Uncovering Patterns: Data Mining the Deforestation Frontier in Brazilian Cerrado

Taise Farias Pinheiro [1], João Felipe Sobrinho Kneipp Cerqueira Pinto [2], Magog Araújo de Carvalho [1], Marcelo Francisco Sestini[1], Danilo Rocco Pettinati[1], Jacqueline Oliveira Santos[1], Sandra Benfica dos Santos[1], Luciana de Souza Soler [1], Andrea Daleffi Scheide[1], Cassiano Gustavo Messias [1], Claudio Aparecido de Almeida [1], Silvana Amaral[1], Thayse Azevedo Moreira[1], Sérgio Lopes Dousseau[1], Diego da Cunha Moraes[2], Jonatas da Silva Costa[2], Joelma da Silva Costa[2], Thiago de Moraes Nisimura[2]

[1] National Institute for Space Research, Monitoring Program by Satellite of the Brazilian Biomes, Brazil
[2] Spatial Coordination of the Amazon, National Institute for Space Research, Brazil
pinheiroftaise@gmail.com

Keywords: Data Mining, Deforestation Patterns, LULC Trajectories, MATOPIBA, Deforestation Hotspot.

Abstract

The Brazilian Cerrado, recognized as the world's most diverse tropical savanna, has undergone significant land clearance for agriculture, cattle ranching, and charcoal production, leading to considerable loss of its natural vegetation. This deforestation disrupts major aquifers, affecting ecosystems and human populations. The Cerrado Deforestation Monitoring Project (PRODES), active since 2017, systematically maps new deforestation using satellite imagery. This study used GeoDMA, a geospatial data mining tool for analyzing complex spatial databases such as those provided by the PRODES. The study area is the Alto Parnaíba region within the MATOPIBA agricultural frontier, characterized by high rates of crop expansion and burned areas. Using PRODES deforestation data from 2000 to 2023, the study categorizes deforestation patterns into geometric, multidirectional, and diffuse types, associated with large-scale agriculture, smallholder farming, and subsistence agriculture, respectively. The analysis reveals that approximately 50% of land use trajectories are small-scale subsistence agriculture, predominantly in northern Alto Parnaíba. Unchanged large-scale farms are concentrated in the southern part of the study area, linked to annual crops. The transition from small to large-scale farms is significant in the western region. Protected areas, especially strictly protected conservation units, show minimal changes in native vegetation. This study highlights the expansion of large-scale agriculture in MATOPIBA, emphasizing the need for targeted interventions to manage land use changes and prevent further deforestation. Future research should extend this approach to other Cerrado ecoregions to improve the detection and analysis of deforestation processes.

1. Introduction

The Cerrado Biome, the most diverse tropical savanna in the world, has experienced large-scale land clearance for agricultural activities, cattle ranching, and charcoal production (Sano et al., 2019b). This widespread deforestation has resulted in a significant loss of the original 2 Mkm² (~ 22 % of the total area of Brazil) of the natural vegetation cover. Nowadays, the Biome displays less than 50% of the natural vegetation cover (INPE, 2023), adversely affecting the biota of the three major habitat types, grasslands, savannas, and forested areas, that characterize the highly heterogeneous Cerrado landscape, home to many endemic species. In addition to being an important ecological, the Cerrado Biome is strategic for water resources (Oliveira et al., 2015). Deforestation disrupts several major aquifers that are crucial for providing water to South America, thereby affecting both local ecosystems and human populations dependent on these water sources.

To address the challenges posed by deforestation, the Cerrado Deforestation Monitoring Project (PRODES) has been actively monitoring deforestation increments since 2017 (INPE, 2023). The procedure consisted of mapping the anthropized area of the Cerrado biome in 2000, and from that reference year onwards systematically identifying all subsequent new deforestations. This monitoring effort supports policy-making by offering valuable insights that help in understanding and managing land use changes.

Extracting meaningful information from the vast databases generated by PRODES requires sophisticated methods. One such method is the application of data mining tools, which are essential for tracking spatial patterns of land use and analyzing the temporal evolution of these patterns. Extensive studies indicate that different actors involved in land use changes can be distinguished by their spatial patterns (Silva et al., 2008). This differentiation is important for exploring the different patterns in the region and how they change over time, which are critical for developing targeted interventions.

