Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogra
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

An Algorithm for Maximum Concordance to Harmonize Legends of Land Use and Land Cover Maps

Um Algoritmo de Máxima Concordância para Harmonizar Legendas de Mapas de Uso e Cobertura da Terra

Sabrina G. Marques[1], Pedro R. Andrade[2], Aline Soterroni[3] e Maria Isabel Sobral Escada[4]

[1]National Institute for Space Research, São José dos Campos, Brazil. sabrina.marques@unesp.br.
ORCID: https://orcid.org/0000-0001-8095-8689
[2]National Institute for Space Research, São José dos Campos, Brazil. pedro.andrade@inpe.br.
ORCID: https://orcid.org/0000-0001-8675-4046
[3]Nature-based Solutions Initiative, Department of Biology, University of Oxford, Oxford, UK. aline.soterroni@biology.ox.ac.uk
ORCID: https://orcid.org/0000-0003-3113-096X
[4]National Institute for Space Research, São José dos Campos, Brazil. isabel.escada@inpe.br.
ORCID: https://orcid.org/0000-0002-5822-8265

Received: 03.2024 | Accepted: 10.2024

Abstract: Land use and land cover maps (LULC) are abstractions of the physical space of a chosen region. Comparison of LULC maps is essential to understand landscape dynamics, alteration patterns, and environmental implications. This article has the objective of propose an algorithm for harmonizing LULC maps based on the spatial distribution of their classes and applies it in a case study to harmonize the maps of Brazil's National Inventory of Anthropogenic Emissions by Sources and Removals of Greenhouse Gases (Fourth Version) and MapBiomas (Collection 7) based on their spatial distribution of LULC classes. The purpose of this paper is to compute the agreement between two initiatives. Furthermore, the results highlight the classes and areas of potential inconsistency or ambiguity, allowing to identify and correct discrepancies, proposing a harmonized legend between then. At the national level, we reached maximum agreement 81% between the two maps. Of the 44 equivalences, the algorithm accurately recognized 36 of the connections between the classes. At the biome level, the algorithm achieved its highest concordance within the Amazonia biome, surpassing Brazil's level by 11%, mainly due to the size and homogeneity of the forest classes. In biomes with a predominance of nonforest vegetation, an increased confusion was observed among the classes 'Grassland', 'Pasture', and 'Forest' was observed between the maps, especially in Pampa and Caatinga.
Keywords: Harmonization algorithm. National Inventory. MapBiomas. Land use and land cover maps.

Resumo: Mapas de uso e cobertura da terra (UCT) são abstrações do espaço físico de uma região escolhida. Comparar esses mapas é essencial para entender a dinâmica da paisagem, padrões de alteração e implicações ambientais. Este artigo tem o objetivo de propor um algoritmo para harmonizar mapas de UCT baseado na distribuição espacial de suas classes e aplicá-lo em um estudo de caso para harmonizar os mapas do Inventário Nacional de Emissões Antrópicas por Fontes e Remoções de Gases de Efeito Estufa do Brasil (Quarta Versão) e MapBiomas (Coleção 7) com base na distribuição espacial das classes de uso e cobertura da terra. Esta investigação visa calcular a concordância entre as duas iniciativas. Além disso, os resultados destacam as classes e áreas de potencial inconsistência ou ambiguidade, permitindo identificar e corrigir discrepâncias, propondo uma legenda harmonizada entre elas. Em nível nacional, alcançamos uma concordância máxima de 81% entre os dois mapas. Das 44 equivalências, o algoritmo reconheceu com precisão 36 dos mapeamentos entre as classes. No nível dos biomas, o algoritmo alcançou sua maior concordância dentro do bioma da Amazônia, superando o nível do Brasil em 11%, principalmente devido ao tamanho e homogeneidade das classes de floresta. Em biomas com predominância de vegetação não florestal, foi observada uma confusão aumentada entre as classes 'Campo', 'Pastagem' e 'Floresta' entre os mapas, especialmente no Pampa e Caatinga.
Palavras Chave: Algoritmo de harmonização. Inventário Nacional. MapBiomas. Mapas de uso e cobertura da terra.

1 INTRODUCTION

The Earth, comprised of a complex network of ecosystems, has been a subject of study and engagement since the beginning of human civilization. The relationship between humans and their environment has significantly shaped cultural, social, and economic practices. However, in the last decades, there has been an observed reversal in this relationship. With the expansion of civilization and the advancement of technology, humanity has transitioned from being mere inhabitants to a dominant force that actively changes and modifies the environment to meet its needs (VERBURG et al., 2013; PIELKE SR. et al., 2011; ELLIS et al., 2013). In the context of climate change, the Agriculture, Forestry, and Other Land Use (AFOLU) sector emerges as a critical component. According to the latest report from the Intergovernmental Panel on Climate Change (IPCC), this sector was responsible for approximately 22% of human-made greenhouse gas (GHG) emissions in 2019. Therefore, precise monitoring through LULC maps is necessary to compile inventories of GHG emissions and removals (SHUKLA et al., 2019).

LULC maps represent the physical space of a chosen region through abstractions that describe the covered areas. They allow a systematic categorization of geographical regions based on specific human uses and natural characteristics. These categorizations represent the spatial distribution of human activities, serving as indicators of human-made pressures on natural ecosystems (JANSEN; GROOM; CARRAI, 2008). In addition, the analytical and symbolic capabilities of LULC maps are indispensable tools in the scientific field. They not only document the current state of the environment but also, when employed for comparisons, provide a perspective for examining human-induced changes over time and their ecological and climatic consequences. As a result, they play a critical role in forming evidence-based decision-making regarding the management and conservation of natural resources (VERBURG et al., 2013).

Comparing LULC maps is a valuable resource for environmental and geographical studies. Sequentially overlaying these maps reveals environmental changes and transformation trends, providing information about deforestation rates, urban expansion, changes in water bodies, and other critical aspects. This comparative analysis is essential for evaluating the impacts of land-use policies and projecting future scenarios (ELLIS et al., 2013).

In Brazil, several initiatives use open data to produce LULC maps, such as MapBiomas (MAPBIOMAS BRASIL, 2021), TerraClass (INPE, 2019), PRODES (INPE, 2021), IBGE (IBGE, 2019), and the National Communications to the United Nations Framework Convention on Climate Change (UNFCCC) (BRASIL, 2021). Although each of these initiatives has different objectives, interests and mapping standards, there are differences in the maps produced for the same area, some of which might be related to the nature of the input data or the methodology developed. This limits the compatibility and comparability of these data. Different maps might have been produced at different intervals, and aggregating this information can allow for more granular time-series analyses.

Harmonization, in the context of LULC, is the process in which the similarities between the definitions of the existing land cover class are emphasized and inconsistencies are reduced (HEROLD et al., 2006). The goal in this case is to have different datasets “harmonized” so that a direct comparison can be made between them. Harmonization does not necessarily eliminate all differences but should eliminate the main inconsistencies.

When the legends are harmonized and equivalences between the class maps are established, it becomes possible to accurately analyze the similarities and differences between these maps. This harmonization allows for the identification of areas where different types of classes are present, as well as zones of concordance, where both maps display the same vegetation type according to the harmonized legend. In addition, it enables the determination of the total concordance between maps, where there is complete alignment in classification across all areas.

