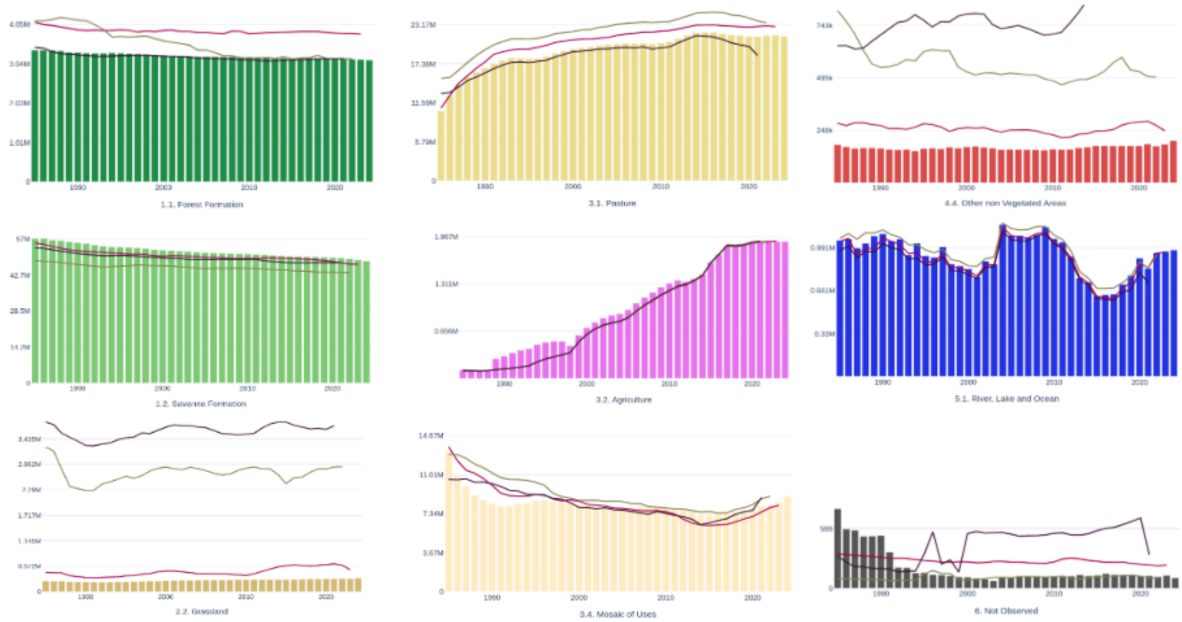

Figure 22. Time series of level 3 classes of MapBiomas Collection 10.0 in the Caatinga biome (1985-2024 years) per area (ha). The green, brown and red lines represent the corresponding areas of collections 7.1, 8.0 and 9.0.

Another way of validating the coverage data was to compare the areas of the classes between collections. For this analysis we used a rule implemented by the Pampa team that accounts for coincidences (Figure 23).

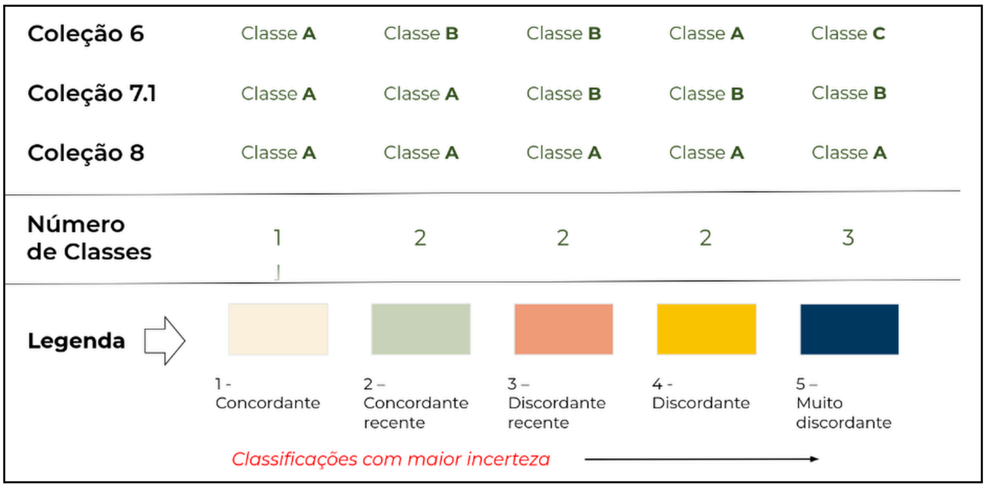

Figure 23: Concordance table between the last 3 collections.

