

OBJECT-BASED IMAGE ANALYSIS AND DATA MINING FOR MAPPING SUGARCANE WITH LANDSAT IMAGERY IN BRAZIL

A. R. Formaggio^{a,*}, M. A. Vieira^a, C. D. Rennó^a, D. A. Aguiar^a, M. P. Mello^a

^a INPE, Instituto Nacional de Pesquisas Espaciais, 12.227-010, São José dos Campos, SP, Brasil
(formag, matheus, camilo, daniel, pupin)@dsr.inpe.br

Commission VI, WG IV/4

KEY WORDS: Sugarcane mapping, Artificial Intelligence, Object-based Image Analysis, Data Mining, Landsat images

ABSTRACT:

Mapping of sugarcane planted area is an important information for decision making, mainly when the search for alternatives to mitigate greenhouse gas emissions has indicated the use of biofuels as a viable option. Thus, the aim of this research was to develop a methodology in order to automate the sugarcane mapping task when remote sensing data are used. Thus the integration of two major approaches of Artificial Intelligence, Object-Based Image Analysis (OBIA) and Data Mining (DM), were tested in a study area located in São Paulo state, which is well representative of the agriculture of large regions of Brazil and other countries. OBIA was used to emulate the interpreter knowledge in the process of sugarcane mapping, and DM techniques were employed for automatic generation of knowledge model. A time series of four Landsat images was acquired in order to represent the wide variability of the patterns during sugarcane crop season. Definiens Developer® multiresolution segmentation algorithm produced the objects and properly trained decision tree applied to the Landsat data for the generation of the thematic map with sugarcane as the main class of interest. An overall accuracy of 94% (Kappa = 0,87) was obtained, showing that OBIA and DM are very efficient and promising in the direction of automating the sugarcane classification process with Landsat multitemporal time series.

1. INTRODUCTION

Agriculture has an important role in the global socioeconomic context, and against this background the sugarcane rises as one of the major crops, with Brazil being the biggest producer and exporter of sugarcane products (Rudorff *et al.*, 2009) stimulated by the growing worldwide demand for biofuels. The area planted with sugarcane has grown 94.3% in the country since year 2000, occupying today approximately 9.4 million hectares (IBGE, 2008).

Customarily, sugarcane mapping with remote sensing (RS) images is made through visual interpretation (Rudorff *et al.*, 2009; Aguiar *et al.*, 2009), but this procedure, albeit capable of providing precise and consistent results, has proved to be costly both in terms of time and the great number of people involved, considering the large extensions of sugarcane crop in Brazil.

In that sense, there is an evident requirement of process automation when RS images must be classified (DeFries & Chan 2000). However, conventional procedures of digital classification have met with great difficulties in the automated recognition of sugarcane patterns, due to the high variability of patterns found along its crop cycle.

In the described context, Knowledge-Based Systems (KBS), a branch of Artificial Intelligence (AI), appear with a large potential. According to Cohen and Shoshany (2005), while conventional systems perform algorithmic processing guided only by statistical variables of data, KBS are computing systems based on knowledge that have the ability to emulate or replicate human judgment in structured or well-known problems. Inside this context, we find the Object-Based Image Analysis (OBIA),

a new approach that makes possible a simulation of visual interpretation through knowledge-modeling. To that end semantic nets are built based on the usage of attributes such as shape, spectral behavior, texture, morphology, and context, among others that may be used in image analysis (Blaschke, 2010).

A task that is certainly the most important for knowledge-based applications such as OBIA, but is often difficult to perform, is the acquisition of knowledge (Witten & Frank, 2005). In most cases the process is typically slow and arduous, because the specialist finds it difficult to express and organize his knowledge in the form of a semantic net; and thus, an interesting and potential help is the adoption of Data Mining (DM) techniques to generate automatically the structure of knowledge.

DM is a special stage within the process known as Knowledge Discovery in Databases (KDD), and it involves the selection and application of intelligent techniques that will extract patterns of interest for the effective production of knowledge (Fayyad *et al.*, 1996).