Harmonization of these LULC maps is challenging due to the different methods, classification systems, and legends adopted by each project. These differences may stem from the choice of satellite imagery, classification methods, field support data, among others. In addition to technical discrepancies, there are practical challenges, such as differences in resolution, projection, and coordinate systems. In addition, harmonizing legends presents excellent challenges due to their nature. Differences in class naming, changes in class definitions, and the

addition or deletion of classes in maps covering the same region at different times or in different initiatives create difficulties in separating actual changes over time from differences in category definitions. Thus, establishing equivalencies between classes from different maps is vital for effective comparisons.

Typically, comparing LULC maps involves constructing a key based on the semantics of each category. Frequently, categories are grouped into broader classifications to minimize discrepancies or are excluded because they lack explanations of similarity. Some classification systems can also standardize the classification scheme and ensure that the maps are rendered in a way that facilitates comparison (CAPANEMA et al., 2019; DI GREGORIO, 2016; REIS et al., 2018; NEVES et al., 2020).

While traditional methods primarily start from the semantics of LULC classes, examining the spatial distribution of categories can yield additional insights. The main objective of this study is to propose an automated legend harmonization algorithm to achieve the highest possible concordance between two maps. The main assumption is that harmonization, which leads to the highest concordance, is a robust initial reference for map comparison, as it can *a priori* indicate potential inconsistencies between classes. Consequently, the proposed algorithm can be an initial automated step in the harmonization process between two maps. Instead of relying solely on the semantics of the classes to initiate the class mapping process, using the proposed algorithm, it is possible to create a legend based on the results already provided by the algorithm.

Among the existing studies in the literature on the harmonization of LULC maps, there is a predominance of methodologies that focus on the semantics of classes for the harmonization process (CAPANEMA et al., 2019; REIS; et al., 2017; DI GREGORIO, 2016). This work stands out for its approach that uses the spatial distribution of classes as the basis for the automatic harmonization of legends. Unlike methodologies that rely exclusively on the semantics of legends, which require a lengthy manual analysis of each map's classes and depend on project-provided descriptions that may not be clear enough, the proposed approach seeks the maximum concordance between maps, providing a reference for comparison. It serves as an initial automated step in the class mapping process.

As a case study, we perform an analysis at the biome level and on a national scale between the MapBiomas and the National Inventory. Ultimately, after generating the mappings through the algorithm, we performed a semantic analysis of the classes and created a concordance map, followed by the development of a harmonized legend that aligns the classifications of the two datasets.

2 HARMONIZATION ALGORITHM

This section presents the proposed algorithm for harmonizing legends from two LULC maps. The algorithm provides information that helps the user to map between classes from two land use maps. It computes a maximum concordance between the two maps, meaning that any class mapping different from the one chosen by the algorithm will result in concordance equal to or lower than the algorithm's result.

The algorithm takes as input two land use maps, M_1 and M_2, and returns to the user a proposed legend based on the identified concordances. M_1 has classes $x_1, x_2, \cdots, x_m$, and M_2 has classes $y_1, y_2, \cdots, y_n$.The algorithm requires that the spatial representations of both maps are compatible, that is, they must be in the same projection, in matrix format, with the same number of rows and columns and spatial resolution. If not, a preprocessing step is required for these maps, which is not performed by the algorithm.

The algorithm consists of three steps. The first step involves creating a cross-tabulation matrix, A. In this matrix, each entry a_{ij} represents the number of pixels where class x_i from map M_1 is concordant with class y_j from map M_2. It is of the form:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \; A \in \mathbb{N}^{m \times n} \tag{1}$$

In the second step, the concordances are determined by row and by column of matrix A, forming two sets, C_R and C_C. These sets are defined using the following functions:

For each row x_i in matrix A, we define the concordance set C_R as:

$$C_R(x_i) = \left\{ y_k \mid n_r(a_{ik}) = \max_{p=1,\ldots,m} \{n_r(a_{ip})\} \right\} \tag{2}$$

where

$$n_r(a_{ik}) = \frac{a_{ik}}{\sum_{p=1}^{m} a_{ip}} \tag{3}$$

and $n_r(a_{ik})$ represents the proportion of pixels in row i that are concordant with column k.

Similarly, for each column y_j in matrix A, we define the concordance set C_C as:

$$C_C(y_j) = \left\{ x_k \mid n_c(a_{kj}) = \max_{p=1,\ldots,n} \{n_c(a_{pj})\} \right\} \tag{4}$$

where

$$n_c(a_{kj}) = \frac{a_{kj}}{\sum_{p=1}^{n} a_{pj}} \tag{5}$$

and $n_c(a_{kj})$ represents the proportion of pixels in column j that are concordant with row k.

In these expressions:

- n is the number of rows in matrix A;
- m is the number of columns in matrix A;
- a_{ij} is the number of concordant pixels between class x_i and class y_j;
- $\sum_{p=1}^{m} a_{ip}$ is the total number of pixels in row i;
- $\sum_{p=1}^{n} a_{pj}$ is the total number of pixels in column j.

The set C_R contains m pairs, $(x_i, c_r(x_i)), i \in \{1, \cdots, m\}$ formed by the classes of M_1 and their respective classes of M_2 that presented the highest concordances according to matrix A. Analogously, set C_C is formed by n pairs, $(c_c(y_i), y_i), i \in \{1, \cdots, n\}$, representing the concordance of the classes of M_2 in relation to the classes of M_1, according to A. After the formation of these two sets, the last step of the algorithm consists of producing the harmonized legend, given by:

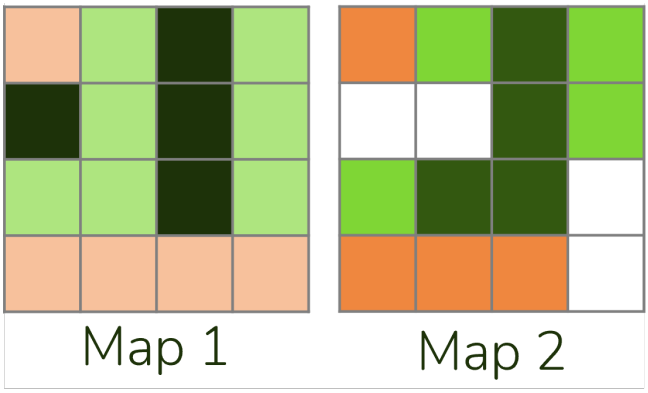
$$C_R \cup C_C \tag{6}$$

containing all matches obtained by the functions c_r and c_c. This set can have a maximum size equal to the sum of the class quantities from both maps and a minimum size equal to the number of classes on the map with the most classes. The union of the sets C_R and C_C encompasses three possible mappings between the classes of maps. The first is the simplest, occurring when there is a mapping by row from class x_i to class y_j, and there is also a mapping by columns between y_j and x_i in the second mapping. Thus, the pair (x_i, y_j) is considered in the harmonization.

The first case is when the mapping from class crs_i to class crs_j occurs in the row and column harmonization. The second case occurs when the mapping of a class exists only in one of the harmonization. For example, when there is a mapping by row from class x_i to class y_j, but class x_i is not mapped in the column harmonization. In this case, the pair (x_i, y_j) is considered in the harmonization. In the last case, there is a mapping by row from class x_i to class y_j, but in the column mapping, class y_j is mapped to class x_k, $k \neq i$, indicating an inconsistency. In any case, the pairs (x_i, y_j) and (x_k, y_j) will be considered in the harmonized legend.

Consider the following example to better contextualize the algorithm's operation. In Figure 1, we present two representations of maps, Map 1 and Map 2.

