Mining – Appendix

Collection 5

Version 1

General coordinator
Pedro Walfir Martins e Souza Filho

Team
Alexandre Filho
Cesar Guerreiro Diniz
Luis Waldir Rodrigues Sadeck
Luiz Cortinhas Ferreira Neto
Maria Luize Silva Pinheiro

1 Overview

Today, Brazil is among the five largest producers of iron ore, niobium, bauxite, and manganese in the world (Bray, 2020), exporting varied mineral inputs, with a high level of purity and internationally recognized quality.

Despite its low representativeness in area, as it is a rare coverage class, the national trend associated with this land use shows a frank expansion, jumping from ~48,000 hectares in 1985 to ~70,000 hectares in 2019. One value ~1.5 times higher than reported in 1985. Three biomes together are responsible for much of the country's mined area, ~85%: Amazon (23 thousand hectares - 32%), Atlantic Forest (~38 thousand hectares - 54 %), Cerrado (~15 thousand hectares - 21%).

In comparison to Collection 4, the Collection 5 mining mapping presents a severe methodological shift. The sample training strategy was modified, leaving the use of "Concession Areas" behind and started to use visually adjusted grids to guide the distribution of training samples over the mining sites. The whole classification process is described below, Figure 1.

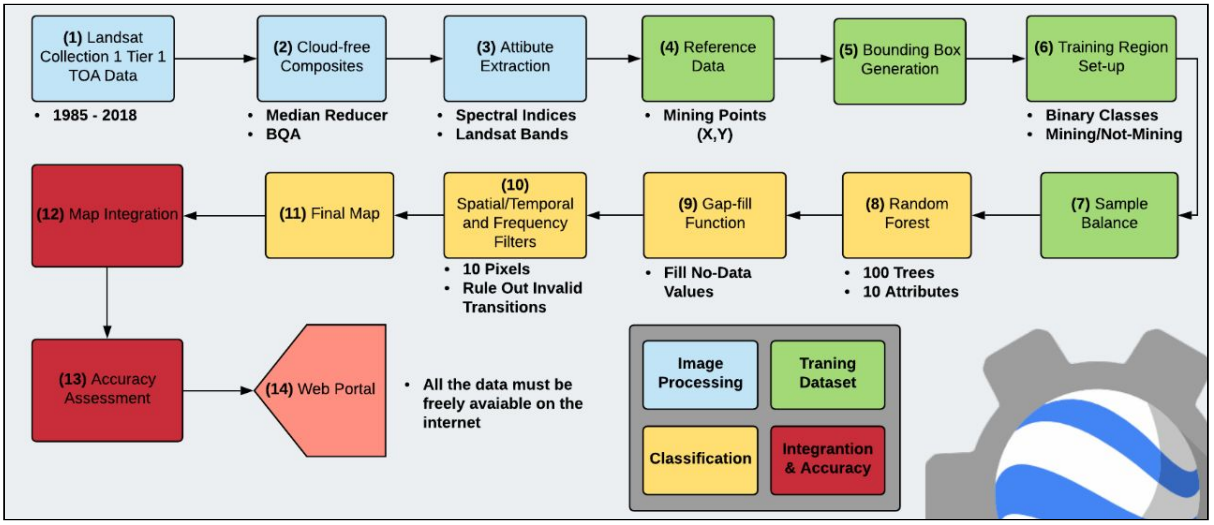

Figure 1 - Data-flow diagram. All processing and analysis occurred inside the Google Earth Engine (GEE) platform. Steps related to image processing are in blue. The steps in green are related to the sample design. Classification procedures are in yellow. The accuracy assessment phase is in red and, finally, the data availability is in salmon. BQA denotes Band Quality Assessment.

2 Landsat image mosaics

The classification of the cross-cutting theme "Mining" used Landsat mosaics that differed from the mosaics used to classify the natural vegetation of the Brazilian biomes. The Mining mosaics were cropped to comprise areas where mining sites are known to exist. These Landsat mosaics are the third generation of the methodology developed specifically for this cross-cutting theme.

2.1 Definition of the temporal period

The Mining annual cloud-free composites are generated by calculating the median pixel value from January 1 to December 31 of each year.

2.2 Mosaic Subsets

The definition of the mining mosaic subset was made based on a set of reference maps highlighting the 300+ most crucial mine site in Brazil. The mining geographical references were extracted from the National Mining Agency – ANM (www.anm.gov.br) and from the Chamber of Commerce and Industry Brazil-Germany – AHK – (https://www.ahkbrasilien.com.br/).

