Anais XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia, Brasil, 16-21 abril 2005, INPE, p. 457-464.

A land cover map for the Brazilian Legal Amazon using SPOT-4 VEGETATION data and machine learning algorithms

João Manuel de Brito Carreiras [1]
José Miguel Cardoso Pereira [1]
Manuel Lameiras Campagnolo [2]
Yosio Edemir Shimabukuro [3]

[1] Instituto Superior de Agronomia, Departamento de Engenharia Florestal
Tapada da Ajuda, 1349-017 Lisboa, Portugal
{jmbcarreiras, jmcpereira}@isa.utl.pt

[2] Instituto Superior de Agronomia, Departamento de Matemática
Tapada da Ajuda, 1349-017 Lisboa, Portugal
mlc@math.isa.utl.pt

[3] Instituto Nacional de Pesquisas Espaciais, Divisão de Sensoriamento Remoto
Avenida dos Astronautas 1758, CEP 12227-010, São José dos Campos, SP, Brasil
yosio@ltid.inpe.br

Abstract. The main purpose of this work is to assess the potential of multitemporal images from the 1-km SPOT-4 VEGETATION (VGT) sensor to obtain a land cover map of the Brazilian Legal Amazon (BLA) for the year 2000, discriminating primary tropical forest, *cerrado* savanna, agriculture/pasture, natural/artificial waterbodies, and secondary succession forest, using a probability-bagging classification tree (PBCT). The 10-fold cross validation procedure yielded an overall sample accuracy of 0.92. Besides, this algorithm allowed us to build a class membership probability map, with ~80% of the pixels with class membership probability greater or equal than 0.8. The estimated total area of agriculture/pasture and secondary succession forest in the BLA was 877,435 km^2 and 57,636 km^2, respectively. Comparison with an existing land cover map indicates that agriculture/pasture occurred primarily in areas previously occupied by primary tropical forest (45.8%) and *cerrado* savanna (32.7%).

Palavras-chave: Brazilian Legal Amazon (BLA), land use/land cover change, SPOT-4 VEGETATION (VGT), machine learning, Amazónia Legal Brasileira, alterações de ocupação do solo, aprendizagem automática.

1. Introdução

In the Amazon Basin, the major tropical forest ecosystem in the world, there is growing concern with deforestation and its influence on the carbon cycle (Skole and Tucker, 1993; Achard et al., 2002). Identification of areas undergoing forest regeneration is also important, as these play an important role as carbon sink (Brown and Lugo, 1990). Due to the extent of the Brazilian Legal Amazon (BLA), as well as restricted accessibility throughout most of the area, remote sensing data have an important role in characterizing land use/land cover change.

The Brazilian Institute for Space Research (Instituto Nacional de Pesquisas Espaciais – INPE) has reported that BLA deforestation in areas of primary tropical forest reached a value of 587,727 km^2 by the end of 2000, including 97,000 km^2 of old deforestation (prior to 1960) in Pará and Maranhão (INPE, 2002). Native vegetation clearing in regions dominated by *cerrado* savanna is not well documented, mainly because spectral differences between natural vegetation and agricultural crops are more subtle than those observed when primary tropical forest is converted to agriculture (Nepstad et al., 1997). Nonetheless, there is evidence from integrated satellite and census data that extensive native vegetation clearing has occurred in regions of *cerrado* savanna between the early 1980s and the mid 1990s (Cardille and Foley, 2003). Also, several studies have shown that secondary succession forest establishes rapidly in abandoned areas (Uhl, 1987; Brown and Lugo, 1990), mainly because the cost of re-clearance is extremely high for most farmers (Moran et al., 1994).

