MAPBIOMAS
COLOMBIA

Mining - Appendix

Collection 4.0

Version 1.0

Responsible Team

Andrés Medina
Karen Andrea Huertas
Emanuel Valero
Luis Medina
Natalia García
Santiago Baquero
Laura Peña
Wendy Jaramillo
Fabián Sánchez

July 2026

Table of Contents

1. Introduction 3
2. Construction and Definition of Landsat Mosaics 4
3. Mining Transversal Methodology 5
3.1. Collection Evolution 5
3.2 Methodological Summary 10
3.2.1 Reference search 11
3.2.2 Training samples (mining/non-mining) 13
3.2.3 Classification Using the Random Forest Algorithm 14
3.2.4 Filters 15
3.2.4.1 Join 15
3.2.4.2 Gap fill 15
3.2.4.3 Spatial Filter 16
3.2.4.4 Temporal Filter 17
3.2.4.5 Frequency Filter 19
3.2.5 Integration **19**
3.3 Special Cases **20**
La Guajira. **20**
Chocó. **20**
Cesar. **21**
Amazonía **22**
4. References **23**

List of Figures

Figure 1: Distribution of the Landsat grid for the mining transversal	4
Figure 2: Comparison of classification regions for Colombia: a) Collection 1; b) Collections 2–4	6
Figure 3 Comparison of mining classification with and without water mask	8
Figure 4: Steps for Mining Classification	10
Figure 5: Mining presence identified by each reference source	11
Figure 6: Mining classification mask for Colombia collection 4	12
Figure 7: Mine and No-Mine Polygons - Open-pit mining - Cerrejón	13
Figure 8: Mine and No-Mine Polygons - Alluvial mining - Taparal - Chocó	13
Figure 9: Mining Classification Polygon. Open-pit mine—El Cerrejón.	14
Figure 10: Mining Classification - Open-Pit Mine - El Cerrejón.	15
Figure 11: Functional Diagram of the Gap Fill Filter	16
Figure 12: Spatial filter functionality diagram	16
Figure 13: Effect of the temporal filter	17
Figure 14: Temporal filter – first and last year rule	18
Figure 15: Diagram of the frequency filter's functionality	19
Figure 16: Classification results for the El Cerrejón coal mine.	20
Figure 17: Mining classification results for part of the department of Chocó.	21
Figure 18: Results of the classification of mining activity in a portion of the department of Cesar.	21
Figure 19: Mining classification results for part of the Amazonía Department.	22

List of Tables

Table 1: Changes in the MapBiomas Colombia Collections	5
Table 2: List of variables included in the feature space used in the classification of the MapBiomas Colombia mining cross-section	7
Table 3: Filter Implementation	9

1. Introduction

The land cover associated with mining is classified as a cross-cutting theme within the MapBiomas methodology, as it generates spectral confusion with other land cover classes, such as urban centers, sandbars, other natural non-vegetated areas, and bare soil areas. This methodology has been previously implemented in several countries such as Bolivia, Ecuador, Peru, and Venezuela, yielding positive results in identifying zones associated with mining activity. In Colombia, this mapping methodology has been implemented since MapBiomas Amazonía Collection 4, presented in 2022, and since MapBiomas Colombia Collection 1, presented in 2023.

This document describes the methodology used in MapBiomas Colombia Collection 4 for the classification of mining zones within the Colombian territory. Mining is defined as all areas involving the extraction of stone or mineral materials—understood as crystalline substances, generally inorganic, with physical and chemical characteristics associated with a specific atomic grouping. This includes construction materials, alluvial materials (such as sands, gravels, and stones near riverbeds and banks), floodplains, and other alluvial terrains, as well as industrial production with clear soil exposure, without differentiating the type of activity (industrial, artisanal, legal, or illegal) (Congreso de Colombia, 2001) (IDEAM, 2010). As was the case in Collection 3 of 2025, areas associated with oil-related activities, such as reservoirs, are included, without taking into account the main infrastructure for this activity.

It is important to note that the original methodology for classifying this cross-cutting theme was based on the prior experience of the MapBiomas Brazil team, specifically Cesar Diniz (Solutions in Geoinformation, Solved, Brazil), who presented the identification of industrial mining in that country to the MapBiomas team. This methodology was implemented in the Google Earth Engine (GEE) tool with minor adjustments to improve performance and product quality.

2. Construction and Definition of Landsat Mosaics

For processing purposes, the MapBiomas methodology divides the project area into a regular grid defined based on the Landsat image grid for the Andes, Caribbean, Insular Caribbean, Pacific, Insular Pacific, Orinoquía, and Inter-Andean Valleys biomes. Each grid cell covers an area of approximately 2.7 Mha. Figure 1 shows the study area of Colombia, comprised of 43 grid cells (Figure 1).

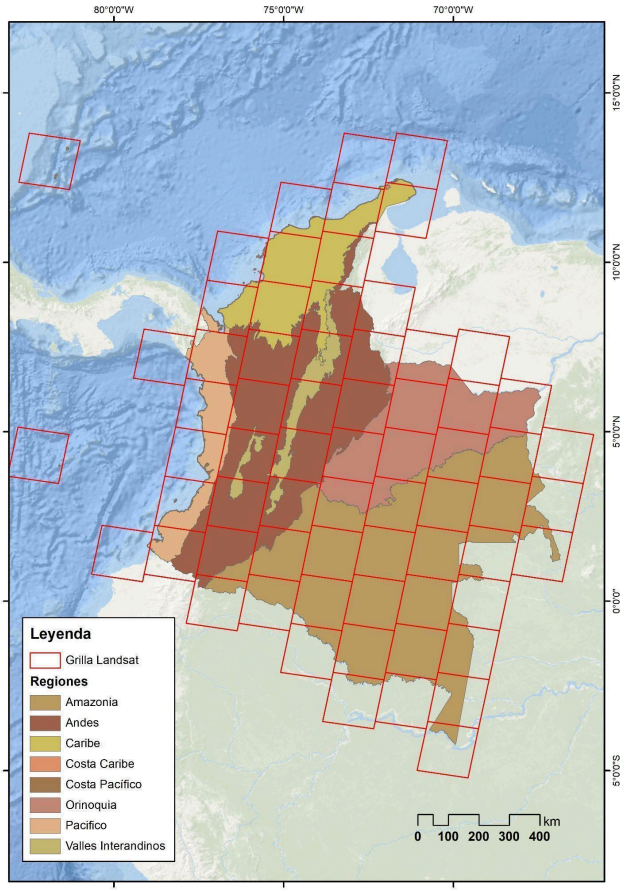


Figure 1: Distribution of the Landsat grid for the mining transversal

Annual Landsat image mosaics are constructed within each of these grid cells. The objective is to generate periodic cloud-free images that allow for classification and change identification[1]. Starting with Collection 3 of 2025, Landsat Collection 2 was used to construct the mosaics; this collection features the most up-to-date reprocessing of Landsat archives, with several improvements in image quality.

[1] For more details, see Section 4 of the document: Appendix - Colombia Algorithm Theoretical Basis Document (ATBD) Collection 4.0

Mosaic construction begins with the selection of Landsat images. A parameterization table is built with the following parameters, stored as metadata in each constructed asset:

- ID: Unique identifier or code for the chart unit/region.
- Year: Year of the series (spanning the period 1985 to 2025) corresponding to each mosaic.
- Start Date / End Date: Period of the year used for selecting images from the GEE Landsat catalog.
- Sensors: Landsat 5 TM, Landsat 7 ETM+, Landsat 8 OLI, Landsat 9, or a combination thereof (e.g., Landsat 5 and Landsat 7).
- Cloud Percentage: Maximum accepted cloud cover percentage for each Landsat image used to build the mosaic. This data is derived from the Landsat image metadata.
- Blacklist: Images excluded from mosaic construction due to quality issues.

Using this information, annual image mosaics are constructed on the Google Earth Engine platform. All generated mosaics are stored as a GEE Image Collection and are available at the path: projects/mapbiomas-mosaics/assets/LANDSAT/LULC/COLOMBIA/mosaics-1. Initially, these present the median values for each Landsat band. Additionally, in Collection 4.0, a topographic correction process was applied to minimize the effects of terrain relief on sensor-recorded reflectance. This process adjusts satellite image reflectance values to compensate for topographic effects such as slopes and aspect. By calculating the solar incidence angle, the reflectance value of each pixel is adjusted to more adequately represent the real surface properties.

Based on these data, additional bands were added to each mosaic, including spectral indices, spectral fractions, texture variables, and topographic-geomorphological indices derived from the Digital Elevation Model (DEM), among others. This aims to improve mining detection capabilities in the country by differentiating the presence and absence of the class of interest. Statistical calculations are applied to each mosaic band to generate specific information on: median, dry season median (25th percentile), wet season median (75th percentile), amplitude (value range), standard deviation, and minimum value. These data compose the variables used as input for the Random Forest classifier.

