Urban Areas - Algorithm Theoretical Basis Document (ATBD)

Collection 9.0

Version 1

Team

Breno Malheiros de Melo	USP/EESC
Edimilson Rodrigues dos Santos Junior	USP/EESC
Eduardo Felix Justiniano	USP/FFLCH
Julia Cansado	USP/FAU
Julio Cesar Pedrassoli	UFBA/POLI
Mayumi Hirye	USP/FAU
Talita Micheleti	USP/FAU

Collaborators

Fabio Mariz Gonçalves	USP/FAU
João Meyer	USP/FAU
Marcel Fantin	USP/IAU
Marcelo Montaño	USP/EESC
Marcos Roberto Martines	UFSCAR

How to cite this document?
MapBiomas. Urban Areas - Algorithm's Theoretical Basis Document (ATBD) (Collection 9, Version 1), 2024. https://brasil.mapbiomas.org/

Summary

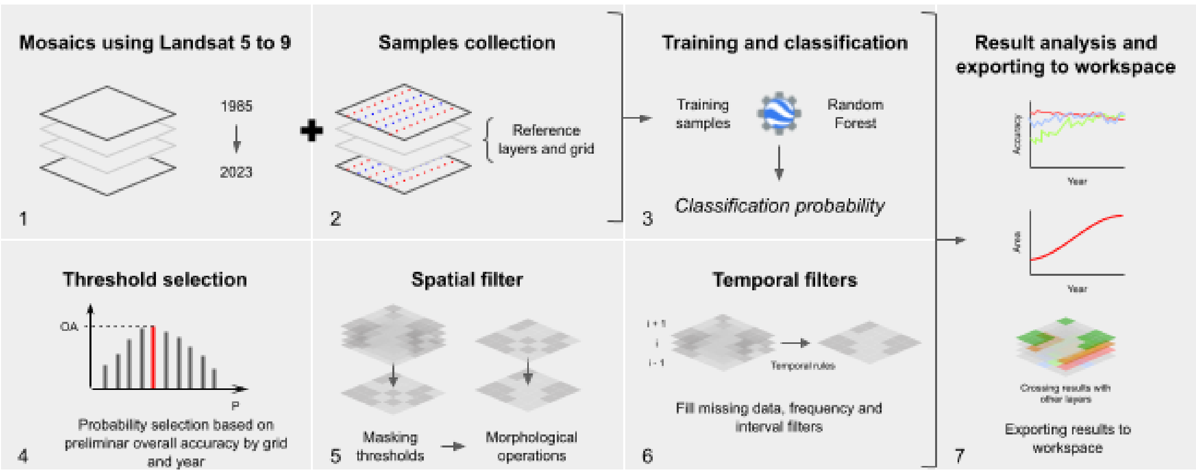
1. Overview

This document presents the methodology developed to map urban areas in Brazilian territory from 1985 to 2023 for Collection 9 of the MapBiomas Project. Following the general methods of MapBiomas, urban areas mapping employs supervised machine learning classification using annual aggregated Landsat images.

Each collection's series of urban area maps has been refined through conceptual and methodological improvements. In addition to updating ancillary datasets and satellite images, new procedures for this collection include sample adjustments, vegetation classification, and complementary analyses. These analyses involve intersecting urban areas with categorical layers such as slope classes, risk areas, and slums.

The methodological procedures involve mosaic production, sample development, training and classification, threshold selection, spatial filtering, temporal filtering, and exporting results to the MapBiomas workspace (Figure 1). After this, all MapBiomas project classes are integrated. The classification stages developed by the urban areas mapping team are detailed in the following sections, with codes available on the MapBiomas GitHub[1].

Figure 1. Basic scheme of urban areas maps production.



Since Collection 6, we have adopted as class name the term “Urban/Urbanized Area” (UA) instead of “Urban Infrastructure” in order to cope with the terminology applied in urban studies, such as IBGE (2017). UA are areas with predominance of significant density of buildings, roads and infrastructure. It should be noted that when making external quantitative comparisons, it is crucial to ensure that the chosen concepts are aligned.

[1] The GitHub contains all the versions of MapBiomas urban areas mapping codes. To the present collection, the following branch must be selected https://github.com/mapbiomas-brazil/urban-infrastructure/tree/mapbiomas90.

2. Landsat image mosaics

Landsat 5, Landsat 7, Landsat 8 and Landsat 9 imagery were used to create annual mosaics for mapping Urban Area in Collection 9.0, according to Table 1.

Table 1. Landsat imagery used in Urban Area mosaics.

Landsat Collection	Sensor	Collection	Level	Bands [wavelength]
Landsat 5 Level 2, Collection 2, Tier 1 [1]	TM	2	Surface Reflectance	SR_B1: Blue [0.45-0.52 µm] SR_B2: Green [0.52-0.60 µm] SR_B3: Red [0.63-0.69 µm] SR_B4: Near Infrared [0.77-0.90 µm] SR_B5 : Shortwave Infrared 1 [1.55-1.75 µm] SR_B7: Shortwave Infrared 2 [2.08-2.35 µm]
Landsat 5 TM Collection 2 Tier 1 Raw Scenes [2]	TM	2	Raw Images	B4: Near infrared [0.76 - 0.90 µm] B5: Shortwave infrared 1 [1.55 - 1.75 µm] B6: Thermal Infrared 1 (resampled from 60m to 30m) [10.40 - 12.50 µm]
Landsat 7 Level 2, Collection 2, Tier 1 [3]	ETM+	2	Surface Reflectance	SR_B1: Blue [0.45-0.52 µm] SR_B2: Green [0.52-0.60 µm] SR_B3: Red [0.63-0.69 µm] SR_B4: Near Infrared [0.77-0.90 µm] SR_B5 : Shortwave Infrared 1 [1.55-1.75 µm] SR_B7: Shortwave Infrared 2 [2.08-2.35 µm]
Landsat 7 Collection 2 Tier 1 Raw Scenes [4]	ETM+	2	Raw Images	B4: Near Infrared [0.77 - 0.90 µm] B5: Shortwave Infrared 1 [1.55 - 1.75 µm] B6_VCID_1: Low-gain Thermal Infrared 1 (resampled from 60m to 30m) [10.40 - 12.50 µm]
Landsat 8 and 9 Level 2, Collection 2, Tier 1 [5]	OLI / TIRS	2	Surface Reflectance	SR_B2: Blue [0.45 - 0.51 µm] SR_B3: Green [0.53 - 0.59 µm] SR_B4: Red [0.64 - 0.67 µm] SR_B5: Near Infrared [0.85 - 0.88 µm] SR_B6: Shortwave Infrared 1 [1.57 - 1.65 µm] SR_B7: Shortwave Infrared 2 [2.11 - 2.29 µm]
Landsat 8 and 9 Collection 2 Tier 1 Raw Scenes [6]	OLI / TIRS	2	Raw Images	B5: Near infrared [0.85 - 0.88 µm] B6: Shortwave infrared 1 [1.57 - 1.65 µm] B10: Thermal infrared 1 (resampled from 100m to 30m) [10.60 - 11.19 µm]

