Data mining applied to temporal dynamics of deforestation pattern: a study case in Southern Amazon forest, Brazil

Mikhaela A. J. S. Pletsch[1]
Thales Sehn Körting[1]
Maria I. S. Escada[1]
Sacha M. O. Siani[1]

[1] Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 15.064 – 91.501-970 – São José dos Campos – SP – Brazil
{mikhaela.pletsch, thales.korting}@inpe.br;
{isabel}@dpi.inpe.br; {sacha}@ortiz.com.br

Abstract. The Amazon forest is one of the most prominent tropical rainforest worldwide. Providing several benefits, measures to its protection have been included in governmental decision making. In this context, the Brazilian initiative known as PRODES was implemented in 1988. Coordinated by the National Institute of Space Research, INPE, this project monitors annually deforestation in Brazilian Amazon. Even though, it is not enough to avoid deforestation and more analysis are required. In such manner, landscape metrics are commonly used to support analysis of deforestation dynamics and spatial patterns. It is just possible, considering that a real landscape reflects certain spatial patterns and structures. Nonetheless, taking into account Amazon extension, classifying landscape manually is considered a very time consuming task. In this manner, the aim of this study is to automate landscape classification, based on an already visually classified area in Southern Amazon forest. After that, we applied the decision tree to 1985 and 2015 data. Although data mining techniques were used, the final classification was not satisfactory for all the applications. Thus, we propose as further researches alternatives to overcome these issues and to validate the process. Finally, a discussion about the algorithm is also held as well as local temporal dynamics of deforestation.

Keywords: . data mining, dynamics of deforestation, landscape metrics

1. Introduction

The Brazilian Amazon comprehends more than 60% of the Amazon Basin, which also means about 35% of the world's rainforest basins. Although rainforests play an important but not completely understood role in ecosystem (FAO, 2011), deforestation is one of the main challenges faced. Characterized as a complex process, deforestation is not triggered by a single factor, but by the combination of several factors, as agricultural expansion and extension of infrastructure (GEIST; LAMBIN, 2001; MARIANO; SIMONASSI et al., 2012). Forest cover has been replacing by other land cover, such as pastures, agricultural crops and settlements. In this context, a possibility to assure environmental conservation is supporting decision making and establishing effective governance (SOARES-FILHO et al., 2005).

The Brazilian initiative PRODES (Legal Amazon Deforestation Monitoring Project) was implemented in 1988 in order to monitor deforestation in Amazon. Coordinated by the National Institute for Space Research – INPE, this project provides the annual rate of Amazon deforestation. Even though, it is not enough to avoid deforestation and more analysis are required to combat deforestation. In this manner, landscape metrics are commonly used to support dynamic and spatial deforestation analysis. It is just possible, considering that a real landscape reflects certain spatial patterns and structures (MCGARIGAL, 2002). Nonetheless, taking into the extension of Amazon account, manually classification of landscape metrics is very time consuming task. Thus, the aim of this study is to automate landscape metrics classification through data mining. Therefore, this work proposes to automate the classification of an area that was already visually classified.

Escada (2003) analyzed *land use/land cover* (LULC) change processes in center-north Rondônia state from 1985 to 2000 through an empiric methodology. For that, distinct data were used, including historical series of remote sensing images, agrarian structure maps, field observations, agricultural census data, plot sizes, way of land appropriation, occupation age and spatial configuration of plots. Comprehending 10 municipalities (Cujubim, Jaru, Vale do Anari, Vale do Paraíso, Machadinho d'Oeste, Ouro Preto d'Oeste, Ariquemes, Rio Crespo, Ji-Paraná, Theobroma) and approximately 15.400 km^2, the area is located in Southern Amazon forest. Predominantly, it is composed of colonization projects coordinated by INCRA (National Institute for Colonization and Agrarian Reform) (ESCADA, 2003). In this manner, the aim of this work is to automate the accumulated deforestation composition of 2000, according to the reference map developed in the aforementioned research. Herewith, cell samples based on deforestation pattern typology were selected. After that, an automatic classification of this year was undertaken as a result of a satisfactory machine learning process. The final decision tree was then applied to the 1985 and 2015, where no more samples were extracted, aiming to automate the classification in the area for any year in the future or in the pass.

