Urban Areas mapping

MapBiomas 10m - Collection 3 (beta)

Version 1

Team authors

Breno Malheiros de Melo	UFSCAR
Clara Dias	USP/FAU
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
Marcelo Montaño	USP/EESC
Marcos Roberto Martines	UFSCAR

April, 2026

1. Overview

This document describes the methodological framework developed to map urban areas across the Brazilian territory as part of the MapBiomas 10 m Collection 3. The proposed approach builds upon the conceptual and methodological foundations established in previous MapBiomas collections, particularly Collection 10 based on Landsat imagery, while incorporating methodological advances enabled by higher spatial resolution data and modern representation learning techniques. Urban areas are defined as portions of the territory characterized by a predominance of built-up structures, roads, and associated infrastructure, reflecting consolidated or expanding urban occupation patterns. This definition remains consistent with that adopted in recent MapBiomas collections and aligns with terminology commonly used in urban and geographic studies in Brazil. Maintaining conceptual continuity ensures comparability between collections and supports long-term analyses of urban expansion and dynamics.

The MapBiomas 10 m Collection introduces methodological improvements by leveraging Sentinel-2 imagery at 10-meter spatial resolution and by adopting latent feature representations derived from Alpha Earth Embeddings. These representations integrate spectral, spatial, and temporal information learned from large volumes of Earth observation data, enabling a richer and more discriminative feature space compared to traditional approaches based solely on spectral bands and indices. This advancement enhances the ability to detect and delineate urban areas, particularly in complex contexts such as peri-urban zones, small and medium-sized cities, and areas of diffuse urban expansion.

Despite these advances, the mapping strategy preserves key elements of the MapBiomas framework to ensure consistency and robustness. The classification process relies on supervised machine learning, using the Random Forest algorithm, and reuses the consolidated training sample dataset developed for the Landsat-based Collection 10. By maintaining the same sample geometry and thematic labeling, differences between products primarily reflect improvements in spatial resolution and feature representation, rather than changes in conceptual definitions or sampling strategies.

2. Classification

2.1 Training samples

The training samples used in this study belong to the same dataset adopted in MapBiomas Collection 10 based on Landsat imagery. The construction strategy of this sample set aims to ensure broad spatial representativeness at the national scale, adopting simple and efficient procedures that enable its application across the entire Brazilian territory. This approach ensures thematic consistency among different collections and allows the reuse of a consolidated training dataset.

The sample base results from the integration of multiple independent auxiliary data sources, combining vector information, remote sensing imagery, and consolidated thematic products. The main sources include the OpenStreetMap (OSM) database, nighttime lights imagery from NOAA, land use and land cover maps from the Third National Inventory of Greenhouse Gas Emissions (MCTI), and global built-up area maps from the Global Human Settlement Layer (GHSL). This methodology enables the generation of a training dataset comprising hundreds of thousands of samples distributed throughout the Brazilian territory, supporting a training process capable of capturing the high spectral and spatial variability associated with different regional urbanization contexts.

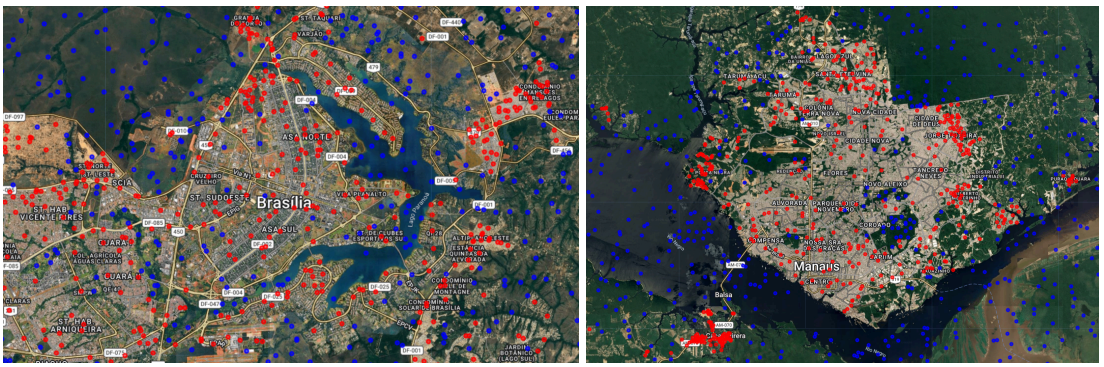

Figure 1. Random points divided by urban areas (red) and non-urban areas (blue)

In the 10 m Collection, these same samples are reused without changes to their geometry or thematic labeling, ensuring conceptual continuity with the Landsat-based collections. The main difference relative to those collections lies in the training process, in which the sample points are associated with a new feature space composed of latent representations derived from the Alpha Earth Embeddings, replacing the Landsat spectral bands and indices traditionally employed. This strategy preserves the

conceptual and thematic basis of the samples while enabling the exploration of a richer and more informative feature space, capable of integrating spectral, spatial, and contextual patterns. As a result, the classification process becomes compatible with the 10-meter spatial resolution without introducing inconsistencies in the definition of the urban areas class or in the logic used to construct the training dataset.