3. Mining Transversal Methodology

3.1. Collection Evolution

The Mining cross-cutting theme has been developed for the three MapBiomas Colombia collections. Each of these collections exhibits methodological differences (Table 1) that reflect variations in the results.

Table 1 - Changes in the MapBiomas Colombia Collections

Parameter	COLLECTION			
	1	2	3	4
Series	1985-2022	1985-2023	1985-2024	1985-2025
Regions	79	53	53	53

Charts	47	47	47	47
Sensor	4,5,7,8 y 9	4,5,7,8 y 9	4,5,7,8 y 9	4,5,7,8 y 9
Mosaics	1955	2067	5155	4246
Variables	22	25	25	25

Among the most significant changes are the following:

Regarding the number of regions defined for mining interpretation, Collection 1 divided the country into 79 regions based on the general MapBiomas Colombia classification regions layer, supported by high-resolution Google Maps basemap imagery. For the second collection, the division and quantity of regions were adjusted to avoid splitting large mining activity regions and to group zones to facilitate identification; thus, Collection 2 managed a total of 53 regions for mining identification. This change is maintained for Collection 3-4, where neither the quantity nor the limits of the classification regions were adjusted.

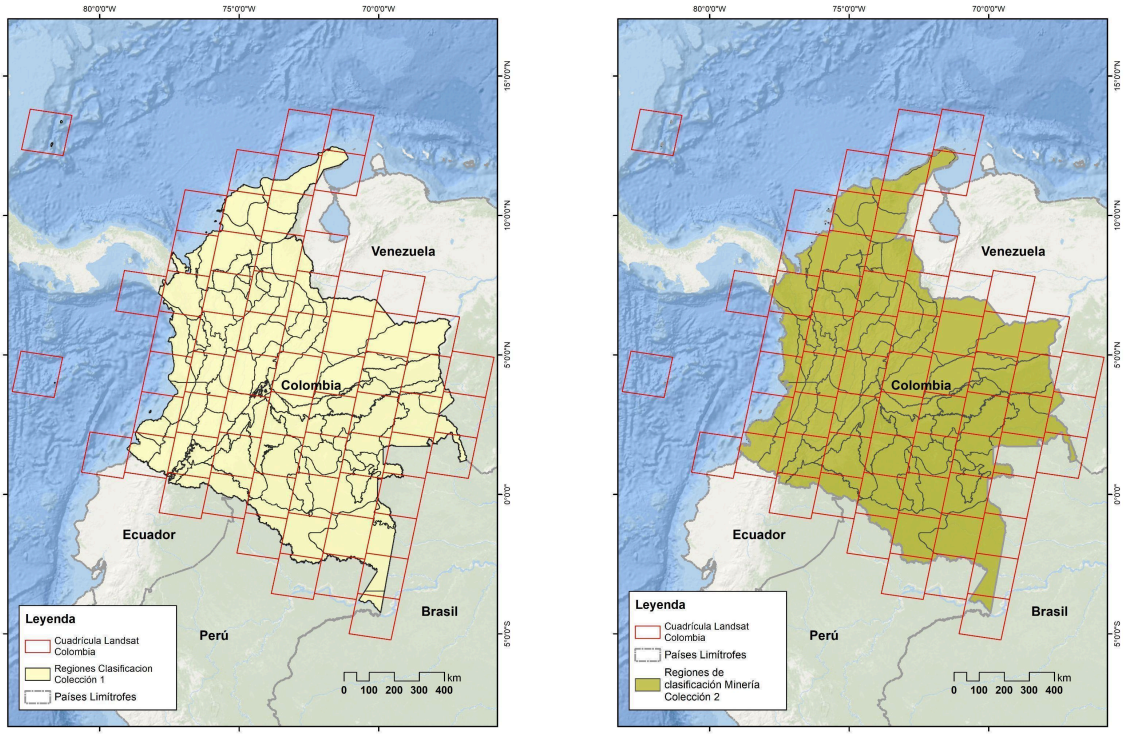


Figure 2: Comparison of classification regions for Colombia: a) Collection 1; b) Collections 2–4

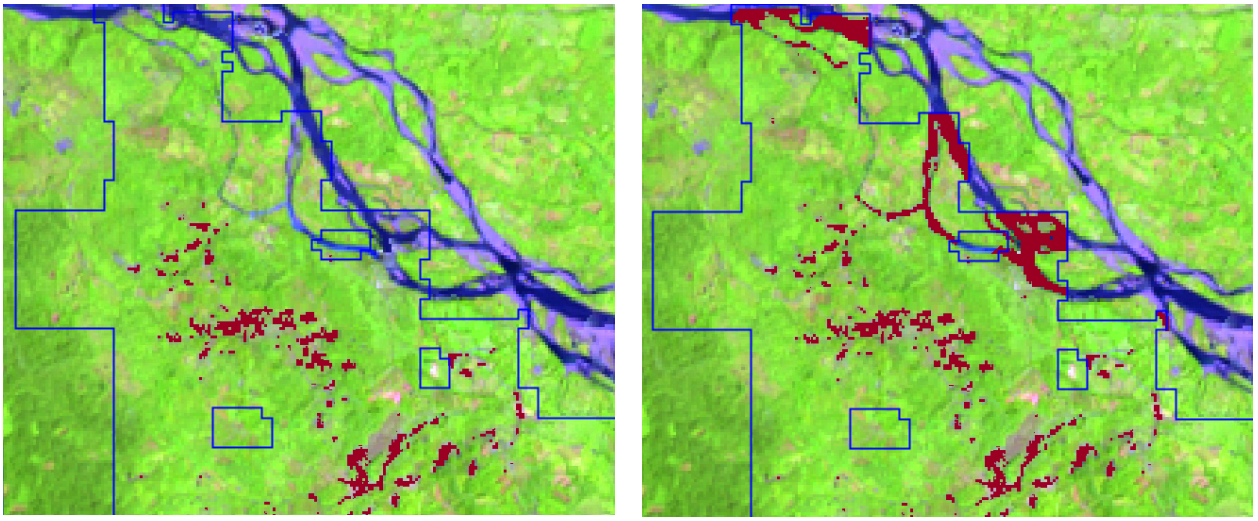
As an initial step for classification, tests were conducted to identify a specific set of variables for mining detection. In MapBiomas Colombia Collection 1, 22 variables were used; for Collection 2, three additional spectral indices were added, allowing for a more detailed classification of this cover. These indices are maintained for Collection 3 and 4. Details of each variable can be found in Table 2.

Table 2 - List of variables included in the feature space used in the classification of the MapBiomas Colombia mining cross-section

ID	Variable	Description / Expression	Reducer	Acronym	Group	Collection
0	blue	Blue band value	median	blue_median	Landsat Band	1-2-3-4
1	green	Green band value	median	green_median	Landsat Band	1-2-3-4
2	red	Red band value	median	red_median	Landsat Band	1-2-3-4
3	red	Red band value	median	red_wet	Landsat Band	1-2-3-4
4	near-infrared (NIR)	Near-Infrared band value (NIR)	median	nir_median	Landsat Band	1-2-3-4
5	near-infrared (NIR)	Near-Infrared band value (NIR)	median	nir_wet	Landsat Band	1-2-3-4
6	Short-wave infrared (SWIR) 1	Short-Wave Infrared SWIR1	median	swir1_median	Landsat Band	1-2-3-4
7	Short-wave infrared (SWIR) 2	Band value SWIR2	median	swir2_median	Landsat Band	1-2-3-4
8	Ndvi	(NIR - Red) / (NIR + Red)	median	ndvi_median	Spectral Index	1-2-3-4
9	Ndvi	(NIR - Red) / (NIR + Red)	median	ndvi_wet	Spectral Index	1-2-3-4
10	wefi	Forest Ecosystem Fraction Index	median	wefi_wet	Spectral Index	1-2-3-4
11	Gcvi wet	(nir/green − 1)	median	gcvi_wet	Spectral Mixture Modeling	1-2-3-4
12	Sefi	Savanna Ecosystem Fraction Index	median	sefi_median	Fraction Index	1-2-3-4
13	Soil	Bare Soil Index	median, amplitude	soil_median	Spectral Mixture Modeling	1-2-3-4
14	Snow	Fractional abundance of snow within pixel	median	snow_median	Fraction MME	1-2-3-4
15	Evi 2	2.5 * ((NIR - Red) / (NIR + 2.4*Red + 1))	median	evi2_median	Spectral Index	1-2-3-4
16	Ndwi	(Green - NIR) / (Green + NIR)	median	ndwi_mcfeeters_median	Spectral Index	1-2-3-4