With this analysis we can infer how much area is being mapped with the same class, how much area is varying between classes from one collection to another and when these disagreements occurred last year. Regions in the north and center of the Caatinga and the Chapada Diamantina show the greatest disagreements (Figure 24). These places of greatest discordance indicate where new samples should be collected and used to find the correct class. They also indicate areas where the pixels in the mosaic have clouds, noise or shadows.

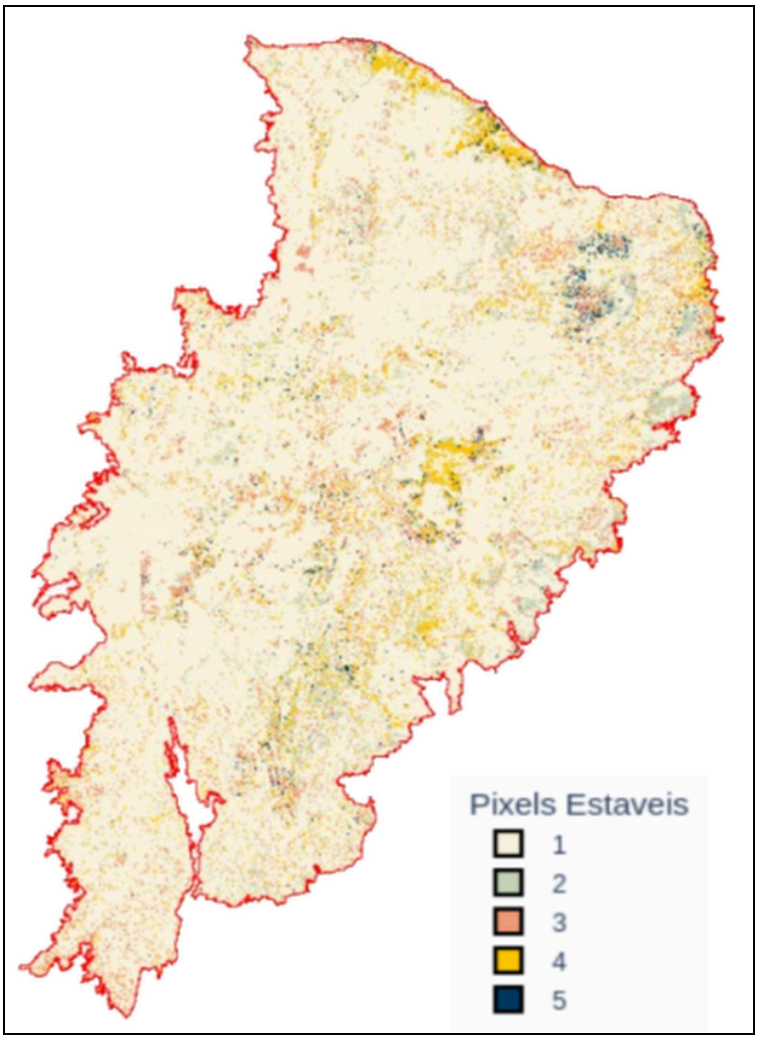

Figure 24: Pixel coincidence model between collections 6.0, 7.1 and 8.0.

The statistics over the series show that on average 76 % of the pixels are coincident between the 3 collections (Figure 25). This percentage is higher than the accuracy of the last three collections, which indicates that this measure does not indicate the quality of the maps, but rather areas with high stability between collections.