Figure 1 – Example of hypothetical maps for illustrative purposes.



Source: Author's production.

Each map consists of a grid of squares, where each square represents a pixel. Map 1 has three classes: x, y, and z, while Map 2 has four classes: a, b, c, and d. It's important to note that both maps have the same number of rows and columns and spatially represent the same area. Therefore, the algorithm will overlay them to perform the comparison.

When providing these maps as input to the algorithm, a cross-tabulation table is generated between the maps. After generating the matrix, the maximum values are identified by row and by column to determine the concordance between the classes. In Figure 2, we present the cross-tabulation matrix obtained between the two maps, highlighting in grey the maximum values per row and per column. For example, in the first row, referring to class x of Map 1, the maximum value identified is 100, which corresponds to class a of Map 2. In the same row, the value 20 is also highlighted in grey, which is the maximum value in the column corresponding to class d of Map 2.

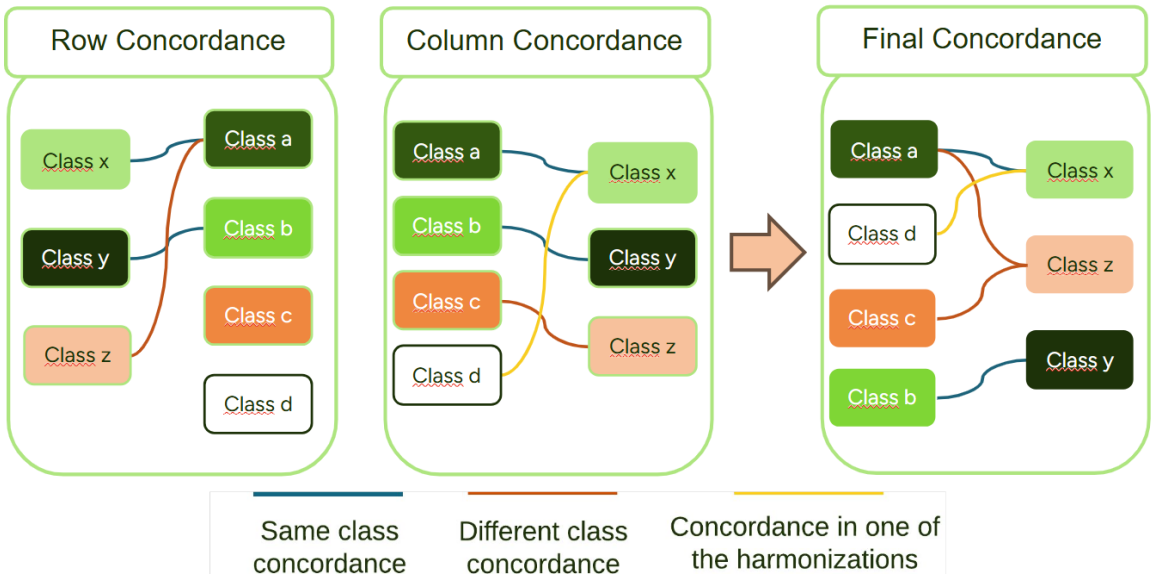
Figure 2 – Example of cross-tabulation between the maps.

	Classe a	Classe b	Classe c	Classe d
Classe x	100	30	2	20
Classe y	15	50	10	3
Classe z	30	20	20	15

Source: Author's production.

In Figure 3, the concordances obtained by row and by column are presented separately, and subsequently, the final concordance obtained from the union of both is shown. Here, it is possible to observe three distinct cases: classes a and x were mapped in both concordances, while class z was mapped as class a in the row concordance and as class c in the column concordance. Additionally, class d was only present in the column concordance, being mapped as class x.

Figure 3 – Concordances obtained by row and by column and the final concordance.



Source: Author's production.

Figure 4 presents the table with the agreements between the map classes.

Figure 4 – Concordances between map classes.

Map 1	Map 2
Class a	Class x, Class z
Class b	Class y
Class c	Class z
Class d	Class x

Source: Author's production.

The algorithm aims to be the first step in the harmonization process of LULC map legends, providing a proposed legend based on the spatial distribution of the classes in the maps, which delivers the highest possible concordance between them. It can capture subtle nuances in class definitions between different maps, reflecting both unidirectional and bidirectional correspondences in class concordances by examining the row and column correspondences. In addition, it highlights potential inconsistencies or ambiguities, allowing users to identify and fix them manually in the next step.

It is important to emphasize that for the algorithm to perform effectively, both classifications should accurately represent the reality. Otherwise, when most of the obtained maps are incorrect, the entire mapping between classes will need to be done manually based on the semantics of the classes.

In practical terms, the automation provided by the algorithm facilitates the integration of data from different sources, optimizing the efficiency of the process, and minimizing errors that can arise from manual approaches. It is an initial step for mapping classes between maps, and it's up to the user to check if the obtained mappings are coherent or if the legend needs to be adapted. It is necessary to separate what are inconsistencies from what are correspondences between classes. It is worth highlighting that since the legend produced by the algorithm provides the combination with the highest concordance between the maps, any changes will result in a lower concordance.

In the goal to make the proposed algorithm accessible and facilitate its use, a Python package named **GeoMapHarmonizer** was developed. The package automates the legend harmonization process for geospatial data, particularly focusing on LULC maps in GeoTIFF format. To handle large datasets, the package applies techniques like dividing maps into smaller blocks and efficiently consolidating results, making it scalable for Big Data applications. Compatibility checks ensure that both maps have the same projection, resolution, and dimensions before applying the harmonization algorithm. The package is built using Python with libraries such as *pandas*, *numpy*, *sklearn*, and *gdal*. The complete package can be found on the GitHub repository.

3 CASE STUDY

As a case study, we use the legend harmonization algorithm presented to compare two maps: MapBiomas and the Brazilian National Inventory. Subsequent to the application of the algorithm, an evaluation of the harmonizations achieved for each biome will be conducted, estimating the maximum concordance for each biome. The key points of the harmonizations obtained will be emphasized, and a comparison of the results will be carried out.

We compare both maps at national and biome level. As the most recent map for the National Inventory is forb year 2016, this will be the reference year for our case study. The National Inventory map is divided by biomes according to the 2004 biome boundaries defined by IBGE, and these are the boundaries considered for biomes in this study. Next, we present an overview of both initiatives, along with their unique features.