[1] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LT05_C02_T1_L2
[2] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LT05_C02_T1
[3] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LE07_C02_T1_L2
[4] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LE07_C02_T1
[5] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2 and https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC09_C02_T1_L2
[6] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1 and https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC09_C02_T1

Mosaics were created following these steps:

1. Filter Landsat Collections scenes by date (year-by-year, from 1985 to 2023) and bounds (Brazilian territory).
2. Mask pixels of clouds and cloud shadows in all scenes, using pixel quality attributes generated by the CFMASK[2] algorithm, using pixel quality, is in QA_PIXEL band.
3. Scale surface reflectance values to 0 to 1, using values of scale (-0,2) and offset (0.0000275) informed in collections' bands description in each reference page.
4. Calculate selected spectral indexes and fractions from spectral mixture analysis for each scene (Table 2).
5. Apply an appropriate reducer to each band/index to obtain one pixel value per year (Table 2).
6. Calculate reduced indexes difference to capture intra-annual changes (Table 2).
7. Composite all bands and indexes to obtain one mosaic per year.

Table 2. List, description, reducer, and script acronym used in Urban Areas mosaic (continue).

Band / Index / Fraction	Description	Reducer	Script acronym
BLUE	Landsat band	median	BLUE_median
GREEN	Landsat band	median	GREEN_median
RED	Landsat band	median	RED_median
NIR	Landsat band	median	NIR_median
SWIR1	Landsat band	median	SWIR1_median
SWIR2	Landsat band	median	SWIR2_median
NDVI	Normalized Difference Vegetation Index	median	NDVI_median
EVI1	Enhanced Vegetation Index 1	median	EVI_median
EVI1	Enhanced Vegetation Index percentiles	10th percentile, 90th percentile	EVI_p10 EVI_p90
EVI1	Enhanced Vegetation Index percentiles difference	difference	EVI_dif9010
EVI2	Enhanced Vegetation Index 2	median	EVI_median
EVI2	Enhanced Vegetation Index percentiles	10th percentile, 90th percentile	EVI2_p10 EVI2_p90
EVI2	Enhanced Vegetation Index percentiles difference	difference	EVI2_dif9010
MNDWI	Modified Normalized Difference Water Index	median	MNDWI_median
NDWI	Normalized Difference Water Index	median	NDWI_median

[2] CFMask is a multi-pass algorithm that uses decision trees to prospectively label pixels in the scene; it then validates or discards those labels according to scene-wide statistics. It also creates a cloud shadow mask by iteratively estimating cloud heights and projecting them onto the ground. Reference: https://www.usgs.gov/core-science-systems/nli/landsat/cfmask-algorithm.

Band / Index / Fraction	Description	Reducer	Script acronym
NDBI	Normalized Difference Built-Up Index	median	NDBI_median
NBR	Normalized Burn Ratio	median	NBR_median
NDRI	Normalized Difference Road Index	median	NDRI_median
BAI	Bare Soil Area Index	median	BAI_median
UI	Urban Index	median	UI_median
NDUI	Normalized Difference Urban Index	median	NDUI_median
BSI	Bare-Soil Index	median	BSI_median
BU	Built-up Index	median	BU_median
GV	Green Vegetation Fraction	median	GV_median
NPV	Non Photosynthetic Vegetation Fraction	median	NPV_median
SOIL	Soil Fraction	median	SOIL_median
CLOUD	Cloud Fraction	median	CLOUD_median
SHADE	Shade Fraction	median	SHADE_median
GVS	Green Vegetation + Soil Fraction	median	GVS_median
NDFI	Normalized Difference Fraction Index	median	NDFI_median
SUBS	Substrate Fraction	median	SUBS_median
VEG	Vegetation Fraction	median	VEG_median
DARK	Dark Fraction	median	DARK_median
EBBI	Enhanced Built-Up and Bareness Index	median	EBBI_median
EBBI	Enhanced Built-Up and Bareness Index percentiles	25th percentile, 90th percentile	EBBI_p25 EBBI_p90
EBBI	Enhanced Built-Up and Bareness Index percentiles difference	difference	EBBI_dif7525
EBBI	Positive part of the Enhanced Built-Up and Bareness Index	median	EBBIsNeg_median
EBBI	Positive part of the Enhanced Built-Up and Bareness Index percentiles	25th percentile, 90th percentile	EBBIsNeg_p25 EBBIsNeg_p75
EBBI	Positive part of the Enhanced Built-Up and Bareness Index percentiles difference	difference	EBBIsNeg_dif7525

3. Classification

3.1. Classification algorithm

The Random Forest algorithm implemented in Google Earth Engine (smileRandomForest) was applied to map Urban Areas in MapBiomas Collection 9.0 using training datasets of points in urban and non-urban areas.

To reduce computational cost, the automatic classification was performed only in "search areas", defined by polygons where urban areas were likely to be found. A uniform hexagonal polygon grid was created over Brazilian territory and intersected with urban census tracts (IBGE, 2021)[3], resulting in a search area of 226 million ha, covering 27% of the Brazilian territory.