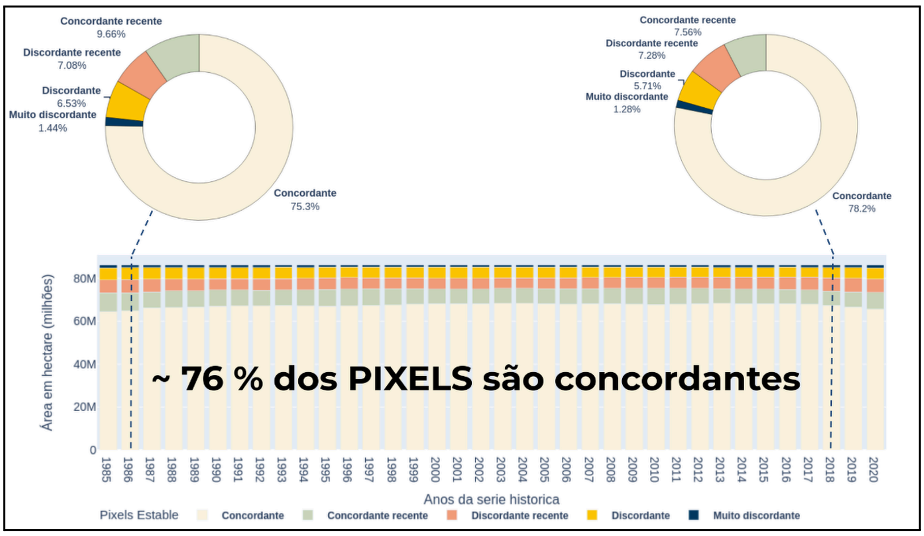

Figure 25: Statistics of the areas of agreement for collections 6.0, 7.1 and 8.0.

Based on this analysis, the question arises as to which classes are affected by large areas of disagreement. Thus, if we analyze the last two collections by cover, then the models would indicate where the pixels are that were classified as savannah in collection 7.1, for example, and are not in collection 8.0, as well as those that are now in collection 8.0 and were not in collection 7.1, figures 26a and 26b.

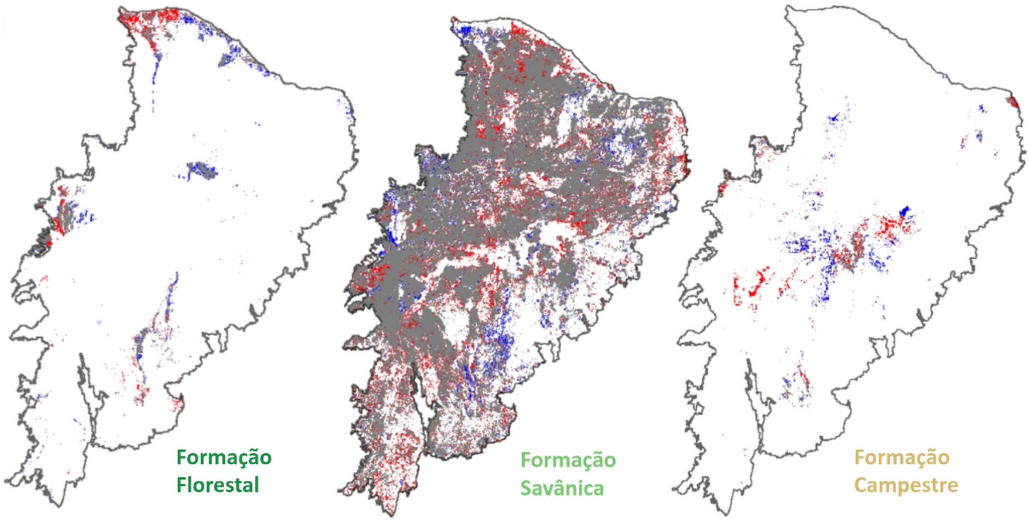

Figure 26a: Cover models for Forest Formation, Savanna Formation and Grassland Formation, for collections 7.1 and 8.0. In gray areas of coincidence, in blue areas only in the 7.1 collection and in red areas that were only mapped for the 8.0 collection.