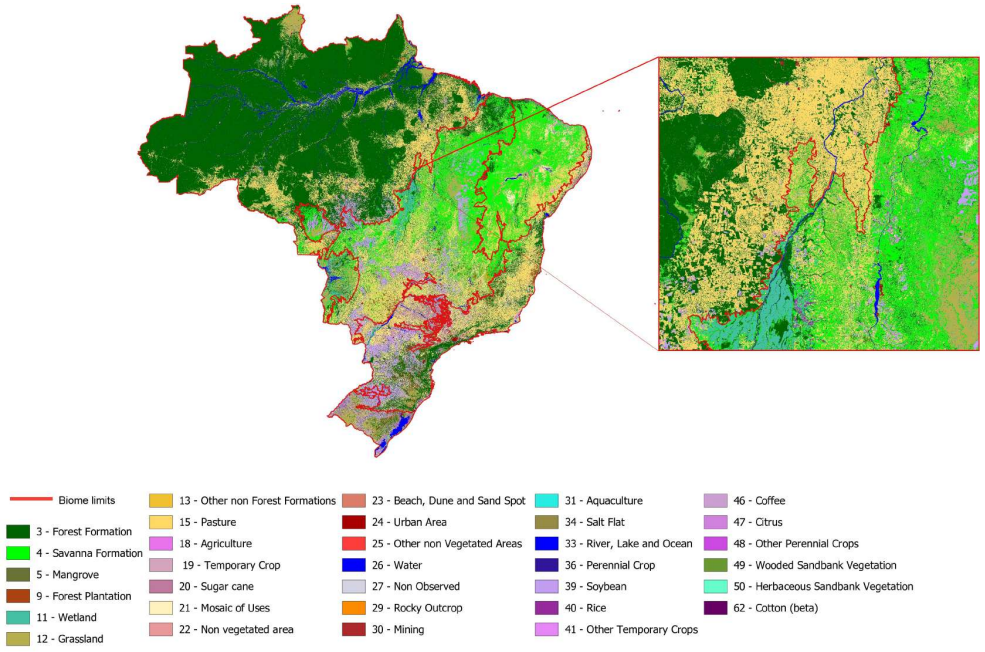
3.1 MapBiomas

The Annual Land Use and Land Cover Mapping Project in Brazil (MapBiomas) originated from an initiative by the Greenhouse Gas Emission Estimates System of the Climate Observatory (SEEG/OC). A collaborative network of co-creators, including NGOs, universities, and technology companies,developed this project. Its objective is to produce annual LULC maps for Brazil using a more cost-effective and rapid methodology (MAPBIOMAS BRASIL, 2021).

The MapBiomas methodology involves a pixel-by-pixel classification of Landsat satellite images with a spatial resolution of 30m. It leverages the Google Earth Engine platform, applies machine learning algorithms, and incorporates insights from a network of local specialists. Data production is categorized by biomes to improve identification of the landscape patterns of the country. The project also addresses transversal themes, including Agriculture, Pasture, Forests, Coastal Zone, Mining, Urban Infrastructure, and other classes encompassing coastal regions (MAPBIOMAS BRASIL, 2022; SOUZA et al., 2020).

The data generated by this project is organized into Collections. In August 2023, MapBiomas released Collection 7.1, with improvements in class classification, presenting LULC maps from 1985 to 2021. MapBiomas offers a detailed legend description, including its equivalence with IBGE, FAO, and IPCC classes. Figure 5 presents an overview of the data from collection 7.

Figure 5 –
LULC map of MapBiomas Collection 7 (Year 2016). The square in red show a small area in more details.

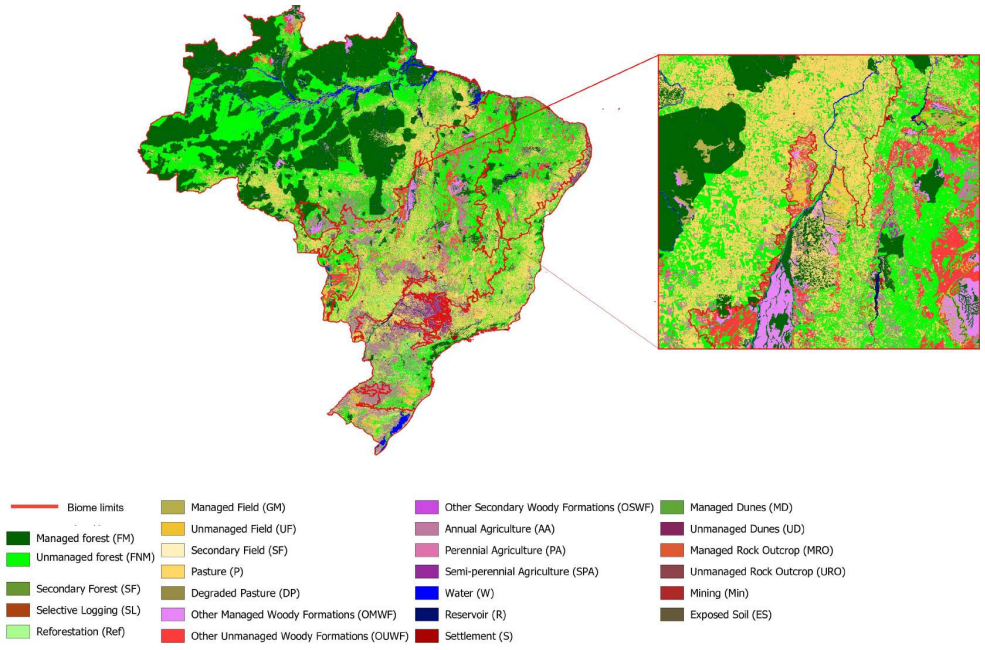


Source: Author's production.

3.2 Brazilian National Inventory

The Brazilian National Inventory of Anthropogenic Emissions by Sources and Removals of Greenhouse Gases (hereafter referred to as National Inventory), whose mission is part of Brazil's Fourth National Communication (4th NC) to the United Nations Framework Convention on Climate Change (UNFCCC). The 4th NC provides anthropogenic emissions of GHGs no longer managed via the Montreal Protocol. The Ministry of Science, Technology and Innovations (MCTI) coordinates and improves the National Inventory. Emission estimates are derived from the LULC map developed by the National Inventory itself. This mapping is produced from images from the Landsat-5/8 satellite and the MSI/Sentinel 2A and 2B sensors at a scale of 1:250,000 and a minimum area of 6 hectares. The adopted methodology includes an object-oriented image segmentation step, a semi-automatic classification and, finally, a visual interpretation (MCTI, 2021; BRASIL, 2021; MCTI, 2020). Figure 6 presents an overview of the produced LULC map of the National Inventory from the 4th NC.

Figure 6 –
LULC map of Brazil's National Inventory (Year 2016). The square in red show a small area in more details.



Source: Author's production.

LULC maps are vector representations that overlay years 1994, 2002, 2005 (only for the Amazon biome), 2010, and 2016, and are divided by means of biomes following the limits established by IBGE in 2004 (MCTI, 2020). LULC maps are available from National Emissions Registry System (SIRENE).

4 RESULTS

Table 1 displays the maximum concordances achieved in each biome[1]. This value is obtained when the harmonized legend produced by the algorithm is applied to both maps, considering the lowest hierarchy level of the classes. Figure 7 shows the harmonization between the National Inventory and MapBiomas, as generated by the algorithm for the entire country. The classes that were identified as equivalent in both row and column harmonization are indicated in blue, while the classes that had different equivalences in row and column are indicated in orange. Lastly, the cases where the class was only identified in one of the harmonizations are indicated in yellow.

