MAPBIOMAS INDONESIA

Algorithm Theoretical Basis Document (ATBD)

Collection 4.1

Version 1

February 2026

Executive Summary

The acceleration of land use change in Indonesia significantly contributes to global carbon emissions, with approximately 86 percent of the country's total emissions linked to land-use change [1]. Understanding the dynamics of land use is critical for balancing economic development and climate change mitigation. Reliable data on drivers and impacts of land change underpin effective policies and strategies for sustainable natural resource management.

Established in 2015 by a consortium of universities, civil society organizations, and technology companies in Brazil, the MapBiomas initiative addresses this need by providing a rapid, collaborative, and cost-efficient approach to generating annual land cover and land use maps from large-scale datasets. Its primary goal is to strengthen local capacity in tropical countries for producing accurate, context-specific remote sensing data to inform decision-making across government, civil society, and the private sector.

By 2026, MapBiomas operates in 17 countries with more than 130 institutions and 500 contributors. In Indonesia, since 2019, Auriga Nusantara and Woods and Wayside International have coordinated the initiative with nine civil society organizations and ten local universities to conduct accuracy assessments.

MapBiomas Indonesia has released five successive map collections. Collection 1 (2000–2019) classified 10 land cover and land use classes, while Collection 2 (2000–2022) introduced rice fields and implemented annual accuracy assessments. Collection 3 (2000–2023) added peat swamp forests and urban areas, and Collection 4 (1990–2024) extended the historical baseline. Collection 4.1 (1990–2024) refines historical data consistency and the representation of transitions from natural to anthropogenic classes, integrating GEDI data and revised transition rules, particularly for natural non-forest vegetation, now considered anthropogenic due to historical human-induced transformations.

This Algorithm Theoretical Basis Document, or ATBD, describes the methodological framework applied in MapBiomas Indonesia Collection 4.1, including datasets and analytical procedures. All datasets, classification outputs, code, statistics, and supplementary analyses are publicly accessible through the MapBiomas Indonesia Platform at plataforma.landy.mapbiomas.id, while the classification algorithms are openly available through the official GitHub repository at https://github.com/mapbiomas-indonesia.

1. Introduction

1.1. Scope

This document presents the theoretical foundation, objectives, and methodological framework applied in the production of annual land cover and land use maps for Indonesia covering the period 1990 to 2024 under MapBiomas Indonesia Collection 4.1. It provides a detailed description of the image processing architecture, classification procedures, the integration approach between basic themes and cross-cut themes, and the accuracy assessment methodology employed. More specific explanations regarding the procedures applied to thematic classes are provided in the annex.

1.2. Overview

The development of each MapBiomas Collection has been supported by several key factors: first, the Google Earth Engine platform, which provides data access, image processing capabilities, standardized algorithms, cloud-based computation, and open access to Landsat time series data; second, the MapBiomas collaborative network, which facilitates knowledge exchange and methodological development among participating institutions and experts; and third, financial support from funding institutions, which has played a critical role in sustaining and advancing the initiative.

Since its initial development, land cover and land use mapping within MapBiomas Indonesia has undergone continuous annual refinement and improvement. These developments are organized into successive versions referred to as "collections." The annual maps are generated using the Landsat satellite image archive available through the Google Earth Engine platform, encompassing the period from 1990 to the present. A comparative summary of the different collections is presented in Table 1.

Table 1. Evolution of MapBiomas Indonesia Collections

Collection	Period	Number of Classes	Method	Overall Accuracy*	Release Date
Collection 1	2000–2019 (20 years)	2 levels / 11 classes	Random Forest + U-Net (Aquaculture)	-	November 2021

Collection 2	2000–2022 (23 years)	2 levels / 11 classes	Random Forest	Level 1: Gb 77.2%; All 17.5%; Ar 5.3% Level 2: Gb 74.5%; All 18.5%; Ar 7.0%	October 2022
Collection 3	2000–2023 (24 years)	2 levels / 13 classes	Random Forest + DNN (Urban Areas)	-	December 2023
Collection 4	1990–2024 (35 years)	2 levels / 13 classes	Random Forest + U-Net (Oil Palm and Pulpwood Plantations)	-	August 2025
Collection 4.1	1990–2024 (35 years)	2 levels / 13 classes	Random Forest + U-Net (Oil Palm and Pulpwood Plantations)	-	February 2026

**Gb = global accuracy; Ar = area disagreement; and All = allocation disagreement.*

MapBiomas Indonesia Collection 4.1 produces several main products.

- The most important product is the annual land cover and land use maps. These maps cover 13 classes and are available for every year from 1990 to 2024. This allows users to see changes over a long period of time.
- The collection also provides Landsat image mosaics and the input variables used for classification. These are created from Landsat 5, 7, 8, and 9 images after preprocessing. The processing system and algorithms are also shared, including Google Earth Engine scripts and source code, so users can understand and reproduce the method.
- In addition, the collection offers analysis products. These include statistics and spatial analysis of land cover change. The results can be linked to administrative boundaries, river basins, protected areas, and other land use categories.
- A mosaic quality layer is also available. This layer shows how many cloud-free observations are available for each pixel every year. It helps users evaluate the reliability of the data in different locations and years.
- Other products include annual and cumulative deforestation maps, as well as maps of secondary vegetation and its age, which are currently provided in a beta version.

- Besides land cover and land use products, MapBiomas Indonesia also develops thematic datasets. For example, MapBiomas Fire Collection 2 provides annual and monthly burned area maps from 2000 to 2024.

1.3. Region of Interest

The name MapBiomas comes from two words, “map” and "biome." A biome is a large geographic area defined by its vegetation characteristics, which are closely related to geomorphology and climate. In countries such as Brazil, biomes are clearly defined and used as the main reference for mapping. Indonesia, however, is an archipelagic country with different geographic characteristics. For this reason, MapBiomas Indonesia does not use a biome-based approach. Instead, it produces annual land cover and land use maps that cover the entire territory of Indonesia without dividing the country by biome.

The working area is defined at the level of major islands or island groups. Mapping is carried out separately for each region, and the results are later integrated into a single national map through a post-processing stage. This approach ensures national consistency while maintaining good analytical detail at the regional level.

Technically, the country is divided into seven regions: (1) Sumatra, (2) Java and Bali, (3) Nusa Tenggara, (4) Kalimantan, (5) Sulawesi, (6) Maluku, and (7) Papua. These seven regions are further divided into 35 subregions. This subdivision improves classification accuracy since each subregion has different landscape characteristics, vegetation types, and land use dynamics.

In addition to the regional approach, mapping is also supported by cross-cut thematic approaches to improve classification quality. These themes include coastal and wetland areas, such as mangroves, aquaculture ponds, and peat swamp forests. Other major land use sectors are also addressed, including agriculture, for example, rice fields, oil palm plantations, and timber plantations, as well as mining areas and settlements.

MapBiomas Indonesia uses administrative boundary maps published by the Geospatial Information Agency at a scale of 1:250,000 as the basis for defining working regions. To ensure that coastal areas are fully represented, a 2 km buffer is applied seaward from the coastline. This ensures that coastal zones and related ecosystems are not cut off during the mapping process.