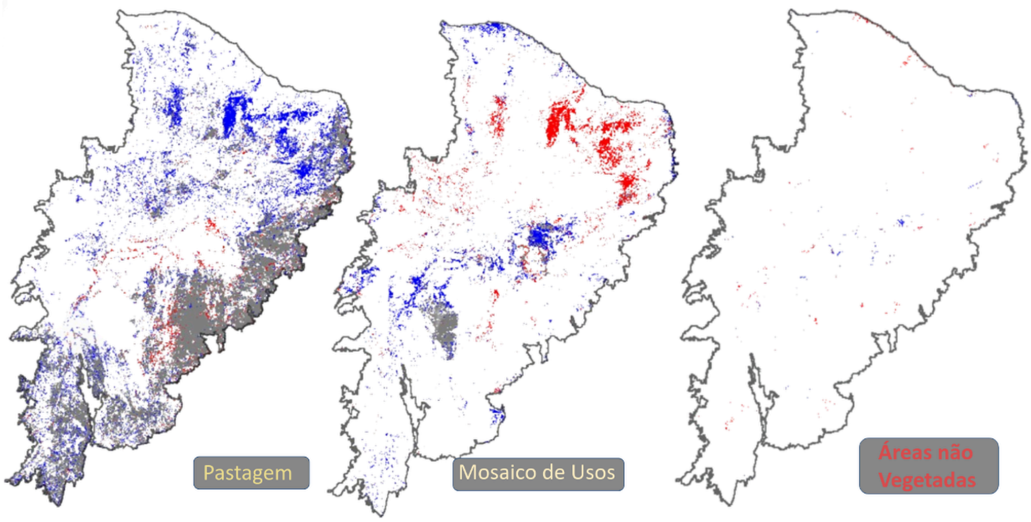

Figure 26b: Cover models for Pasture, Mosaic of Uses and Non vegetated area, for collections 7.1 and 8.0. In gray areas of coincidence, in blue areas only in the 7.1 collection and in red areas that were only mapped for the 8.0 collection.

Land covers with a greater presence in the caatinga, such as savannah Formation, have more than 80% agreement between the last two collections, see figure 27.

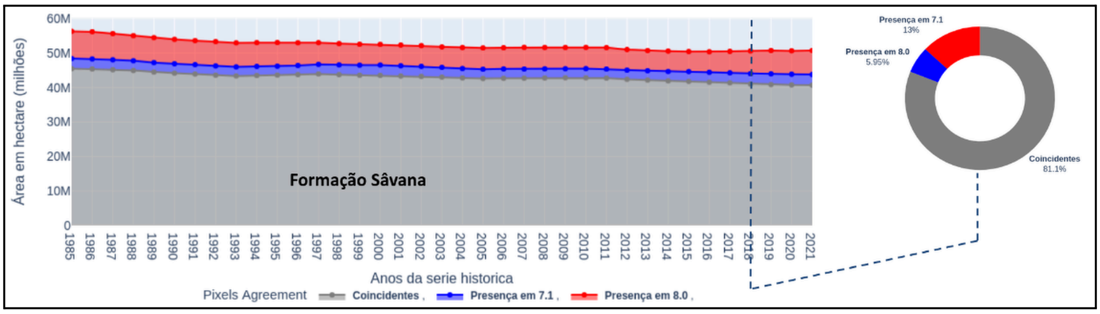

Figure 27: Areas of concordant pixels between collections 7.1 and 8.0.

5.2 Validation with reference maps

The reference product is IBGE's Monitoramento da Cobertura e Uso da Terra do Brasil, edition 2018/2020 (published in 2022), from the 2000–2020 monitoring series. The product uses a regular 1 km grid (one 1 km² cell) and covers the entire Brazilian territory; consequently, it covers the complete terrestrial extent of the Caatinga biome. Access date: 6 August 2026. IBGE reports no internal spatial gaps in the national land grid. Cells outside Brazil and ocean areas outside the land product are out of scope, and the 1 km grid necessarily generalises biome boundaries and features smaller than a cell; these are scale and mask limitations, not missing Caatinga subareas.

Inter-collection coincidence analysis (agreement between consecutive collections) is also used to locate the most unstable areas, which guide new sample collection.

Before comparison, IBGE classes are harmonised to the MapBiomas legend (Figure 28) and coincidence is calculated for matching reference years. The analysis is a thematic and scale comparison not an independent 30 m pixel-accuracy estimate because the IBGE 1 km grid aggregates heterogeneous land-cover patterns.

Codigo_Classe	Class_mapbiomas	Nome
1	33	Área artificial
2	21	Área Agrícola
3	21	Pastagem com Manejo
4	21	Mosaico de Ocupações em Área Florestal
5	21	Silvicultura
6	3	Vegetação Florestal
9	10	Área Úmida
10	4	Vegetação Campestre
11	21	Mosaico de Ocupações em Área Campestre
12	33	Corpo d'água Continental
13	33	Corpo d'água Costeiro
14	22	Área Descoberta

Figure 28: Harmonized legend between IBGE and MapBiomas classes.