Considering the above mentioned situation, in regard to the need of obtaining information on sugarcane in extensive areas, we observe the importance of developing and testing methodologies based on AI techniques that will be able to support the automation of the process of classification of RS images.

Thus, the objective of the present work has been to propose and evaluate a method to map sugarcane through the integration of DM and OBIA, starting from a time series of Landsat data. Both

* Corresponding author.

DM and OBIA are founded on knowledge; however, in general terms and for the methodology considered in the present research, it is adequate to mention that, while DM deals with knowledge discovery, OBIA shall be in charge of representing the acquired knowledge.

The class of interest to be mapped in the images of the present research was concentrated in “areas planted with sugarcane and available for harvesting”. Therefore, the proposed approach should be able to model the process of identification and extraction of information from a time series of Landsat images, having as final objective to automatically find/determine the area with sugarcane available for harvesting in a given region of interest and in a given crop year.

2. METHOD

2.1 Study Area

The area studied in the survey includes the municipalities of Ipuã, Guará and São Joaquim da Barra, all three located in the North of the State of São Paulo (Fig. 1).

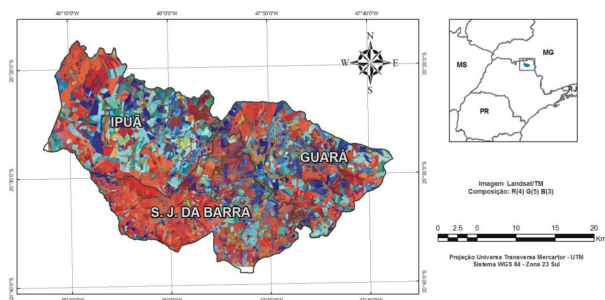


Figure 1: Localization of the Study Area, highlighting the municipalities of Guará, Ipuã, and São Joaquim da Barra (central coordinates: 47°54'W x 20°30'S), in the State of São Paulo (SP), Brazil.

The three appointed municipalities have together a total area of 124,100 ha, and the region has been selected because it represents well the farming conditions of a great part of the Southeastern and Southern regions in Brazil, and of several countries with tropical agriculture. Major crops in the region include: cotton, peanut, rice, sugarcane, bean, manioc, corn, soybean, sorghum, tomatoes, banana, coffee, and orange. However, soybean, sugarcane, and corn crops are predominant and occupy the largest part of planted areas in the region (IBGE, 2008).

2.2 Dataset and pre-processing

Four dates have been selected for images captured by sensors TM/Landsat-5 and ETM+/Landsat-7, to make up the time series used to map the area with sugarcane available for harvest: September 20, 2000 (TM/Landsat-5), October 22, 2000 (TM/Landsat-5), February 19, 2001 (ETM+/Landsat-7), and March 23, 2001 (ETM+/Landsat-7), corresponding to key periods for the extraction of sugarcane-related information.

The images have been georeferenced through the use of 21 control points collected in the field with a double-frequency

GPS (Global Position System) receptor. Coordinates of those points have been processed and referenced to the WGS84 system through the utilization of two landmarks of the Brazilian Network of Continuous Monitoring (*Rede Brasileira de Monitoramento Contínuo*, UBER / MGUB) located in the city of Uberlândia, State of Minas Gerais (MG). Georeferencing has been applied to each image by adjusting the coordinates with the use of a first-degree polynomial, and with interpolation applying the nearest neighbour technique. All procedures have produced values of RMS (Root Mean Square) under 0.5 pixel.

All four used Landsat dates have been normalized using the technique known as Iteratively Re-weighted Multivariate Alteration Detection (IR-MAD), proposed by Schroeder *et al.* (2006).

In order to validate results obtained in the present survey, a reference map has been produced by an expert interpreter through visual interpretation of the images kept in the database, and through fieldwork.