The main objective of this work was to evaluate the capability of a time series of monthly composite images of the 1-km SPOT-4 VEGETATION (VGT) sensor for the year 2000, to produce a land cover map of the BLA. We were interested in discriminating primary tropical forest, *cerrado* savanna, agriculture/pasture, natural/artificial waterbodies, and secondary succession forest, using probability-bagging classification trees (PBCT). A special emphasis was placed on agriculture/pasture and secondary succession forest classes, due to their influence in the carbon cycle. An important contribution of this work was to map agriculture/pasture in areas previously occupied by primary tropical forest and *cerrado* savanna. The dataset used in this study has the potential for detecting deforested areas occurring in *cerrado* savanna regions, as phenological differences may be more evident on a seasonal basis. Accuracy assessment of the resulting land cover map was performed quantitatively, with an error matrix and with a derived map of class membership probability.

2. Study area

The BLA is a politically defined region of Brazil and encompasses the states of Acre, Amapá, Amazonas, Mato Grosso, Pará, Rondônia, Roraima, Tocantins, and a part of Maranhão (west of 44º W). Most of the BLA is included in the Amazon river basin, with the exception of southern Mato Grosso and western Maranhão, included in the Paraguai and Parnaíba river basins, respectively. This region covers an area of approximately 5,000,000 km^2, consisting primarily of closed tropical forest, but also including large areas of flooded forest and *cerrado* savanna (Goulding et al., 2003).

3. Data

The original dataset used in this study was a set of daily 1-km SPOT-4 VGT images spanning the entire year 2000, and covering the BLA (3360x2800 pixels). The dataset over the BLA reached only to 45º W, thus missing the portion of the state of Maranhão between 44º and 45º W. We used the S1 product, consisting of 1-km georeferenced, calibrated, atmospherically corrected surface reflectance data (Passot, 2000). The high temporal resolution and the low cost of SPOT VGT data renders them appropriate for land use/land cover change monitoring at regional to global scales. We have further combined SPOT-4 VGT S1 daily images into 12-monthly composite images from January to December 2000. The compositing algorithm is described in Carreiras and Pereira (2004). Training and testing data were collected over the study area using high-resolution remote sensing imagery as ancillary data. A set of 19 Landsat 5 TM and Landsat 7 Enhanced TM Plus (ETM+) scenes from the years 2000 and 2001 were used. Areas of secondary succession forest were selected using two dates per scene. Landsat 5 TM and 7 ETM+ were obtained from INPE, in the scope of the Prodes Digital Project, and from the Global Land Cover Facility (GLCF), University of Maryland (USA).

The location and distribution of agriculture/pasture and secondary succession forest was assessed by comparison with an existing 59-class vegetation map of the BLA, prepared by the Instituto Brasileiro de Geografia e Estatística (IBGE) - Mapa de Vegetação do Brasil (IBGE, 1988). Those 59 classes were further aggregated into four major classes: primary tropical forest, *cerrado* savanna, transition forest, and other vegetation types.

4. Methods

4.1. Image classification

Our main focus was on the correct evaluation of the spatial distribution of agriculture/pasture and secondary succession forest. However, we also mapped areas of primary tropical forest, *cerrado* savanna, and natural/artificial waterbodies. Areas that were deforested and currently support agriculture or pasture were easily identified in Landsat TM/ETM+ data, due to their spectral characteristics and regular field geometry. Areas of forest regeneration were identified based on two dates of Landsat TM/ETM+ data; an area was considered to support secondary succession forest if in the first date it had been subjected to deforestation, and in 2000/2001 it no longer supported agriculture or pasture. Each training sample polygon extracted from the VGT imagery comprises a maximum of four pixels.

Classification trees are a non-parametric method. A classification tree partitions the space of all possible spectral signatures x, starting with the whole spectral space (the root of the tree) and successively splitting that space in subsets, such that each subset is more likely to be assigned to one of the land cover classes than the subset from which it is split (Breiman et al., 1984). All subsets of the spectral space are represented by nodes in a tree, and each split corresponding to the descendents of a node. Since each node of the tree represents an element P_j of a partition of the space of all possible x, one can estimate $P(class_i/x \in P_j)$ for all terminal nodes and all classes, and assign the node to the class with the highest probability. The estimate of $P(class_i/x \in P_j)$ is simply the proportion of pixels that belong to $class_i$ among all the training sample pixels that are in P_j. In the Classification and Regression Trees (CART) algorithm of Breiman et al. (1984) heuristic techniques are used to find a tree structure that discriminates the classes (i.e. which terminal nodes have a high proportion of sample individuals of some class) but is not overfitted to the training sample (i.e. the tree should not be “too large”).