Geospatial Data Mining and Analysis (GeoDMA) is a powerful approach that combines geospatial analysis and data mining to uncover hidden trends and relationships within large-scale geospatial datasets (Korting et al., 2008). GeoDMA provides a comprehensive framework for analyzing complex spatial data, and performing the complete data mining process, including attribute selection, training, and classification.

Since the 2010s, deforestation in the Cerrado biome has been particularly pronounced in the MATOPIBA region, a continuous zone in the northern part of the biome. MATOPIBA holds high agricultural potential and is characterized by large soybean cultivation fields. The rapid agricultural expansion in this area has led to significant environmental changes. This paper focuses on using GeoDMA to assess spatial and temporal land use patterns in Alto Parnaíba, a region within MATOPIBA that is experiencing high rates of crop expansion and burned areas (Sano et al., 2019a). By employing GeoDMA, this study

aims to outline the evolution of deforestation spatial patterns over time, spanning from the year 2000 to the present day. Additionally, the paper examines the underlying processes driving land use changes in the region, providing a comprehensive understanding of the dynamics at play.

2. Study Area

The Alto Parnaíba Ecoregion covers an area of approximately 171 thousand km², corresponding to 8.2% of the original two million km² of the Cerrado biome. Approximately 20% of the original natural area has been deforested. The Alto Parnaíba embraces portions of the states of Maranhão (62%), Piauí (37%), and Tocantins (1%) (Fig. 1). Alto Parnaíba is part of the MATOPIBA region, a new agricultural frontier in the Cerrado. Among the main aspects that led to the emergence of this frontier is the availability of relatively extensive areas with flat terrain, which facilitates the introduction of intensive mechanization; the presence of soils with physical properties favourable for mechanization; a rainfall regime that allows for rainfed agriculture; the lower cost of land compared to other grain-producing regions in the Cerrado; and the strategic location for grain export.

Biophysical characteristics, as described by Sano et al. (2019a), include a mean annual precipitation of 1160 mm, ranging from 800 mm in the east to 1400 mm in the west. The mean annual temperature is 25.3°C. The topography consists of depressions with elevations ranging from 85 m to 415 m, decreasing from south to north, bordering the lowest elevations in the Cerrado. The dominant soil types in the study area are Ferralsols (highly weathered, deep >2 m, well-drained soils) and Cambisols (slightly to moderately weathered).

Vegetation is characterized by a mosaic of diverse types, ranging from grassland-savannic to forest formations, corresponding to a gradient of biomass. The savanna formations are mostly associated with tree steppe-like savanna and forested steppe-like savanna.

The Alto Parnaíba Ecoregion contains large areas of annual croplands, primarily in the south (Sano et al. 2019). The region also records frequent burned areas, most likely related to soil preparation for planting crops (Silva et al. 2021). This ecoregion has been identified as a very high conservation priority and a high sustainable use priority.

3. Methods

3.1 Deforestation dataset

We used the existing 30-m resolution deforestation maps from the Cerrado Deforestation Monitoring Project (PRODES) for the 2000 year to mask out previously deforested land (i.e., the anthropized area of the Cerrado biome in 2000 and water surfaces). This 2000 deforestation mask corresponds to an area of 41.341 Km². The method consists of vector delineation by visual inspection based on Landsat 5 images (2000 and 2002), Landsat 7 (2004–2010), RESOURCESAT 2 (2012), and Landsat 8 (after 2013). The PRODES minimum mapping unit is 1 hectare. PRODES data is available on the TerraBrasilis web portal, a platform developed to provide access, query, analysis, and dissemination of spatial data generated by government environment monitoring programs such as PRODES (Assis et al., 2019).

Data mining was applied to the PRODES database for the period from 2002 to 2023. To evaluate the historical evolution of deforestation patterns, this database was divided into three periods: 2002-2009, 2010-2015, and 2016-2023, aligning with the phases of the PPCerrado actions, a strategy for the prevention and control of deforestation in the Cerrado biome.