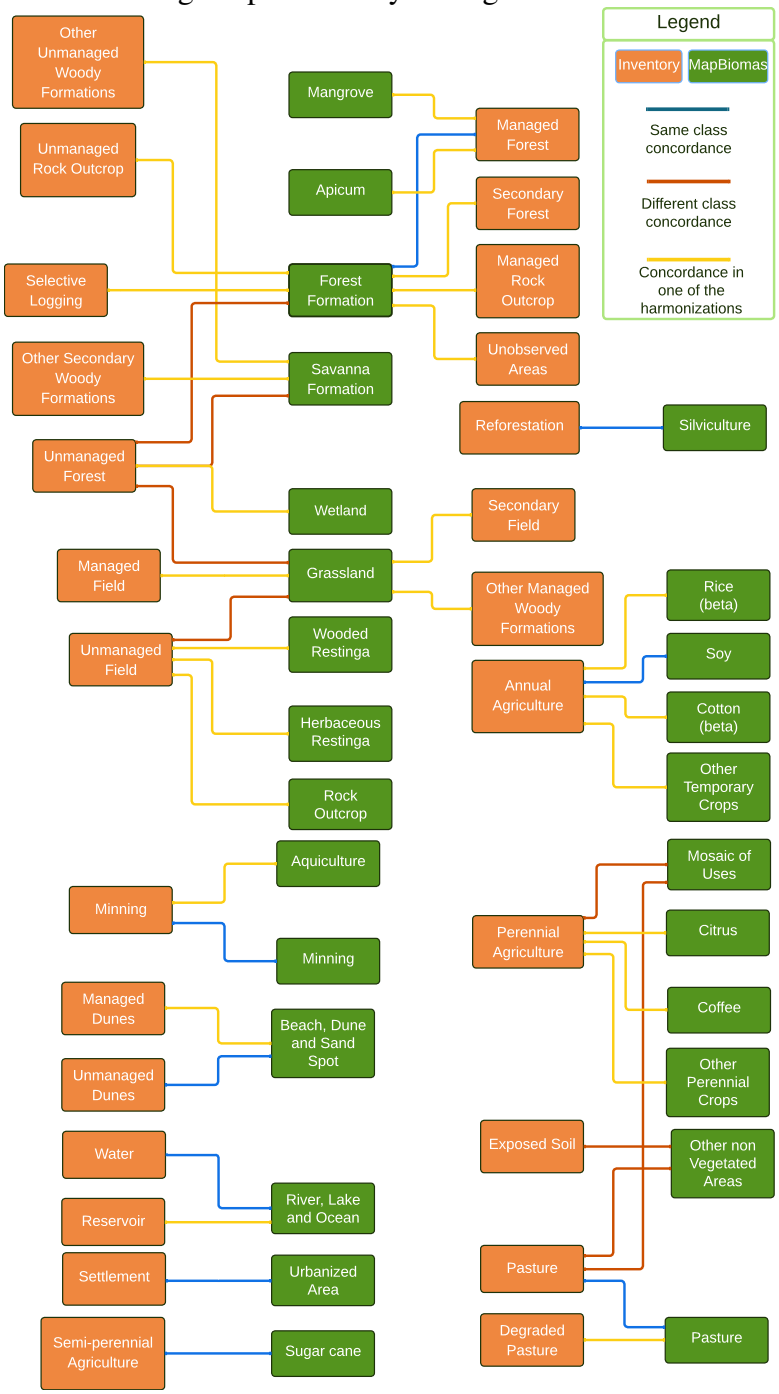
[1] The charts and others harmonizations for the biomes can be viewed in detail on the project's GitHub page.

Table 1 – Maximum concordance obtained in each of the harmonizations and the area of each applied region.

	Area (km^2)	**Maximum concordance (%)**
Amazon	4,253,027	92.39%
Caatinga	843,615	75.27%
Cerrado	1,983,655	74.33%
Atlantic Forest	1,116,119	77.86%
Pampa	203,965	79.32%
Pantanal	150,972	55.51%
Brazil	8,604,500	81.03%

Source: Author's production.

Figure 7 –
Harmonized legend produced by the algorithm for all of Brazil.



Source: Author's production.

The Amazon biome has the largest area among all listed biomes, totaling 4,253,027 km^2, with the highest concordance of 92.39%. Much of this is due to the vast expanse of classes defined as forest, which favors the overlap between them and, consequently, their correct identification. All forest classes of the National Inventory ('Managed Forest/I'[2], 'Unmanaged Forest/I', 'Secondary Forest/I', and 'Selective Logging/I') were mapped to 'Forest Formation/M'. More granular classes, such as 'Beach, Dune, and Sand Spot/M', 'Other non-Vegetated Areas/M', 'Rice (beta)/M', and 'Unmanaged Dunes/I' were incorrectly matched as 'Forest'. In contrast, 'Perennial Agriculture/I' was identified as 'Pasture/M' in the harmonization. The small area of these classes in the Amazon biome leads to a low impact on the overall harmonization. However, it raises points of attention, especially considering classes related to pasture and agriculture being identified as forest since, on a small scale, they can have implications for conservation policies or zoning.

In the Cerrado biome, the harmonization produced a concordance of 74.33%. It is important to highlight that the 'Managed Forest/I' class was associated with the 'Aquaculture/M'. Additionally, the majority of other classes in the National Inventory were predominantly grouped under the 'Savanna Formation/M', with 33% of the entire concordance area. This includes the 'Secondary Field/I' class that was incorrectly associated. In addition, in this biome, another 32% of the concordance area was labeled as 'Pasture/I', with 11% of this total mapping of the 'Mosaic of Uses/M' class as 'Pasture/I'.

For the Caatinga biome, the produced legend achieved a concordance of 75.27%. In this biome, some mappings stood out between the maps: the 'Mosaic of Uses/M' class was incorrectly mapped as 'Unmanaged Forest/I', just as the classes of 'Unmanaged Field/I' and 'Secondary Field/I' were also incorrectly mapped as 'Savanna Formation/M'. The classes for water and agriculture were mostly correctly mapped. From this, it can be inferred that in this biome, the forests and grassland classes showed a lot of confusion between the maps, which might indicate that the semantic definitions of these classes may be very similar between the initiatives, especially when considering that this biome is characterized by shrub and herbaceous vegetation.

In the Atlantic Forest biome, where a concordance of 77.86% was observed, there was a trend to group various classes from the National Inventory into the 'Forest Formation' category of MapBiomas. In this biome, the harmonized legend showcases 41 unique class combinations, a testament to the biome's diversity brought about by the conversion of primary lands. This has led to the formation of a landscape featuring small forest blocks interspersed with converted regions, a stark departure from the Amazon's vast forest blocks that create a more uniform composition. Most of the classes do not have the same harmonization by row and column. It should be noted that classes such as 'Managed Field/I' and 'Secondary Field/I' were labeled as 'Forest Formation/M', along with the 'Managed Dunes/I' class. The 'Herbaceus Restinga' from MapBiomas was identified as 'Pasture' from the National Inventory.

The Pampa biome presents a 79.32% concordance between the two maps. Most of the classes in the National Inventory were labeled as 'grasslands, which could suggest that MapBiomas overestimates the grasslands classes in this region, given that 15% of the entire biome was labeled by the pair 'Pasture/I' and 'Grassland/M'. This also happened with Unmanaged Forest/I, where 8% of the total area was labeled as 'Grassland/M'.