The greatest coincidences are found in areas to the west and north of the Caatinga where there are large expanses of savannah and in the south-east of the Caatinga where there is a high presence of pasture (Figure 29).

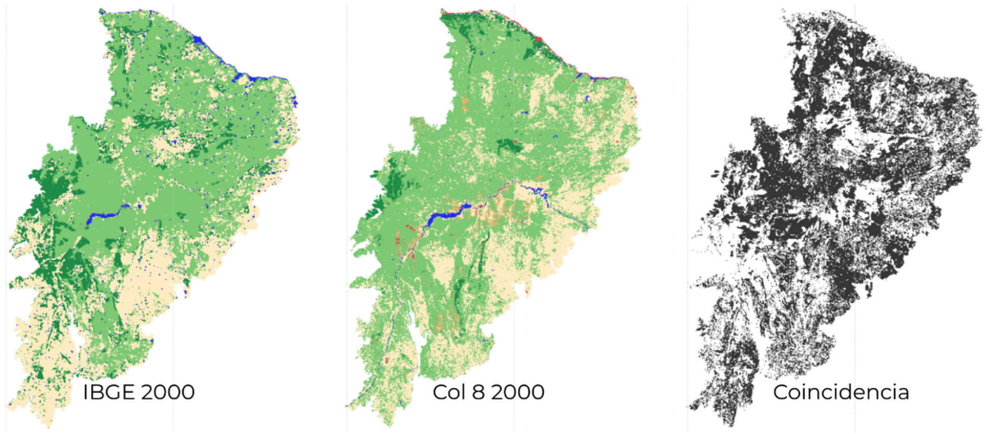

Figure 29: Coincidence Map of Caatinga using the IBGE LULC and MapBiomas LULC for the year 2000.

The comparison uses the seven IBGE reference years 2000, 2010, 2012, 2014, 2016, 2018 and 2020. Differences of approximately 50% relative to MapBiomas in each assessed year must be interpreted as thematic, legend and resolution disagreement rather than incomplete Caatinga coverage, because the IBGE layers are national and spatially complete over the biome.

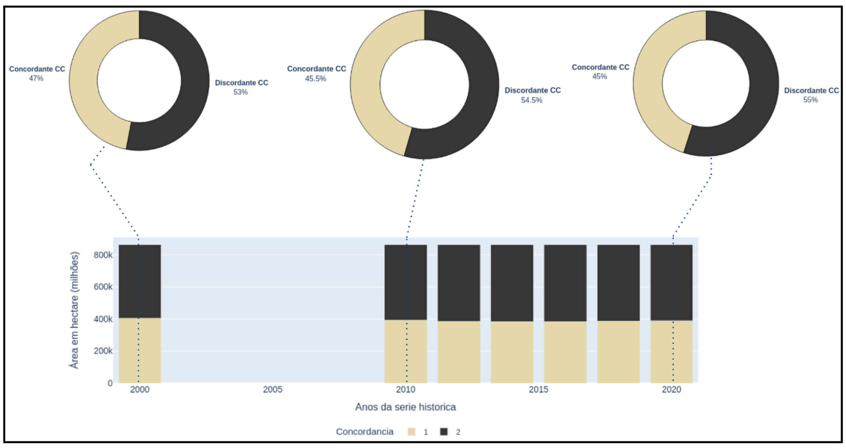

Figure 30. Areas of concordance between the national, spatially complete IBGE 1 km monitoring grids and MapBiomas 30 m maps for 2000, 2010, 2012, 2014, 2016, 2018 and 2020.

6. Practical Considerations and Limitations

Collection 11 is processed independently for 49 hydrographic basins and 41 annual mosaics. This regionalized design controls spectral heterogeneity but creates a large number of basin-year models and increases operational complexity. GTB training is sequential and can trigger Google Earth Engine memory or timeout constraints; the documented safeguards include reduced tree counts for recurrent overflow basins, simplified settings for specific difficult basin-years and, where samples are insufficient, the controlled reuse of samples from a neighbouring basin. These exceptions must remain versioned and auditable because they can affect local comparability. Users should interpret fine-scale or class-specific change together with accuracy estimates, reference-map agreement and inter-collection stability, particularly in areas affected by cloud, terrain shadow, seasonal spectral convergence or sparse training data.