2.3 Knowledge discovery process

Knowledge discovery has been performed based on what is proposed by Hsu *et al.* (2002). Starting from the time series of images defined for classification of sugarcane, the first step it to build a training set characterizing the discovery of characteristics of the studied object. Later, attributes defined in the training set are processed by the structural classifier in the Data Mining stage, and finally the acquired knowledge is evaluated.

2.3.1 Building the training set

Basically, this stage is constituted by the extraction or discovery of characteristics of the main object of this study, that is, sugarcane. The training set is a subset of the time-series images formed by sample objects (training instances) selected from segments of those images. From the subset is created a Decision Tree (DT) that is exercised during the Data Mining stage in order to enable the discovery of patterns representative of sugarcane in the used time series.

The most important stages in the construction of the training set are: (1) Segmentation of images; (2) Definition of the classes of interest; (3) Selection of training instances; and (4) Discovery of attributes.

In this work we have adopted the multiresolution segmentation algorithm developed by Baatz and Schäpe (2000), that was implemented in the Definiens Developer® platform to define the objects. Then we have determined that the classes of interest would be: 1) sugarcane, and 2) others; considering that the first represents areas planted with sugarcane that will be available for harvest in the respective harvest-year, while the class “others” includes all other types of land use that may be present in the studied region.

The next stage in the discovery of characteristics is the selection of a set of sample objects. Through a process of cognitive evaluation, an interpreter with expertise in sugarcane mapping analyses such objects and associates them to one of the classes of interest previously defined. Working this way, a total amount of 396 objects have been selected, among which 184 objects represented sugarcane patterns, and the remaining ones represented patterns referring to the “others” class.

To complete construction of the training set, the next stage has been the extraction of attributes. For each sample object selected in the previous stage, attributes defined by the platform Definiens Developer® have been extracted. Several types of attributes that can be used in image analysis are implemented inside that environment; equations describing each of those attributes may be accessed in DEFINIENS (2006). Still in this stage, we may highlight the inclusion of the attribute *NDVI* (*Normalized Difference Vegetation Index*) proposed by Rouse *et al.* (1973), since this vegetation index senses the amount and vigor of phytomass existing in the scene under study.

2.3.2 Data Mining

Once the training set has been conveniently organized, it will be used in the next stage to mine the different patterns associated to both sugarcane and others patterns.

The DM stage involves selection and application of intelligent techniques in order to extract patterns of interest for the effective production of knowledge (Fayyad *et al.*, 1996).

In the present work, the C4.5 algorithm developed by Quinlan (1993) has been selected to generate the model of knowledge. That algorithm generates Decision Trees (DT) from a given training set, and because this phase of the methodology is performed inside the WEKA computing environment (Weka, 2010), the training set created in the previous stage has been imported to this environment. And starting from there, a series of experiments followed.

First, a model of knowledge (Decision Tree) was generated from the original training set. Later, in order to establish whether the algorithm would be able to distinguish objects from the training set without the attributes used in the first model, the most important attributes have been deleted from original data. Then a second model has been generated, and its efficacy in distinguishing objects has been tested. Such tests are made until the best structure of knowledge (DT) is achieved for the problem raised.

2.3.3 Interpretation and evaluation of knowledge

The stage of interpretation and evaluation of knowledge is when the patterns identified during the Data Mining stage are interpreted and evaluated. First a subjective analysis was made of the findings brought by DM. In that stage a coherence of the generated model of knowledge was observed. To that end a few pieces of relevant information were taken into consideration: 1) the size of the decision tree; 2) the attributes and thresholds defined; 3) branches that have presented a greater amount of errors.

A cross validation (Good, 2001) has been applied to the model of knowledge to obtain a statistical parameter that is more strict in regard to the generated Decision Tree. This way, when the DT was to be exercised, the training set has been partitioned into two subsets, and thus the exercise with the decision tree has been made with half the instances of the training set, while a cross validation test was performed with the other half.