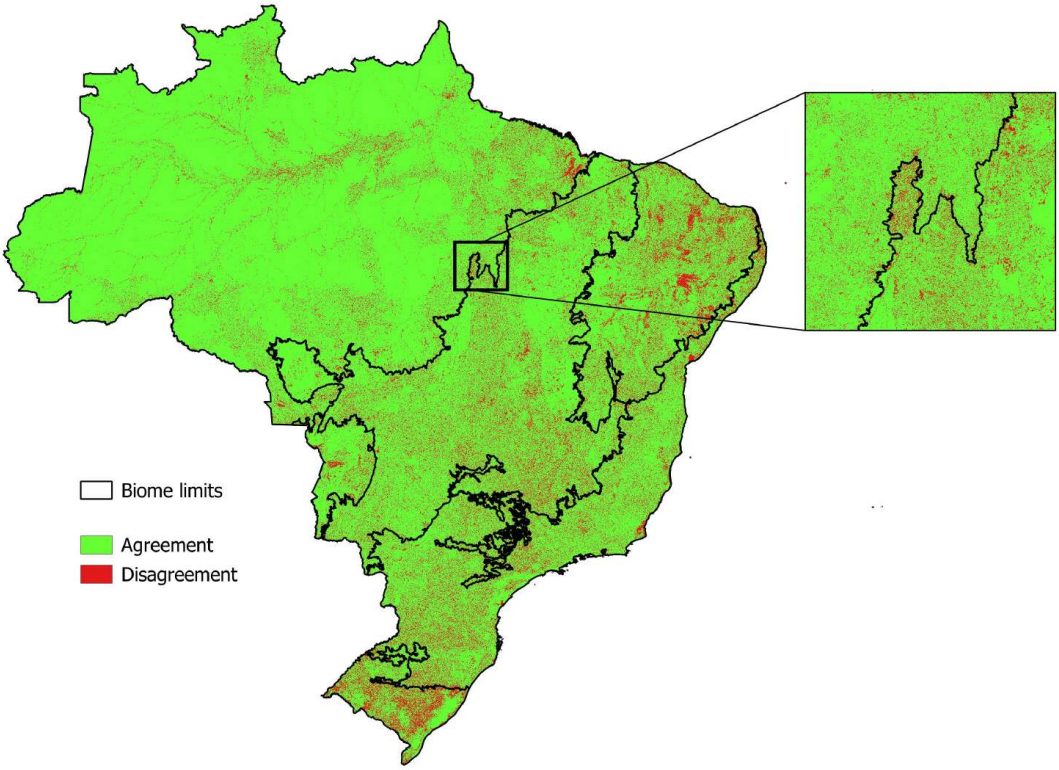
The Pantanal showed the lowest concordance among all biomes, registering only 55.51%. This discrepancy may be attributed to the unique spatial distribution of classes in this biome. The predominance of certain classes in different areas could have influenced a lower concordance between the initiatives. In particular, the classes related to 'Forest' were correctly mapped, except for the 'Secondary Forest/I' class, which was identified as 'Grassland/M'. Another highlight was the classes 'Other Unmanaged Woody Formations/I' and 'Wetland/M' identified as equivalents, representing 9% of the entire equivalence area of the biome.

When analyzing the concordances obtained at the national level and displayed in Figure 7, which presents a maximum concordance of 81%, there are some interesting trends and characteristics. The Amazon biome was the only one that showed a concordance higher than Brazil's by 11%, mainly due to the size and homogeneity of the forest classes. It is evident that, at national level, extensive forested and agricultural areas exhibit relatively strong correspondence between the two maps. This alignment is a positive indicator for macroecological assessments and large-scale policy considerations. On the other hand, this general accuracy

[2] We use /I for classes of Inventory and /M for MapBiomas.

should not overshadow the particularities of the biomes and the idiosyncrasies of data in more specific areas.

Figure 8 –
Concordance map obtained from the legend provided by the algorithm. The square in black show a small area in more details.

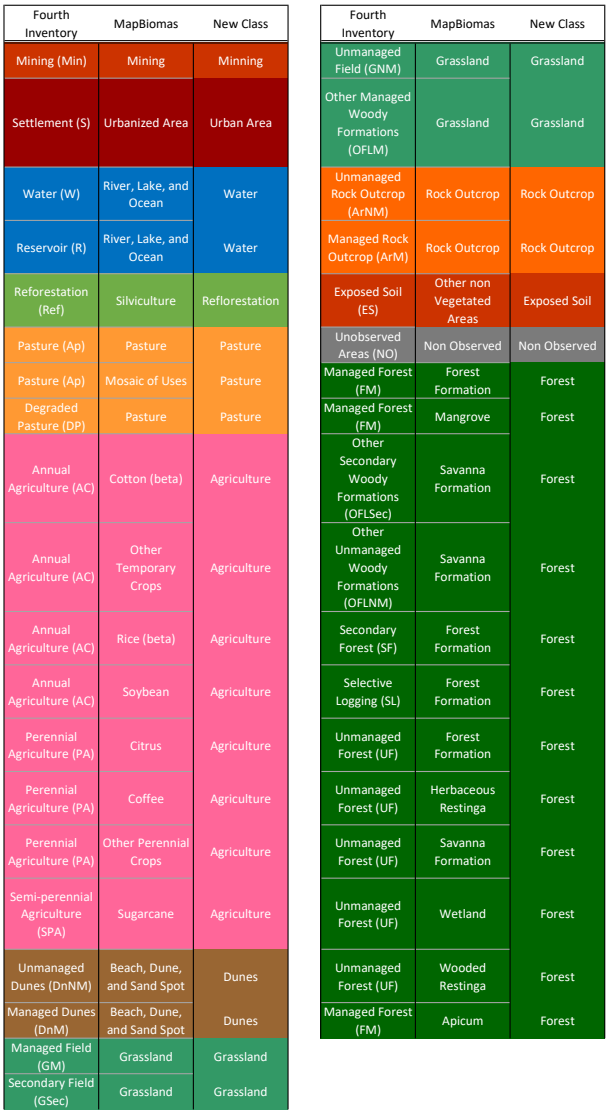


Source: Author's production.

In the harmonizations of the Amazon, Cerrado, Atlantic Forest, Pantanal, and throughout Brazil, the 'Mosaic of Uses/M' class was identified as 'Pasture/I', indicating that most of this class overlaps with the pasture class from the National Inventory and could be attributed to this class in the final harmonization for the sake of accuracy. Meanwhile, for Brazil, Caatinga, Atlantic Forest, and Pampa, the classes 'Unmanaged Rock Outcrop/I' and 'Forest Formation/M' were associated, raising an alert given their semantic differences. Similarly, this also occurs between 'Unmanaged Forest/I' and 'Rock Outcrop/M'. The classes 'Managed Forest/I' and 'Cotton (beta)/M' were incorrectly associated in three of harmonizations obtained. This might occur due to the small area encompassing the 'Cotton (beta)/M' class, which is more subject to erroneous overlaps. This can also occur with more emphasis on transition areas between biomes, which is more difficult to classify accurately due to the more significant variability in native vegetation.

Certain relationships become more evident when examining all the obtained harmonizations. The 'Annual Agriculture/I' and 'Soybean/M' classes were correctly identified in all seven harmonizations, indicating a good match between the two maps regarding annual agricultural areas dedicated to soy. Similarly, the class 'Pasture' in both maps was correctly associated in all cases. For 'Reforestation/I' and 'Silviculture/M', both maps have a good match for reforestation or silviculture areas, correctly identifying them in all regions. The 'Reservoir' and 'Water' classes of the National Inventory were also attributed in all harmonizations to the 'River, Lake and Ocean/M' class. This also occurred between 'Settlement/I' and 'Urbanized Areas/M', as well as 'Unmanaged Forest/I' and 'Forest Formation/M'.

Figure 9 –
Harmonized legend built from the algorithm.