The references formerly constituted georeferenced dots (location points), centered on an area inside or near to a mining site. The dots were visually analyzed and converted to bounding boxes, covering the entire mining site. Figure 2 shows the location of 300+ mining, as well as a zoom-in one of its adjusted bounding boxes.

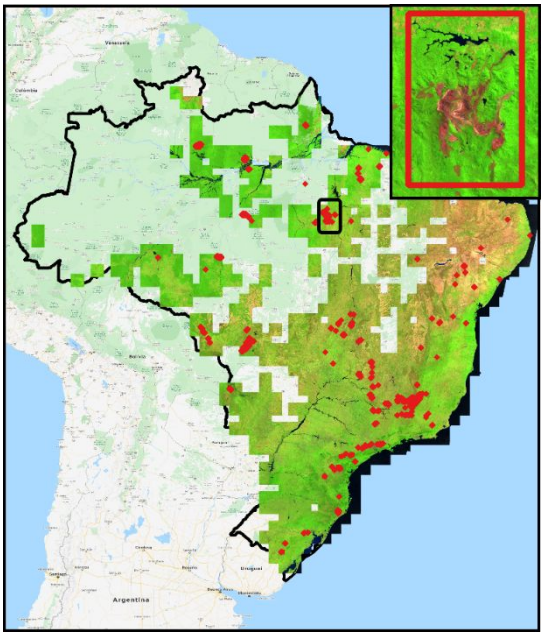

Figure 2 – In red (dots), the location of 300+ mining. In the top-right corner, zoom in on one of the visually adjusted bounding boxes.

2.3 Image selection

The cloud/shadow removal script takes advantage of the Landsat Collection 1 Level-1 QA band and the GEE median reducer. In Collection 1 Tier 1 data, each pixel in the QA band contains unsigned integers values that represent a particular surface, atmospheric, and sensor conditions that may affect the overall usefulness of a given pixel. When effectively used, QA values can improve data integrity by indicating which pixels might be affected by instrument artifacts or subject to cloud contamination (USGS, 2017). In conjunction with that, GEE can be instructed to pick the median pixel values in a stack of images. By doing so,

GEE rejects values that are too high (e.g., clouds) or too low (e.g., shadows) and picks the median pixel value in each band, over time, Figure 3.

Figure 3 - Left, Collection 2 "cloud-free composite". Right, Collection 3 "cloud-free.

3 Classification

The automatic classification of the Landsat mosaics was performed entirely on the Google Earth Engine platform, based on the Random Forest classifier (Breiman, 2001).

3.1 Classification scheme

Each class was classified separately, in a binary fashion. For this matter, the classification process was carried out, always considering only two possible classes for each pixel, the interest class (mining) or non-interest class (non-mining).

For the supervised classification of the Landsat mosaics, we selected training points based on the reference map and the MapBiomas 4.0 Collection. The details of the parameters used in the Random Forest classifier, the reference maps used for each interest class, and the feature space produced for each classification are presented in the sections to follow.

3.2 Mining Feature Space

A Landsat median composite, from January 1 to December 31, was clipped by the "Reference grid vector," obtained through visual analysis of the reference data. Subsequently, spectral indices were also used to help within the identification of surface mineral activity and added to the set of classification inputs, see Table 1.

Table 1 – Indices and Landsat Bands used to classification

Band Name	Expression	Reducer	Reference
NDSI	(GREEN – SWIR2) / (GREEN + SWIR2)	Median	Deng, 2015
NDVI	(NIR - RED) / (NIR + RED)	Median	Tucker, 1979
SWIR2	Landsat Swir2 band median value	Median	USGS
SWIR1	Landsat Swir1 band median value	Median	USGS
NIR	Landsat Nir band median value	Median	USGS
RED	Landsat Red band median value	Median	USGS
GREEN	Landsat Green band median value	Median	USGS

3.3 Classification algorithm, training samples, and parameters

When lacking a reference map that precisely matches the annual mosaic to be classified, the closest possible reference was used. When no reference map was available, then the results of the classification of the previous year were used to guide the training of the subsequent one. Table 2 shows the Random Forest parameters used to classify each one of the years.

Table 2 - Random Forest parameters used to classify each one of the years.