Urban materials are reported to be highly spatially and temporally diverse. In time, diversity is related to the urbanization process itself. For example, in 1985, the streets of Humaitá (located inAmazonas State) were sparse and predominantly unpaved. Today, these same streets are paved and form part of a denser urban environment. Additionally, roofs and pavements display varied spectral behavior depending on their materials, colors, aging, and coating (such as algae, lichen, dirt, dust, rubber tire marks, etc.) (HEROLD et al., 2004).

To cope with the diversity of urban cover types, different random forest classifiers were built. We divided Brazil territory into 558 tiles that correspond to charts with a scale of 1:250.000, derived from the International Map of the World (IMW). Tiles with no search area were discarded, resulting in 522 valid tiles. Then, a specific classifier was trained to each of these tiles of each year of the 39 years of the Collection 9.0.

Random Forest parameters were set to 500 trees and 20 minimum leaf populations. The classification result is an image assigning to each pixel its probability of being urban.

3.2. Training Samples

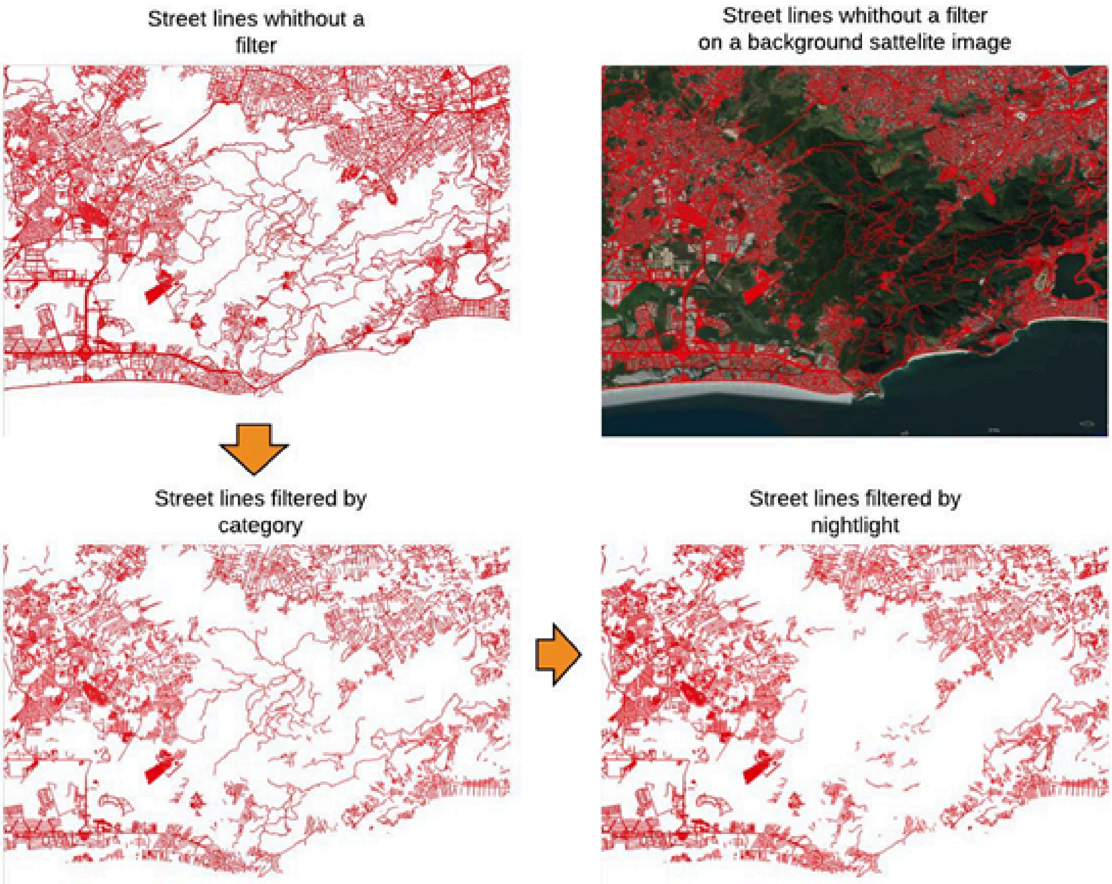
Training samples were obtained from OpenStreetMap database (OpenStreetMap, 2018) combined with nightlight images (NOAA), land cover and land use maps from the Third National Inventory (MCTI, 2015) and built-up maps of the Global Human Settlement Layer (GHSL), provided by Joint Research Center (JRC) (Corbane et al., 2019).

Firstly, a preliminary urban mask was built based on pathways from OpenStreetMap database, representing all roads, streets, sidewalks, and unknown roads already registered

[3] Census tracts are classified according to their situation. To build the search areas, we considered tracts in the following situations: (1) urban area with high density of buildings, (2) urban area with low density of buildings and (3) urban nucleus.

by OpenStreetMap users. Pathways within the urban patches or conglomerates of specific categories (residential, service, path, and living street) were selected. Then, pathways outside urban areas were removed using a nightlight image (NOAA) (Figure 2). For specific years, pathways were also filtered by existing data: built-up surfaces maps of the Global Human Settlement Layer (GHSL) for 1985, and urban area mappings of the Third National Inventory (MCTI, 2015), for 1994, 2002, and 2010. A buffer of approximately 100 meters distance of each pathway transformed these filtered pathways into areas.

Figure 2. Example of filters used on the vector layer of OpenstreetMap in Rio de Janeiro - RJ Brazil.



Secondly, an explorative classification of urban areas using indexes of normalized difference of vegetation and water (NDVI and NDWI) was produced to mask water and vegetation (Figure 3).

The final urban mask was obtained by the intersection of the preliminary mask, derived from OpenStreetMaps polylines and filtered by ancillary data, with the explorative classification (Figure 4 and Figure 5). The final non-urban mask is the symmetrical difference of the final urban mask.

Random points were generated in each of the 522 tiles observing the extension of urban area and non-urban area, according to the final urban mask (Figure 6). The final sample dataset comprises an average of 2,000 urban and non-urban points per tile in the initial years, increasing to 4,000 points in the final years of the time series.

Figure 3. Explorative classification results for Rio de Janeiro - RJ, Brazil.