Classification trees are sensitive to small perturbations in the training set, which may originate large changes in the resulting classifiers (Breiman, 1996). Therefore, these unstable methods can have their accuracy improved with a perturbing and combining technique, that is, by generating multiple perturbed versions of the classifier (a.k.a. ensemble, or committee) and combining those into a single predictor (Breiman, 1998). These methods can be divided in two types: those that adaptively change the distribution of the training set based on the performance of previous classifiers (e.g. boosting) and those that do not (e.g. bagging) (Bauer and Kohavi, 1998). In this study we will only focus on the bagging algorithm applied to CART. In bagging, each sub-classifier c_i ($i=1 \ldots n$) is run on n different b_i bootstrap samples of the original m training set observations. Each b_i is generated by uniformly sampling m observations from the training set with replacement. The final classifier C is built from c_i, whose output is the class most frequently predicted by its sub-classifiers, with ties broken arbitrarily (Breiman, 1996). Although the main purpose of bagging was to build a strong classifier by means of variance reduction (Breiman, 1996), some variants of bagging have also proven adequate for the estimation of class membership probability (Provost and Domingos, 2003). Probability-bagging classification trees (PBCT) is one of such variants, so that instead of returning a classification, each sub-classifier returns a probability distribution for the classes in each terminal node (Bauer and Kohavi, 1998; Provost and Domingos, 2003). Subsequently, the PBCT algorithm averages the probability for each class over all sub-classifiers, and predicts the class with the highest probability. However, Provost and Domingos (2003) note that these probability estimators of class membership are not unbiased. Nevertheless, those estimates can be useful in land cover mapping, by assigning to each pixel a relative degree of classification confidence. In this study we used 25 bootstrap replicates to build a PBCT, evaluated with a 10-fold cross validation approach.

4.2. Accuracy assessment

The error matrix of each classifier is used as an approximation for the accuracy of the land cover map. The difficulty in obtaining up-to-date independent information of spatial distribution of land cover classes in the BLA led us to choose a complementary approach. As mentioned before, the application of the PBCT algorithm can provide information of class membership probability. Therefore, application of this classifier to the entire BLA can provide a map of class membership probability, that is, in each pixel, the highest averaged probability over the sub-classifiers derived from the 25 replicates of the training samples.

5. Results and discussion

5.1. Image classification

A total of 8386 pixels of known land cover, corresponding to 2264 sampling polygons, were identified in the Landsat TM/ETM+ imagery and overlaid on the SPOT-4 VGT monthly composites. The precision of the PBCT algorithm was evaluated using 10-fold cross validation, with the corresponding error matrix (Table 1).

Table 1 – Confusion matrix for the PBCT algorithm, using the 10-fold cross validation approach.

	Predicted class (# pixels)						
Observed class (# pixels)	(1)	(2)	(3)	(4)	(5)	Total	Omission error
Primary tropical forest (1)	3457	60	37	0	24	3578	0.034
Cerrado savanna(2)	94	692	99	1	0	886	0.219
Agriculture/pasture (3)	70	98	3133	0	36	3337	0.061
Natural/artificial waterbodies (4)	0	0	0	351	0	351	0.000
Secondary succession forest (5)	97	4	54	0	79	234	0.662
Total	3718	854	3323	352	139	8386	
Commission error	0.070	0.190	0.057	0.003	0.432		

The natural/artificial waterbodies, agriculture/pasture, and primary tropical forest classes, displayed commission and omission errors below 0.08. The *cerrado* savanna class had higher commission and omission errors, around 0.20. The secondary succession forest class was the most problematic, with highest commission (0.432) and omission (0.662) errors. The major confusion occurs between the secondary succession forest class and the primary tropical forest and agriculture/pasture classes. This is understandable, since from a spectral standpoint, secondary succession forest is a transitional class between agriculture/pasture and primary tropical forest.