1.1. Landscape Ecology and Deforestation Patterns

In the 1980s, efforts aiming the comprehension of spatial heterogeneity emerged as the research area called *landscape ecology* (TURNER, 2005). Its scope is to elucidate patterns and dynamics of real landscapes, through the investigation of natural and unnatural triggers in various scales (MCGARIGAL, 2002; DIBARI, 2007; TURNER, 2005). As an auxiliary tool to understand the relationship between processes and patterns, landscape metrics were developed (TURNER, 2005), once it is considered as a summary of landscape arrangements (MCGARIGAL, 2002).

In scientific community many studies aimed to point out agents of deforestation by means of spatial structure and size of deforested areas. Husson, Jeanjean e Puig (1995) cataloged a typology of forest and non-forest interfaces and fragmentation patterns that occurred in tropical forests. Mertens e Lambin (1997) also recognized some of the main deforestation patterns, including geometric, island, corridor, diffuse, fishbone and patchy deforestation patterns.

2. Materials and Methods

Provided by Escada (2003), the input data is composed by deforestation polygons from 1985 to 2000. The accumulated deforestation composition of 2015 was provided by PRODES/INPE. The software TerraView was used as well as its plugins, Fill Cells and GeoDMA (Geographic Data Mining Analyst). Fill Cells was applied to extract maximum and minimum polygons areas. After that, we implemented the three accumulated deforestation polygons, 1985, 2000 and 2015 in GeoDMA environment. This plugin is based on the concepts proposed by (SILVA et al., 2005) to identify deforestation patterns in Amazon area. Coded in C++, it provides an interface to the user and geographic information data stored in databases (KÖRTING; FONSECA; CÂMARA, 2013). GeoDMA can deal with distinct geospatial data, but for this study grid cell and landscape-based features were used. The following steps of the methodology presented in Figure 1 are explained below.

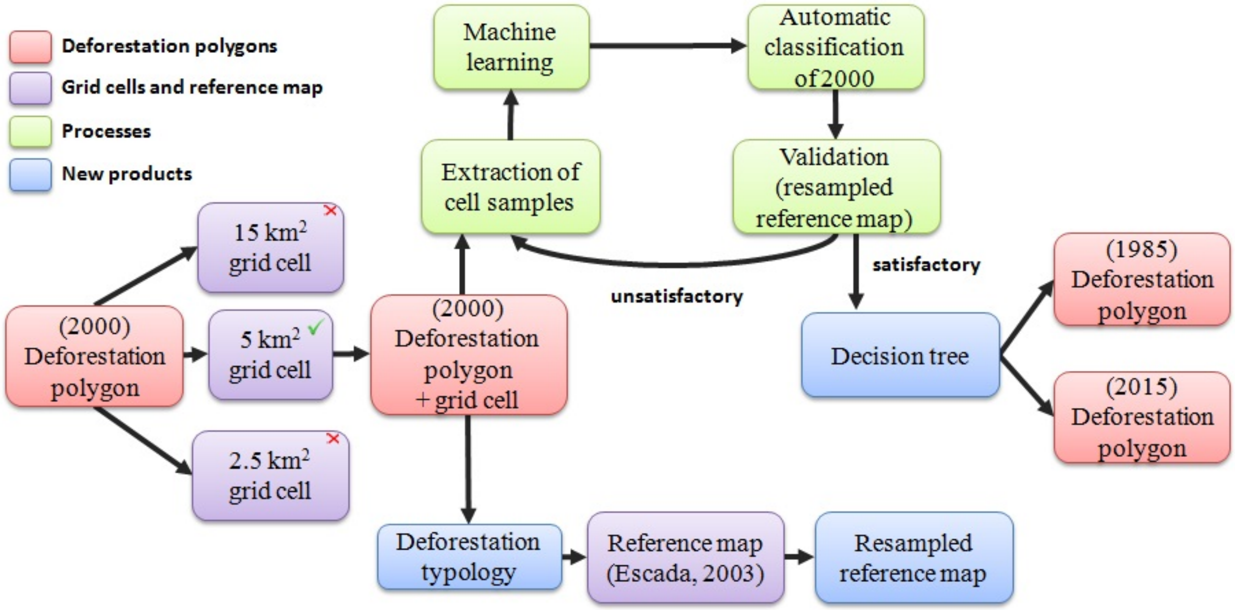

Deforestation polygons
Grid cells and reference map
Processes
New products
Machine learning
Automatic classification of 2000
Extraction of cell samples
Validation (resampled reference map)
unsatisfactory
satisfactory
15 km²
grid cell
(2000) Deforestation polygon
5 km²
grid cell
(2000) Deforestation polygon + grid cell
2.5 km²
grid cell
Decision tree
(1985) Deforestation polygon
(2015) Deforestation polygon
Deforestation typology
Reference map (Escada, 2003)
Resampled reference map