Table 2. Land Cover and Land Use Characteristics by Island

Region	Main Biophysical Characteristics	Land Use Characteristics	Mainland Change Pressures	Implications for Mapping
Sumatra	Lowland tropical rainforest, montane forest, and extensive peat swamp forest in the eastern part	Oil palm plantations and industrial timber plantations dominate lowlands, along with dryland agriculture and rice fields	Large-scale plantation expansion, peat degradation, and land fires	High transition dynamics from natural forest to anthropogenic classes, and potential spectral confusion between secondary natural vegetation and young plantations
Kalimantan	Extensive lowland forest cover, swamp forest, and peat swamp forest	Oil palm plantations, industrial timber plantations, and mining, especially in lowland and coastal areas	Plantation expansion, coal mining, and infrastructure development	Need to distinguish natural forest, plantation forest, and degraded vegetation, and ensure temporal consistency in areas with gradual clearing
Sulawesi	Complex topography dominated by mountains with fragmented forests	Mixed dryland agriculture, cocoa and coconut plantations, and nickel mining	Smallholder agricultural expansion and mining	Landscape fragmentation increases classification complexity and requires a region-based approach to improve accuracy
Maluku	Relatively intact tropical forest on large islands, with hilly topography	Small-scale agriculture such as coconut, nutmeg, and clove	Local-scale plantation and mining expansion	Forest cover relatively stable, but traditional agricultural mosaics may create spectral variability
Papua	Very extensive natural forest cover, including lowland, swamp, and montane forests	Limited land use, with local agriculture and plantation expansion in some areas	Plantation expansion, infrastructure development, and new land clearing	It's an important reference area for natural forest classes, and requires attention to early change detection

Java Bali	Highly transformed landscape, with remaining forests in mountainous areas	Dominated by rice fields, intensive agriculture, and settlements	Urbanization, agricultural intensification, and infrastructure development	High complexity due to land fragmentation and small patch size, requiring consistent temporal resolution
Nusa Tenggara	Dominated by savanna, shrubland, and seasonal forest under dry climate	Dryland agriculture and grazing land	Climate variability, fire, and grazing pressure	High seasonal dynamics affect vegetation spectral response, making temporal metrics important for classification

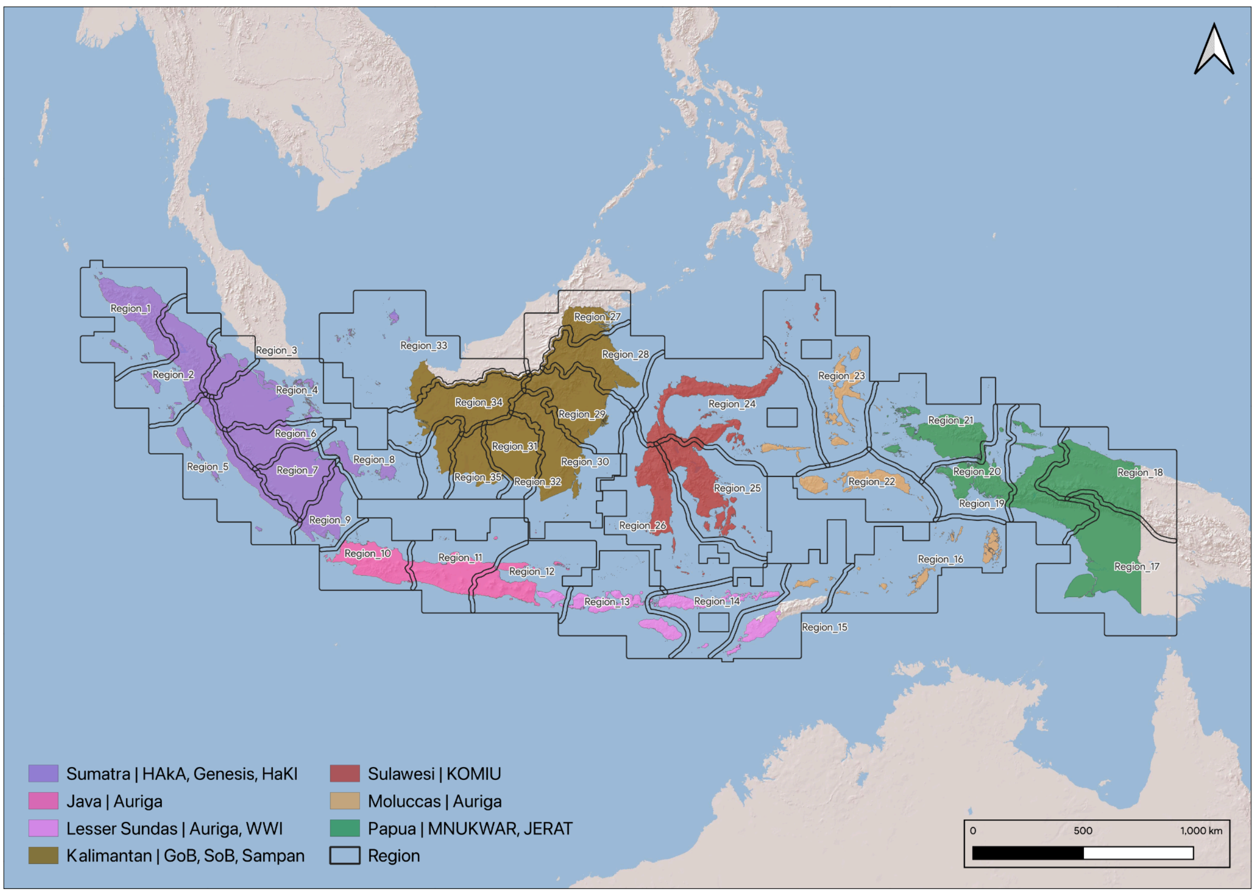


Figure 1. Territorial Coverage and Network of MapBiomas Indonesia

1.4. Key Application

MapBiomas Indonesia is designed as a platform to monitor land cover and land use dynamics. It can be used to:

- Map and measure transitions between land cover and land-use classes over time
- Calculate forest loss and forest gain
- Monitor water resources and their interaction with land cover and land use classes
- Track the expansion of agricultural land and industrial plantations
- Monitor infrastructure development
- Support monitoring of protected areas
- Monitor land-based commodity concessions and permits
- Provide data for spatial and regional land use planning

2. Basic Information and Background

2.1. MapBiomas Indonesia Network

MapBiomas is a multi-institution network composed of universities, civil society organizations, and technology developers, each with specific roles in developing land cover and land use maps in different regions. MapBiomas Indonesia was initiated by a civil society network in Indonesia, coordinated by Auriga Nusantara, with institutional support from Woods and Wayside International and technical and technological support from MapBiomas Brazil.

The development of each MapBiomas Collection has been supported by several key factors: first, the Google Earth Engine platform, which provides data access, image processing capabilities, standardized algorithms, cloud-based computation, and open access to Landsat time series data; second, the MapBiomas collaborative network, which facilitates knowledge exchange and methodological development among participating institutions and experts; and third, financial support from funding institutions, which has played a critical role in sustaining and advancing the initiative.

In Indonesia, nine civil society organizations contribute and play roles in regional technical coordination:

Sumatra

- Hutan, Alam dan Lingkungan Aceh, HAkA, covering Northern Sumatra
- Hutan Kita Institute, HAKI, covering Central Sumatra
- Genesis, covering Southern Sumatra

Kalimantan

- Sampan, covering West Kalimantan
- Save Our Borneo, SOB, covering Central and South Kalimantan
- Green of Borneo, GoB, covering East and North Kalimantan

Sulawesi

- Kompas Peduli Hutan, KOMIU, covering Sulawesi

Papua

- MNUKWAR, covering West Papua
- JERAT Papua, covering Papua

Java, Bali, Nusa Tenggara, Maluku, and cross-cut themes

- Auriga Nusantara
- Woods and Wayside International

Since Collection 2, the MapBiomas Indonesia network has also involved ten local universities in the accuracy assessment process:

- Universitas Syiah Kuala, Aceh
- Universitas Bengkulu
- Universitas Lampung
- Universitas Muhammadiyah Palangkaraya, Central Kalimantan
- Universitas Mulawarman, East Kalimantan
- Universitas Tadulako, Central Sulawesi
- Universitas Papua
- Universitas Indonesia
- Universitas Gadjah Mada
- Institut Pertanian Bogor

MapBiomas Indonesia also has an independent Scientific Advisory Committee, or SAC, currently composed of

- Prof. Projo Danoedoro, Universitas Gadjah Mada
- Dr. Arief Darmawan, Universitas Lampung
- Dr. Nur Hygiawawti Rahayu, BAPPENAS
- Dr. Hasriani Muis, Universitas Tadulako
- Francina Frenshegty Kesaulija, Universitas Papua

2.2. Data Source

The imagery datasets used by MapBiomas Indonesia from Collection 1 to Collection 4 were derived from Landsat sensors, including Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and Operational Land Imager and Thermal Infrared Sensor (OLI TIRS). These sensors are on board the Landsat 5, Landsat 7, Landsat 8, and Landsat 9 satellites. All Landsat images have a spatial resolution of 30 meters and are accessible through the Google Earth Engine platform. The data are produced by NASA and USGS, and one example of the available collections is cataloged as ee.ImageCollection("LANDSAT/LT04/C02/T1_L2").