Application of the PBCT classifier to the entire dataset resulted in a land cover map of the BLA for the year 2000 (Figure 1). The class represented in each pixel is that with the highest probability, averaged over the 25 sub-classifiers. The area and proportion of each land cover class per state is presented in Table 2. The state of Maranhão had the highest proportion of agriculture/pasture, with 60.0% (119,831 km^2), followed by Mato Grosso with 35.4% (320,523 km^2), and Tocantins with 34.3% (95,661 km^2). The states of Amazonas (1.2%), Amapá (8.6%), and Acre (10.4%), displayed a relatively low proportion of agriculture/pasture. Secondary succession forest was concentrated in Pará with 36,478 km^2 (2.9%), Maranhão with 11,286 km^2 (5.7%), Amazonas with 3413 km2 (0.2%), and Mato

Grosso with 2961 km^2 (0.3%). Deforestation started long ago in the states of Maranhão and Pará (Nepstad et al., 1997) and a part of this area appears to have been abandoned, allowing for the regrowth of secondary succession forest.

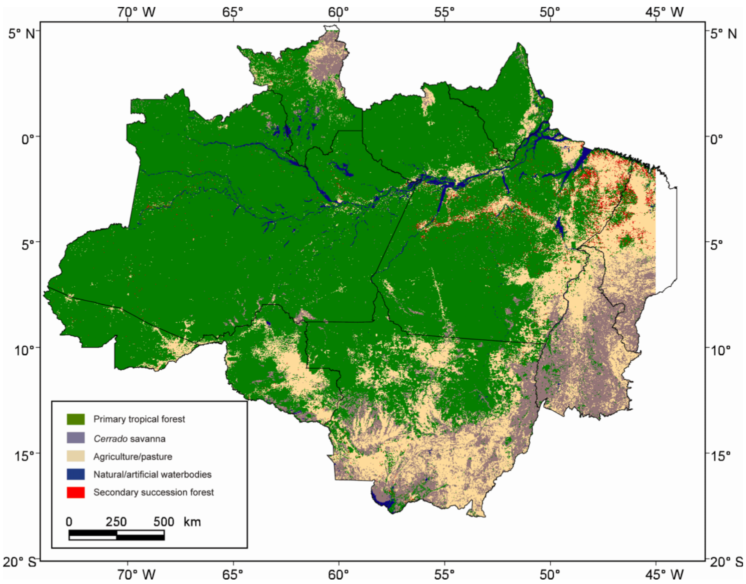

Figure 1 – Land cover map of the BLA for the year 2000, derived from the PBCT algorithm.

Table 2 – Distribution of land cover classes obtained from the PBCT algorithm, per state of the BLA, for the year 2000. Total/state land cover class percentage is indicated in parentheses.

	Land cover classes [km^2(%)]					
State	Primary tropical forest	*Cerrado* savanna	Agriculture/ pasture	Natural/artificial waterbodies	Secondary succession forest	Total
Acre	141,593(89.2)	59(0.0)	16,440(10.4)	0(0.0)	681(0.4)	158,773
Amapá	124,915(87.4)	2707(1.9)	12,314(8.6)	2502(1.8)	474(0.3)	142,911
Amazonas	1,538,919(96.1)	12,205(0.8)	18,613(1.2)	28,879(1.8)	3413(0.2)	1,602,030
Maranhão *	19,447(9.7)	48,336(24.2)	119,831(60.0)	717(0.4)	11,286(5.7)	199,617
Mato Grosso	368,962(40.8)	208,367(23.0)	320,523(35.4)	4104(0.5)	2961(0.3)	904,917
Pará	955,362(76.5)	24,440(2.0)	200,172(16.0)	32,996(2.6)	36,478(2.9)	1,249,449
Rondônia	154,546(64.3)	16,627(6.9)	67,447(28.1)	754(0.3)	1008(0.4)	240,383
Roraima	161,815(72.1)	31,124(13.9)	26,434(11.8)	3912(1.7)	1141(0.5)	224,425
Tocantins	15,733(5.6)	166,251(59.7)	95,661(34.3)	802(0.3)	193(0.1)	278,640
Total	3,481,292(69.6)	510,116(10.2)	877,435(17.5)	74,666(1.5)	57,636(1.2)	5,001,145