Figure 1: Method flowchart

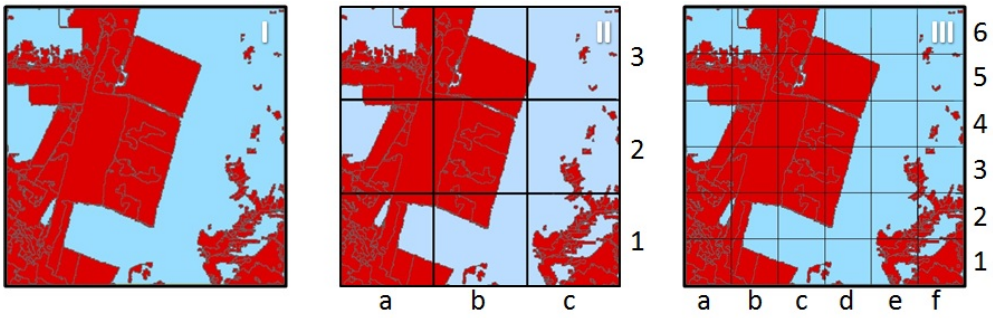


Figure 2: Grid cell analysis - I: 15 km^2, II: 5 km^2, III: 2.5 km^2

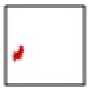
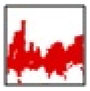
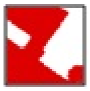
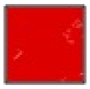
2.1. Grid Cell Selection

The main reference map used in this study was developed by (ESCADA, 2003) for the year 2000. In this manner, firstly, three different grid cells (15 km^2, 5 km^2 and 2.5 km^2) were created based on PRODES deforestation polygon for the same year as the reference map, 2000. This approach aimed to visually identify the most suitable pixel dimension for the area. In this manner, we applied the three different grid cells to analyze its influence in local landscape patterns. The Figure 2 presents an example area, where the second grid was the most adequate, once it fits most objects in cells, does not sharply segment the polygons and presents more homogeneous areas. Thus, the second grid was selected to the next steps. Consequently, we concluded that extreme dimensions, too big or too small, may not be satisfactory for this kind of study.

2.2. Deforestation Pattern Typology

According to the grid cell selection, a deforestation pattern typology was created (Table 1). The reference map was resampled based on the aforementioned typology, which was used to visually

Table 1: Deforestation pattern typology

Pattern	Spatial structure	Plot size	Occupation stage	Description
Diffuse		Small	Early	Normally found in the surroundings of deforested areas, this pattern normally represents small farmers and the beginning of intense deforestation
Directional		Small	Intermediate	Found along roads occupation, this pattern is represented predominantly by small farmers. It also can be related to INCRA's settlement projects
Geometric		Medium and big	Early and intermediate	Predominantly medium and big farmers, the geometric format may indicate the use of bigger machines to deforest
Consolidated		Small, medium and big	Advanced	Consolidated occupation area with small forest remnants and it also indicate older pioneer settlements with land concentration and/or large farms

validate the decision tree. Although Escada (2003) have identified 10 land use patterns, for this work these patterns were assembled in 4 types, considering the frequency of patterns and the sensibility of GeoDMA to run data mining. The pattern diffuse was not considered a particular pattern by Escada (2003). Nonetheless, here we considered it as an extra class, since its presence on the borders of consolidated areas may be a trigger to more deforested areas.

2.3. Processes in GeoDMA

Based on polygons from 2000 and on the deforestation pattern typology, the most appropriate samples were visually selected in order to support GeoDMA machine learning and training. In this step, approximately, 60% of the selected samples are assigned to the training and 40% for internal validation. GeoDMA provides the supervised classification with decision trees algorithm (version C4.5 (QUINLAN, 1993)). After the automatic classification of 2000, the result of it was visually validated based on the final classification developed by Escada (2003) for the area in the same year.

In the case the result of the automatic classification of 2000 was not satisfactory as expected, based on the reference map, we collected new or more samples.

2.4. Decision Tree

After analyzing the behavior of the decision tree using training samples obtained from polygons of year 2000, the most suitable decision tree was finally applied to 1985 and 2015.