In Collection 1 and Collection 2, MapBiomas Indonesia used Top of Atmosphere, or TOA, reflectance data from USGS Landsat Collection 1 Tier 1. Starting from Collection 3, Landsat mosaics have been processed using USGS Landsat Collection 2 Tier 1 data.

Tier 1 datasets have undergone radiometric calibration and orthorectification using ground control points and digital elevation models. These processes ensure accurate pixel positioning and adequate atmospheric correction, resulting in more consistent and reliable data for long-term land cover and land-use time series analysis.

2.3. Google Earth Engine

Google Earth Engine, or GEE, is a cloud-based computing platform developed by Google for large-scale geospatial data analysis and visualization. Its primary advantage lies in the ability to process massive volumes of satellite imagery and geographic information without requiring complex local computing infrastructure, since all processing is performed within a cloud environment. Through JavaScript- and Python-based application programming interfaces, or APIs, GEE enables users to develop customized analytical workflows and applications. Its visualization tools also provide a robust environment for exploring and interpreting spatial data.

The MapBiomas initiative uses Google Earth Engine as its main technological platform for remote sensing data processing and analysis. With access to both historical and near-real-time satellite archives, the platform supports detailed temporal analysis of land cover and land-use dynamics. The cloud-based computing system allows efficient management of large datasets and facilitates the production of high-resolution maps that consistently represent landscape changes over time.

In addition, the MapBiomas team has developed specific classification algorithms that run automatically through the Google Earth Engine API to identify and map different land cover classes. This automated workflow enables regular map updates with high levels of consistency and accuracy. The resulting maps can be visualized interactively and shared with the public, reinforcing the principles of transparency, collaboration, and open data that underpin the implementation of MapBiomas.

3. MapBiomas Indonesia Method

3.1. Method Description

The MapBiomas Collection 4.1 land cover and land use maps were produced through several main stages, as illustrated in Figure 2. The first stage involved the generation of annual Landsat image mosaics using predefined time windows designed to optimize spectral contrast. This approach improves the separability of land cover and land use classes according to regional characteristics and thematic priorities.

The next stage consisted of building the input variables, or feature space, derived from multiple spectral attributes of Landsat bands. These attributes were used to train annual random forest classification models based on specifically collected training samples, adjusted to data availability and statistical requirements. This process resulted in annual land cover and land use maps covering the entire study area. For specific classes such as oil palm plantations and timber plantations, identification was conducted using a deep learning U-Net model.

Following the initial classification, spatial and temporal filters were applied to reduce noise and improve interannual consistency. The classification outputs from different regions and thematic groups were then integrated hierarchically using expert-based priority rules. Additional filtering was conducted on the integrated map to generate the final product.

The final stage involved accuracy assessment using 12,957 independent validation samples per year for the period 2000 to 2022. The assessment followed international best practice standards and was used to quantify the reliability of the maps. In addition to the annual maps, statistical analyses were generated, including interclass transitions and their spatial distribution across administrative units such as provinces, regencies, and other territorial categories.

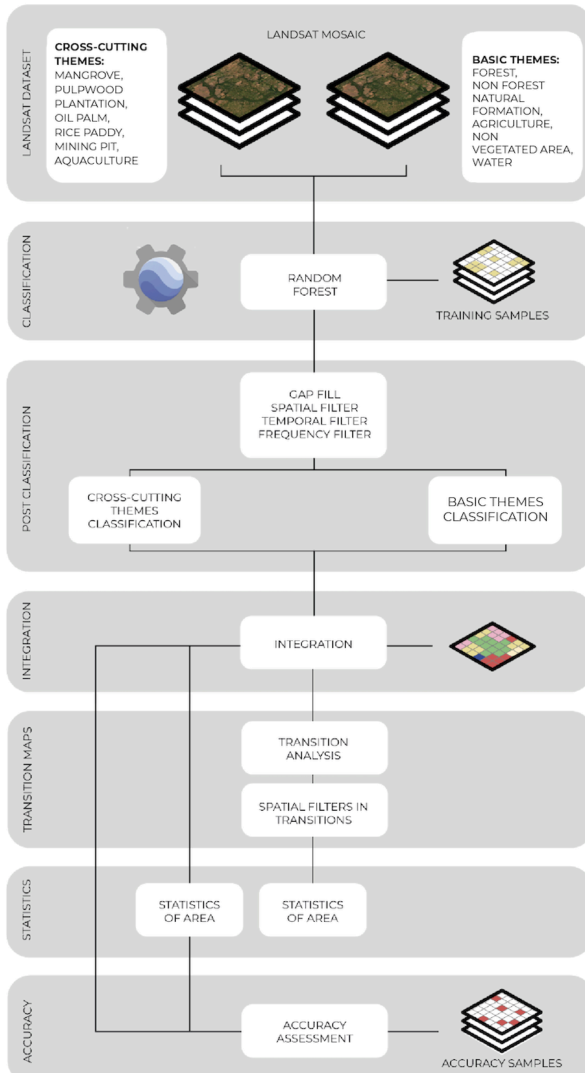


Figure 2. General workflow of the MapBiomas algorithm implementation in Google Earth Engine.

3.2. Landsat Mosaics

The Landsat mosaic is a compilation of annual Landsat imagery from 1990 to 2024 used as the primary basis for land cover mapping. The main objective of this process is to generate annual composite images that represent both dry and wet seasons. This approach enhances spectral contrast between forest and non-forest classes, reduces cloud contamination, and ensures that the most representative imagery is used for each year. As a result, the annual mosaics consistently capture land cover dynamics over time.

In the mosaic production process, the spatial reference unit is the Worldwide Reference System 2, WRS2. This grid system is used by Landsat satellites to organize image coverage according to path and row. Each path-row combination represents an area of approximately 183 by 170 kilometers, equivalent to about 31,110 square kilometers. The system is consistently applied across Landsat 5, 7, 8, and 9 sensors, enabling the integration of multisensor data into a harmonized time series.

The use of patch-row units as the basis for mosaic processing ensures spatial consistency across years and sensors. It also facilitates data management at national and regional scales. This framework provides a stable and reliable foundation for producing annual mosaics suitable for long-term land cover and land-use change analysis.

Table 3. Path-row numbers by region

Region	**path-row**	**Total of mosaic (35 years)**
Sumatra	51	1.785
Kalimantan	44	1.540
Sulawesi	21	735
Maluku	38	1.330
Papua	41	1.435
Jawa-Bali	19	665
Nusa Tenggara	23	805
Total	**237**	**8.295**