2.2 Feature Space

The core component of the feature space applied to urban areas mapping in the MapBiomas 10 m Collection consists of latent representations generated by the AlphaEarth Foundations model, made available in Google Earth Engine as the Satellite Embedding dataset (GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL). This model was developed as a self-supervised geospatial representation layer that integrates spectral, spatial, and temporal signals from multiple remote sensing sources, transforming them into 64-dimensional vectors per pixel.

From a conceptual perspective, each pixel in this annual embedding collection corresponds to a point on a normalized 64-dimensional surface, where its position in the latent space synthesizes information learned from time series of observations acquired by different sensors (including optical, radar, and other environmental variables) and temporal patterns. This compact representation implicitly encodes spatial characteristics, seasonal variations, and contextual relationships among observation modalities, potentially providing richer information than attributes derived directly from traditional spectral bands or indices.

2.3 Classification strategy

Urban areas classification adopts the Random Forest algorithm as the main classification model. Although the use of simpler classifiers, such as k-Nearest Neighbors (k-NN), is often recommended in embedding-based contexts, the nature of the training sample production process for urban areas introduces unavoidable levels of noise into the training dataset, a scenario in which Random Forest demonstrates greater robustness.

Random Forest constitutes a supervised ensemble learning algorithm in which multiple independent decision trees are constructed from random subsets of the samples (bootstrap) and random subsets of the feature space. This mechanism reduces the model’s dependence on individual observations and limits the impact of

isolated labeling errors, since incorrectly labeled samples tend to influence only a fraction of the trees. The majority voting process across trees dilutes these effects, allowing the model to emphasize dominant and consistent patterns in the feature space while reducing the influence of outliers and noisy samples. In contrast, proximity-based methods such as k-NN classify each observation based on the labels of the nearest samples in the feature space. Under this approach, incorrectly labeled samples exert direct and localized influence on the classification decision, increasing susceptibility to systematic errors. This effect becomes particularly relevant in transition zones between classes, such as peri-urban areas or urban expansion zones, where spatial and spectral variability is high and class separation is inherently ambiguous.

Beyond classifier selection, the spatial subdivision strategy plays a fundamental role in the quality of the results. Given the territorial extent of Brazil, avoiding excessively generalist models becomes essential. To address this issue, the territory is subdivided into regular spatial units associated with cartographic sheets and search areas, following the approach adopted in MapBiomas Collection 10. This strategy constrains training and classification to more homogeneous regional contexts, favoring model adaptation to local urbanization patterns.

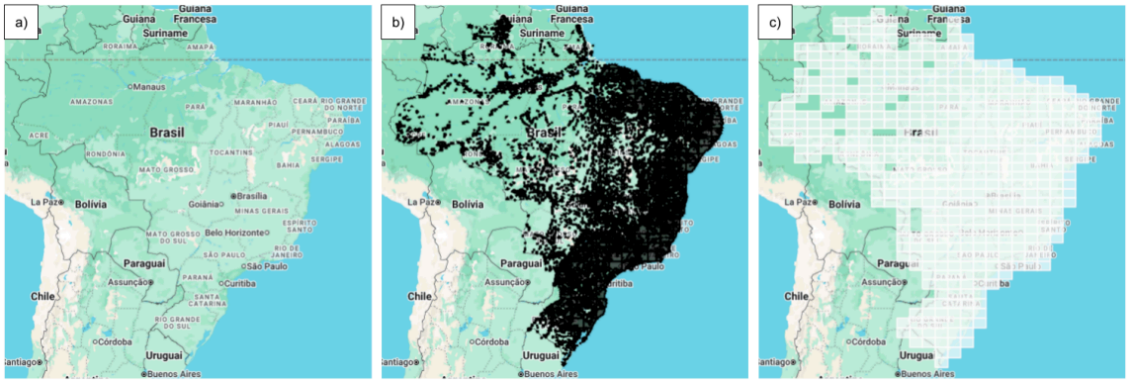

Figure 2. Search area and regular sheets (processing tiles). a) Brazil. b) Search area - where urban areas can be found. c) Processing units - regular grid defining the tiles for processing the classification.