Fourth Inventory	MapBiomas	New Class
Mining (Min)	Mining	Minning
Settlement (S)	Urbanized Area	Urban Area
Water (W)	River, Lake, and Ocean	Water
Reservoir (R)	River, Lake, and Ocean	Water
Reforestation (Ref)	Silviculture	Reflorestation
Pasture (Ap)	Pasture	Pasture
Pasture (Ap)	Mosaic of Uses	Pasture
Degraded Pasture (DP)	Pasture	Pasture
Annual Agriculture (AC)	Cotton (beta)	Agriculture
Annual Agriculture (AC)	Other Temporary Crops	Agriculture
Annual Agriculture (AC)	Rice (beta)	Agriculture
Annual Agriculture (AC)	Soybean	Agriculture
Perennial Agriculture (PA)	Citrus	Agriculture
Perennial Agriculture (PA)	Coffee	Agriculture
Perennial Agriculture (PA)	Other Perennial Crops	Agriculture
Semi-perennial Agriculture (SPA)	Sugarcane	Agriculture
Unmanaged Dunes (DnNM)	Beach, Dune, and Sand Spot	Dunes
Managed Dunes (DnM)	Beach, Dune, and Sand Spot	Dunes
Managed Field (GM)	Grassland	Grassland
Secondary Field (GSec)	Grassland	Grassland

Fourth Inventory	MapBiomas	New Class
Unmanaged Field (GNM)	Grassland	Grassland
Other Managed Woody Formations (OFLM)	Grassland	Grassland
Unmanaged Rock Outcrop (ArNM)	Rock Outcrop	Rock Outcrop
Managed Rock Outcrop (ArM)	Rock Outcrop	Rock Outcrop
Exposed Soil (ES)	Other non Vegetated Areas	Exposed Soil
Unobserved Areas (NO)	Non Observed	Non Observed
Managed Forest (FM)	Forest Formation	Forest
Managed Forest (FM)	Mangrove	Forest
Other Secondary Woody Formations (OFLSec)	Savanna Formation	Forest
Other Unmanaged Woody Formations (OFLNM)	Savanna Formation	Forest
Secondary Forest (SF)	Forest Formation	Forest
Selective Logging (SL)	Forest Formation	Forest
Unmanaged Forest (UF)	Forest Formation	Forest
Unmanaged Forest (UF)	Herbaceous Restinga	Forest
Unmanaged Forest (UF)	Savanna Formation	Forest
Unmanaged Forest (UF)	Wetland	Forest
Unmanaged Forest (UF)	Wooded Restinga	Forest
Managed Forest (FM)	Apicum	Forest

Source: Author's production.

Figure 8 shows the map highlighting areas of concordance between the maps based on the legend provided by the harmonization algorithm. The main regions with lower concordance between the maps are clearly visible. In the Pampa, Pantanal, and Caatinga biomes — those with the lowest percentage of concordance — the largest areas of divergence are concentrated, while in the other biomes, these areas are smaller and more scattered. When comparing the maps, it becomes apparent that MapBiomas tends to overestimate forested areas, diverging from the National Inventory, which shows a greater extent of pastureland and other non-forest formations, especially in this regions.

In Figure 9, we have the harmonized legend and a semantic analysis of the classes between the MapBiomas maps and the National Inventory based on the proposed harmonization algorithm. The harmonization generated by the algorithm and the harmonization that combines the semantic analysis and the algorithm are largely aligned for most classes. However, there are some areas of divergence, particularly in the nuances of forest formations and pastures. This is mainly due to the characteristics of the classes assigned to each biome, since both initiatives define the classes of native vegetation, especially the forest and grassland classes, according to the characteristics of each biome. This causes discrepancies between classes and confuses pasture and grassland classes, given

their structure and similar characteristics in some biomes. The same applies to some forest classes, which, in biomes is characterized by non-forest vegetation such as grassland, pasture and rock outcrop. The different classifications used by the initiatives lead to some confusion between these forests and grasslands, as well as between grasslands and pasture classes.

This preliminary analysis using the algorithm provides a broader perspective on classes in a legend that lack direct equivalents, such as the "Agriculture and Pasture Mosaic" class in MapBiomas. Rather than disregarding such areas in a study, it is possible to analyze which class they are predominantly mapped to and consider remapping accordingly. For instance, in (CAPANEMA et al., 2019), the 'Mosaic of Uses' class was excluded from the study, and a similar approach was taken in (NEVES et al., 2020), where the 'Mosaic of Uses', 'Secondary Vegetation', 'Mining', and 'Annual Deforestation' classes were disregarded due to the lack of equivalent categories. In these cases, applying the harmonization algorithm beforehand could help identify spatial correspondences and, when appropriate, incorporate these classes into the final harmonization.

5 CONCLUSION

The legend harmonization algorithm provides a first automated step for the class mapping process, a frequent challenge in LULC studies. One of the main strengths of this method is its comprehensive approach, ensuring a clear equivalence for every class in every map. This approach has to be complemented by a double check, where classes are compared in rows and columns, ensuring that all map classes will have an equivalent.

The integrity and precision of LULC maps are essential for understanding landscape dynamics, land alteration patterns, and their environmental implications. By comparing and harmonizing LULC maps from different initiatives, this study emphasized the importance of robust and comprehensive approaches, such as the legend harmonization algorithm presented.

The harmonization between the maps of both initiatives showed a good concordance rate with some reservations, especially when considering the Pantanal biome. It was possible to observe excellent mappings for significant classes such as forests and reforestation, urban areas, pastures, and water. When analyzing the harmonization at Brazil level, it is possible to notice that the main class confusions that occurred in each biome diminish when aggregating all areas, in addition to reinforcing the classes that were similarly mapped in all biomes.

In biomes with a predominance of non-forest vegetation, it was noticeable that there was an increased confusion among the grassland, pasture, and forest classes between the maps, especially in Pampa and Caatinga. Therefore, greater attention is needed in these cases when adapting to a coherent harmonization between the maps. The proposed legend, obtained from the algorithm's results, addresses the discrepancies between the classes identified during the initial concordance and may aid future studies.

In practical terms, the automation provided by the algorithm facilitates the integration of data from different sources, optimizing the efficiency of the process and minimizing errors that can arise from manual approaches. This optimization saves time and improves data interpretability, establishing a common standard that benefits researchers, decision-makers, and other stakeholders.

It is possible to assess changes over time and the influence of land use policies and practices by highlighting the similarities and differences. Moreover, this comparison becomes even more relevant in the absence of inventories in subsequent years. It allows for extrapolating trends and analysing carbon emissions by biome, ultimately providing insights for land-use planning and decision-making processes.

Authors Contribution

S.G.M.: Conceptualization, Methodology, Software, Validation, Writing - Original Draft, Writing - Review & Editing; P.R.A.: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing. A.C.S.: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing. M.I.S.E: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing.

Conflicts of Interest

The authors state that there is no conflict of interest.

References

BRASIL. **Quarta Comunicação Nacional do Brasil à Convenção Quadro das Nações Unidas sobre Mudança do Clima**. [S.l.: s.n.], 2021. P. 620. ISBN 9786587432182.

CAPANEMA, Vinicius et al. Comparação Entre Os Produtos Temáticos De Uso E Cobertura Da Terra Do Terraclass Amazônia E Mapbiomas: Teste De Aderência Entre Classes. **Anais do XIX Simposio Brasileiro de Sensoramento Remoto**, p. 724–727, Abril 2019.

DI GREGORIO, Antonio. **Land Cover Classification System: Classification concepts**. Rome: Food and Agriculture Organization of the United Nations (FAO), 2016. P. 40. ISBN 978-92-5-109017-6. Available from: <https://www.fao.org/3/x0596e/x0596e00.htm>.