17	Mndwi	(Green-NIR)/(Green +NIR)	median	mndwi_median	Water Index	1-2-3-4
18	slope	Terrain slope	Identity	slope	Geomorphometric	1-2-3-4
19	Slppost	Stratified Slope	–	slppost	Static/Topographic Vars	1-2-3-4
20	Elevation	DEM	–	elevation	Static/Topographic Vars	1-2-3-4
21	Shade_mask2	Shadow Map	–	shade_mask2	Static/Topographic Vars	1-2-3-4
22	FM	SWIR1 / NIR	median	ferrous_median	Spectral Index	2-3-4
23	CM	SWIR1 / SWIR2	median	clay_median	Spectral Index	2-3-4
24	IO	Red / Blue	median	iron_oxide_median	Spectral Index	2-3-4

Additionally, for Collection 2, a water mask was included based on the MapBiomas Colombia Collection 1 classification; due to the lack of available data for 2023, the Global Surface Water classification was used. This mask optimized mining mapping by excluding water bodies and sandbars from the classification. For Collection 3, the Collection 2 mask is included as a reference point to identify mining zones, and a buffer between 30 to 100 meters was generated to include areas of mining expansion. However, these areas are manually adjusted in zones showing greater growth in mining activity. Figure 3 below shows an example of the water mask implementation for 2023. Similarly, for Collection 4, mining data generated by the MapBiomas Alerta Colombia project for the years 2020–2025 are included.



a) With water mask b) Without water mask

Figure 3 Comparison of mining classification with and without water mask

Within the classification process, Collection 2 made the first approximation for identifying areas associated with salt extraction, located mainly in the Caribbean region; this overview continues in Collections 3 and 4.

In the post-classification process, the order and quantity of executed filters were changed. For Collection 1, the filter order was: gap fill, spatial, and temporal ; in Collection 2, the frequency filter was added, and the filter order was left to the discretion of each interpreter. This approach was maintained for Collections 3 and 4; Table 3 shows the order of filters implemented in Collection 4.

Table 3 - Filter Implementation

Region	Order of Implemented Filters							
	1	2	3	4	5	6	7	8
30202	Spatial	GapFill	Frequency	Temporal	Temporal	Frequency	Frequency	Temporal
30203	Spatial							
30204	Spatial	GapFill	Frequency	Temporal				
30205	Join	Spatial						
30207	Join	Spatial	Temporal	Frequency	Frequency			
30208	Join	Spatial						
30210	Join	Spatial	Temporal					
30213	Spatial	Temporal						
30216	Spatial	Temporal	GapFill	Frequency	Frequency			
30218	Join	Spatial	GapFill	Temporal	Temporal	Temporal	Frequency	Temporal
30221	Join	Spatial	Temporal	Frequency	Temporal	Temporal		
30301	Join	GapFill	Spatial	Temporal	Spatial			
30302	Join	Spatial	Frequency	Temporal				
30303	Spatial	Temporal						
30304	Join	Spatial	Temporal					
30401	Join	Spatial	GapFill	Temporal				
30402	Join	Spatial	Frequency	Temporal	Spatial			
30403	Join	GapFill	Temporal	Spatial	Frequency			
30404	Join	GapFill	Spatial	Temporal	Spatial			
30405	Join	GapFill	Spatial	Temporal	Spatial			
30406	Join	GapFill	Temporal	Spatial	Temporal	Spatial	Frequency	
30407	Join	Spatial	Temporal					
30408	Join	Spatial	Temporal					
30409	Join	GapFill	Spatial					
30410	Join	GapFill	Spatial	Frequency				
30411	Join	Temporal	Spatial					
30412	Join	Spatial	GapFill	Temporal	Spatial			
30413	Join	GapFill	Temporal	Spatial				
30414	Join	Spatial						
30501	Join	GapFill	Spatial	Temporal	Spatial			
30502	Join	GapFill	Temporal	Spatial				

30503	Join	GapFill	Spatial	Frequency	Temporal	Frequency	Frequency
30504	Join	Spatial	GapFill	Frequency	Temporal		
30505	Join	GapFill	Spatial	Temporal			
30506	Join	Spatial	Temporal	Frequency			
30601	Join	GapFill	Spatial	Temporal	Frequency		
30602	Join	Spatial	Temporal				
30603	Spatial	Temporal					
30604	Join	GapFill	Temporal	Frequency	Spatial		
30605	Join	GapFill	Spatial	Temporal	Spatial		
30606	Join	GapFill	Temporal	Spatial			
30607	Join	GapFill	Spatial	Temporal			
30608	Join	Spatial					

3.2 Methodological Summary

To generate the classification, a methodological framework was developed consisting of five stages, three of which correspond to the classification process and the remaining two to the post-classification stage; the results were stored in Google Earth Engine. The process involved identifying mining areas using official external sources, collecting samples, and classifying them in each region, then integrating all the mapped areas into a single layer, and finally combining them with the general MapBiomas Colombia map. Each step associated with the classification is described below (Figure 4).

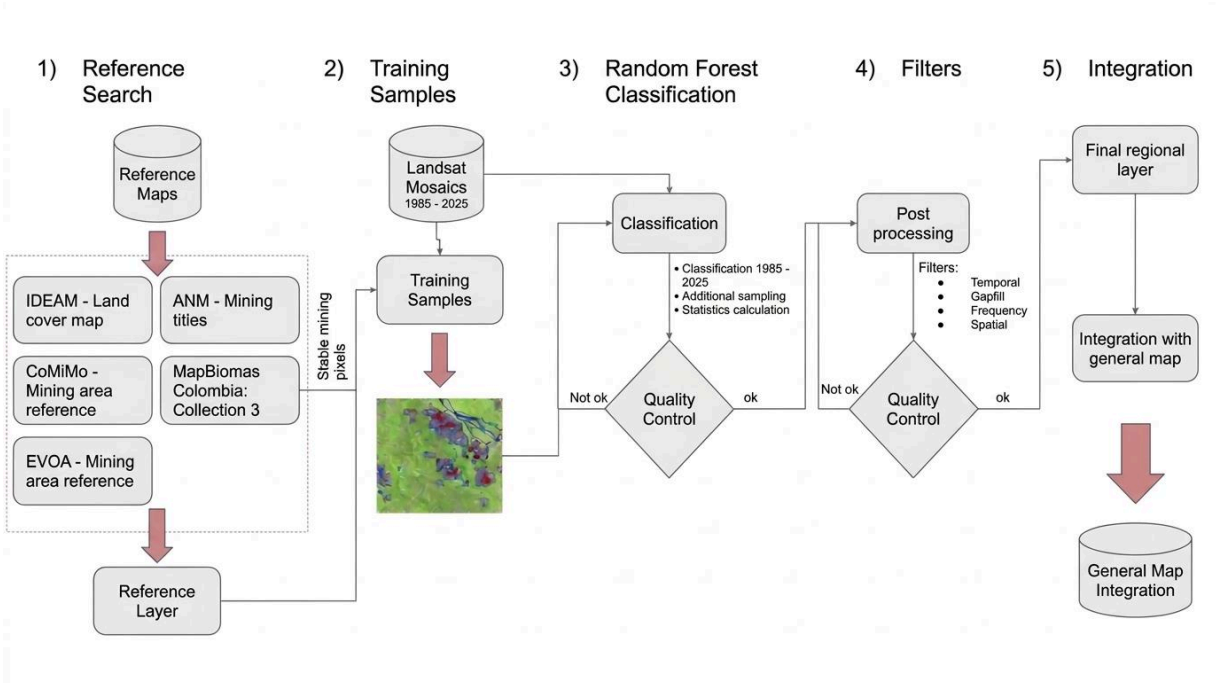
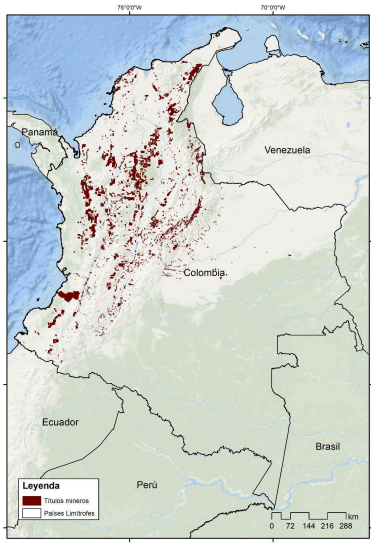


Figure 4: Steps for Mining Classification

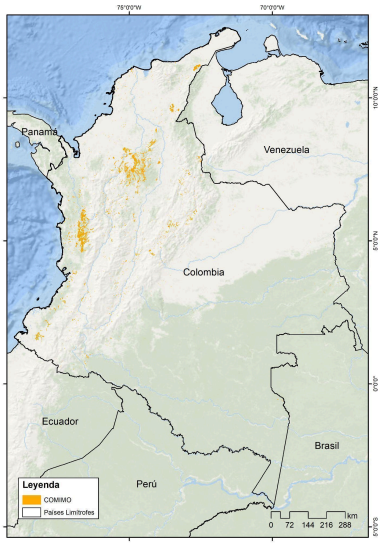
3.2.1 Reference search

A search for official auxiliary data sources reporting mining activity across different time ranges was conducted, including COMIMO (Colombian Mining Monitoring), EVOA (alluvial gold mining), Corine Land Cover 2018, the mining titles (Figure 5), which were used to map Collections 1, 2, and 3. Similarly, data provided by MapBiomas Alerta 2020–2025 on mining areas are incorporated to generate a reference mask for identifying mining activities, which serves as the starting point for the process of identifying and classifying mining in the country (see Figure 6).