The model training was implemented using a moving window approach: a block of nine grid units (3×3) was used, where the central unit served as the classification target, and the nine neighboring units provided the training data. For evaluation purposes, the final validation was conducted using MapBiomas validation samples

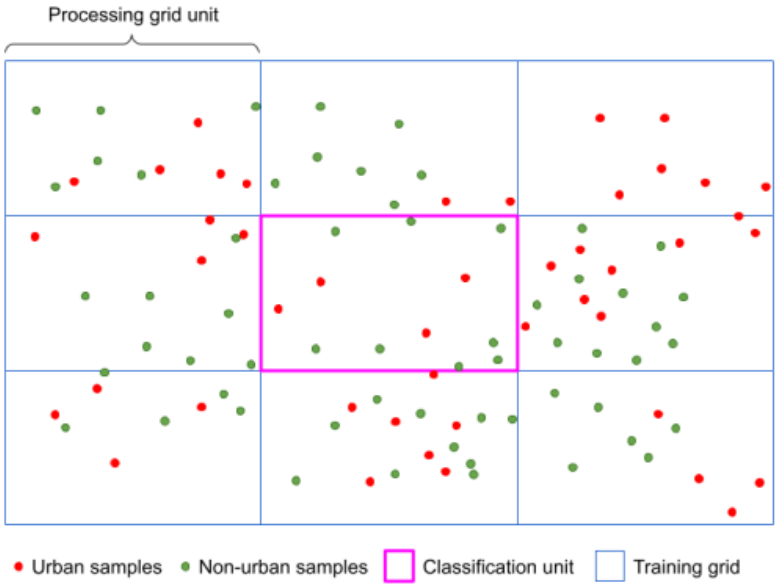

Figure 3. Processing grid for classification scheme

The output of the classification step consists of probability images, in which each pixel expresses the probability of belonging to the urban areas class. The decision threshold used to convert these probabilities into a binary classification (urban and non-urban) is defined automatically based on a preliminary accuracy assessment conducted for each cartographic sheet. As a result, the process generates not only a set of probability images but also threshold values specific to each sheet and year, ensuring greater spatial and temporal consistency in the final classification.

3. Post-classification

3.1 Spatial Filter

To refine urban areas (UA) classification, data on urbanized areas and informal settlements from the 2022 census were integrated with the Index of Roads and Infrastructure (IRS) developed by Justiniano et al. (2022). This combined dataset supports the delineation of the maximum plausible spatial extent of urban pixels, based on recent historical patterns of occupation and infrastructure density. By constraining the classification to areas with consistent evidence of urbanization, this approach reduces commission errors and improves the accuracy and spatial coherence of urban areas mapping.

3.2. Temporal Filter

Temporal filters (TF) were applied as rules to check classification consistency over time, observing the conceptual aspects delimited to the mapped category. For this purpose, the sequence of filters indicated and described in Table 1 was developed. General rules (GR), Consolidation rule (CR) for middle years, and specific rules for the first years (FYR) and last years (LYR). The goal is to perform a series of corrections based on the original pixel history classification and subsequently consolidate these corrections using the consolidation rule.

Table 1. Descriptions of Rules.

Rule	Years (i)	Kernel						Conditionals
		i-2	i-1	i	i+1	i+2	i+3	
GR	2018 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
FYR	2017			x	x			If the pixel under analysis is classified as ‘UA’ within two years of the kernel, then the ‘UA’ is validated
LYR	2024		x	x				If the pixel under analysis is classified as ‘UA’ within two or more years, then the ‘UA’ is validated
CR	2018 to 2023		x	x	x			It consolidates the results based on the corrections indicated by the previous rules.

3.3 Morphological filter

Isolated pixels or small clusters, which frequently correspond to classification artifacts, are addressed through spatial filtering based on morphological operations. In urban areas, small clusters of zero-value pixels may represent internal features such as squares, parks, or water bodies, whereas in non-urban areas, isolated one-value pixels often correspond to agricultural or rural structures. The morphological filtering applies circular kernels, using closing operations to fill small internal gaps (clusters smaller than 10 pixels) and opening operations to remove isolated noise (clusters smaller than 10 pixels), thereby improving the spatial coherence of the final classification

4. Validation strategies

For each MapBiomas collection, the classification methodology is updated and the entire time series is reprocessed, resulting in a recalculation of the mapped class areas. As a result, variations in the total mapped urbanized area are expected between collections, as illustrated in Figure 4. For 10m Collection 3, the total mapped urbanized area starts at approximately 3.5 million hectares in 2017 and increases over time, reaching around 4.3 million hectares by 2024.