The most adequate decision tree (Figure 3) is structured according to 4 main features: patch size coefficient of variation, number of patches, area weighted mean patch fractal dimension, maximum area of a polygon (Table 2). These features are the basis for the final classification and data mining,

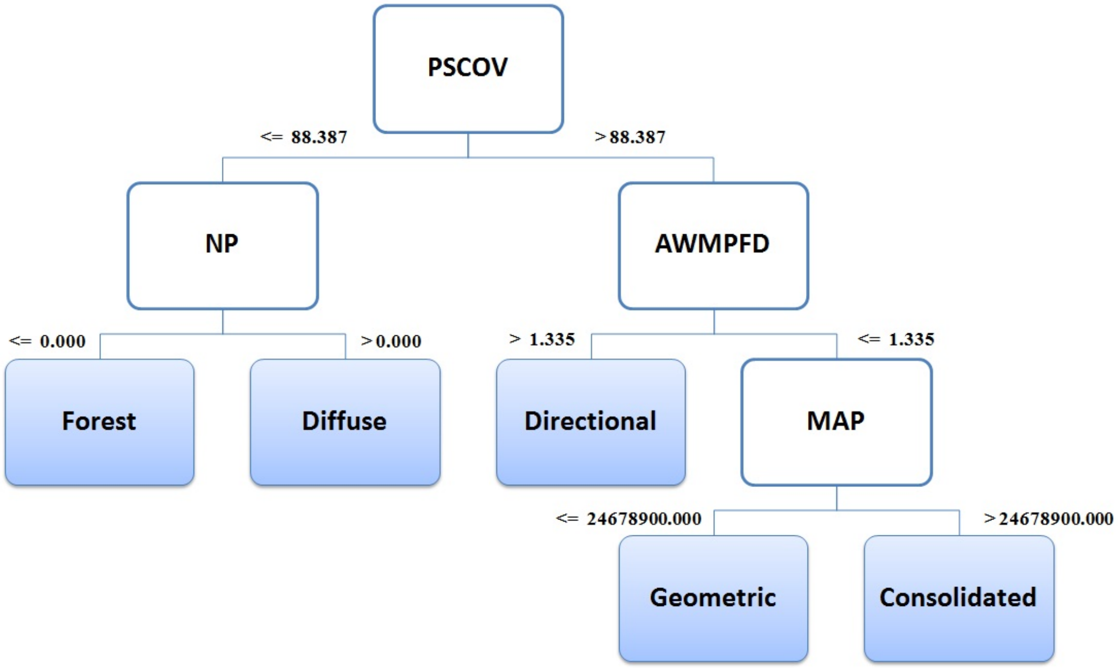


Figure 3: Decision tree

which were already applied to the year 2000. Considering that, no more samples are extract from the following years and the final classification was undertaken for 1985 and 2015. Furthermore, it is remarkable the although GeoDMA is based on many types of features, for this region just four features were selected.

3. Results and Discussion

3.1. Automated and Visually Classification Comparison

The result of the automated classification as well as the referenced map can be seen in Figure 4. Because the consolidated pattern was seen just in few cells, the automation was not able to recognize this arrangement in Southern of the area in 2000. In other words, there were not enough samples to support machine learning process. In such a way, it is the case either to consider samples not completely consolidated in the extraction of cell samples step or to remake to deforestation pattern typology considering not completely consolidated cells. Moreover, directional, geometric and non-deforested areas were satisfactory identified.

3.2. Temporal Dynamics of Deforestation

The final automated classification is presented in Figure 5. The temporal dynamics of deforestation reveal the notable land cover changed in 30 years. It also shows how the landscape metrics changed along the time, where directional and geometric patterns tend to become consolidated, but the velocity of change is not fast enough to transform diffused to consolidated areas.

Although the overall accuracy was visually high, the metrics used are rustic and could not

Table 2: Landscape-based features

Feature	Description	Formula
PSCOV	Patch Size Coefficient of Variation calculates the ratio between the features PSSD and MPS	$\frac{PSSD}{MPS} * 100$
PSSD	Patch Size Std is the root mean squared error (deviation from the mean) in patch size. This is the population standard deviation, not the sample standard deviation	$\sqrt{\frac{\sum_{j=1}^{n}(a_j - MPS)^2}{n} 10^{-4}}$
MPS	Mean Patch Size equals the sum of the areas (m2) of all patches of the corresponding patch type, divided by the number of patches of the same type	$\frac{\sum_{j=1}^{n} a_j}{n} 10^{-4}$
NP	Number of Patches equals the number of patches inside a particular landscape	n
AWMPFD	Area-weighted Mean Patch Fractal Dimension	-
MAP	Maximum Area of a Polygon	-