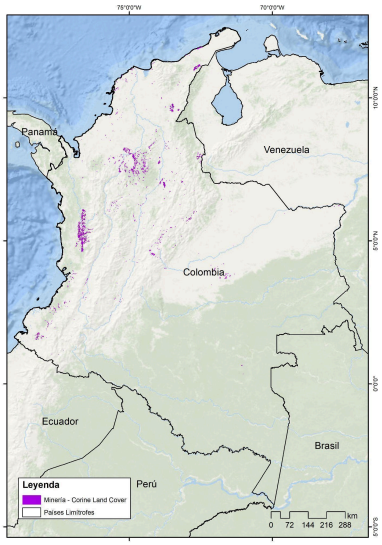
Mining Titles



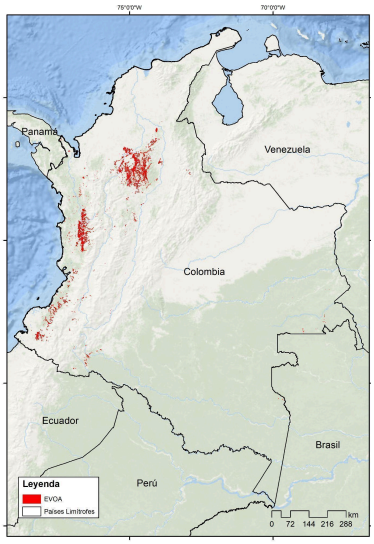
COMIMO



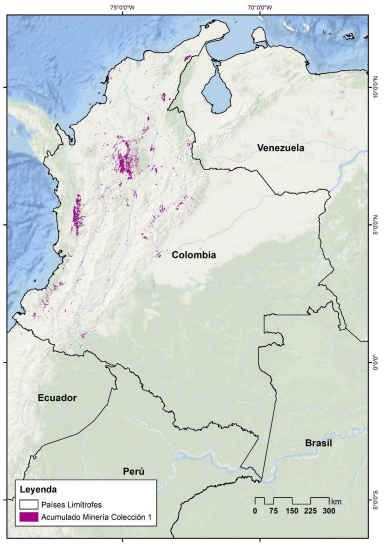
Corine Land Cover



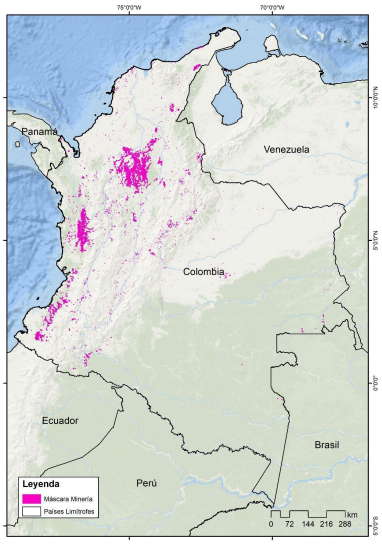
EVOA



MapBiomas Collection 1



MapBiomas Collection 2



MapBiomas Collection 3 MapBiomas Alerta 2020-2025

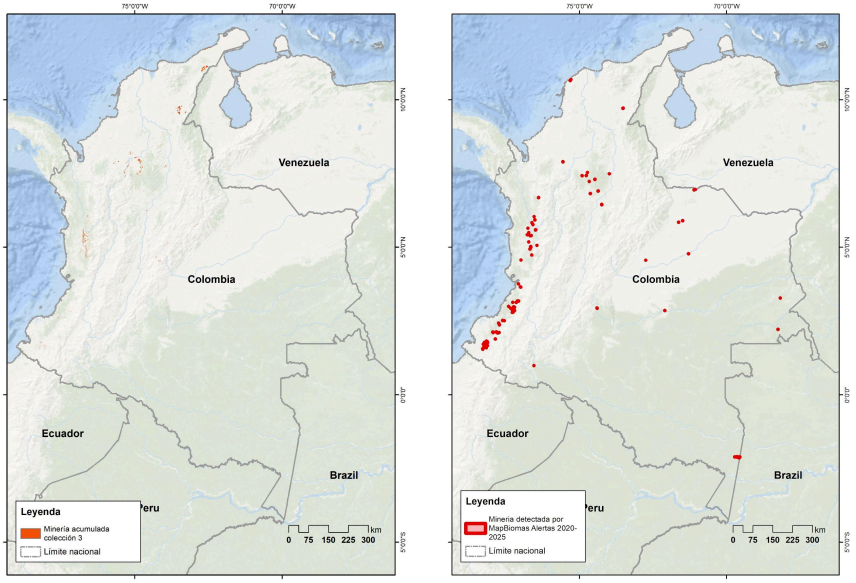


Figure 5: Mining presence identified by each reference source

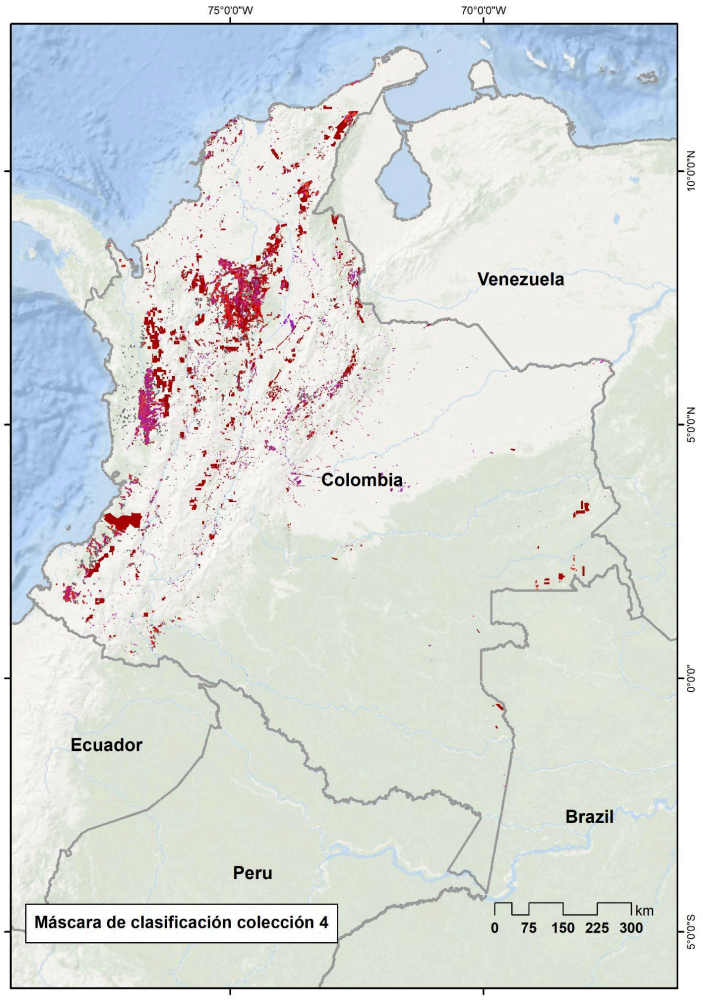


Figure 6: Mining classification mask for Colombia collection 4

3.2.2 Training samples (mining/non-mining)

Within the polygons that delineate the mining areas of interest (Figure 9), two types of training polygons were manually defined and identified: one defined as “Mine” (ID 30), which strictly contains the zones that are part of the mining operation or the mine itself, and the second polygon encompasses the “unobserved” class, or those areas not associated with mining activity (ID 27) (Figures 7 and 8). The training samples taken within the mask are used to generate the classification of the mining zone. Anything outside the delimited zone is not classified, as it is considered a non-mining zone and/or assigned to another class within the MapBiomas legend.