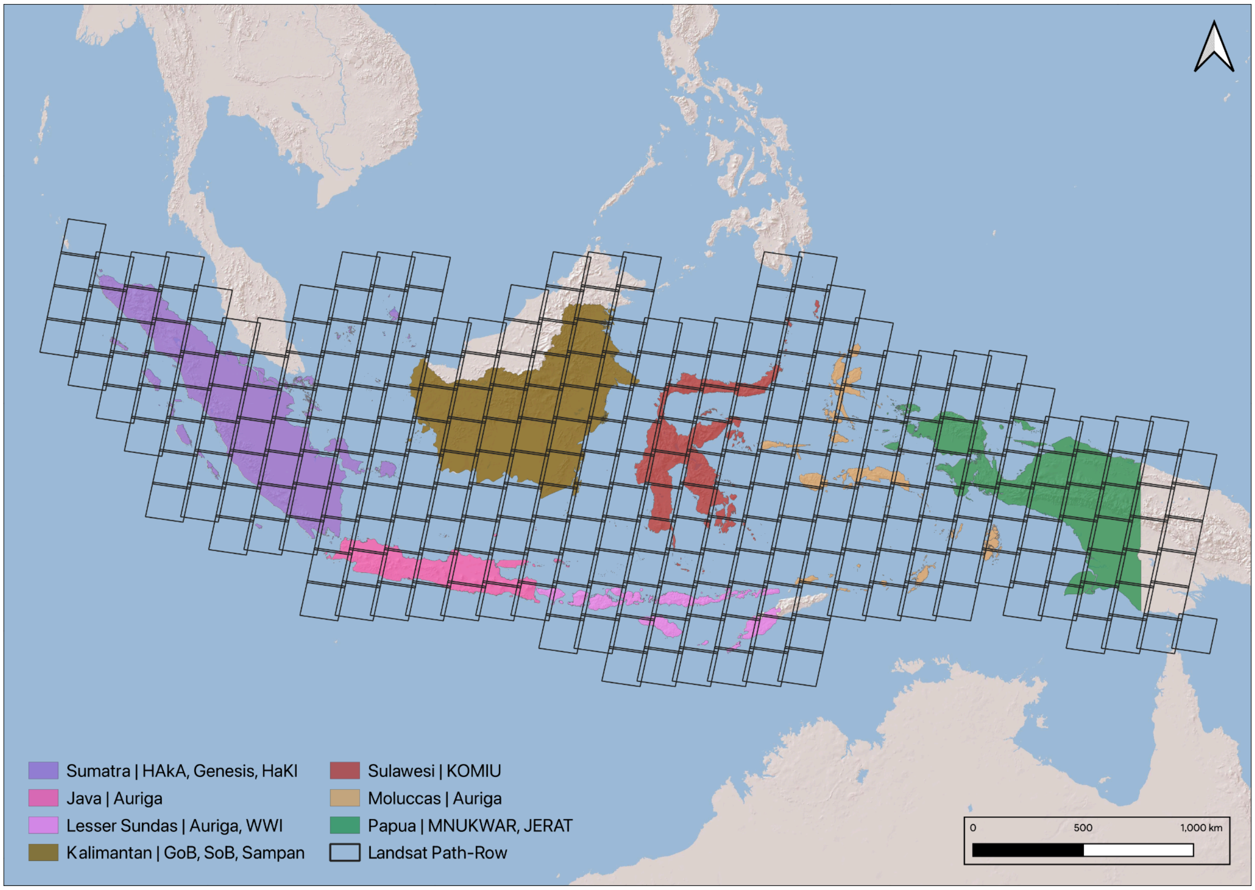


Figure 3. Landsat path row scene coverage used by MapBiomas Indonesia.

In the Landsat mosaic production process within MapBiomas, the median value is used to select the most representative pixel value from a set of images acquired within a given time period. For each pixel location, multiple values are usually available from different acquisition dates. Extremely high values are often associated with clouds, while very low values typically result from cloud shadows or atmospheric disturbances. The median is selected because it lies at the center of the value distribution and is therefore less sensitive to these extreme values.

Technically, the procedure follows several steps. First, multiple Landsat images are collected for a defined period, such as one year or a specific season. Second, pixels identified as clouds or shadows are removed using the Quality Assessment band. Third, for each spectral band and pixel location, the median of the remaining values is calculated. This median value is then assigned as the final pixel value in the mosaic.

To further reduce noise and improve mosaic quality, a dedicated tool was developed to evaluate each image individually and exclude non-informative scenes, for example, those with excessive cloud cover. This approach enhances mosaic quality by producing imagery that is relatively cloud-free, radiometrically stable, and temporally consistent. Such consistency is critical for land cover classification using algorithms such as random forest, since the quality of input data strongly influences classification accuracy.

3.3. Feature Space

The variables in the feature space are derived from the annual median mosaics and used as inputs for the classification process. All Landsat bands and their derived variables are compiled into a raster file containing a total of 156 bands. These components include the original Landsat spectral bands, multiple spectral indices, and fraction and texture information derived from these bands, as well as indices based on spectral fractions.

To generate annual pixel values, a set of statistical reducers is applied to the available image collection. These reducers capture both annual and seasonal spectral dynamics, thereby enriching the representation of land cover and land use characteristics.

The reducers include the following.

- Median, defined as the median value of all available pixels in the annual mosaic at a given location.
- Dry season median, calculated from pixels within the 25th percentile of NDVI values, representing the lowest NDVI conditions associated with the dry season.
- Wet season median, calculated from pixels within the 75th percentile of NDVI values, representing the highest NDVI conditions associated with the wet season.
- Amplitude, defined as the range between the maximum and minimum pixel values within the annual mosaic.
- Standard deviation, representing the dispersion of all available pixel values in the annual mosaic at a given location.
- Minimum: the lowest pixel value recorded in the annual mosaic at a specific location.
- Maximum: the highest pixel value recorded in the annual mosaic at a specific location.

The combination of original bands, spectral indices, fractions, textures, topographic variables, and multiple statistical reducers forms a multidimensional feature space. This structure enables the classification algorithm to distinguish land cover and land use classes with greater precision in annual time series analyses, consistent with established practices in Landsat-based mapping.

Table 4. List, Description, and References of Bands, Fractions, and Indices in the Feature Space.

	Name	Formula	Statistical Reducer			
			median	**median_dry**	**median_wet**	**stdDev**
Band	blue	B1 (L5-L7); B2 (L8)				
	green	B1 (L5-L7); B3 (L8)				
	red	B1 (L5-L7); B4 (L8)				
	nir	B1 (L5-L7); B5 (L8)				
	swir1	B1 (L5-L7); B6 (L8)				
	swir2	B7 (L5); B8 (L7); B7 (L8)				

Index	ndvi	(nir - red)/(nir+red)				
	evi	2.5 * (((nir - red)) / ((nir + (6 * red)) - (7.5 * blue) + 1))				
	evi2	(2.5*(nir-red)/(nir+2.4*red+1)				
	cai	(swir2/swir1)				
	ndwi	(nir-swir1)/(nir+swir1)				
	gcvi	(nir/green-1)				
	pri	(blue-green/(blue+green)				
	savi	(1+L)*(nir-red)/(nir+red+0,5)				
	ndsi	(swir1 - nir)/ (swir1 + nir)				
	mmri	((green - swir1)/ (green + swir1)) - abs((nir - red)/(nir + red))) / (abs((green - swir1)/ (green + swir1)) + abs((nir - red)/(nir + red)))				
	ndbi	(swir1 - nir)/(swir1 + nir)				
	nbr	(nir-swir2)/(nir+swir2)				
	tgi	(120*(red-blue))-(190*(red-green))				
	total nitrate	('2.71828**(a - b * (c / (green + red + blue)))				
	nir_entropy					
	green_entropy					
	nir_contrast					
	nir_correlation					
Fraction	gv	*Fractional abundance of green vegetation within the pixel*				
	npv	*Fractional abundance of non-photosynthetic vegetation within the pixel*				
	soil	*Fractional abundance of soil within the pixel*				
	cloud	*Fractional abundance of cloud within the pixel*				
	shade	100-(gv+npv+soil+cloud)				
MEM Index	gvs	gv/(gv+npv+soil+could)				
	ndfi	(gvs-(npv+soil))/(gvs+(npv+soil))				
	sefi	((gv+npv_s-soil)/(gv+npv_s+soil)				
	wefi	((gv+npv)-(soil+shade))/((gv+npv)+(soil+shade))				

	fns	((gn+shade)-soil)/((gv+shade)+soil)				
Slope		ALOS DSM: GLobal 30m				

3.4. Classification

3.4.1. Classification System

Land cover refers to the physical materials and objects that occupy the surface of the Earth, such as vegetation, water, soil, or built-up areas. Land use, in contrast, relates to human activities and the way land is managed or utilized for specific purposes. It reflects anthropogenic interactions with the land surface.

A classification system is a conceptual framework that includes class names, codes, definitions, and clear diagnostic criteria to distinguish different types of land cover and land use. It also defines the relationships among classes, forming an organized and internally consistent structure.

Through such a classification system, the complex and heterogeneous conditions observed in the field are simplified into a set of well-defined classes. This simplification allows the diversity of land cover and land use to be represented accurately while remaining operational for mapping and spatial analysis.

3.4.2. Legend

MapBiomas Indonesia adopts the land cover and land use classification concept developed by the Food and Agriculture Organization of the United Nations. This framework applies a hierarchical structure to accommodate multiple levels of categorization. The structure begins with broad, general classes, which are then systematically subdivided into more detailed subclasses. This hierarchical approach enables flexibility in analysis, allowing users to work at aggregated or more detailed levels depending on their objectives.