* the state of Maranhão is only included west of 45° W

If the combined area mapped as agriculture/pasture and as secondary succession forest in 2000 (935,071 km^2) can be viewed as a proxy for the total deforested area in the BLA by 2000, then this number is almost twice the cumulative deforestation value of 587,727 km^2 reported by INPE (2002) up to the year 2000. Discrepancies between this study and INPE's estimate are more evident in those states where the proportion of *cerrado* savanna is higher (Tocantins, Maranhão, Mato Grosso, and Pará). Consequently, our results bring further evidence of large-scale deforestation in *cerrado* savanna. The estimated value of 57,636 km^2 for the overall extent of secondary succession forest in the BLA is significantly lower than those available in the literature, namely the 157,973 km^2 from Lucas et al. (2000) for early 1990s. These authors mapped forest regeneration with an unsupervised classification algorithm and NOAA AVHRR data, relying in ancillary information to label the resulting clusters. The fact that our study does not consider the part of Maranhão between 44º and 45º W could explain a fraction of that difference, as this state has one of the highest rates of forest regeneration. Perhaps, most important is the fact that mitigation measures implemented by some state governments to abate deforestation of primary tropical forest and *cerrado* savanna (e.g. Fearnside, 2003) could result in the re-deforestation of secondary succession forest, thus reducing its extent.

The incidence of agriculture/pasture and secondary succession forest was assessed by comparison with a 59-class vegetation map of Brazil (IBGE, 1988), which was condensed into four classes. This analysis indicates that in Amapá, Maranhão, Mato Grosso, Roraima, and Tocantins, establishment of agriculture/pasture in regions of *cerrado* savanna has been as significant as in areas of primary tropical forest. Our analysis indicates that only 45.8% (401,866 km^2) of the areas in the BLA with agriculture/pasture in 2000 were established in areas previously occupied by primary tropical forest. A large amount of the area with agriculture/pasture in 2000 was located in regions formerly occupied by *cerrado* savanna (32.7%, 286,921 km^2); the remaining was established in areas of transition between the previous classes (18.9%, 165,835 km^2), and, to a much lesser extent, in other vegetation types (2.6%, 22,813 km^2). The distribution of the areas of secondary succession forest in the BLA in 2000 indicates that the majority of this class occurred in areas previously occupied by primary tropical forest (87.8%, 50,604 km^2); the exceptions are the states of Mato Grosso, Tocantins, and Amapá, where forest regeneration appears in areas originally covered by transition forest, *cerrado* savanna, and other vegetation types, respectively. This is not surprising, as farmers often abandon deforested areas due to intense vegetation regrowth in areas of primary tropical forest (Moran et al., 1994; Nepstad et al., 1997).

Therefore, the abovementioned discrepancies between this study and the numbers of INPE (2002) are largely explained by the extent of agriculture/pasture established in areas of *cerrado* savanna, not considered in the analysis carried out by INPE.

5.2. Accuracy assessment

A map of class membership probability for the class with the highest probability was derived from the output of the PBCT algorithm (Figure 3). Used in combination with the land cover map of Figure 2, it gives useful information about the relative degree of membership of the most probable class.