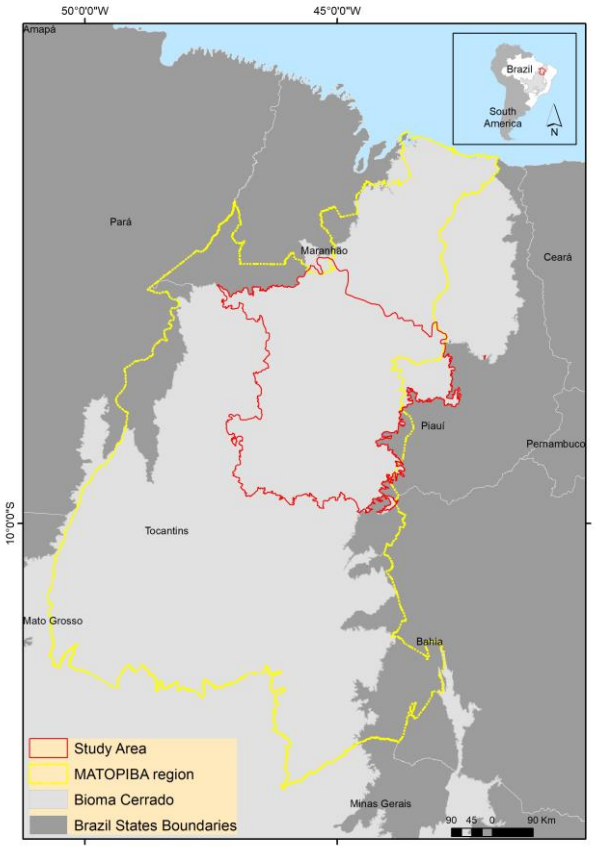

Figure 1. Study Area

3.2 From pattern to process

This article aimed to assess land use patterns and their associated processes. The methodology consists of the following: i) The definition of a spatial pattern typology. We proposed a typology associated with three different types of human occupation. Figure 2 shows the shape of the deforestation spatial patterns and associated characteristics (shape, color, context, size, and texture) in the OLI-Landsat sensor (color compositions using three bands in false color with band composition R5-G6-B4). The Geometric Regular typology is related to large clearings commonly associated with large-scale clearings for agriculture. The Multidirectional Unordered is related to small to medium clearings frequently linked to smallholder livestock farmers. The Diffuse pattern is characterized by small clearings generally associated with smallholder subsistence agriculture.

The patterns were represented by 1x1 km cells that encompass a set of deforestation polygons, that are described through landscape metrics. ii) Building a training set of spatial patterns for the dataset by selecting representative cells for each typology; iii) Employing a structural classifier (decision tree based on the C4.5 algorithm) to identify patterns.

By aligning the spatial objects with a reference set of spatial patterns, we reveal the semantic relationships and spatial arrangements in each period. Finally, we performed the analysis of the land use trajectories. We used cells to group the successive transitions between land use patterns into different trajectories. All possible trajectories were reduced to the four major trajectories shown in Table 1.

The grouping of trajectories into types aimed to answer the main questions of this study, which are: 1) to identify the main land use change trajectories, and 2) to identify the main actors involved in the land cover changes. These trajectories help to understand the dynamics of land-use changes and the transition patterns between different scales of farming and vegetation status over the analyzed period.