Within this system, categories can be defined based on one or more characteristics, including spectral response and image texture derived from remote sensing data. Each class is distinguished using consistent and measurable diagnostic criteria, ensuring clarity and reproducibility in the classification process.

The MapBiomas Indonesia Collection 3 classification system consists of two category levels. Level 1 includes five main classes, namely, forest, non-forest natural vegetation, agriculture, non-vegetation, and water. Level 2 comprises 13 classes derived from these five primary categories, providing a more detailed representation of land cover and land use variability across Indonesia.

Table 5. Land Cover and Land Use Classes of MapBiomas Indonesia Collection 4.1

Natural/Anthropic	EN	ID	pixel id	Hexadecimal code	Color	Level
	1. Forest	**1. Hutan**	**1**	#1f8d49		**1**
Natural	1.1. Forest Formation	1.1. Formasi Hutan	3	#1f8d49		2
Natural	1.2. Mangrove	1.2. Mangrove	5	#04381d		2
Natural	1.3 Peat swamp forest	1.3 Hutan rawa gambut	76	#2f7360		2
	2. Non-Forest Natural Formation	**2. Tumbuhan Non-Hutan**	**10**	#d6bc74		**1**
Natural	2.1. Other Natural Vegetation	2.1. Tumbuhan Non-Hutan Lainnya	13	#d89f5c		2
	3. Agriculture	**3. Pertanian**	**18**	#E974ED		**1**
Anthropic	3.1. Rice Paddy	3.1. Sawah	40	#c71585		2
Anthropic	3.2. Oil Palm	3.2. Sawit	35	#9065d0		2
Anthropic	3.3. Pulpwood Plantation	3.3. Kebun Kayu	9	#7a5900		2
Anthropic	3.4. Other Agriculture	3.4. Pertanian Lainnya	21	#ffefc3		2
	4. Non-Vegetated Area	**4. Non-Vegetasi**	**22**	#d4271e		**1**
Anthropic	4.1. Mining Pit	4.1. Lubang Tambang	30	#9c0027		2
Anthropic	4.2. Urban Area	4.2. Permukiman	24	#d4271e		2
Natural	4.3. Other Non-Vegetation	4.3. Non-Vegetasi Lainnya	25	#db4d4f		2
	5. Water Body	**5. Tubuh Air**	**26**	#2532e4		**1**
Anthropic	5.1. Aquaculture	5.1. Tambak	31	#091077		2
Natural	5.2. River, Lake, Ocean	5.2. Sungai, Danau, Laut	33	#2532e4		2
Not defined	**6. Not observed**	**6. Citra Tertutup Awan**	**27**	#ffffff		1

Collection 4.1 divides land cover and land use into 13 classes. The structure consists of five basic theme classes, namely forest formations, non forest vegetation, other agriculture, other non vegetation, and rivers, lakes, and sea. The remaining classes are defined as cross cut themes, including mangrove, peat swamp forest, paddy field, oil palm, timber plantation, mining area, settlement, and aquaculture pond.

Basic theme classes are generally dominated by natural characteristics or represent broad structural categories at a higher hierarchical level. Mapping of these basic classes is conducted on a regional basis, covering Sumatra, Kalimantan, Sulawesi, Papua, and the combined region of Java, Bali, Nusa Tenggara, and Maluku. This regional approach accounts for differences in biophysical conditions and landscape dynamics across Indonesia.

In contrast, cross cut theme classes represent land cover and land use types that are typically dynamic, anthropogenic in nature, or require specific approaches for identification and mapping. These classes often exhibit more complex spectral patterns or rapid changes driven by human activities. The mapping stages for these thematic classes are described in greater detail in the appendix, including the methodology, supporting variables, and classification strategies applied.

3.4.3. Training Sample

The land cover classification process begins with a clear understanding of the spectral variability associated with each class. This step is essential to ensure that training samples are collected appropriately. The samples consist of pixels that remain stable within the same class throughout the time series of the previous collection, referred to as stable samples. Their spectral response consistently represents the same land cover type over the entire period.

The algorithm first identifies stable pixels within the selected area. From this layer, random points are generated and spatially balanced according to the area of each class. The number of points per class is defined by the interpreter, and the location of each point is used as input to train the random forest classifier.

Optionally, the interpreter may exclude classes considered temporally unstable, as they may introduce noise in the initial classification. There is also an option to include specific classes in the allocation of points on stable pixels or to manually adjust the samples by removing or adding locations based on an evaluation of their temporal stability. These additional points are referred to as complementary samples.

Stability assessment is conducted by comparing the spectral response of the same pixel across all mosaics in the time series for which data are available. This multitemporal evaluation provides a basis for selecting complementary samples using a temporal window approach. As a result, the training dataset can be refined, leading to improved classification accuracy. The entire process is implemented using the geometry creation tools available in Google Earth Engine.

3.4.4. Random Forest

Random Forest [23] is the machine learning algorithm selected by MapBiomas Indonesia as the classifier for basic theme classes. It is a supervised learning method that constructs an ensemble of decision trees to learn patterns from labeled training samples. The algorithm aggregates the predictions of multiple trees and assigns the final class based on majority voting.

Within Google Earth Engine, Random Forest is implemented as an integrated machine learning algorithm for geospatial analysis. It operates by building a set of decision trees derived from the training dataset. At each node of a tree, a random subset of predictor variables is selected, and the optimal split is determined to best separate the classes. Each tree produces one classification vote for a given element. The final class is defined by the majority vote across all trees. In the context of satellite imagery, each pixel is assigned to the class that receives the highest number of votes.

This ensemble learning approach is known for its robustness and high accuracy, particularly in conditions characterized by substantial variability and data noise. The use of multiple trees reduces overfitting and improves generalization performance compared to a single decision tree.

Proper configuration requires the user to define key parameters, including the number of trees, referred to as "nTrees"; the number of randomly selected variables considered at each node, referred to as "mTry"; and the set of predictor variables used as input features. In Collection 4.1, the number of trees per region ranges from 50 to 100. The predictor variables are derived from annual mosaics and from the training data generated during the sample identification process.

3.4.5. Classification

MapBiomas Indonesia applies a supervised classification approach to generate land cover and land use data. Supervised classification is a technique that partitions the spectral domain into regions that can be associated with specific land cover classes according to the objectives of the application [21]. This approach requires previously labeled reference samples to train the classifier, which is then used to assign classes to data with unknown labels [22].

The samples are analyzed using the Random Forest algorithm, considering variables defined as spectral parameters. The classification process is executed on the Google Earth Engine platform and is performed annually for each region. The resulting classifications are subsequently reviewed and evaluated, including adjustments to the samples within each class, until the output is considered the most representative final classification.

In Collection 4.1, Random Forest is also applied to cross-cut theme classes such as mangrove, peat swamp forest, paddy field, settlement, mining area, and aquaculture pond. For timber plantation and oil palm classes, a deep learning-based U-Net classification method is used. A more detailed explanation of the U-Net method is provided in the appendix.

3.4.6. Post-Classification

Post-classification procedures are applied to land cover and land use classification results to correct remaining errors. Because the method is pixel-based and covers a long temporal period, the outputs may contain spatial and temporal biases. These errors are commonly associated with residual cloud shadows, missing image sections due to gaps in Landsat mosaics, and small scattered pixels.

To improve year-to-year consistency in land cover and land use classification, a series of post-classification filters is implemented.

- Gap fill is used to replace missing data by incorporating information from the closest available time period.
- Temporal filtering is applied to ensure logical class transitions between consecutive years, reducing implausible temporal changes.

- Frequency filtering evaluates the dominance of a class over a defined period, helping to correct unstable or inconsistent classifications.
- Spatial filtering aims to reduce isolated pixels and smooth the spatial pattern of classes on the map.

All post-classification processes are implemented using Google Earth Engine.

3.4.6.1. Gap Fill

Gap fill is applied to replace missing data in classification outputs. In long time series, certain years may contain data gaps due to persistent cloud cover. These gaps are filled using classification results from the closest available year. The system first searches for the nearest previous year. If no valid data are available, it uses the closest subsequent year in temporal order.