2.3.4 Application of the Decision Tree

Once the Decision Tree had been trained and tested, we advanced to the stage of representation of the acquired

knowledge. This phase configures the classification of time-series images, and its end product is a thematic map with areas planted with sugarcane available for harvest.

OBIA congregates the characteristics that are necessary to represent knowledge, and for that reason it uses a formalism that gathers a set of nodes connected by means of a set of archs (Bittencourt, 2006), resembling the DT structure. That characteristic has enabled the DT to be replicated within the Definiens Developer® platform, considering the structure and relationship of the archs and nodes defined in the stage of knowledge discovery.

In the final stage of the methodology, the algorithm of hierarchical classification implemented in Definiens platform has been executed. The algorithm performs this classification across a structure of classes and superclasses, through a relation of dependence subjected to the acquired knowledge (Definiens, 2006). The process classifies the image top-down, that is, it first separates classes that are more general or have a higher separability, and then separates subclasses or classes with a higher uncertainty, classifying parts of the image in sequence.

The hierarchical classification closes the classification methodology proposed in this work. The final result is the thematic map that identifies areas planted with sugarcane that shall be available for harvest.

2.3.5 Validation of the classification

The classification generated according to the methodology proposed in the present research – integrating DM and OBIA – has been evaluated using as reference the map resulting from visual interpretation and fieldwork. The purpose of it has been to make a direct comparison between visual interpretation and the classification method tested in this research. This assessment of the quality of the classification has been performed in a quantitative way, through indexes of global accuracy (GA) and Kappa (κ), both extracted from the confusion matrix. Details about these tests can be found in Congalton and Green (1999).

Confusion matrices have been generated through the intersection of sample points from the theme map and the reference map. In total, 500 points have been generated randomly and independently.

3. RESULTS AND DISCUSSION

3.1 Decision Tree

The algorithm C4.5 (Weka, 2010), using the previously defined training set, has generated the model of knowledge represented by the Decision Tree.

DT training has been performed with half the instances of the training set, while a cross validation test was performed with the other half. We verified that the DT achieved a success rate of 96.97% in the classification of tested data, reaching the Kappa index of 0.94. Among the 396 instances of the training set, only 12 have been classified incorrectly.

The attributes selected and used to describe sugarcane patterns in the present case were: *NDVI_March*, *NDVI_September*, *CLCM*, *Homegeneity*, *September_bandTM4*, *Mean*

February_bandTM3, GLCM Homogeneity October_bandTM5, GLCM Homogeneity October_bandTM3, GLCM Homogeneity February_bandTM4, NDVI_February, Mean October_bandTM5, Mean February_bandTM5. Detailed descriptions of those attributes can be found in Definiens (2006).

3.2 Classification of the time series

Starting from the trained model, and exploring the potential of the Definiens platform in structuring knowledge through networks, we have then proceeded to sugarcane classification.

The DT has been replicated inside that environment with the purpose of obtaining the final thematic map. This process is characterized as the stage of representation of knowledge, when the user has visual contact with the result obtained with DM. Figure 2 presents the classified map of the areas with sugarcane available for harvest, which has been obtained according to the methodology proposed in this research.

The map presents the spatial distribution of sugarcane plots in the study area. The first partitioning made by the DT used the attribute *NDVI_mar*. This first node represents the root of the DT and virtually separates all the possible soybean areas from the possible sugarcane areas present in the study area. In the March image a good part, if not all the soybean, had already been harvested, leaving on soil surface just the straw left after harvesting. Because sugarcane was in its maximum vegetative strength at the time, through the attribute *NDVI_mar* the classifier could quite well separate one crop from the other.

However, in the first moment it was not yet possible to separate sugarcane completely from the other classes. For example, forest regions also present high values of NDVI. Therefore, algorithm C4.5 uses another strategy to separate these two patterns. Through the texture attribute *GLCM Homogeneity September_bandTM4*, the structural classifier separates part of the riparian vegetation pattern from the potential sugarcane areas (Figures 3-a, 3-b).