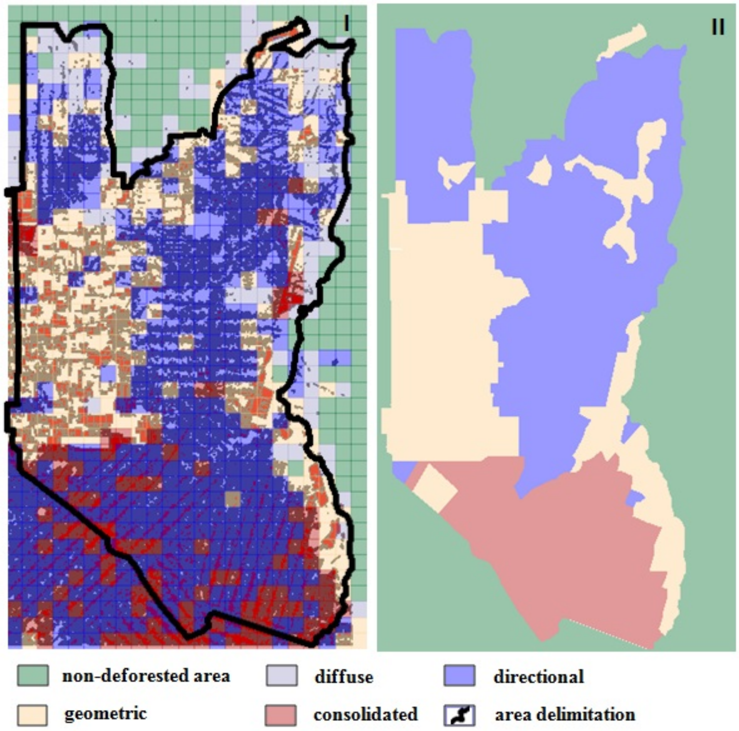


Figure 4: Comparison of the results for 2000 - I. automated classification; II. visually classification