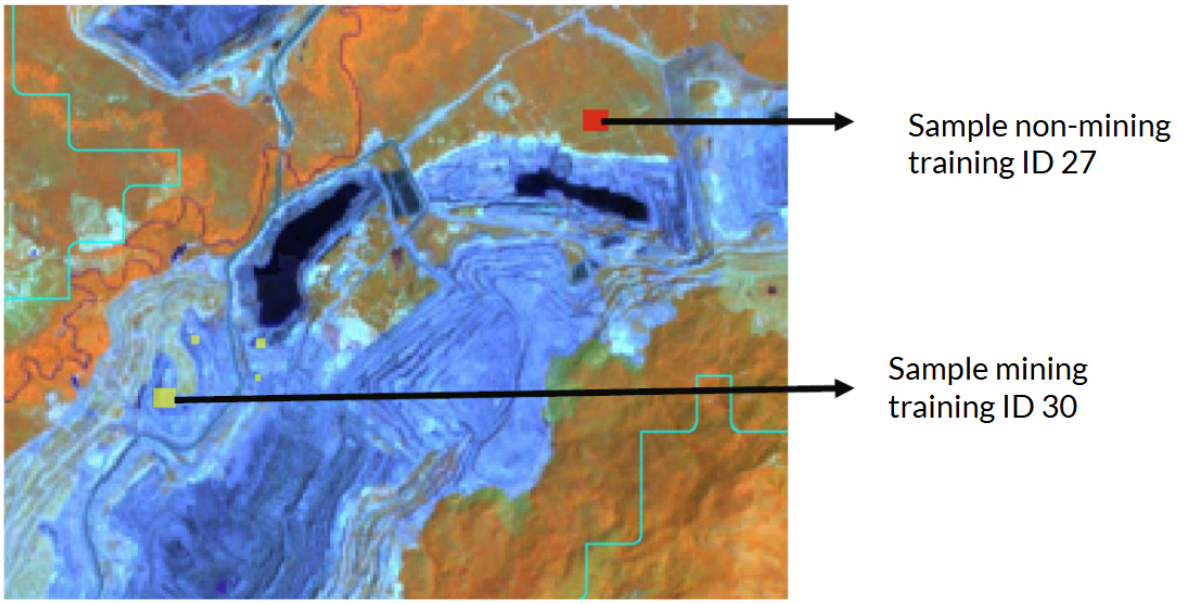


Figure 7: Mine and No-Mine Polygons - Open-pit mining - Cerrejón

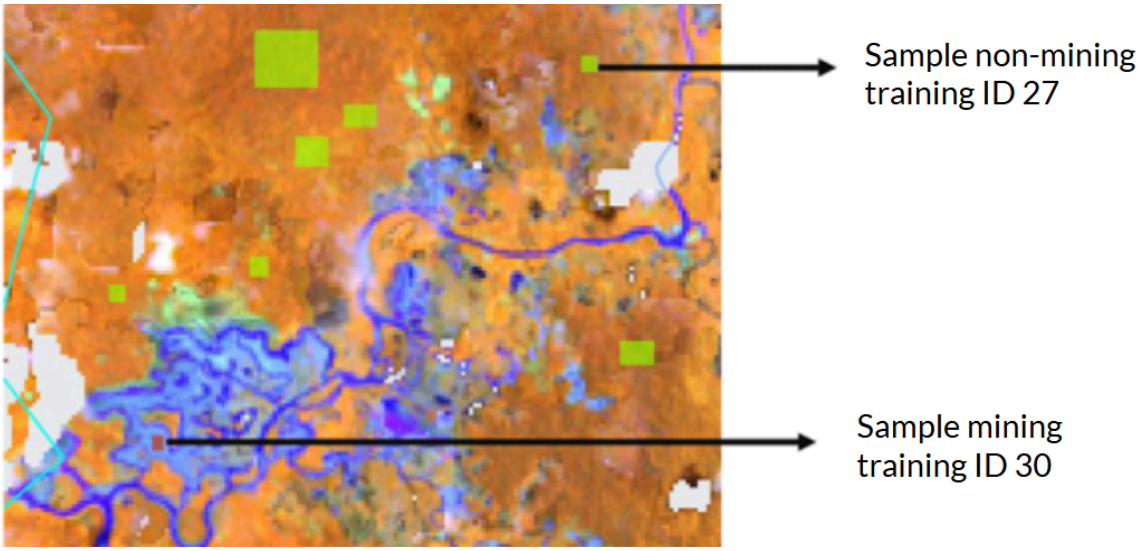


Figure 8: Mine and No-Mine Polygons - Alluvial mining - Taparal - Chocó

3.2.3 Classification Using the Random Forest Algorithm

The classification process is based on the Random Forest algorithm, a machine learning method that allows for the identification and spatial delineation of areas with mining activity in the study regions. Random Forest is based on the construction of a set of decision trees, each of which performs successive binary splits based on the predictor variables to assign a class to the observations. The construction of these trees incorporates random components that promote their diversity and reduce the correlation among them. Finally, the individual predictions are combined using a majority voting mechanism, based on which a class is assigned to each pixel (Breiman, 2001). The classification of mining areas was carried out by implementing this algorithm on the Google Earth Engine platform.

The classification was performed within the areas of interest previously delineated based on the reference sources (Figure 9). Additionally, through a visual inspection of the mosaics, potential new mining fronts were identified that were not referenced in those sources. These areas were manually delineated by digitizing new polygons, which were incorporated into the analysis areas to be included in the classification process. In this way, the goal was to expand the spatial identification of areas with mining activity beyond the records available in the reference sources. The procedure was carried out within each of the study regions (Figure 2B).

As a result, a binary-classified raster image was obtained, in which each pixel represents the presence or absence of areas with characteristics associated with mining. Pixels classified as indicating the presence of mining are shown in red (Figure 10), and the results are spatially restricted to the areas previously delineated by the classification mask.

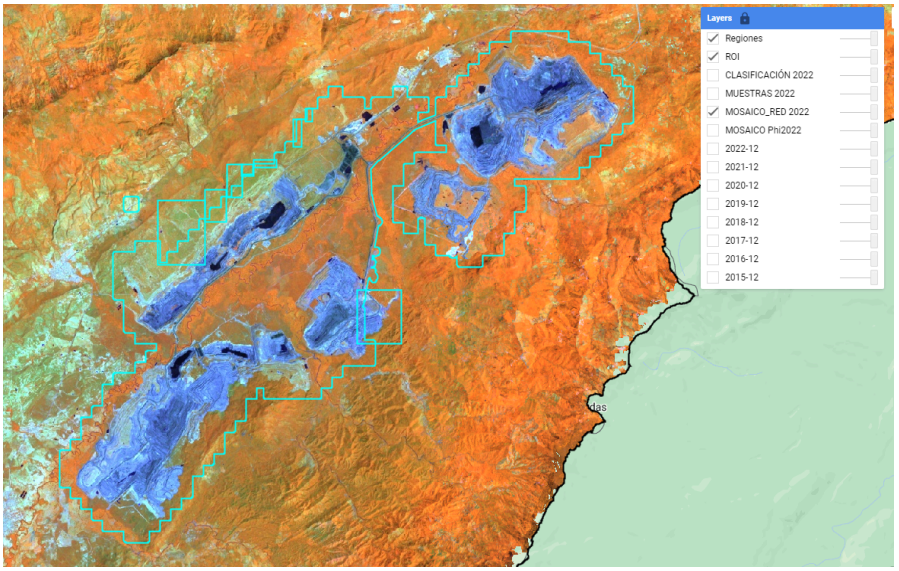


Figure 9: Mining Classification Polygon. Open-pit mine—El Cerrejón.

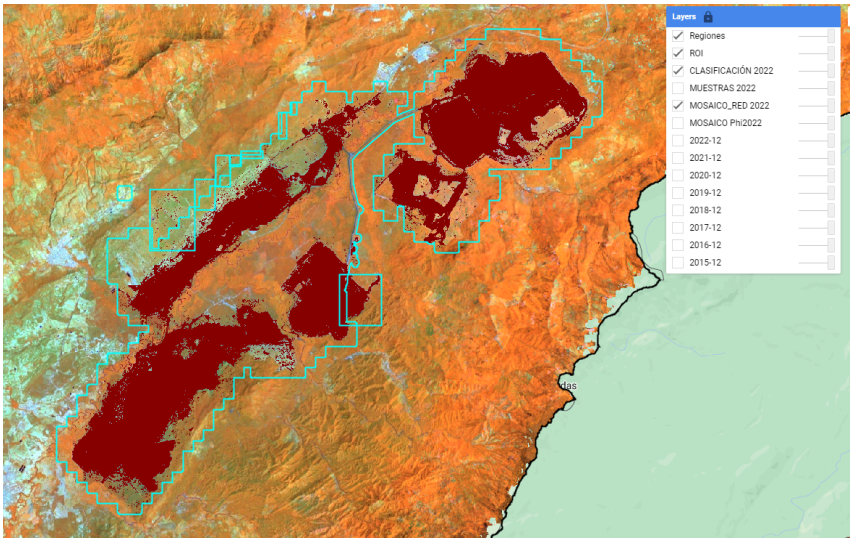


Figure 10: Mining Classification - Open-Pit Mine - El Cerrejón.

3.2.4 Filters

A series of filters was applied to the results of the Random Forest classification to reduce temporal inconsistencies, data gaps, overestimation noise, and other classification issues. To this end, gap-fill, spatial, temporal, and frequency filters were used. Additionally, a "Join" was included to merge the classified or reclassified years in Collection 4 with the previously processed years from Collection 3.