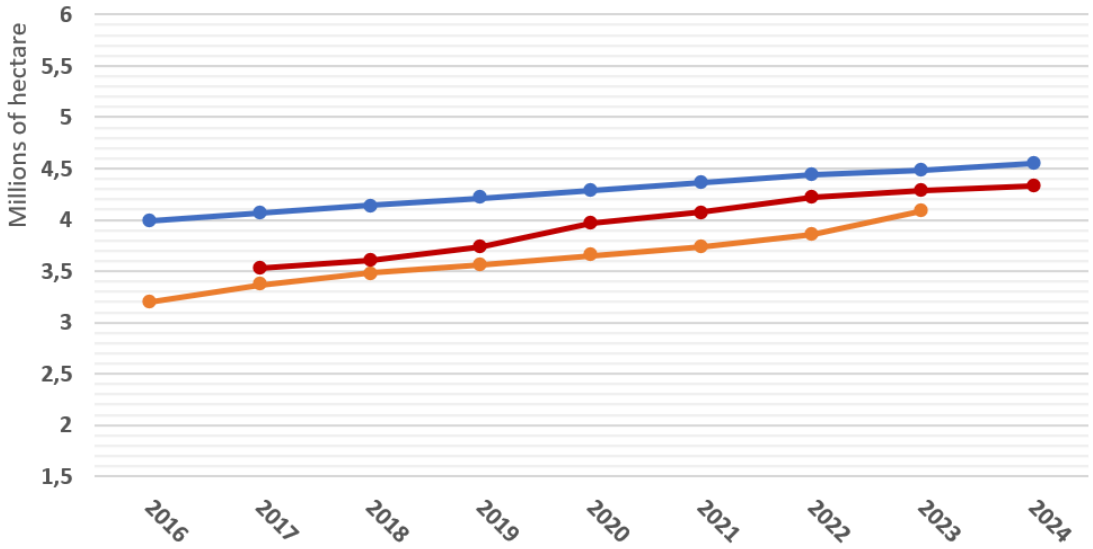

Figure 4. Total urban area in millions of hectares mapped by year in 10 m Collections 2, 3 and 30m collection 10

4.1 Accuracy Analysis

Following MapBiomas LULC validation strategy, the error assessment analysis was conducted using ~75,000 samples per year, labeled according to MapBiomas LULC classes by experts after the visual interpretation of Landsat data, MODIS-NDVI times series, and high-resolution imagery from Google Earth (when available). The accuracy analysis was based on Stehman (Stehman, 2014; Stehman; Foody, 2019) using the population error matrix and the global, user, and producer accuracies.