The primary objective of the Gap Fill filter is to address empty areas in mosaics that result from the absence of valid information within a given time range. This situation occurs when no suitable imagery is available for a particular year or region. In Landsat mosaics, there are periods when several consecutive years lack usable imagery or when empty areas appear due to cloud masking during image preprocessing.

To manage these conditions, the algorithm detects areas without information in the mosaic. Based on the available time series, each gap is filled using data from the nearest preceding year. If this is not possible, the algorithm uses data from the nearest subsequent year. The procedure continues until valid data are identified and assigned.

Before applying this filter, the time series from previous years must be carefully reviewed. The algorithm may fill gaps with data that do not accurately represent actual conditions in the study area, potentially introducing misclassification.

Although the gap-fill filter reduces the impact of missing data and improves temporal continuity, it has inherent limitations. Real land cover changes that occurred during the filled year cannot be detected. Such changes will only become visible in the following year when valid data are available. As a result, change detection analyses may be affected in areas where gap filling has been applied.

3.4.6.2. Spatial filter

The spatial filter is applied to improve classification quality by correcting isolated pixels or pixels that differ from the surrounding group. This procedure reduces errors such as residual shadows or small remnants of other land cover classes that are not consistent with their spatial context. It also helps clarify and simplify specific details in the mosaic, resulting in a more coherent spatial pattern.

In its implementation, a minimum mapping unit of 0.5 hectares is used, equivalent to five connected pixels. Pixels must be spatially connected to one another to be considered a valid group representing a class. If a pixel does not meet the connectivity criterion, it is treated as isolated and subsequently reassigned according to the dominant class in its neighborhood.

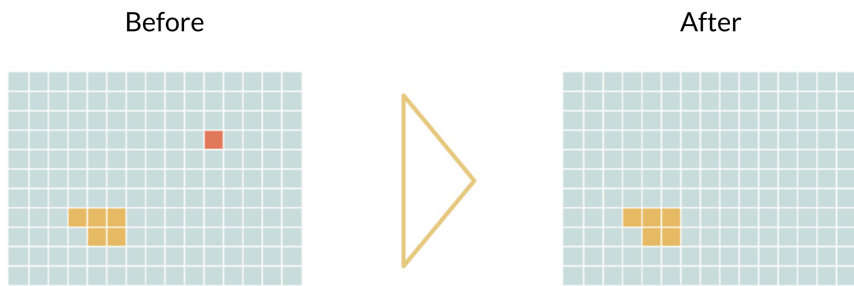


Figure 4. Illustration of the spatial filter mechanism

For example, in the initial image, there is one isolated pixel. After the spatial filter is applied, that pixel is replaced by the most dominant class in its surrounding area because it is not connected to any larger group of pixels. As a result, the pixel changes class to match the dominant class in its neighborhood.

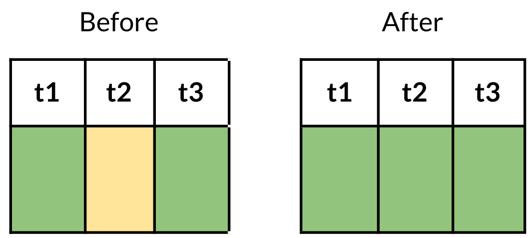
3.4.6.3. Temporal Filter

Temporal filtering is applied to correct pixels identified as noise due to inconsistencies in land cover classification across a time series. Such inconsistencies generally arise when gradual real-world changes are represented as abrupt shifts in annual classifications or when limitations and distortions in image mosaics affect classification results in specific years. The temporal filter consists of three types, designed for middle years, the first year, and the last year in the series.

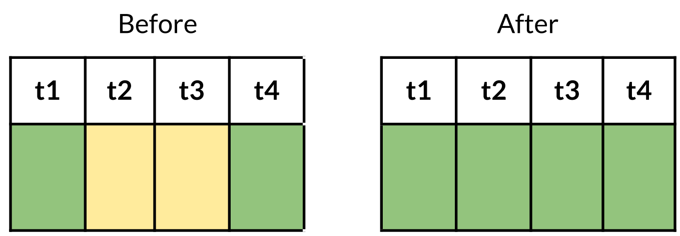
The temporal filter for middle years operates by comparing pixel values at the beginning and end of a defined time window and then evaluating the consistency of the class in the intervening year or years, for example, within the 1990 to 2024 period. Under a 3-year rule, if a pixel is classified as forest in year t1 and year t3 but as agriculture in year t2, this pattern is considered temporally inconsistent. The classification at t2 is then adjusted to match the surrounding stable class, ensuring temporal coherence. The same principle applies to 4-year and 5-year windows, where stability is assessed over longer sequences to reinforce consistent trajectories.

This rule is not applied to the first and last years of the time series, for example, 1990 and 2024, because there is no data available on one side for comparison. In these cases, a specific approach is used that relies only on the nearest available year.

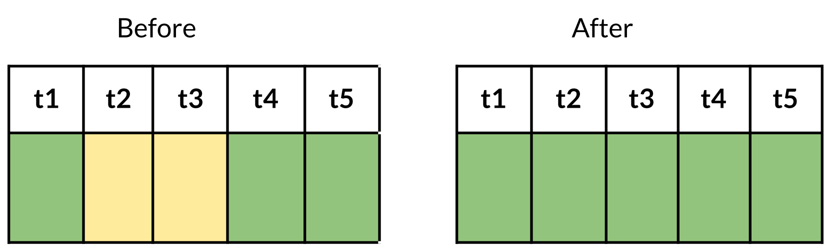
Although temporal filtering improves the stability of the time series and reduces unrealistic fluctuations, it has limitations. Genuine short-term land cover changes may be interpreted as noise and consequently removed. As a result, rapid or temporary land cover dynamics may not be fully captured in the final dataset.



Temporal filter - 3 years



Temporal filter - 4 years



Temporal filter - 5 years

Figure 5. Illustration of the temporal filter mechanism

The first-year filter is applied to correct inconsistencies caused by missing or incomplete information in 1990. In this context, certain transitions observed in subsequent years are considered unrealistic for the evaluation area. For example, a change from agriculture to forest immediately after 1990 may be treated as implausible and corrected to maintain logical temporal continuity.

The last-year filter addresses misclassified pixels at the end of the time series, in this case, 2024. A pixel initially classified as forest may be corrected to agriculture if the preceding years consistently indicate agricultural land use. The objective is to ensure that the final year does not introduce abrupt, unsupported changes that disrupt the overall temporal pattern.

These filters are flexible and allow the evaluator to define which classes should be corrected or excluded from correction. This is necessary because some land cover classes can legitimately undergo rapid transitions, such as forest converting to an agriculture or grassland mosaic. Therefore, not all class changes are treated as errors.

In addition to class-based rules, specific years can also be excluded from filtering. This is typically applied to certain middle years when the time series remains stable over several periods but shows inconsistency in isolated segments. The primary objective of these temporal filters is to produce a stable and logically consistent time series, minimizing abrupt fluctuations that could distort statistical reporting or subsequent classification analysis.

3.4.6.4. Frequent Filter

The frequency filter evaluates how often a given land cover class appears throughout the entire analysis period. Classes that occur below a defined threshold are considered unstable and are subject to correction. For example, if the analysis spans 35 years and a class appears in fewer than 4 years, which corresponds to approximately 10 percent of the period, it may be filtered out.

The primary objective of this mechanism is to reduce bias caused by false positives, where a class is incorrectly detected in isolated years. By removing sporadic occurrences, the filter preserves dominant and temporally consistent land cover patterns.

In thematic maps, the frequency threshold is adjusted according to the characteristics of each class. Certain classes are naturally more dynamic and may require a lower or higher tolerance. Detailed class-specific configurations are provided in the appendix.

3.4.6.5. Incident Filter

The incident filter is designed to remove pixels that change class excessively during the analysis period. A pixel is considered unstable when it changes more than eight times and is connected to fewer than six neighboring pixels of the same class. Such pixels are interpreted as noise rather than meaningful land cover dynamics.