Parameter	Value
Number of trees	100
Number of points	100000
Number of Variables	20 (Coastal Zone) and 14 (Mining)
Classes	2 (binary classification)

As in any supervised method, the Random Forest classifier needs to rely on a training dataset. For the mining recognition, the training data was obtained from MapBiomas Collection 4.0 and the reference sites downloaded from the Chamber of Commerce and Industry Brazil-Germany – AHK – (https://www.ahkbrasilien.com.br/), posteriorly converted to grids, Figure 4.

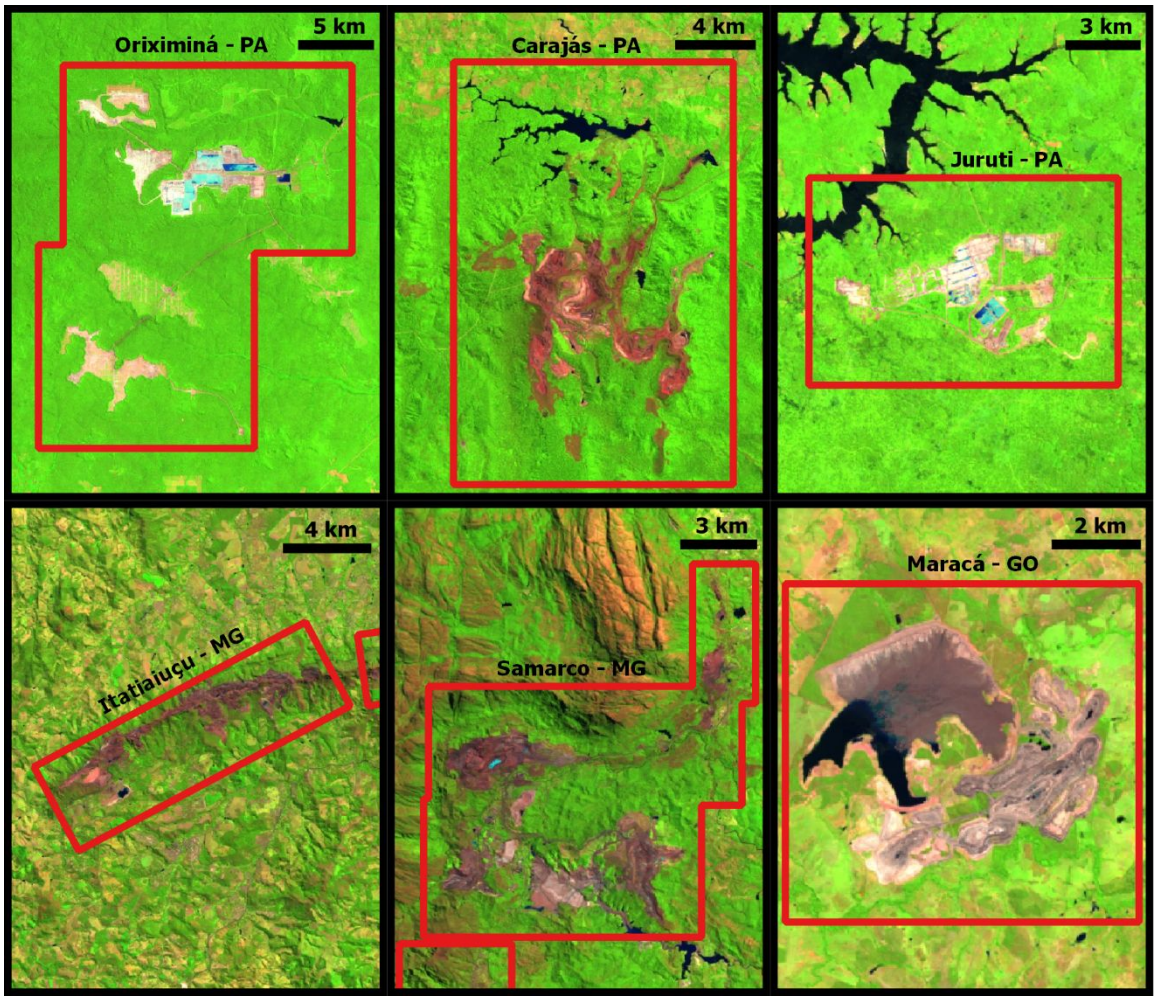

Figure 4 – The training samples were collected only inside of the Reference Grids (red polygons). From left-to-right and top-to-bottom, the mine grids of Oriximina-PA, Carajas-PA, Juruti-PA, Itatiaiuçu-MG, Samrco-MG, Maracá-GO.

4 Post-classification

Due to the pixel-based nature of the classification method and the very long temporal series, a chain of post-classification filters was applied. The post-classification process includes the application of a gap-fill, a temporal, a spatial, and a frequency filter.