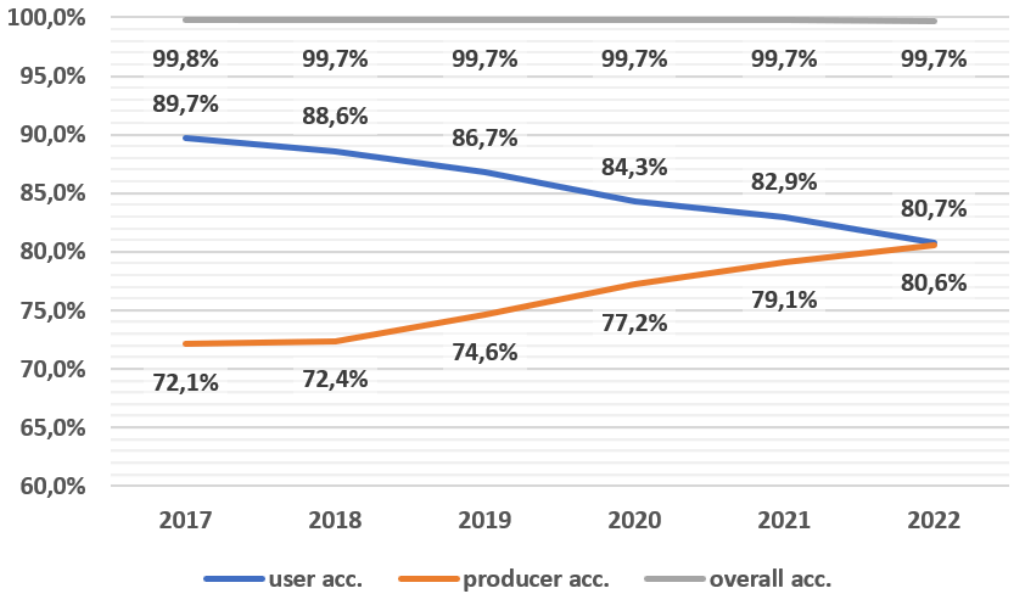

Figure 5. Accuracy results by year in collection 3

From 2017 to 2022, there is a gradual decrease in user's accuracy (from 89.7% to 80.7%), indicating an increase in commission errors. In practical terms, this means that a larger proportion of areas classified as urban are not actually urban. This pattern is expected for rare classes, where even small classification errors can significantly affect precision, often due to spectral confusion with similar targets such as bare soil or other anthropogenic surfaces.

In contrast, there is a consistent increase in producer's accuracy (from 72.1% to 80.6%), reflecting a reduction in omission errors. This suggests that the model is becoming more effective at detecting true urban areas over time, capturing a larger share of actual occurrences. For rare classes, this improvement is particularly important, as underdetection is typically one of the main challenges.

The overall accuracy remains extremely high and stable (~99.7%), but this metric is heavily influenced by the dominance of non-urban classes across the territory. Therefore, it does not adequately represent the performance for the urban class, reinforcing the importance of class-specific accuracy metrics.

4.2 Comparison with reference maps

MapBiomas Collection 3 were compared Brazil Urbanized Areas (2019) produced by IBGE, Instituto Brasileiro de Geografia e Estatística (IBGE, 2022). This map is a visual interpretation of urban features, identified according to the elements of specific shape (geometry of objects) and pattern (spatial arrangement), using Sentinel 2 imagery, with spatial resolution of 10m, supplemented by higher-resolution data where necessary. It is available in shapefile format at IBGE's website. The mapped urban land use types include: "Urbanized Area," categorized into two classes — high density and low density —, "Other Urban Facilities," and "Vacant Urbanized Areas."

The comparison between Collection 3 and the Urbanized Areas (2019) product shows varying spatial distributions of urban mapping across Brazilian biomes, with overlap values ranging from 53.9% in the Caatinga to 76.6% in the Pantanal, and 63.3% at the national level. In all biomes, the proportion of areas mapped only by the Urbanized Areas product is consistently higher (20.0~37.1%) than those mapped exclusively by Collection 3 (2.1~8.9%), indicating systematic differences in the spatial extent identified by each dataset. The Pantanal stands out with the highest overlap (76.6%) and the lowest share of areas mapped only by the Urbanized Areas product (20.0%). This pattern is consistent with the Pantanal being the biome with the lowest urban coverage in Brazil, where the limited spatial extent of urban areas results in a greater concentration of agreement between the datasets. In contrast, biomes such as the Caatinga and Amazon show lower overlap and higher proportions of non-coincident areas, reflecting a more distributed pattern of mapped urban extents between the two products.

Table 2. Comparison of Urban Area Mapping between MapBiomas Collection 3 and the Urbanized Areas (IBGE, 2022) for the Year 2019.

Biome	Only Mapped by Collection3	Only Mapped by Urbanized Areas (2019)	Overlap
Amazon	5,7%	32,2%	62,2%
Caatinga	8,9%	37,1%	53,9%
Cerrado	7,9%	24,3%	67,8%
Atlantic Forest	4,3%	31,6%	64,1%
Pampa	2,1%	31,4%	66,5%
Pantanal	3,5%	20,0%	76,6%
Brazil	**5,7%**	**30,9%**	**63,3%**

5. References

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., & Kohli, P. (2025)**. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data**. arXiv. https://arxiv.org/abs/2507.22291.

IBGE. Censo Demográfico 2010 – Aglomerados Subnormais. Rio de Janeiro, RJ: IBGE, 2020.

IBGE, Coordenação de Meio Ambiente (org.). **Áreas urbanizadas do Brasil: 2019.** Rio de Janeiro, RJ: IBGE, 2022.

IBGE. **Malha de Setores Censitários**. Brasil: 2020. Malha censitária.

JUSTINIANO, Eduardo Felix 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, p. 102791, 2022.

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

STEHMAN, Stephen V. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. **International Journal of Remote Sensing**, v. 35, n. 13, p. 4923–4939, 2014.

STEHMAN, Stephen V.; FOODY, Giles M. Key issues in rigorous accuracy assessment of land cover products. **Remote Sensing of Environment**, v. 231, p. 111199, 2019.