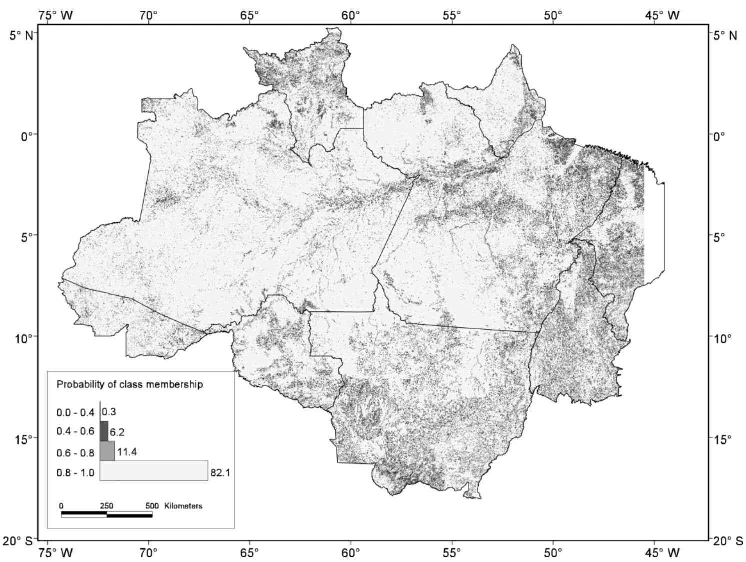

Figure 2 – Map of classes of class membership probability for the land cover map of the BLA obtained from the PBCT algorithm. The colour of each bar represents a given class of class membership probability, and the size is proportional to their extent in the map (ahead is the percentage of each class in the BLA).

It is evident that natural/artificial waterbodies and primary tropical forest have the highest degree of class membership, i.e. they are the classes mapped with the highest degree of confidence. Some areas associated with *cerrado* savanna and agriculture/pasture display lower class membership probability. The class with highest commission and omission errors, secondary succession forest, also has low class membership probability, meaning that allocation of a pixel to this class is done with a relatively low degree of confidence. The lower probability of class membership in the state of Roraima may be due to residual cloud contamination in the monthly composite images. These class membership probability estimates were compared with class-specific accuracy (i.e. commission error). The class-specific mean membership probability is highly inversely correlated with classification errors, i.e. higher classification errors tend to have lower class membership probability. Also, the overall mean probability of class membership for the entire BLA was 0.90 (0.148 standard deviation), which is similar to the overall accuracy (0.92). Consequently, the estimates of class membership probability add supplementary information regarding map quality.

6. Conclusions

The PBCT algorithm performs well in most of the mapped land cover classes, with the exception of the secondary succession forest class, which has higher commission and omission errors. This class is extensively confused with primary tropical forest and agriculture/pasture. The higher amount of aggregated agriculture/pasture and secondary succession forest classes in 2000 obtained in this study (935,071 km^2), when compared with deforestation estimates up to 2000 from INPE (2002) (587,727 km^2) is mostly due to agriculture/pasture occurring in areas previously occupied by *cerrado* savanna. It appears that agriculture/pasture establishment in areas previously occupied by cerrado savanna was as important as in areas of primary tropical forest. The possibility of obtaining maps of class membership probability derived from PBCT further improved the characterization of the proposed 5-class land cover map. We have shown that higher class membership probability is associated with lower classification errors. This approach adds supplementary information for accuracy assessment of land cover maps. This study demonstrated the utility of the SPOT-4 VGT sensor in predicting the extent of primary tropical forest, *cerrado* savanna, natural/artificial waterbodies, and agriculture/pasture in the BLA with a reasonable degree of accuracy. The land cover map produced may be useful for analysis of regional carbon and water fluxes, and for evaluating impacts of land use/land cover change on biotic diversity and soil degradation.

Acknowledgments

João M. B. Carreiras work was partially developed at Instituto Nacional de Pesquisas Espaciais (INPE, Brazil), as a contribution to the GLC 2000 project and to the LBA Experiment in Amazonia. This work was funded by a doctoral grant from the Ministério da Ciência e Tecnologia, Fundação para a Ciência e a Tecnologia, Portugal (Ref. PRAXIS XXI/BD/21507/99). VEGETATION images were made available in the framework of the GLC 2000 and GBA 2000 projects of the Joint Research Centre (JRC) of the European Commission. We acknowledge Instituto Brasileiro de Geografia e Estatística (IBGE) for the Mapa de Vegetação do Brasil (Vegetation Map of Brazil), distributed by the University of New Hampshire, EOS-WEBSTER Earth Science Information Partner (ESIP). We acknowledge the Brazilian Institute for Space Research and the Global Land Cover Facility (University of Maryland, USA) for the Landsat TM and ETM+ used in this study.