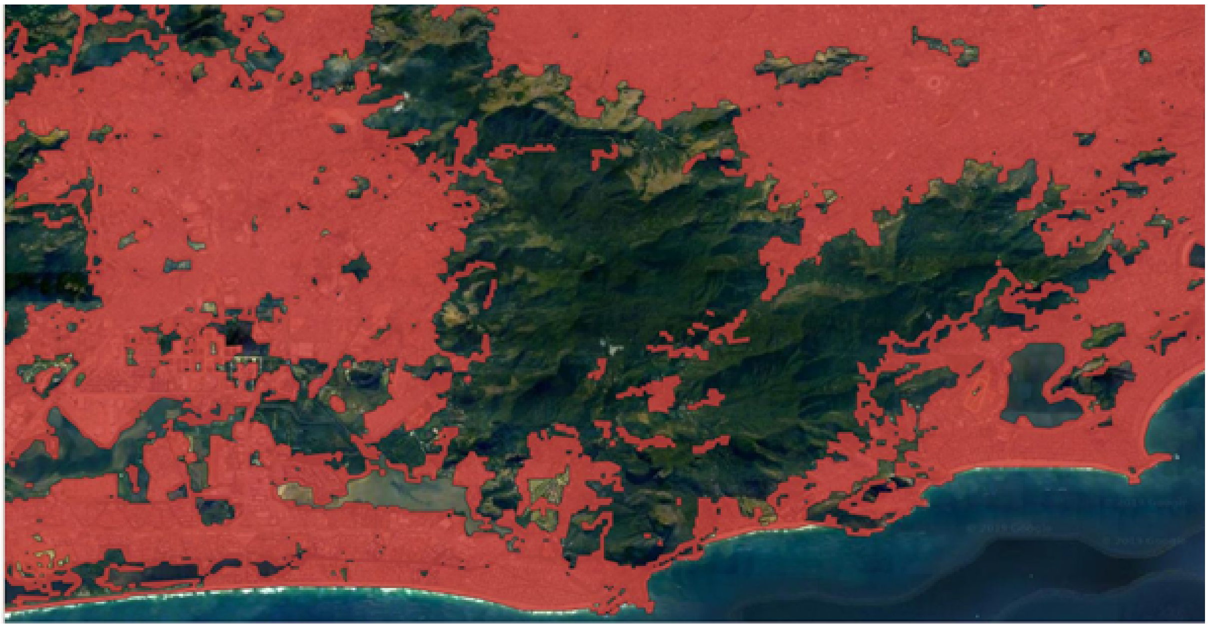
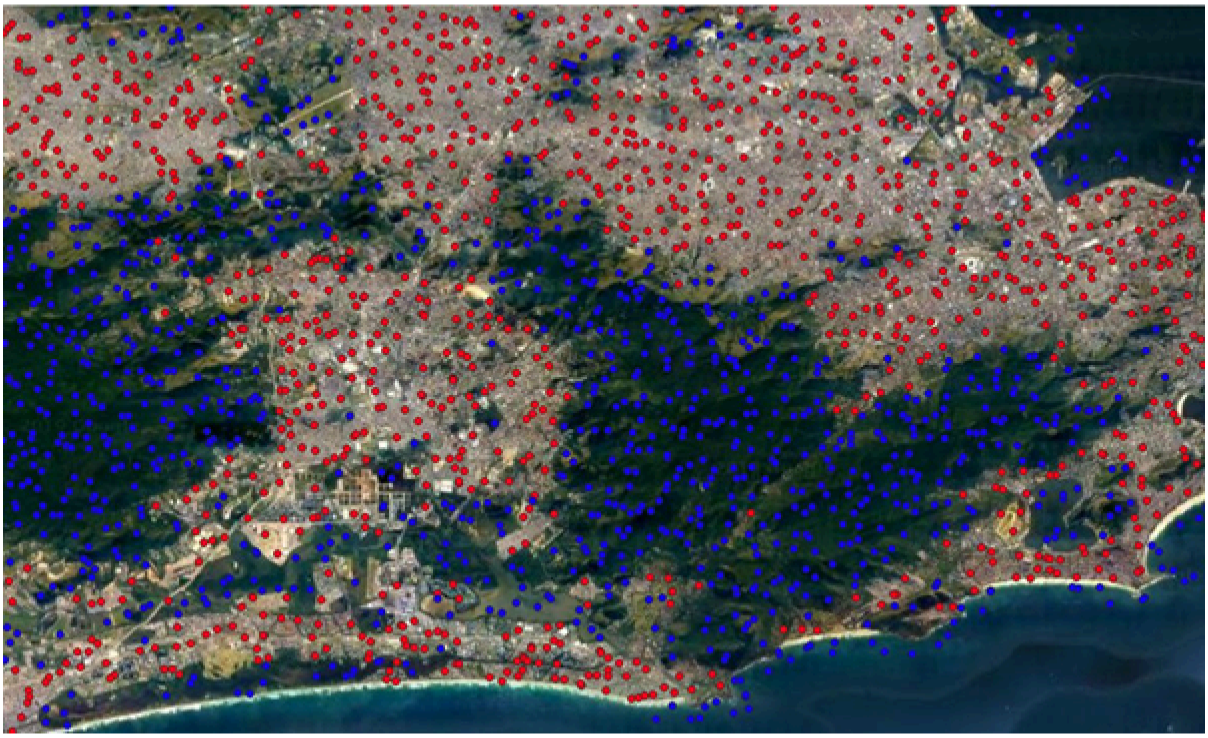


Figure 4. Final urban mask for Rio de Janeiro- RJ, Brazil.

Figure 5. Final non urban mask (orange color) for Rio de Janeiro - RJ, Brazil.

Figure 6. Random points divided by urban areas (red) and non-urban areas (blue).



Random forest algorithm is sensitive to imbalanced training data set (Breiman, 2001). Since its primary goal is to minimize the overall error rate, results tend to focus on accurately predicting the majority class, non-urban areas, causing a decrease in accuracy for the minority class, urban areas. To avoid this, we empirically determined a balance of urban and non-urban points to each tile and sequence of years (1985-1993, 1994-2001, 2002-2009, 2010-2017 and 2018-2023), overestimating the urban areas in relation to non-urban areas. Considering all tiles, the average balance of samples was as follows: between 1985-1993, 1 urban sample to 2.7 non-urban samples; between 1994-2009, 1 urban sample to 2 non-urban samples, between 2010-2017, 1 urban sample to 1.9 non-urban samples; and in the final years (2018-2023), 1 urban sample to 1.7 non-urban samples.