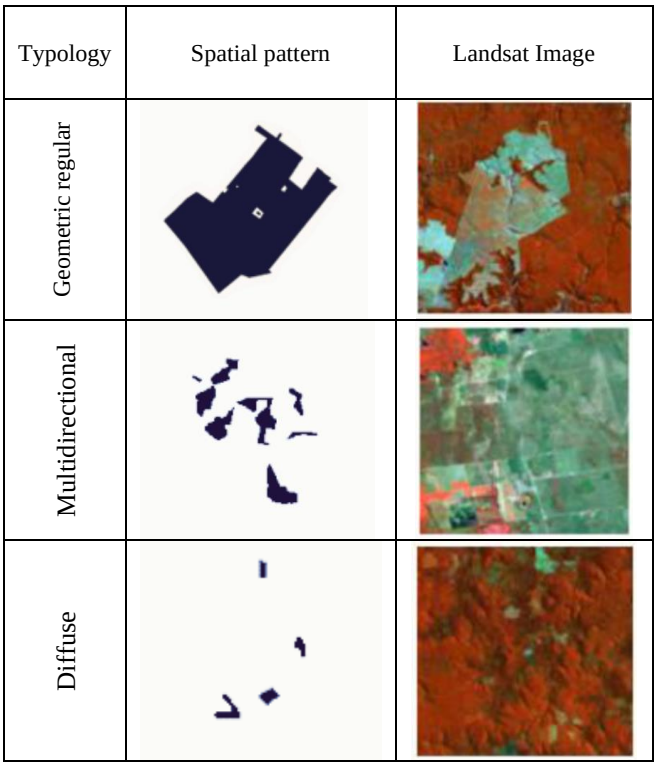

Figure 2. Spatial patterns of deforestation and associated semantics.

3.3 Accuracy of pattern classification

Expert interpreters evaluated the accuracy of the spatial pattern classification. We randomly selected from the automatic classification 100 cell samples of each pattern, which the interpreters manually classified according to the typology presented in Figure 2. We compared both classifications, calculated the confusion matrix, and determined the overall accuracy as well as Kappa statistics. The Kappa coefficient measures the agreement between two classifications, with values greater than 0.61 considered substantial.

Trajectories		
CLASS	DESCRIPTION	TRANSITION OVER PERIODS
No changes	This trajectory refers to areas with pristine vegetation that remained unchanged throughout the analysis period according to the PRODES database. PRODES monitors the suppression of forest, savanna, and grassland-shrub formations in the Cerrado Biome. Degradation events, such as those caused by logging and fire, are not considered	t1 - t1+n - tf vegetation- vegetation - vegetation
Small-small	This trajectory describes areas where land use was consistently associated with small or medium-scale farming throughout the entire analyzed period;	small/medium - small/medium - small/medium
Small-large	This trajectory indicates a pattern where areas initially used by small-scale farmers were converted into areas used by large-scale farmers over time;	Small/medium-small/medium/large-large
Large-large	This trajectory represents areas where large-scale clearings or land conversions occurred from the beginning of the analysis period, indicating continuous large-scale agricultural or land-use activities	large - large - large

Table 1. Description of the main land use change trajectories for the period from 2002 to 2023 (t=time, ti=initial time, tf=final time)

4. Results and Discussion

4.1 Classification of the patterns of deforestation

The C4.5 algorithm automatically generated a decision tree with three metrics and four levels (Fig. 3). The decision tree used the patch area metric (defined by the internal area of the landscape objects contained within a cell) to distinguish smaller objects from larger ones.

4.2 Accuracy Assessment

The confusion matrix shows the results of interpreter classification in columns and the results of data mining classification in rows (Table 2). The main diagonal, presented as boldface in the matrix, indicates the cases correctly allocated. The overall accuracy between the manual and automatic classification was estimated at 80%. Most of the error, revealed in the confusion matrix, is associated with the misclassification of the Multidirectional pattern as Geometric or Diffuse. These misclassifications are likely because Multidirectional is a transition class, exhibiting spatial characteristics observed in both Geometric and Diffuse patterns. The accuracy of the confusion matrix was also expressed by the kappa coefficient and was estimated as 70%.

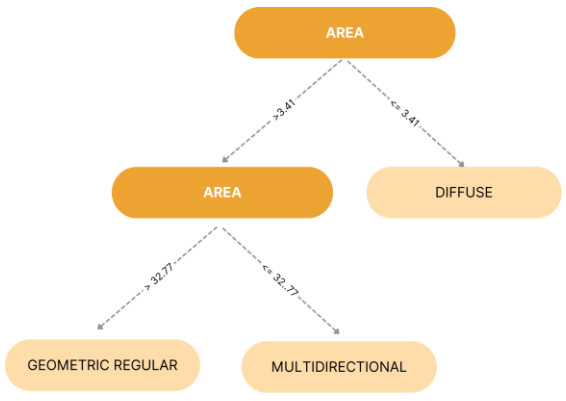

Figure 3. Decision tree for the deforestation spatial patterns

	Patterns	Reference			
		Geometric	Diffuse	Multidirecional	Total
Classification	Geometric	**71**	0	5	71
	Diffuse	0	**90**	15	90
	Multidirectional	29	10	**80**	161
	Producer accuracy	71%	90%	80%	
	User´s Accuracy	93%	86%	67%	
	Kappa coeficient		70%		
	Overall accuracy		80%		