4.1 Gap-Fill filter

The chain starts by filling in possible no-data values. In a long time-series of severely cloud-affected regions, such as tropical coastal zones, it is expected that no-data values may populate some of the resultant median composite pixels. In this filter, no-data values ("gaps") are theoretically not allowed and are replaced by the temporally nearest valid classification. In this procedure, if no "future" valid position is available, then the no-data value is replaced by its previous valid class. Up to three prior years can be used to fill in persistent no-data positions. Therefore, gaps should only exist if a given pixel has been

permanently classified as no-data throughout the entire temporal domain. To keep track of pixel temporal origins, a mask of years was built, as shown in Figure 5.

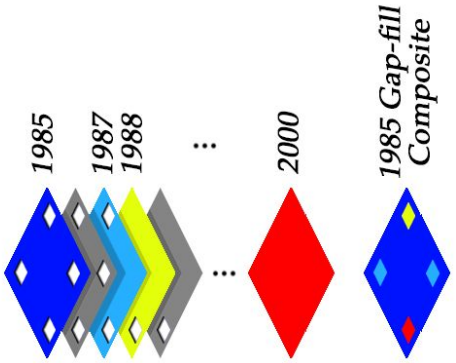

Figure 5 – Gap-filling mechanism. The next valid classification replaces existing no-data values. If no "future" valid position is available, then the no-data value is replaced by its previous valid classification, based on up to a maximum of three (3) prior years. To keep track of pixel temporal origins, a mask of years was built.

4.2 Temporal filter

After gap filling, a temporal filter was executed. The temporal filter uses sequential classifications in a 3-year unidirectional moving window to identify temporally non-permitted transitions. Based on a single generic rule (GR), the temporal filter inspects the central position of three consecutive years ("ternary"), and if the extremities of the ternary are identical but the center position is not, then the central pixel is reclassified to match its temporal neighbor class, as shown in Table 3.

Table 3 *- The temporal filter inspects the central position of three consecutive years, and in cases of identical extremities, the center position is reclassified to match its neighbor. T1, T2, and T3 stand for positions one (1), two (2) and three (3), respectively. GR means "generic rule," while Mi and N-Mi represent mining and non-mining pixels.*

Rule	Input (Year)			Output		
	T1	T2	T3	T1	T2	T3
GR	Mi	N-Mi	Mi	Mi	Mi	Mi
GR	N-Mi	Mi	N-Mi	N-Mi	N-Mi	N-Mi

4.3 Spatial filter

Next, a spatial filter was applied. To avoid unwanted modifications to the edges of the pixel groups (blobs), a spatial filter was built based on the "connectedPixelCount" function. Native to the GEE platform, this function locates connected components (neighbors) that share the same pixel value. Thus, only pixels that do not share connections to a predefined number of identical neighbors are considered isolated, as shown in Figure 6. In this filter, at least ten connected pixels are needed to reach the minimum connection value. Consequently, the minimum mapping unit is directly affected by the spatial filter applied, and it was defined as 10 pixels (~1 ha).

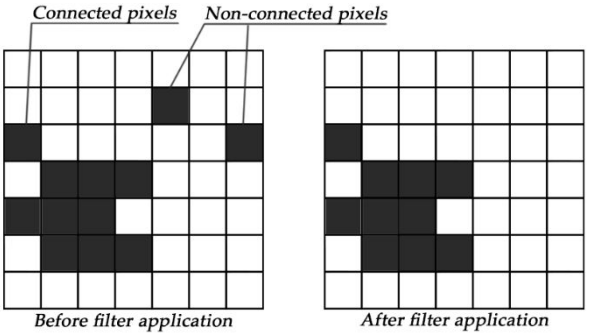

Figure 6 – *The spatial filter removes pixels that do not share neighbors of identical value. The minimum connection value was 10 pixels.*

4.4 Frequency filter

The last step of the filter chain is the frequency filter, as shown in Figure 7. This filter takes into consideration the occurrence frequency of a given class throughout the entire time series. Thus, all class occurrences with less than 10% temporal persistence (3 years or fewer out of 33) are filtered out and incorporated into the non-class binary. This mechanism contributes to reducing the temporal oscillation of the classification signal, decreasing the number of false positives, and preserving consolidated class pixels.