3.2.4.1 Join

For Collection 3 and 4, a preliminary step was implemented prior to the post-classification filters. This step consists of combining the years classified for the current collection (e.g., 2025) with those already classified and processed in the previous collection (e.g., 1985–2024). The result of this step is a single stack of images that are subsequently processed using the traditional filters. Additionally, this step also allowed us to modify the classification of specific areas in the years processed prior to 2025, with the aim of improving classification quality across the entire time series.

3.2.4.2 Gap fill

Due to its geographic location and topography, Colombia experiences high cloud cover throughout the year. In addition, because of banding issues with the Landsat 7 sensor, the mosaics exhibit data loss or pixels with no data. Because of this, it is necessary to fill in these gaps in data or missing data using gap filling, which allows us to replace these gaps with the value of the temporally closest pixel.

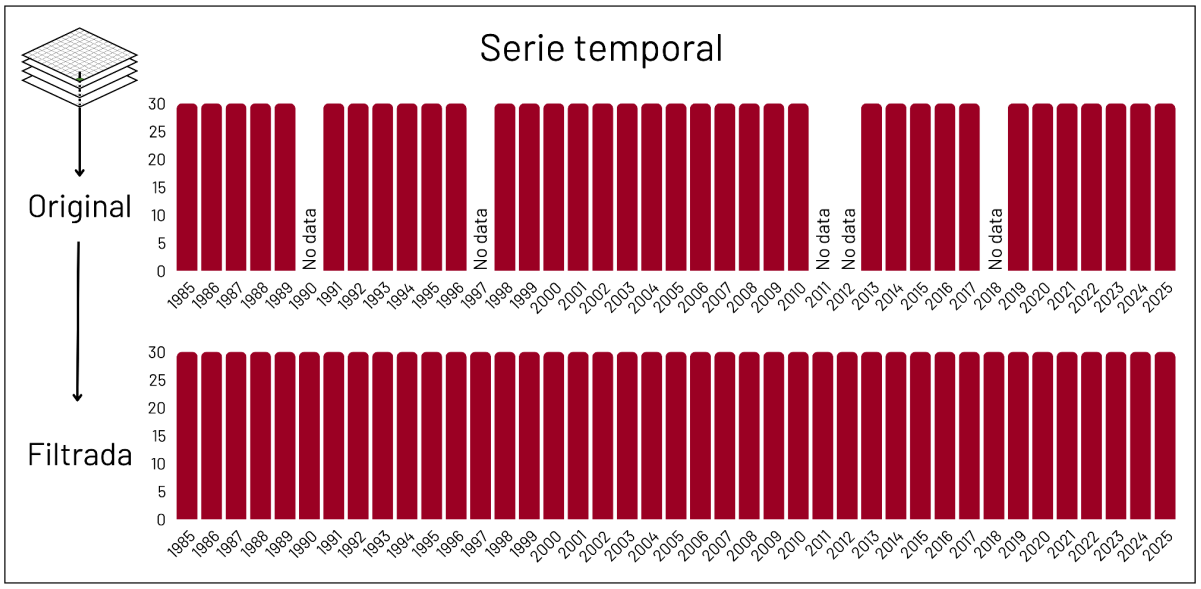


Figure 11: Functional Diagram of the Gap Fill Filter

3.2.4.3 Spatial Filter

The spatial filter is based on the "connectedPixelCount" function in Google Earth Engine, which identifies groups of neighboring pixels that share the same value. Using this function and considering a predefined minimum number of identical pixels, those that do not present connectivity (different values) are considered isolated pixels and were reclassified (Figure 12). For both classes (ID: 27, ID: 30), the minimum connection threshold was set at 3 pixels (approx. 0,3 ha).

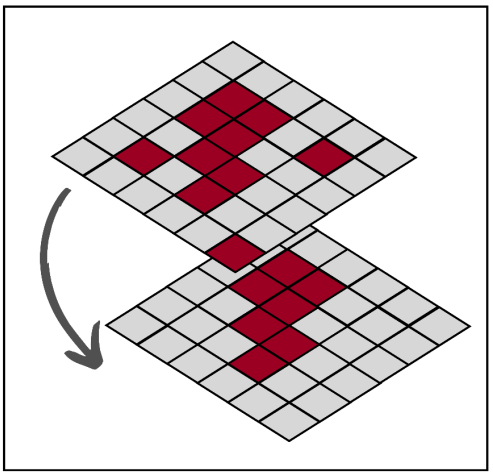


Figure 12: Spatial filter functionality diagram

3.2.4.4 Temporal Filter

This filter consists of replacing pixels identified as noise or salt-and-pepper effect, which represent inconsistencies generated in land cover classification due to gradual year-to-year changes and the absence or distortion of information within a mosaic. In this context, three filters are applied within the temporal filter: intermediate year, first year, and last year filters, which are described below.

Intermediate year: This filter is executed by considering the first and last pixel of the time series, modifying the classification of pixels located in the intermediate years to match those classified at the temporal extremes (1985–2025). In the lower part of Figure 13, it can be observed that under the 3-year rule, a pixel is classified as presence of the class of interest in green (forest, agriculture, mining, etc.) in years t1 and t3, while for t2 the absence of the class of interest is observed in beige. In this sense, after applying the filter, abrupt changes in classification are stabilized by modifying the intermediate pixel to match its initial and final counterparts. Similarly, this filter is applied to 4- and 5-year temporal series, as well as in the inverse manner, as illustrated in the upper part of Figure 13. It is important to clarify that for the starting and ending years of the analyzed period, in this case 1985 and 2025, the intermediate-year rule is not applied due to the absence of adjacent years for comparison.

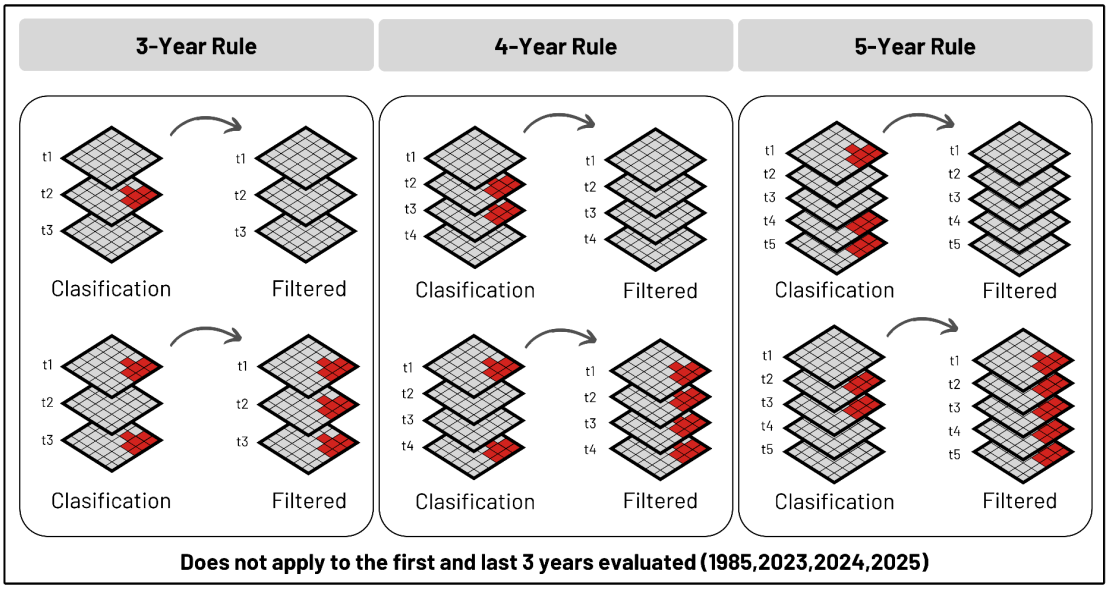


Figure 13: Effect of the temporal filter

Intermediate year between the first and last years: The first-year rule corrects inconsistencies in the time series caused by missing data for the year 1985. The example below shows a transition from the absence of areas in the class of interest (mining) in the first year to the presence of mining areas in the first year.

The last-year rule corrects the misclassified pixel at the end of the time series (2025) due to abrupt changes such as the one mentioned above. In Figure 14, we can see how the filter acts on the pixel classified at the end of the time series as a mining area (in green); after the filter is applied, it is reclassified as a non-mining area, thereby homogenizing the series.