Table 2. Confusion matrix of the classification of deforestation patterns

4.3 Spatial-temporal pattern analysis

Figure 3 illustrates the area (Km²) of different deforestation patterns (Geometric, Multidirectional, Diffuse) and the remaining native vegetation for the periods 2002-2009, 2010-2015, and 2016-2023. The vegetation cover area has decreased over the three periods, with a significant drop in the 2016-2023 period. We identified the expansion of the ‘Diffuse” deforestation pattern throughout the landscape over time, from just below 13.000 Km² to approximately 14.000 Km². The Diffuse pattern is typically related to smallholder subsistence agriculture, and the increase across the landscape can be linked to migration towards the deforestation frontier. The region has exhibited population growth since 2000 (IBGE, 2020).

The Multidirectional deforestation pattern has been reduced over time, from slightly above 10.500 K² in the 2002-2009 period to approximately 9.500 Km² in the 2016-2023 period. The classification unveiled a notable rise in the geometric pattern, particularly in the 2016-2023 period, which is a pattern associated with modern production activities and establishing an agricultural frontier in the region.

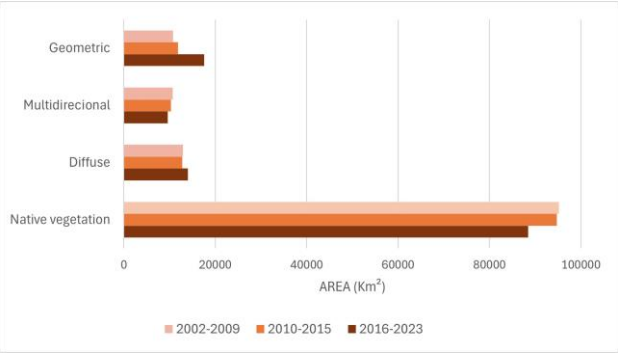

Figure 3. The area (Km²) of deforestation patterns compared to the native vegetation within each period.

From the grouping of land use patterns into trajectories, two main phenomena become evident. Temporal analysis showed that approximately 50% of land use trajectories are associated with small-small trajectories (Fig. 4), meaning small-scale farmers maintain subsistence activities throughout the analysed period. This trajectory is pronounced in the northern part of Alto Parnaiba (Fig. 5). Cattle ranching is extensive in northern Alto Parnaiba and uses low levels of technology and soil fertilization.
The remaining trajectories, approximately 25%, involve small-scale farms transitioning to large-scale operations, while another 25% are associated with unchanged large-scale operations (Fig. 4).

There is a notable concentration of unchanged large-scale farms in the southern part of the study area, near the Chapadão do São Francisco Ecoregion, which is associated with annual crops. Annual croplands comprise over 8.5% of the Cerrado and are concentrated in specific areas, including the plateaus of Paraná Guimarães, western Chapadão do São Francisco, eastern Chapada dos Parecis, and southern Alto Parnaíba (Sano et al. 2019)

The transition from small-scale to large-scale farms is primarily concentrated in the western region, possibly correlating with the expansion of large-scale production in that direction. This might reflect the presence of areas with plateaus like those observed in the Chapadão do São Francisco. The large-scale production in Chapadão do São Francisco is the result of the flat surface, high precipitation (mean annual rainfall=1296 mm), and presence of Ferralsols that led to the expansion of annual cropland (Sano et al. 2019a).

Approximately 30% of cases where no changes in native vegetation were observed within protected areas. This is notable, particularly inside Indigenous land and strictly protected conservation units such as Mirador State Park and Uruçuí Ecological Station.