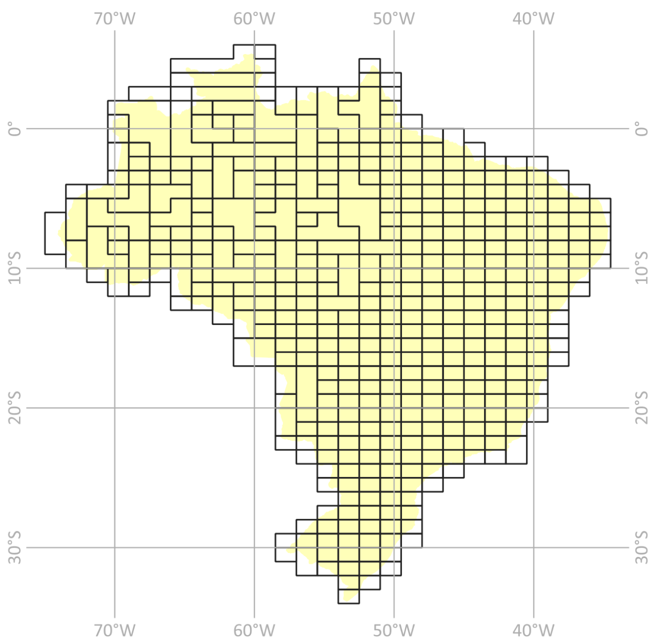
3.3. Feature space

The feature space that characterizes Urban Areas for MapBiomas Collection 9.0 is the dataset of urban and non-urban points trained with 43 variables from Landsat image mosaics (summarized in Table 1, Section 2), calculated for each tile.

Datasets of urban and non-urban samples were used with the assumption that once a point was urban, it remained urban for the following years. Therefore, images of 1985 up to 1993 were used to train the dataset of 1985, resulting in one feature space per year per tile. Likewise, images of 1994 up to 2002 were used to train the dataset of 1994, images of 2003 up to 2009, to train the dataset of 2003, images of 2010 up to 2017, to train the dataset of 2010 and images of 2018 up to 2023, to train the dataset of 2018.

A neighborhood approach was adopted in tiles without feature space due to the lack of urban samples or lack of cloud-free data. In these cases, the nearest tile with a feature space was used, resulting in 421 tiles with feature spaces for each year (Figure 7).

Figure 7. Feature space's tiles.



4. Spatial and Temporal Filters

The spatial and temporal filters were configured to enhance the classification considering the diversity of materials and features of the urbanized areas across various Brazilian municipalities from 1985 to 2022. These procedures resulted in a binary raster which indicated urban and non-urban areas.

4.1 Spatial Filter

The classifier can assign high values of UA probability to non-urban areas such as mining, sands, rural structures, and others. Conversely, the presence of trees and squares in the urban context can result in low probability values for those pixels. Setting universal probability thresholds for assigning a pixel to a UA would lead to errors due to the varying characteristics of different cities. To address this, the thresholds for defining UA were set within the spatial filter code, which combined other data and contexts. The layers used in this process are presented in Table 3.

Table 3. Layers and thresholds for the spatial filter.

Layer	Description	Threshold criteria	Why use it?
IRS	The Index of Roads and Infrastructure (IRS) defines urban limits according to roads and infrastructure density. Quantitative layer.	Values greater than or equal to 500 (JUSTINIANO et al., 2022).	Provides a general mask layer identifying where urban areas must be.
VIIRS	Visible Infrared Imaging Radiometer Suite (VIIRS) defines general regions where urban areas can be found according to night light values. Quantitative layer.	Greater than or equal 1.	
Census tracts	Census tracts from official data define urban limits according to census criteria and official organism. Qualitative layer.	Tracts with urban characterization (types 1, 2 and 3) (IBGE, 2020).	
Slums	Slums, provided by official data, define regions where there are human populations with specific vulnerabilities around or within urban perimeter. Qualitative layer.	All the regions were considered.	
UA (probability)	Urban areas (UA) classification determines the probability of urban areas based on RF applied to time series data. Quantitative layer.	Defined using 'best threshold algorithm'.	Provides the urban classification

Given that UA typically have higher population densities and light emission at night, raster files from the Visible Infrared Imaging Radiometer Suite (VIIRS), aboard the Suomi NNP satellite's Day Night Band were used. The threshold value of the VIIRS band and the annual UA probability values were established through an algorithm that calculates the best threshold for each grid per year of the time series (see urban areas GitHub).

The result of applying filters based on threshold values per year is a raster with values zero and one, where the latter value is associated with the urban area. This raster often reveals isolated pixels or small clusters of pixels with values different from their surroundings. In the urban areas, small clusters of isolated zero-value pixels are typically associated with squares, boulevards, water, trees, and other urban elements. In non-urban areas, isolated one-value pixels may correspond to agricultural structures, summer homes, and other non-urban structures.

Spatial filters also perform morphological operations to refine the classification. Circular kernels with one pixel as a neighborhood are used to conduct these operations. Morphological closing operations remove holes with fewer than 60 clustered pixels, while morphological opening operations eliminate noise consisting of fewer than 5 pixels.

4.2 Temporal filter

Temporal filters (TF) were applied to ensure classification consistency over time, taking into account the conceptual aspects defined for the mapped category. The sequence of filters, indicated and described in Table 4, was developed for this purpose. General rules (GR) were established for the middle years, while specific rules were defined for the first years (FYR) and last years (LYR) of the time series for each TF. Temporal consistency was determined based on results obtained by pixel over a period of 3 to 5 years (kernel) for the immediately preceding TF results.

Table 4. Descriptions of TFs used.