References

Bauer, E.; Kohavi, R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. **Machine Learning**, v. 36, p. 105-142, 1998.

Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. **Classification and Regression Trees**. Pacific Grove, California: Wadsworth, 1984. 519 p.

Breiman, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123-140, 1996.

Breiman, L. Arcing classifiers. **The Annals of Statistics**, v. 26, n. 3, p. 801-849, 1998.

Brown, S.; Lugo, A.E. Tropical secondary forests. **Journal of Tropical Ecology**, v. 6, n. 1, p. 1-32, 1990.

Cardille, J.A.; Foley, J.A. Agricultural land-use change in Brazilian Amazônia between 1980 and 1995: evidence from integrated satellite and census data. **Remote Sensing of Environment**, v. 87, p. 551-562, 2003.

Carreiras, J.M.B.; Pereira, J.M.C. SPOT-4 VEGETATION multitemporal compositing for land cover change studies over tropical regions. **International Journal of Remote Sensing** (aceite), 2004.

Fearnside, P.M. Deforestation control in Mato Grosso: a new model for slowing the loss of Brazil's Amazon forest. **Ambio**, v. 32, n. 5, p. 343-345, 2003.

Goulding, M.; Barthem, R.; Ferreira, E. **The Smithsonian Atlas of the Amazon**. Washington, USA: Smithsonian Books, 2003. 255 p.

IBGE. **Mapa de Vegetação do Brasil, escala 1:5000000**. Brasília, DF, Brazil: Ministério da Agricultura, Instituto Brasileiro de Desenvolvimento Florestal, Secretária de Planejamento e Coordenação da Presidência da Republica, Fundação Instituto Brasileiro de Geografia e Estatística, 1988.

INPE. **Monitoring of the Brazilian Amazonian forest by satellite, 2000-2001**. São José dos Campos, SP, Brazil: Instituto Nacional de Pesquisas Espaciais, 2002. 21 p.

Lucas, R.M.; Honzák, M.; Curran, P.J.; Foody, G.M.; Milne, R.; Brown, T.; Amaral, S. Mapping the regional extent of tropical forest regeneration stages in the Brazilian legal Amazon using NOAA AVHRR data. **International Journal of Remote Sensing**, v. 21, n. 15, p. 2855-2881, 2000.

Moran, E.F.; Brondizio, E.; Mausel, P.; Wu, Y. (1994). Integrating Amazonian vegetation, land-use, and satellite data. **Bioscience**, v. 44, n. 5, p. 329-338, 1994.

Nepstad, D.C.; Klink, C.A.; Uhl, C.; Vieira, I.C.; Lefebvre, P.; Pedlowski, M.; Matricardi, E.; Negreiros, G.; Brown, I.F.; Amaral, E.; Homma, A.; Walker, R. Land-use in Amazonia and the *Cerrado* of Brazil. **Ciência e Cultura Journal of the Brazilian Association for the Advancement of Science**, v. 49, n. 1/2, p. 73-86, 1997.

Passot, X. VEGETATION image processing methods in the CTIV. In: VEGETATION 2000 - 2 years of operation to prepare the future, 3-6 April 2000, Lake Maggiore, Italy. **Proceedings**. Ispra, Italy: SAI/JRC, 2000, p. 15-21.

Provost, F.; Domingos, P. Tree induction for probability-based ranking. **Machine Learning**, v. 52, n. 3, p. 199-215, 2003.

Skole, D.; Tucker, C. Tropical deforestation and habitat fragmentation in the Amazon: Satellite data from 1978 to 1988. **Science**, v. 260, p. 1905– 1910, 1993.

Uhl. C. Factors controlling succession following slash-and-burn agriculture in Amazonia. **Journal of Ecology**, v. 75, n. 2, p. 377-407, 1987.