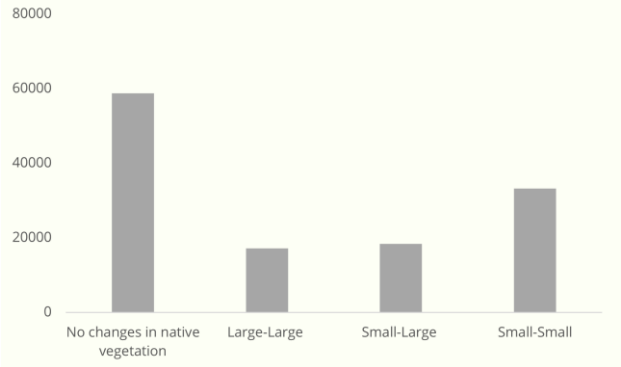

Figure 4. Trajectories of deforestation pattern in the Alto Parnaíba Ecorregion

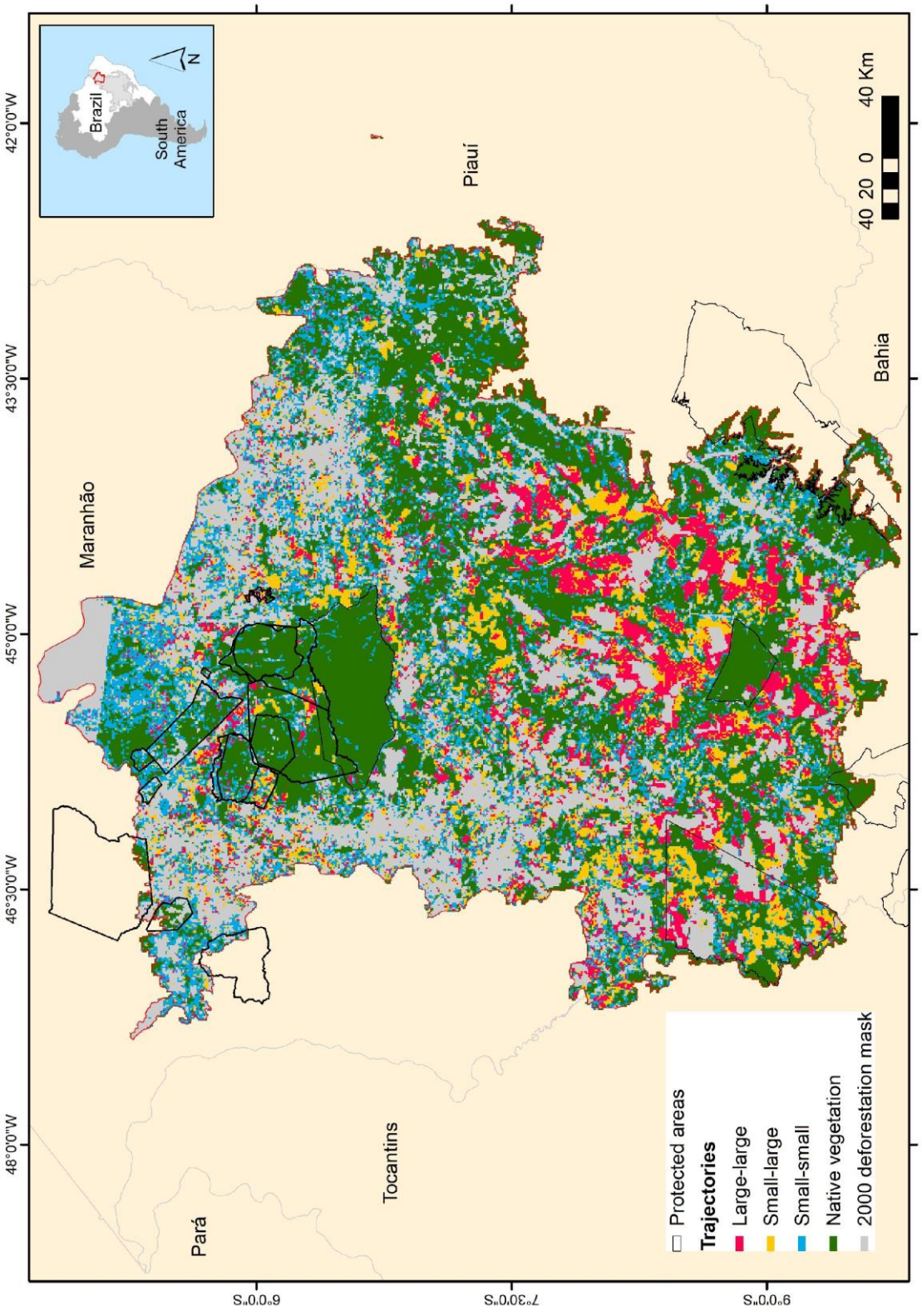

Figure 5. The land use cell trajectories illustrate the deforestation process in the Alto Parnaíba Ecoregion

5. Conclusions

Data Mining based on PRODES Cerrado deforestation product supports the extraction of spatial-temporal configurations from these databases. The methodology incorporates data mining, and landscape ecology techniques to achieve robust results in pattern detection. It addresses the challenge of identifying land use change patterns in a large database in a rapidly evolving region.