TF	Scope	Type	Years	Kernel						Conditionals
				i-2	i-1	i	i+1	i+2	i+3	
1	It acts on the pixels that were classified as ‘UA’ in the SF results and sets the mask for the filters up to TF2.	FYR	1985 and 1986			x	x	x		If the pixel under analysis is classified as ‘UA’ within two or more years of the kernel, then the ‘UA’ is validated to the next step.
		GR	1987 to 2021	x	x	x	x	x		If the pixel under analysis is classified as ‘UA’ within three or more years of the interval, then the ‘UA’ is validated to the next step.
		LYR	2022 to 2023	x	x	x				If the pixel under analysis is classified as ‘UA’ within two or more years of the interval, then the ‘UA’ is validated to the next step.
2	It acts on pixels that have been validated as ‘UA’ in the TF1 results.	GR	1985 to 2020			x	x	x	x	If the pixel under analysis is classified as ‘UA’ within two or more years of the range, then the ‘UA’ is validated to the next step.
		LYR	2021 and 2022	x	x	x	x			If the pixel is classified as ‘UA’ within two or more years of the range, then the ‘UA’ is validated to the next step.
		LYR	2023	x	x	x				If the pixel is classified as ‘UA’ within two or more years of the range, then the ‘UA’ is validated to the next step.
3	Extends the filter mask and acts on pixels not classified as urban in TF2.	FYR	1985			x				The results obtained for TF2 are assumed.
		GR	1986 to 2022		x	x	x			If the pixel is classified as ‘UA’ for i-1 and i+1, then ‘UA’ is assigned.
		LYR	2023		x	x				If the pixel is rated 'UA' for i-1, then 'UA' is assigned.
4	Area Consolidation Filter.	FYR	1985			x	x			If a pixel under analysis is rated 'UA' for i and not rated for i+1, then it becomes unranked.
		FYR	1986			x				The results obtained for TF3 are assumed.
		GR	1986 to 2023			x	x			If a pixel under analysis is classified as 'UA' for i, then for i+1 it will also be.

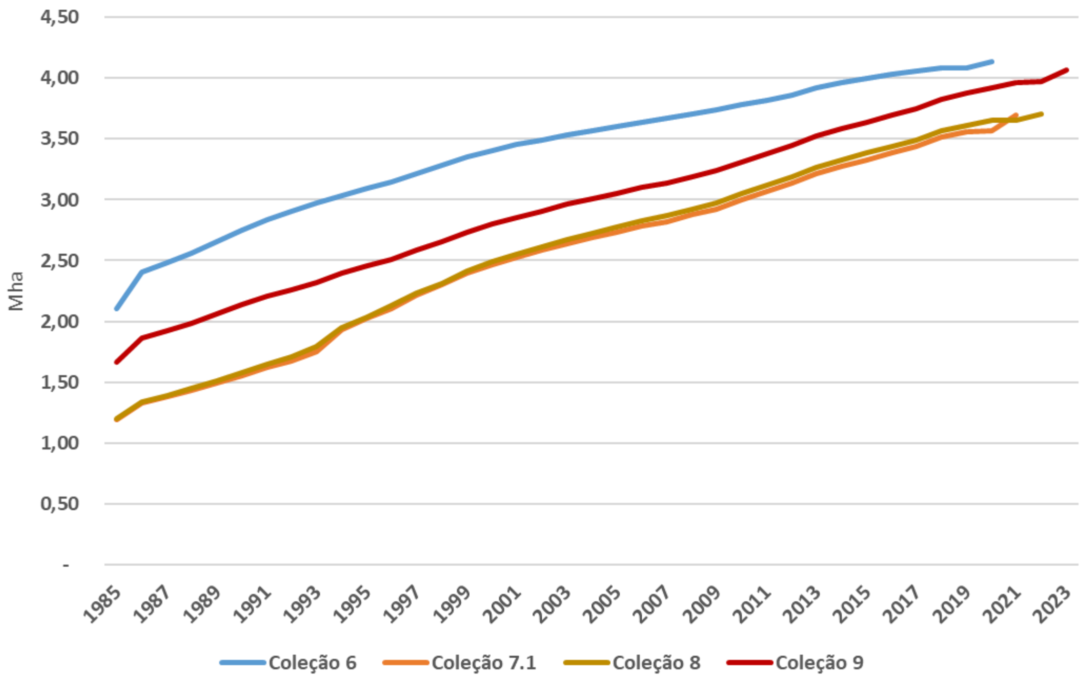
FYR = firsts years rules; GR = general rule; LYR = lasts years rules.

5. Comparison between Collections

5.1 Area

Figure 7 presents urban growth comparing the current collection to the previous ones. The results suggest that the criteria adopted in this collection have been more comprehensive over time, leading to a more homogeneous and conceptually compatible classification process. This is particularly perceived by the area curve from 1985 to 1994, where the actual procedures impacted the results avoiding unexpected changes. Also, an average increase of 17% is noticeable when the curve is compared with results from the Collection 8.0, improving the results related to omission errors - except when compared to Collection 6.0.

Figure 7. Comparison between growth of urbanized areas for Collections



5.2 Performance

The validation analysis was performed using point samples collected by the Laboratory of Image Processing and Geoprocessing (LAPIG), University of Goiás-GO, Brazil. According to the reference data, (i) the samples of Urban Areas (UA) that were correctly classified are the True Positives; (ii) the samples of UA that were classified as not UA are the False Negatives; and (iii) the samples of non-UA that were classified as UA are the False Positives.

The comparison of accuracy results between Collection 6.0, Collection 7.1, Collection 8.0, and Collection 9.0 are shown in Figure 9 and Figure 10. Figure 9 shows that omission errors in the latest collection were smaller than in the previous ones, except for Collection 6.0. However, the commission errors were larger and more concentrated in the first 15 years of the time series.

Figure 9. Omission (Producer's accuracy).