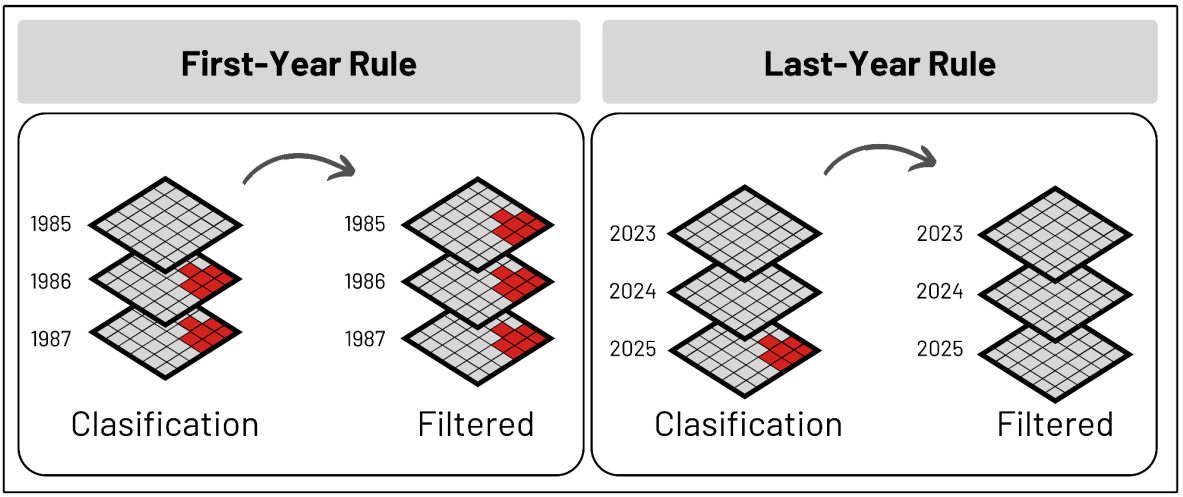


Figure 14: Temporal filter – first and last year rule

Due to the nature of the mining class—which can appear and disappear in certain areas because it may be illegal and lack the permits required for continuous operation over time—the analyst evaluates in which situations to apply this filter and its respective configuration in order to avoid overestimating or underestimating the class throughout the time series. In conclusion, the main objective of this set of filters is to obtain a stable time series, free of inconsistent jumps that could affect statistical reports and the classification itself. Similarly, years can be excluded from the time series to better align its behavior with what is observed in the mosaics; this occurs primarily in intermediate years, when a series of years is stable but specific other periods are not.

3.2.4.5 Frequency Filter

The frequency filter allows us to fill in gaps in the classification based on the temporal recurrence of pixels identified as mining across the entire image series. To do this, we calculate the frequency with which each pixel was classified as mining during the analysis period. When this frequency reaches or exceeds a pre-established threshold (for example, 60% of the observations), the gaps associated with that pixel are filled in with the mining category. In this way, the temporal consistency of the classifications is leveraged to reduce spatial discontinuities and improve the continuity of the areas identified as mining.

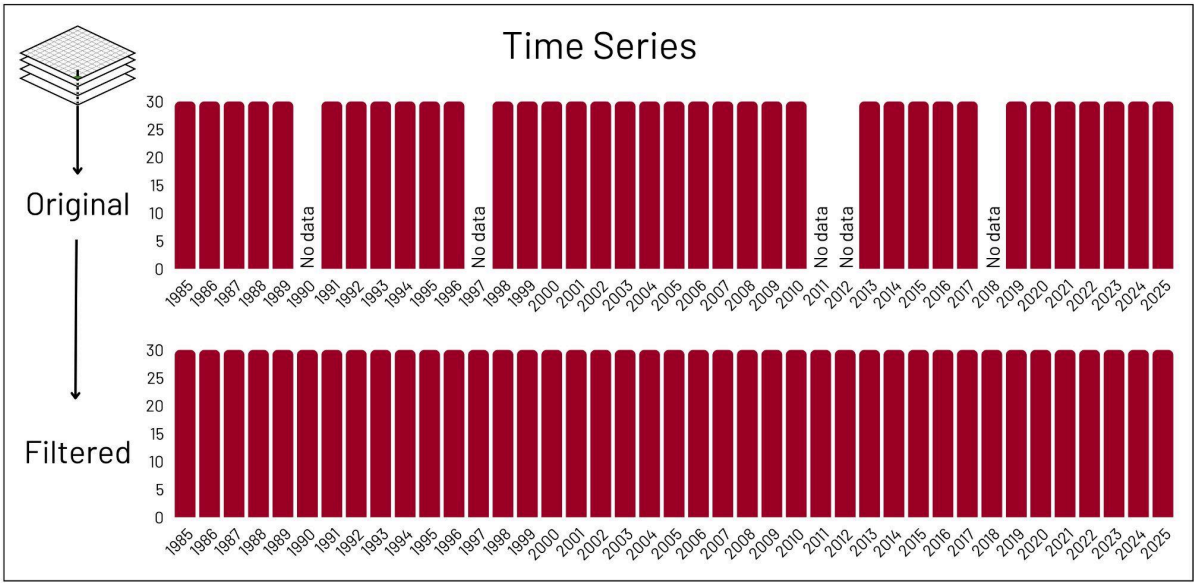


Figure 15: Diagram of the frequency filter's functionality

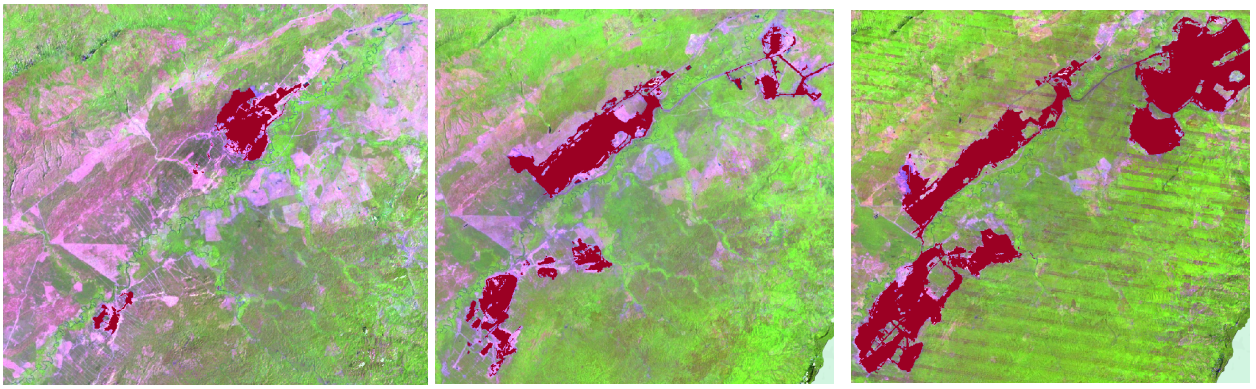
3.2.5 Integration

With the final classification of mining areas completed, an initial integration of all regions at the national level was carried out, and this result was incorporated into the general classification map along with the other mapped classes. This integration was based on a series of precedence or order rules, in which the mining cross-section has a priority level of 4, below Glacier and Snowfields, Rocky Outcrop, and Hypersaline Tidal Flat.

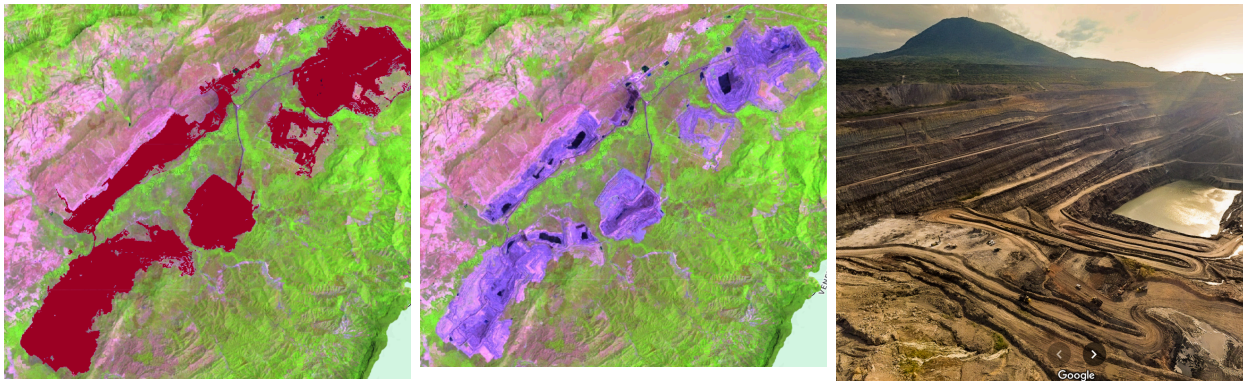
Considering the integration order presented above, the resulting output is the final map showing the presence of mining across the different areas of the country.

3.3 Special Cases

La Guajira.