Pixels that meet both conditions are replaced using the class value from the same spatial position in other years, typically based on the most consistent or dominant class in the temporal sequence. The filter reduces spurious transitions that may arise from classification errors, residual noise, or local artifacts in the mosaics. As with the frequency filter, the application of the incident filter can be adjusted according to the characteristics of each class. Some classes naturally exhibit higher variability and may require tailored thresholds.

MapBiomas Indonesia has produced four sets of annual digital land cover and land use maps, referred to as Collections. The methodologies and Landsat-based classification algorithms applied in each collection have progressively evolved.

3.4.7. Improvement of Collection 4.1

Collection 4.1 was developed to enhance the consistency of historical land cover and land use data from 1990 to 2024 and to improve the representation of transitions from natural to anthropogenic classes. It provides annual maps and associated datasets covering the entire territory of Indonesia, refining the earlier Collection 4.0 released in August 2025.

Key improvements in Collection 4.1 include the implementation of a new set of temporal filters and the integration of auxiliary data from the Global Ecosystem Dynamics Investigation (GEDI) to improve the quality of land cover and transition maps. Transition rules were also revised to better characterize the natural non-forest vegetation class, which predominantly results from the historical conversion of forest formation areas due to human activities such as land clearing for oil palm plantations and pulpwood estates. For the purpose of transition analysis, this class is classified as anthropogenic.

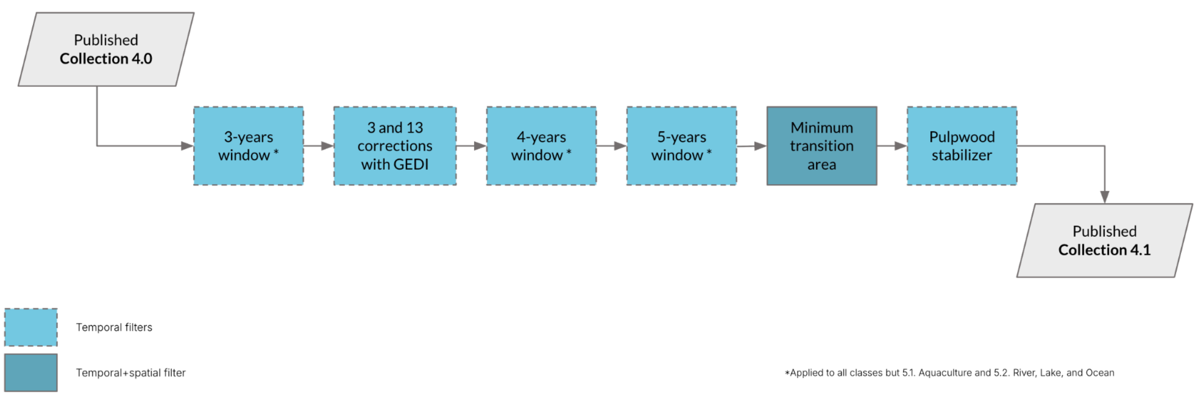


Figure 6. Method workflow of Collection 4.1

This workflow illustrates the steps taken to transform MapBiomas Indonesia Collection 4.0 into Collection 4.1. The process begins with the published Collection 4.0 and applies a series of temporal filters, spatial filters, and corrections to improve historical data consistency and the accuracy of land cover transitions.

1. 3-year window filter: smooths short-term temporal variability in all classes except aquaculture and water bodies.
2. Corrections with GEDI: integrates Global Ecosystem Dynamics Investigation data to refine forest and natural vegetation non-forest characterization. GEDI data were used to correct classification noise between the Forest Formation class (3) with canopy height greater than 15 meters and the Other Natural Vegetation class (13). GEDI's canopy height measurements provide accurate information on forest structure, allowing pixels that were misclassified due to spectral similarity or transitional noise to be reassigned correctly.

Areas with canopy height above 15 meters that were previously labeled as Other Natural Vegetation are corrected to Forest Formation
3. 4-year and 5-year window filters: further temporal smoothing to reduce classification noise.
4. Minimum transition area: a combined temporal and spatial filter to remove small, likely spurious transitions.
5. Pulpwood stabilizer: adjusts classification in pulpwood plantations to ensure consistency over time.

4. Integration

The basic theme map and the cross-cut theme map are integrated on a pixel-by-pixel basis through an overlay process. Each pixel from both maps is compared spatially at the same location. The final class assignment follows a predefined class hierarchy and prevalence rule.

In practice, this means that when two classes overlap at the same pixel, the class with higher priority in the hierarchy prevails. The hierarchy is defined in advance to ensure consistency and logical ordering among land cover and land use categories.

This approach ensures that thematic layers are combined systematically, avoiding ambiguity when classes from different themes intersect. The result is a unified map in which each pixel reflects the dominant class according to the established prevalence rules.

5. Validation Strategy

The validation strategy for the land cover and land use maps is based on two complementary approaches. First, a comparative analysis is conducted between Collection 4 maps and available reference maps. Second, a statistical accuracy assessment is performed using independent samples that represent the entire territory of Indonesia for specific time periods.

Since Collection 2, accuracy assessment has relied on independent samples as reference data. The validation process has been carried out by ten universities distributed across Indonesia, ensuring geographic representation and institutional independence.

5.1. Validation with Reference Maps

Each land cover and land use map is validated against available reference maps using a spatial agreement test. This analysis evaluates the level of spatial correspondence between Collection 4 maps and the reference datasets. The reference maps used are listed in the reference section, while the comparative results are presented in the appendix.

5.2. Validation with Independent Samples

Independent sample validation is conducted periodically, every two collections. It was first implemented in Collection 2—a total of 12,957 independent samples were used to validate the annual MapBiomas Indonesia land cover and land use maps. The number and spatial distribution of samples were determined using statistical sampling techniques.

Each sample point was visually interpreted using Landsat imagery and high-resolution imagery from Google Earth. For every sample, land use and land cover classes were assigned annually from 2000 to 2022. Detailed information on the number and spatial distribution of samples is provided in the appendix.

The validation process involved 15 teams, each consisting of three members, for a total of 45 interpreters. These independent teams were affiliated with ten universities across Indonesia, including Universitas Syiah Kuala, Universitas Bengkulu, Universitas Lampung, Universitas Muhammadiyah Palangkaraya, Universitas Mulawarman, Universitas Tadulako, Universitas Papua, Universitas Indonesia, Universitas Gadjah Mada, and Institut Pertanian Bogor.