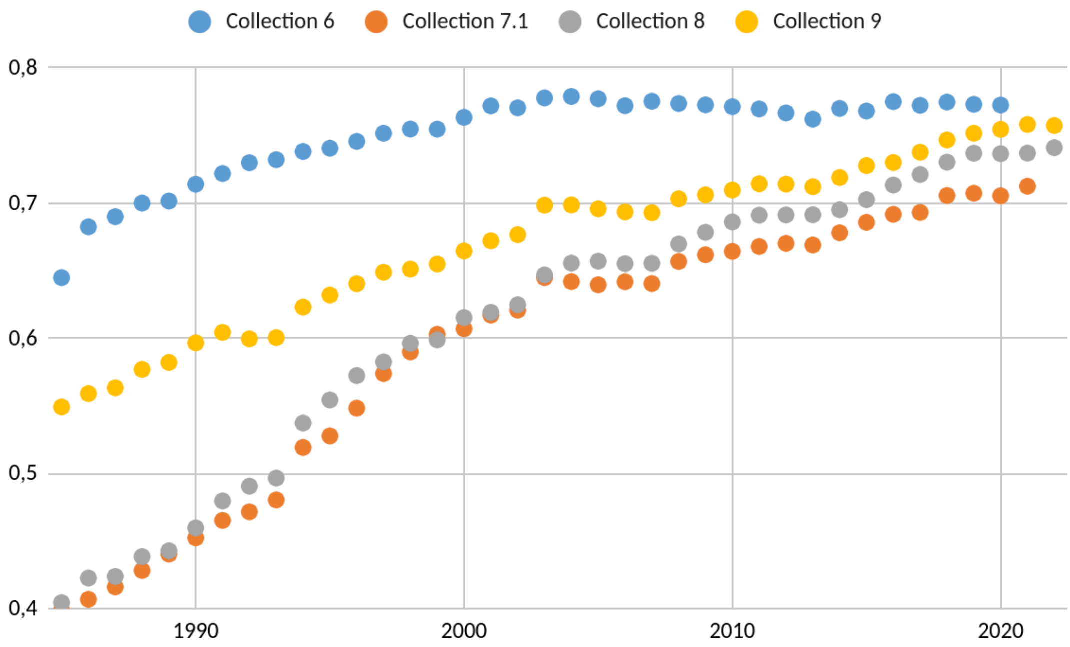
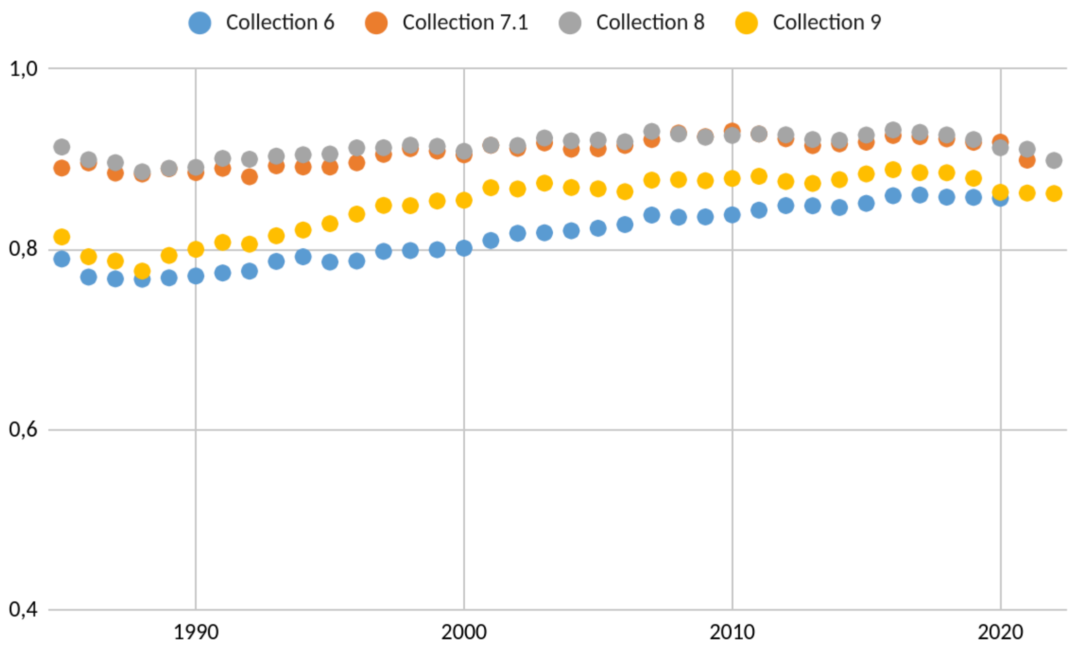


Figure 10. Commission (User's accuracy).



6. Reference Maps

MapBiomas Collection 9.0 were compared to two urban area maps: (1) the World Settlement Footprint (WSF) produced by Deutsches Zentrum für Luftund Raumfahrt (DLR) (MARCONCINI et al., 2020) and (2) Brazil Urbanized Areas produced by IBGE, Instituto Brasileiro de Geografia e Estatística (IBGE, 2022).

WSF is a 10m resolution binary mask outlining the extent of human settlements globally derived by means of 2014-2015 multitemporal Landsat-8 and Sentinel-1 imagery, using different classification schemes based on Support Vector Machines. It is available at Earth Engine Data Catalog[4].

Quantitative analysis (Table 5) shows that the urbanized areas mapped for 2015 by MapBiomas Collection 9.0 amount to 3,691,485 hectares, which is 8.3% more than the corresponding data reported by the WSF for the same reference year. This discrepancy can be attributed to the higher resolution of the WSF data (10 meters) compared to the MapBiomas data (30 meters). The overlap between these two mappings is 81% when considering the entirety of Brazilian territory. By biome, it is observed that the Caatinga has the lowest overlap at 75%, whereas the Amazon and Cerrado biomes exhibit overlaps exceeding 83%.

[4] https://developers.google.com/earth-engine/datasets/catalog/DLR_WSF_WSF2015_v1

Table 5. Quantitative analysis between MapBiomas Collection 9.0 and WSF 2015.