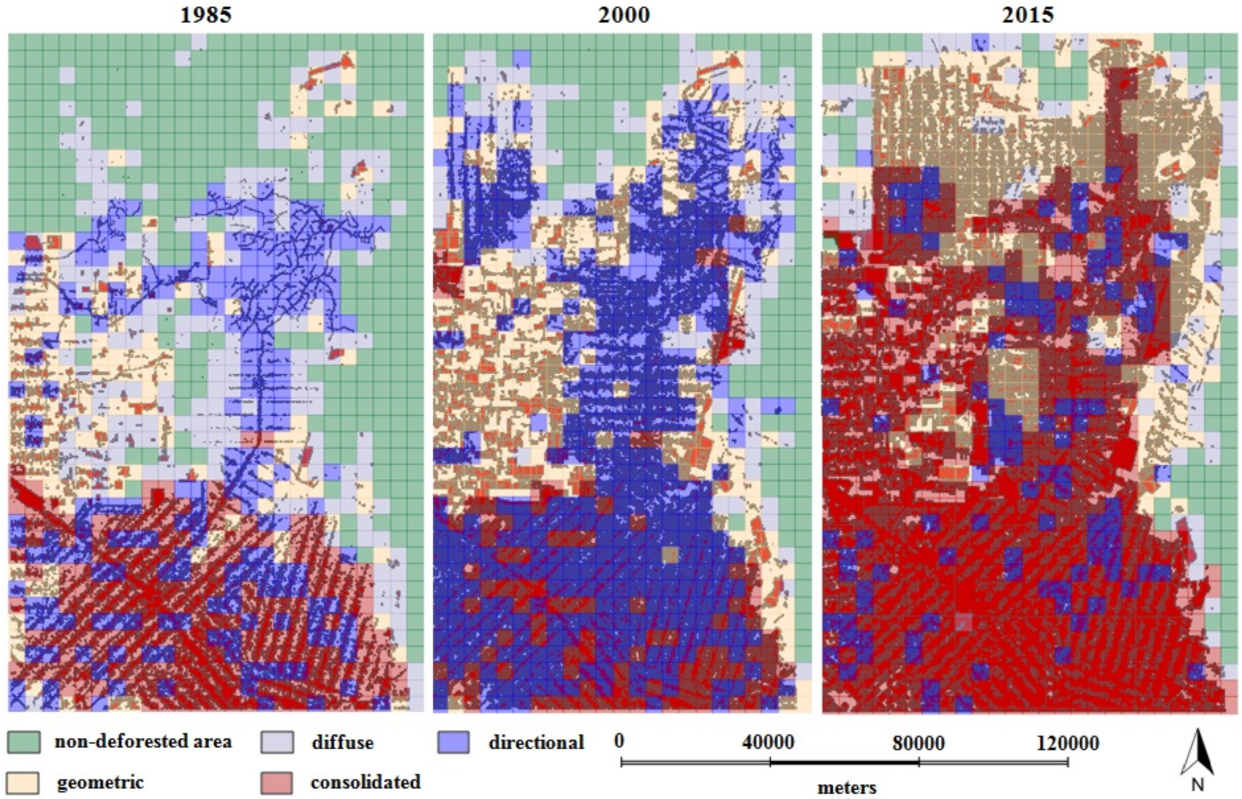


Figure 5: Temporal dynamics of deforestation

identify very specific agents of deforestation, considering that just samples from 2000 were taken into account. The classification for 1985 was adequate, however, the consolidated areas changes to directional in 2000 and then switches again to consolidated areas in 2015, which can not be considered satisfactory. Besides the aforementioned misclassification in 2000, the remain patterns were acceptable. In 2015, there was a clearly misclassification between geometric and directional patterns. This fact may be occurred by virtue of the selected landscape-based features.

4. Conclusions and Further Research

Landscape metrics assist the identification of forest deforestation agents through remote sensing data. However, those metrics varies along the time and this phenomenon may influence negatively machine learning step, causing not satisfactory classifications too in the future or in the past. That means new classes may arise along the time and will not be considered in the model, since they were not selected as training samples for that machine learning process. In this context, an alternative is to extract samples from all the desired years in order to semi-automate the process.

The achieved results were not satisfactory for all the classifications, which requires more researches in order to improve the decision tree for different regions and years. Further researches are also necessary aiming to validate extracted samples through random points selection and creation of confusion matrix.

Finally, remote sensing has becoming an important auxiliary tool, but more methods are required. Thus, we consider this kind of analysis an alternative opportunity to support more refined temporal dynamics of deforestation. With positive results for all three classifications, statistic modeling studies could be carried out to generate prognosis for the future, supporting governmental decision making and protection measures.

Acknowledgments

The authors would like to thank National Counsel of Technological and Scientific Development (CNPq) for the financial support.

References

DIBARI, J. N. Evaluation of five landscape-level metrics for measuring the effects of urbanization on landscape structure: the case of tucson, arizona, usa. *Landscape and Urban Planning*, Elsevier, v. 79, n. 3, p. 308–313, 2007.

ESCADA, I. S. E. *Evolução de padrões da terra na região centro-norte de Rondônia*. Tese (Doutorado), 2003.

FAO, I. The state of forests in the amazon basin, congo basin and southeast asia. *Rome: FAO-ITTO*, 2011.

GEIST, H. J.; LAMBIN, E. F. *What Drives Tropical Deforestation?* [S.l.], 2001.

HUSSON, A.; JEANJEAN, H.; PUIG, H. Study of forest non-forest interface typology of fragmentation of tropical forest. *TREES Series B: Research Report*, n. 2, 1995.

KÖRTING, T. S.; FONSECA, L. M. G.; CÂMARA, G. Geodma—geographic data mining analyst. *Computers & Geosciences*, Elsevier, v. 57, p. 133–145, 2013.

MARIANO, F. Z.; SIMONASSI, A. G. et al. Causas do desmatamento no brasil e seu ordenamento no contexto mundial. *Brazilian Journal of Rural Economy and Sociology (RESR)*, Sociedade Brasileira de Economia e Sociologia Rural, v. 50, n. 1, 2012.

MCGARIGAL, K. Landscape pattern metrics. *Encyclopedia of environmetrics*, Wiley Online Library, 2002.

MERTENS, B.; LAMBIN, E. F. Spatial modelling of deforestation in southern cameroon: spatial disaggregation of diverse deforestation processes. *Applied Geography*, Elsevier, v. 17, n. 2, p. 143–162, 1997.

QUINLAN, R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann: Morgan Kaufmann, 1993.

SILVA, M. P. S. et al. Mining patterns of change in remote sensing image databases. In: IEEE. *Fifth IEEE International Conference on Data Mining (ICDM'05)*. [S.l.], 2005. p. 8–pp.

SOARES-FILHO, B. S. et al. Cenários de desmatamento para a amazônia. *Estudos Avançados*, SciELO Brasil, v. 19, n. 54, p. 137–152, 2005.

TURNER, M. G. Landscape ecology: what is the state of the science? *Annual review of ecology, evolution, and systematics*, JSTOR, p. 319–344, 2005.