By analyzing different periods, patterns classification highlighted the expansion of large-scale agriculture frontier over time. The results enable the detection of pattern changes crucial for assessing, managing, or preventing deforestation processes.

Further experiments are needed to understand the spatial-temporal pattern changes throughout the Cerrado Biome. Developing specific deforestation patterns for the other ecoregions could improve the detection and analysis of deforestation processes.

Acknowledgments

Thanks to the National Council of Technological and Scientific Development - CNPq project number 422354/2023-6 (MONITORAMENTO E AVISOS DE MUDANÇAS DE COBERTURA DA TERRA NOS BIOMAS BRASILEIROS – CAPACITAÇÃO E SEMIAUTOMATIZAÇÃO DO PROGRAMA BIOMAS/BR), supported by the National Institute for Space Research (INPE).

References

Assis, L.F., Ferreira, K.R., Vinhas, L., Maurano, L., Almeida, C., Carvalho, A., Rodrigues, J., Maciel, A., Camargo, C., 2019. TerraBrasilis: A Spatial Data Analytics Infrastructure for Large-Scale Thematic Mapping. ISPRS Int J Geoinf 8, 513. https://doi.org/10.3390/ijgi8110513

INPE, 2023. The Cerrado Deforestation Monitoring Project - PRODES Cerrado.

Korting, T.S., Fonseca, L.M.G., Escada, M.I.S., Silva, F.C., Silva, M.P., 2008. GeoDMA - A Novel System for Spatial Data Mining, in: IEEE International Conference on Data Mining Workshops. IEEE, Pisa, Italy, pp. 975–978. https://doi.org/10.1109/ICDMW.2008.22

Oliveira, P.T.S., Wendland, E., Nearing, M.A., Scott, R.L., Rosolem, R., da Rocha, H.R., 2015. The water balance components of undisturbed tropical woodlands in the Brazilian cerrado. Hydrol Earth Syst Sci 19, 2899–2910. https://doi.org/10.5194/hess-19-2899-2015

Sano, E.E., Rodrigues, A.A., Martins, E.S., Bettiol, G.M., Bustamante, M.M.C., Bezerra, A.S., Couto, A.F., Vasconcelos, V., Schüler, J., Bolfe, E.L., 2019a. Cerrado ecoregions: A spatial framework to assess and prioritize Brazilian savanna environmental diversity for conservation. J Environ Manage 232, 818–828. https://doi.org/10.1016/j.jenvman.2018.11.108

Sano, E.E., Rosa, R., Scaramuzza, C.A. de M., Adami, M., Bolfe, E.L., Coutinho, A.C., Esquerdo, J.C.D.M., Maurano, L.E.P., Narvaes, I. da S., Filho, F.J.B. de O., da Silva, E.B., Victoria, D. de C., Ferreira, L.G., Brito, J.L.S., Bayma, A.P., de Oliveira, G.H., Bayma-Silva, G., 2019b. Land use dynamics in the Brazilian Cerrado in the period from 2002 to 2013. Pesqui Agropecu Bras 54. https://doi.org/10.1590/S1678-3921.pab2019.v54.00138

Silva, M.P., Câmara, G., Escada, M.I.S., Cartaxo, R., 2008. Remote sensing image mining: detecting agents of land use change in tropical forest areas. Int J Remote Sens 29, 4803–4822. https://doi.org/10.1080/01431160801950634

Silva, P.S., Nogueira, J., Rodrigues, J.A., Santos, F.L.M., Pereira, J.M.C., DaCamara, C.C., Daldegan, G.A., Pereira, A.A., Peres, L.F., Schmidt, I.B., Libonati, R., 2021. Putting fire on the map of Brazilian savanna ecoregions. J Environ Manage 296. https://doi.org/10.1016/j.jenvman.2021.113098