7. Concluding Remarks and Perspectives

Collection 11 extends the Caatinga annual LULC series to 1985–2025 and consolidates a basin-specific workflow that combines seasonal Landsat mosaics, feature selection, probabilistic sample refinement, GTB classification and a chained post-classification sequence. The direct treatment of Pasture, Temporary Crop/Agriculture and Perennial Crop, together with structure-preserving spatial filtering and explicit disagreement metrics, strengthens methodological transparency relative to previous collections.

8. References

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

GAO, B.-C. NDWI a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, v. 58, n. 3, p. 257–266, 1996.

IBGE. *Biomas e sistema costeiro-marinho do Brasil*. Rio de Janeiro: IBGE, 2019.

IBGE*. Vegetação RADAM. Disponível em: <ftp://geoftp.ibge.gov.br/informacoes_ambientais/acervo_radambrasil/vetores/>. Acesso em: 30 maio. 2018.*

IBGE *Mapa de Biomas do Brasil – primeira aproximação. Rio de Janeiro, 2020, disponível em: https://www.ibge.gov.br/geociencias/informacoes-ambientais/15842-biomas.html?=&t=download s, acessado em julho de 2020;*

LAWRENCE, R.; BUNN, A.; POWELL, S.; ZAMBON, M. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, v. 90, n. 3, p. 331–336, 2004.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: *IEEE International Conference on Data Mining (ICDM)*, 8., 2008. *Proceedings*. IEEE, 2008. p. 413–422.

PATIL, D.; PATIL, K.; NALE, R.; CHAUDHARI, S. Semantic segmentation of satellite images using modified U-Net. In: *IEEE Region 10 Symposium (TENSYMP)*, 2022. *Proceedings*. IEEE, 2022. p. 1–6.

PONTIUS JR., R. G.; MILLONES, M. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, v. 32, n. 15, p. 4407–4429, 2011.

RAMEZAN, C. A. Transferability of recursive feature elimination (RFE)-derived feature sets for support vector machine land cover classification. *Remote Sensing*, v. 14, n. 24, p. 6218, 2022.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. *Proceedings*. Cham: Springer, 2015. p. 234–241.

ZHANG, R.; MA, J. Feature selection for hyperspectral data based on recursive support vector machines. *International Journal of Remote Sensing*, v. 30, n. 14, p. 3669–3677, 2009.

MAPBIOMAS. *Algorithm Theoretical Basis Document (ATBD) General*.

AGÊNCIA NACIONAL DE ÁGUAS (ANA). *Bases hidrográficas Ottobacias*.

INSTITUTO NACIONAL DE METEOROLOGIA (INMET). Dados históricos de precipitação.

COCHRAN, W. G. Sampling Techniques. 3rd ed. New York: Wiley, 1977.

IBGE. Monitoramento da cobertura e uso da terra do Brasil: 2018/2020. Rio de Janeiro: IBGE, 2022. ISBN 978-85-240-4544-8. Available at: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?id=2101966&view=detalhes. Accessed: 6 Aug. 2026.

IBGE. Monitoramento da Cobertura e Uso da Terra do Brasil. Product page. Available at: https://www.ibge.gov.br/geociencias/cartas-e-mapas/informacoes-ambientais/15831-cobertura-e-uso-da-terra-do-brasil.html. Accessed: 6 Aug. 2026.

MAPBIOMAS. Accuracy Assessment Appendix Collection 10. Version 1.1, Aug. 2025. Supplied PDF.

MAPBIOMAS. Caatinga Collection 11 spectral-index formula catalogue. Available at: https://github.com/mapbiomas/brazil-caatinga/blob/main/lulc_30m_landsat/collection_11/src/samples_process/README_indices_espectrais.md. Accessed: 6 Aug. 2026.

FEYISA, G. L. et al. Automated Water Extraction Index. Remote Sensing of Environment, 2014. doi:10.1016/j.rse.2013.08.029.

LOUHAICHI, M. et al. Spatially located platform and aerial photography for documentation of grazing impacts. Geocarto International, 2001. doi:10.1080/10106040108542184.

SOUZA JR., C. et al. Combining spectral and spatial information to map canopy damage from selective logging and forest fires. Remote Sensing of Environment, 2005. doi:10.1016/j.rse.2005.03.011.

TUCKER, C. J. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment, 1979. doi:10.1016/0034-4257(79)90013-0.