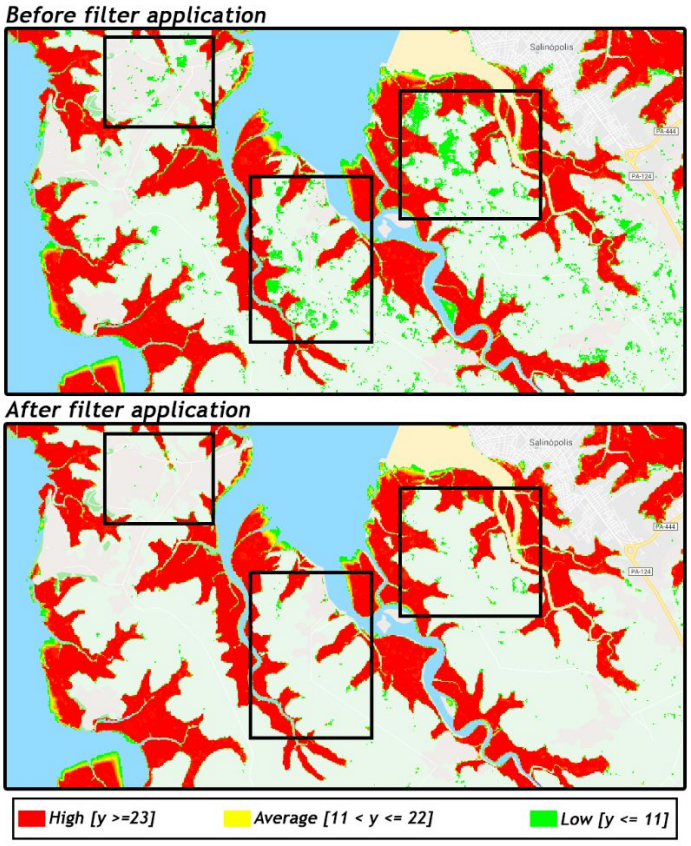

Figure 7 – Red, yellow and green represent classified pixels with high (23 or more years, y >=23), average (between 11 and 22 years, 11 <= y <= 22), and low (ten years or less, y < 11) occurrence frequencies, respectively. The top image shows classified pixels before applying the frequency filter. The bottom image shows pixels after using the frequency filter. The

black boxes are centered on areas that have been significantly affected by the filter. Note that all classification occurrences with less than 10% temporal persistence (3 years in 33 possible years) were filtered out.

4.5 Integration with biomes themes

After the application of the filter-chain, the cross-cutting themes and the biomes data were integrated. This integration was guided by a set of specific hierarchical prevalence rules (Table 4). As output of this step, a final vegetation cover/land use map for each chart of the MapBiomas project.

Table 4 - Prevalence rules for combining the output of digital classification with the cross-cutting themes in Collection 5.

ID	COLLECTION 5 CLASSES	PREVALENCE ID
1	**1. Forest**	-
2	1.1. Natural Forest	-
3	1.1.1. Forest Formation	17
4	1.1.2. Savanna Formation	18
5	1.1.3. Mangrove	3
9	1.2. Forest Plantation	7
10	**2. Non-Forest Natural Formation**	-
11	2.1. Wetland	19
12	2.2. Grassland Formation	20
32	2.3. Salt Flat	5
29	2.4. Rocky Outcrop	13
13	2.3. Other Non-Forest Natural Formation	21
14	**3. Farming**	-
15	3.1. Pasture	14
18	3.2. Agriculture	-
19	3.2.1 Annual Crop	12
39	3.2.1.1 Soy Bean	9
20	3.2.1.2 Sugar Cane	8
36	3.2.1.3 Perennial Crop	11
21	3.3. Mosaic of Agriculture and Pasture	22
22	**4. Non-Vegetated Area**	-
23	4.1. Beach and Dune	2
24	4.2. Urban Infrastructure	6

30	4.3. Mining	1
25	4.4. Other Non-Vegetated Area	15
26	**5. Water**	-
33	5.1. River, Lake and Ocean	16
31	5.2. Aquaculture	4
27	6. Non Observed	-

References

Bray, E. L. (2020). Mineral Commodity Summaries - Bauxite and alumina. *U.S. Geological Survey, Mineral Commodity Summaries*. https://pubs.usgs.gov/periodicals/mcs2020/mcs2020-bauxite-alumina.pdf

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

USGS. (2017). *LANDSAT COLLECTION 1 LEVEL 1 PRODUCT DEFINITION*. 26. https://landsat.usgs.gov/sites/default/files/documents/LSDS-1656_Landsat_Level-1_Product_Collection_Definition.pdf.