Year: 2015 Biome	**MapBiomas Collection 9.0** (in ha)	**WSF 2015** (in ha)	**Overlap: MapBiomas Collection 9.0 and WSF 2015** (in ha)	**Overlap in relation to WSF 2015** (in %)	**MapBiomas Collection 9.0 in relation to WSF 2015** (in %)
Amazon	415.601	310.982	259.461	83,4%	133,6%
Caatinga	387.973	337.939	252.974	74,9%	114,8%
Cerrado	861.114	724.210	606.411	83,7%	118,9%
Atlantic Forest	1.895.438	1.910.032	1.539.472	80,6%	99,2%
Pampa	126.005	120.837	99.858	82,6%	104,3%
Pantanal	5.354	4.596	3.729	81,1%	116,5%
Brazil	3.691.485	3.408.597	2.761.906	81,0%	108,3%

Brazil Urbanized Areas is a visual interpretation of urban features, identified according to the elements of specific shape (geometry of objects) and pattern (spatial arrangement). It is based on Sentinel 2 imagery, with spatial resolution of 10m. It is available in shapefile format at IBGE's website[5]. The mapped urban land use types include: "Urbanized Area," categorized into two classes — high density and low density —, "Other Urban Facilities," and "Vacant Urban Development." Considering the definition of "Urban Areas" adopted by MapBiomas, the comparative analysis was carried out by evaluating the following classes from the IBGE mapping: "Urbanized Area - High Density," "Other Urban Facilities," and "Vacant Urban Development".

The comparison with IBGE's 2019 data points to an underestimation of MapBiomas Collection 9.0 urban area. For the year 2019, IBGE reports 4.3 million hectares, whereas MapBiomas reports 3.9 million hectares. This underestimation is reported in MapBiomas accuracy assessment. This discrepancy is speculated to be attributed to the different classification methods: visual interpretation used by IBGE versus semi-supervised classification employed by MapBiomas.

The overlap between MapBiomas Collection 9 and IBGE 2019 is higher compared to that with WFS 2015, at 86.4% versus 81% respectively. However, the overlap by biome shows consistent results in both comparative analyses: the Caatinga biome has the lowest overlap, while the Cerrado and Amazon biomes exhibit the highest overlap.

---

5 https://www.ibge.gov.br/geociencias/cartas-e-mapas/redes-geograficas/15789-areas-urbanizadas.html?=&t=acesso-ao-produto

Table 6. Quantitative analysis between MapBiomas Collection 9.0 and IBGE 2019 (selected classes).

Year: 2019 Biome	**MapBiomas Collection 9.0** (in ha)	**IBGE 2019** (in ha)	**Overlap: MapBiomas Collection 9.0 and IBGE 2019** (in ha)	**Overlap in relation to IBGE 2019** (in %)	**MapBiomas Collection 9.0 in relation to IBGE 2019** (in %)
Amazon	431,610	397,735	315,081	79.2%	108.5%
Caatinga	447,490	531,115	340,936	64.2%	84.3%
Cerrado	915,453	857,191	697,174	81.3%	106.8%
Atlantic Forest	2,001,045	2,332,943	1,745,393	74.8%	85.8%
Pampa	129,413	165,387	119,470	72.2%	78.2%
Pantanal	5,467	4,687	3,663	78.2%	116.6%
Brazil	3,930,477	4,289,058	3,221,718	86.4%	91.6%

7. References

BAGAN, H., BORJIGIN, H., & YAMAGATA, Y. (2018). Assessing nighttime lights for mapping the urban areas of 50 cities across the globe. **Environment and Planning B: Urban Analytics and City Science**, 239980831775292. DOI: 10.1177/239980831775292

BREIMAN, L. (2001). Random Forests. **Machine Learning**, *45*(1), 5–32. DOI: 10.1023/A:1010933404324

CORBANE, C. et al. **GHS built-up grid, derived from Landsat, multitemporal (1975-1990-2000-2014), R2018A**. European Commission, Joint Research Centre (JRC), 2019. doi:10.2905/jrc-ghsl-10007 PID: http://data.europa.eu/89h/jrc-ghsl-10007.

GOLDBLATT, R. et al. Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover. **Remote Sensing of Environment**, v. 205, p. 253–275, 1 fev. 2018.

HEROLD, M. et al. Spectrometry for urban area remote sensing—Development and analysis of a spectral library from 350 to 2400 nm. **Remote Sensing of Environment**, v. 91, n. 3-4, p. 304-319, 2004. DOI: 10.1016/j.rse.2004.02.013.

Instituto Brasileiro de Geografia e Estatística (IBGE). **Áreas urbanizadas do Brasil: 2015**. Rio de Janeiro: IBGE, 2017.

Instituto Brasileiro de Geografia e Estatística (IBGE). **Malha de Setores Censitários**. Rio de Janeiro: IBGE, 2021. Available at: https://www.ibge.gov.br/geociencias/downloads-geociencias.html?caminho=organizacao_do_territorio/malhas_territoriais/malhas_de_setores_censitarios__divisoes_intramunicipais/2021/Malha_de_setores_(shp)_Brasil

Instituto Brasileiro de Geografia e Estatística (IBGE). Áreas urbanizadas do Brasil: 2019. IBGE, Coordenação de Meio Ambiente: Rio de Janeiro, 2022. Available at: https://www.ibge.gov.br/geociencias/cartas-e-mapas/redes-geograficas/15789-areas-urbanizadas.html?=&t=acesso-ao-produto

Instituto Geológico (IG) / Secretaria do Meio Ambiente do Estado de São Paulo (SMA). **Unidades Homogêneas de Uso e Cobertura da Terra.** 2014. Available at: http://s.ambiente.sp.gov.br/cpla/UHCT_112015_v2.zip

JUSTINIANO et al. Proposal for an index of roads and structures for the mapping of non-vegetated urban surfaces using OSM and Sentinel-2 data. **International Journal of Applied Earth Observation and Geoinformation**, v. 109, 11 p., mai. 2022.

LIANG, J. et al. Modeling urban growth sustainability in the cloud by augmenting Google Earth Engine (GEE). **Computers, Environment and Urban Systems**, v. 84, p. 101542, 1 nov. 2020.

LIU, X.et al. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. **Nature Sustainability** (2020). DOI:10.1038/s41893-020-0521-x

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., Paganini, M., Strano, E. (2020). Outlining where humans live, the World Settlement Footprint 2015. **Scientific Data**, 7(1), 1-14. doi:10.1038/s41597-020-00580-5

MONTEIRO, A. et al. SIG Contribution in the Making of Geotechnical Maps in Urban Areas. **IOP Conference Series: Materials Science and Engineering**, v. 245, p. 022029, out. 2017.

OSM. **OpenStreetMap (Standard)**, 2021.