ELLIS, Erle C.; KAPLAN, Jed O.; FULLER, Dorian Q.; VAVRUS, Steve; GOLDEWIJK, Kees Klein; VERBURG, Peter H. Used planet: A global history. **Proceedings of the National Academy of Sciences**, v. 110, n. 20, p. 7978–7985, 2013. DOI: 10.1073/pnas.1217241110. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1217241110. Available from: <https://www.pnas.org/doi/abs/10.1073/pnas.1217241110>.

HEROLD, Martin; WOODCOCK, Curtis E.; DI GREGORIO, Antonio; MAYAUX, Philippe; BELWARD, Alan S.; LATHAM, John; SCHMULLIUS, Christiane C. A joint initiative for harmonization and validation of land cover datasets. **IEEE Transactions on Geoscience and Remote Sensing**, v. 44, n. 7, p. 1719–1727, 2006. ISSN 01962892. DOI: 10.1109/TGRS.2006.871219.

IBGE. **Cobertura e Uso da Terra**. [S.l.: s.n.], 2019. Acesso em: 24 abr. 2022. Available from: <https://www.ibge.gov.br/geociencias/informacoes-ambientais/cobertura-e-uso-da-terra.html>.

INPE. **Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite**. [S.l.: s.n.], 2021. Acesso em: 24 abr. 2023. Available from: <http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes>.

_____. **TerraClass**. [S.l.: s.n.], 2019. Acesso em: 24 abr. 2023. Available from: <http://www.inpe.br/cra/projetos_pesquisas/dados_terraclass.php>.

JANSEN, Louisa J.M.; GROOM, Geoff; CARRAI, Giancarlo. Land-cover harmonisation and semantic similarity: some methodological issues. **Journal of Land Use Science**, Taylor Francis, v. 3, n. 2-3, p. 131–160, 2008. DOI: 10.1080/17474230802332076. eprint: https://doi.org/10.1080/17474230802332076. Available from: <https://doi.org/10.1080/17474230802332076>.

MAPBIOMAS BRASIL. **MapBiomas Brasil**. [S.l.: s.n.], 2021. Acesso em: 24 abr. 2022. Available from: <https://mapbiomas.org/>.

_____. **MapBiomas General "Handbook" Algorithm Theoretical Basis Document (ATBD) Collection 6**. 1. ed. [S.l.: s.n.], 2022. P. 0–48. Available from: <https://mapbiomas-br-site.s3.amazonaws.com/Metodologia/ATBD_Collection_6_v1_January_2022.pdf>.

MCTI. **Comunicações Nacionais do Brasil à Convenção-Quadro das Nações Unidas sobre Mudança do Clima**. [S.l.: s.n.], 2021. Acesso em: 24 abr. 2022. Available from: <https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/sirene/publicacoes/comunicacoes-nacionais-do-brasil-a-unfccc>.

MCTI. **Quarto Inventário Nacional de Emissões e Remoções Antrópicas de Gases de Efeito Estufa – Relatório de Referência Setor Uso da Terra, Mudança do Uso da Terra e Florestas**. Brasília: [s.n.], 2020. P. 312.

NEVES, Alana Kasahara; KÖRTING, Thales Sehn; FONSECA, Leila Maria Garcia; ESCADA, Maria Isabel Sobral. Assessment of TerraClass and MapBiomas data on legend and map agreement for the Brazilian amazon biome. **Acta Amazonica**, v. 50, n. 2, p. 170–182, 2020. ISSN 00445967. DOI: 10.1590/1809-4392201900981.

PIELKE SR., Roger A.; PITMAN, Andy; NIYOGI, Dev; MAHMOOD, Rezaul; MCALPINE, Clive; HOSSAIN, Faisal; GOLDEWIJK, Kees Klein; NAIR, Udaysankar; BETTS, Richard; FALL, Souleymane; REICHSTEIN, Markus; KABAT, Pavel; NOBLET, Nathalie de. Land use/land cover changes and climate: modeling analysis and observational evidence. **WIREs Climate Change**, v. 2, n. 6, p. 828–850, 2011. DOI: https://doi.org/10.1002/wcc.144. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.144. Available from: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.144>.

REIS, Mariane S et al. Towards a Reproducible LULC Hierarchical Class Legend for Use in the Southwest of Pará State, Brazil : Data-Driven Hierarchies. **Land**, v. 7, 65 2018. DOI: 10.3390/land7020065.

REIS; Mariane Souza; ESCADA; Maria Isabel Sobral; SANT'ANNA; Sidnei João Siqueira; DUTRA, Luciano Vieira. Harmonização de legendas formalizadas em Land Cover Meta Language-LCML. **Anais do XVIII Simposio Brasileiro de Sensoramento Remoto**, v. 4, n. 1, p. 1–23, 2017.

SHUKLA, Priyadarshi R et al. Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems, 2019.

SOUZA, Carlos M.; Z. SHIMBO, Julia; ROSA, Marcos R.; PARENTE, Leandro L.; A. ALENCAR, Ane; RUDORFF, Bernardo F. T.; HASENACK, Heinrich; MATSUMOTO, Marcelo; G. FERREIRA, Laerte; SOUZA-FILHO, Pedro W. M.; OLIVEIRA, Sergio W. de; ROCHA, Washington F.; FONSECA, Antônio V.; MARQUES, Camila B.; DINIZ, Cesar G.; COSTA, Diego; MONTEIRO, Dyeden; ROSA, Eduardo R.; VÉLEZ-MARTIN, Eduardo; WEBER, Eliseu J.; LENTI, Felipe E. B.; PATERNOST, Fernando F.; PAREYN, Frans G. C.; SIQUEIRA, João V.; VIERA, José L.; NETO, Luiz C. Ferreira; SARAIVA, Marciano M.; SALES, Marcio H.; SALGADO, Moises P. G.; VASCONCELOS, Rodrigo; GALANO, Soltan; MESQUITA, Vinicius V.; AZEVEDO, Tasso. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing**, v. 12, n. 17, 2020. ISSN 2072-4292. DOI: 10.3390/rs12172735. Available from: <https://www.mdpi.com/2072-4292/12/17/2735>.

VERBURG, Peter H; ERB, Karl-Heinz; MERTZ, Ole; ESPINDOLA, Giovana. Land System Science: between global challenges and local realities. **Current Opinion in Environmental Sustainability**, v. 5, n. 5, p. 433–437, 2013. Human settlements and industrial systems. ISSN 1877-3435. DOI: https://doi.org/10.1016/j.cosust.2013.08.001. Available from: <https://www.sciencedirect.com/science/article/pii/S1877343513000936>.

Biography of the main author

Sabrina G Marques, born in Presidente Prudente in 1998, holds a degree in Mathematics from São Paulo State University in Presidente Prudente and a master's degree in Applied Computing from the National Institute for Space Research in São José dos Campos, where she investigated carbon emissions in the land use and land cover sector, as well as techniques for map harmonization. She currently works as a data scientist at Bradesco Seguros in the Customer Relationship Management area in Barueri. She is interested in integrating data science with sustainability and technological innovation.

Esta obra está licenciada com uma Licença Creative Commons Atribuição 4.0 Internacional – CC BY. Esta licença permite que outros distribuam, remixem, adaptem e criem a partir do seu trabalho, mesmo para fins comerciais, desde que lhe atribuam o devido crédito pela criação original.