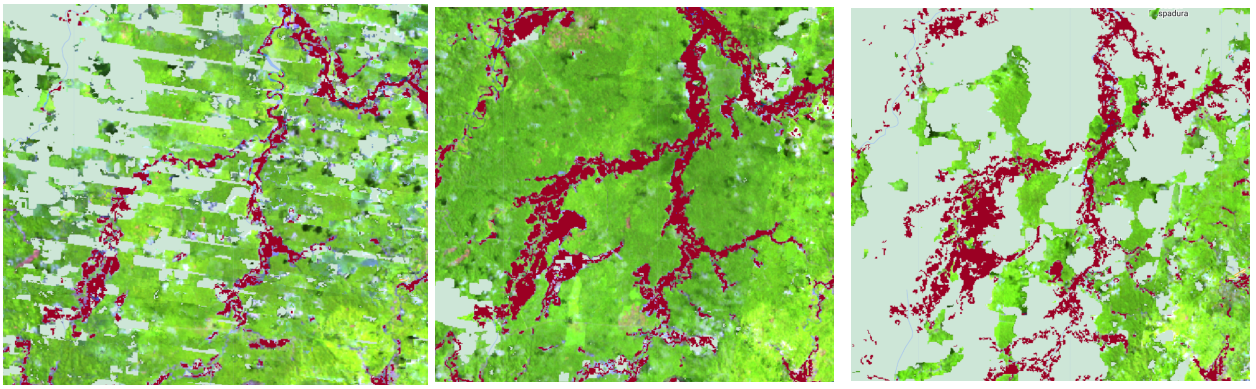
1986	2000	2010



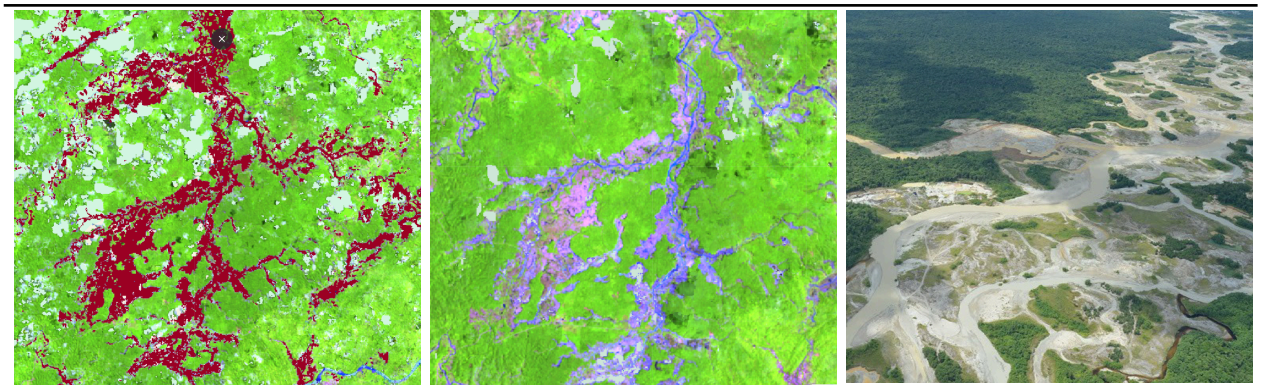
2025	2025. RGB 654 L8	Landscape

Figure 16: Classification results for the El Cerrejón coal mine.
Albania, Barrancas and Hatonuevo - Guajira, Scale: 2 km, Coordinates: -72.6306 W, 11.075 N.

Chocó.



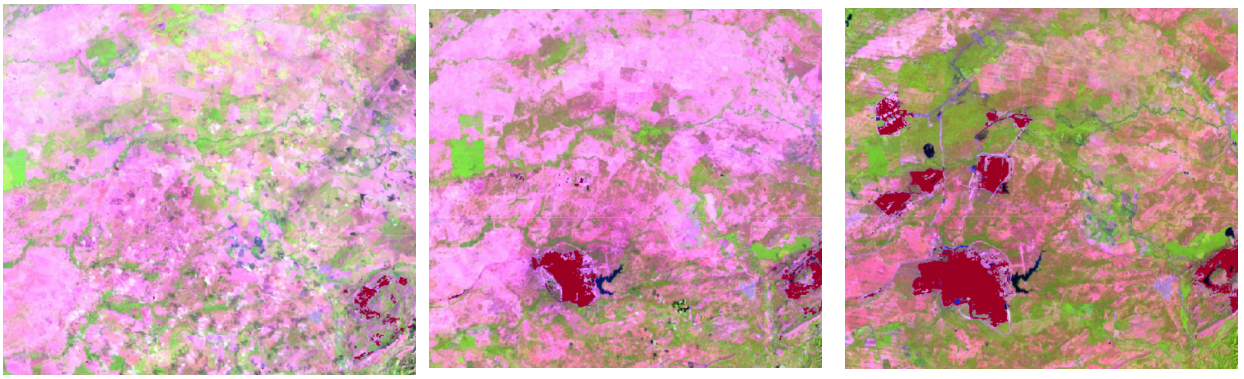
2010	2014	2018



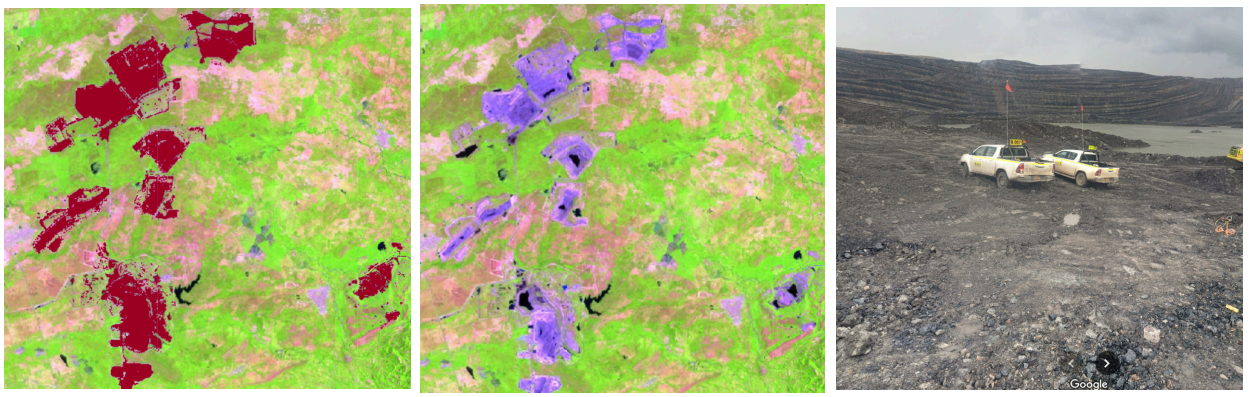
2025	2025. RGB 654 L8	Landscape

Figure 17: Mining classification results for part of the department of Chocó.
Itsmina - Chocó, Scale: 1 km, Coordinates: -76.72422 W, 5.20808 N

Cesar.



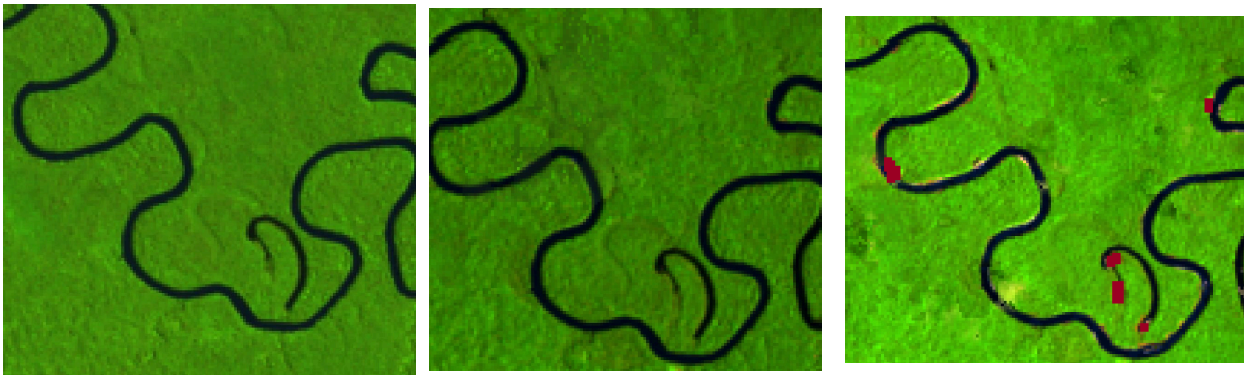
1990	2000	2010



2025	2025. RGB 654 L8	Landscape

Figure 18: Results of the classification of mining activity in a portion of the department of Cesar.
El Paso - Cesar, Scale: 5 km, Coordinates: -73.4725 W, 9.6784 N.

Amazonía.



1985	2021	2023



2025	2025. RGB 654 L8	Paisaje

Figure 19: Mining classification results for part of the Amazonía Department. Brazil-Colombia border, Scale: 200 m, Coordinates: -69.58185 W, -2.12594 S.

4. References

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Congreso de Colombia. (2001). LEY 685 DE 2001 Código de Minas. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=9202

IDEAM. (2010). Leyenda Nacional de Coberturas de la Tierra. Metodología CORINE Land Cover adaptada para Colombia Escala 1:100.000.