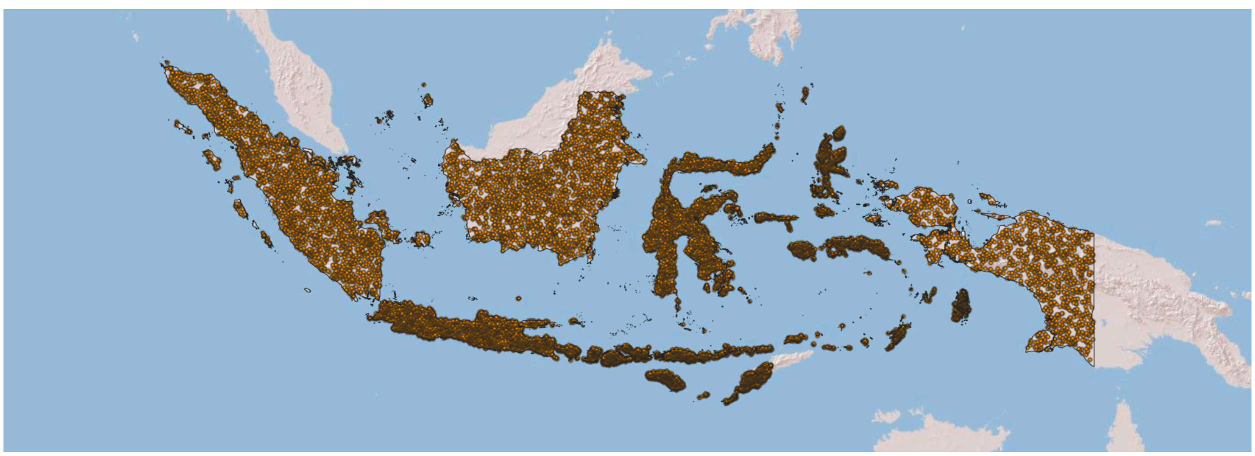
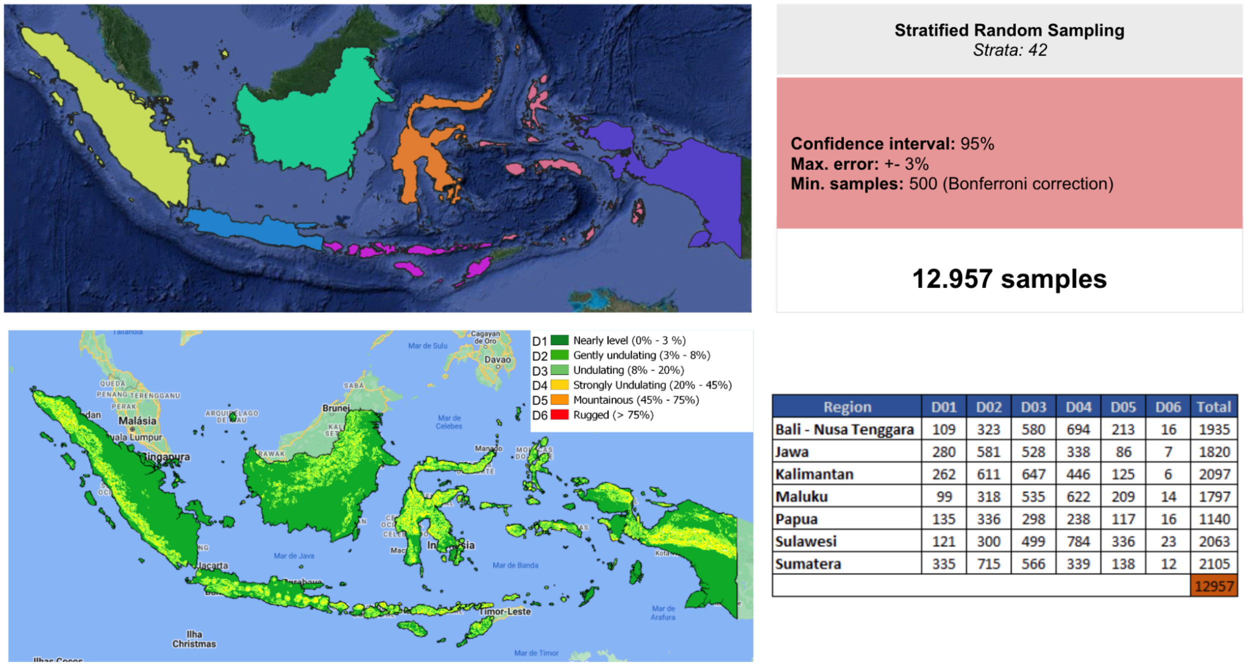


Figure 7. Distribution of validation samples

Within each team, all interpreters independently examined the same sample points and assigned land cover classes. A sample was defined as reference data when at least two interpreters reached the same classification decision.

After the team-level interpretation, all independent samples were reviewed and consolidated by a supervisor of ten experts. These experts evaluated the consistency and quality of the interpreted samples before officially establishing them as reference data.

The entire validation process was conducted using the Temporal Visual Inspection, or TVI, platform developed by LAPIG at the Federal University of Goiás. The platform enables synchronized temporal visualization of satellite imagery and classification results, supporting systematic and consistent interpretation across years.



Region	D01	D02	D03	D04	D05	D06	Total
Bali - Nusa Tenggara	109	323	580	694	213	16	1935
Jawa	280	581	528	338	86	7	1820
Kalimantan	262	611	647	446	125	6	2097
Maluku	99	318	535	622	209	14	1797
Papua	135	336	298	238	117	16	1140
Sulawesi	121	300	499	784	336	23	2063
Sumatera	335	715	566	339	138	12	2105
							12957

Figure 8. Sample design

The sampling design follows a stratified random sampling approach. The minimum analysis units are defined by seven geographic regions, Sumatera, Kalimantan, Sulawesi, Jawa, Bali and Nusa Tenggara, Maluku, and Papua, combined with six slope classes derived from SRTM data. This configuration results in a total of 42 strata, obtained from the combination of regions and slope classes.

The confidence interval adopted is 95 percent, with a minimum error margin of plus or minus 3 percent. To control the overall error across multiple regions, a Bonferroni correction is applied. Under this criterion, the minimum number of samples per region is set at 500.

The general statistical formulation used to estimate the required sample size per region can be expressed as:

$$n = \max_{p \cdot q} \left(\frac{N z_{\gamma}^{2} pq}{(N-1)E^{2} + z_{\gamma}^{2} pq} \right),$$

Equation 1. Sample size per region

where n is the sample size, N is the population size or total number of points, E is the maximum acceptable margin of error, p is the estimated proportion of the attribute of interest, q = 1 − p, and Z is the standard normal distribution factor corresponding to the confidence level 1 − αg.

The adjusted significance level αg is calculated using the Bonferroni correction, where αg = α / (k − 1), 1 − α is the desired confidence level, and k is the number of land-cover classes. This correction controls the overall type I error rate when multiple class-level comparisons are performed.

Accuracy analysis is conducted using a confusion matrix that compares mapped classes with reference sample classes [10], [11]. From this matrix, overall accuracy, user's accuracy, and producer's accuracy are derived. Overall land cover and land use accuracy are calculated annually and also disaggregated by class and by region.

6. Reference

[1] Carbon Brief. 2021. Analysis: Which countries are historically responsible for climate change? https://www.carbonbrief.org/analysis-which-countries-are-historically-responsible-for-climate-change/

[2] Carlos M. Souza Jr. 2020. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. Remote Sens. 2020, 12(17), 2735; https://doi.org/10.3390/rs12172735.

[3] Kementerian Dalam Negeri Indonesia. 2022. Keputusan Menteri Dalam Negeri Nomor 050-145 Tahun 2022 Tentang Pemberian dan Pemutakhiran Kode, Data Wilayah Administrasi Pemerintahan, dan Pulau Tahun 2021.

[4] J. R. G. TOWNSHEND and C. O. JUSTICE, "Selecting the spatial resolution of satellite sensors required for global monitoring of land transformations," Int. J. Remote Sens., vol. 9, no. 2, pp. 187–236, Feb. 1988.

[5] A. V. Egorov, D. P. Roy, H. K. Zhang, Z. Li, L. Yan, and H. Huang, "Landsat 4, 5, and 7 (1982 to 2017) Analysis Ready Data (ARD) observation coverage over the conterminous United States and implications for terrestrial monitoring," Remote Sens., vol. 11, no. 4, 2019.

[6] M. A. Wulder et al., "Current status of Landsat program, science, and applications," Remote Sens. Environ., vol. 225, pp. 127–147, 2019.

[7] https://www.usgs.gov/landsat-missions/landsat-satellite-missions#multimedia

[8] Google Earth Engine. 2022. https://earthengine.google.com/

[9] Olofsson et al. (2014). Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment Volume 148, 25 May 2014, Pages 42-57.

[10] Stephen V. Stehman. 2014. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes, International Journal of Remote Sensing, 35:13, 4923-4939, 11 July 2014

[11] Stephen V. Stehman, Giles M. Foody. 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sensing of Environment Volume 231, 15 September 2019, 111199.

[12] H. Yang, S. Li, J. Chen, X. Zhang, and S. Xu, “The standardization and harmonization of land cover classification systems towards harmonized datasets: A review,” ISPRS Int. J. Geo-Information, vol. 6, no. 5, pp. 1–16, 2017.

[13] A. Di Gregorio and L. J. M. Jansen, “Land cover classification system (LCCS): classification concepts and user manual for software version 1.0.” FAO, 2000.

[14] M. Shinozuka and B. Mansouri, “4 - Synthetic aperture radar and remote sensing technologies for structural health monitoring of civil infrastructure systems,” in Woodhead Publishing Series in Civil and Structural Engineering, V. M. Karbhari and F. B. T.-S. H. M. of C. I. S. Ansari, Eds. Woodhead Publishing, 2009, pp. 113–151.