In order to classify sugarcane regions, the structural classifier has used a combination of the attribute *Mean February_bandTM3* followed by the attribute *GLCM Homogeneity_February_bandTM4*.

In the sequence, our classifier selects the texture attribute *GLCM Homogeneity_February_bandTM4* to finally classify sugarcane regions (Figure 3). The texture of plots together with the presence of roads inside the farms grants sugarcane crops a quite peculiar texture characteristic (Figure 3-c). It has enabled the structural classifier to identify, through the attribute *GLCM Homogeneity_February_bandTM4*, virtually all the regions with sugarcane crop in the region under study.

However, not all sugarcane regions had been apt to be identified up to that moment. Highlighted among these are the plots where sugarcane planting presented flaws, and/or there is a pattern of low-yield plants. These regions have shown a wrinkled texture that made it impossible for the classifier to identify them as sugarcane.

In order to classify those regions, the algorithm has then created a new branch. In this last partitioning, the DT separates sugarcane regions from a few soybean plots that have not been

harvested in March, and therefore have not been identified through NDVI in the first partition of the DT.

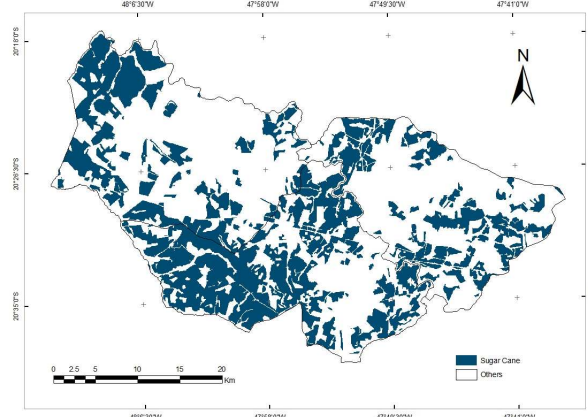


Figure 2: Classified map of the sugarcane areas available for harvest, obtained according to the methodology proposed in the present research, that combines OBIA and DM.

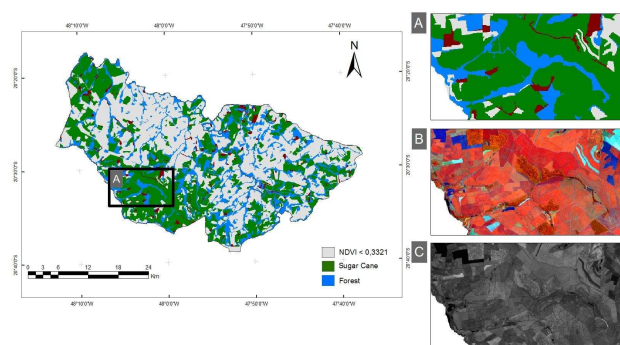


Figure 3: Identification of sugarcane areas: (A) detail of classification; (B) detail of the March image in composição R(4) G(5) B(3); (C) image of band 3 of the TM sensor, illustrating roads inside the plots.

To solve this problem, the classifier used the spectral attribute *Mean February_bandTM5*. Through calculation of the spectral mean of objects in the month of February, the DT could separate the few soybean regions that had not been identified up to that moment. In the region of medium-wave infrared – band 5 of the TM – electromagnetic radiation incident upon vegetation is fairly much absorbed by water (Ponzoni and Shimabukuro, 2007). Therefore, we may conclude that the DT uses that attribute because in the month of February, the content of water in soybean leaves is much lower than sugarcane's.

In the month of February, while soybean is in its senescence period, sugarcane is close to its maximum vegetative strength, and consequently the water content in sugarcane leaves is greater than in soybean leaves; therefore sugarcane shall present lower reflectance values in the mentioned band of the electromagnetic spectrum. It enables the DT to separate, using the spectral attribute *Mean_February_bandTM5*, the remaining soybean plots (Figure 4).

3.3 Validation of the classification

The confusion matrix (Table 1) has been generated after crossing the sample points with the reference map and the sugarcane map generated through the proposed methodology.

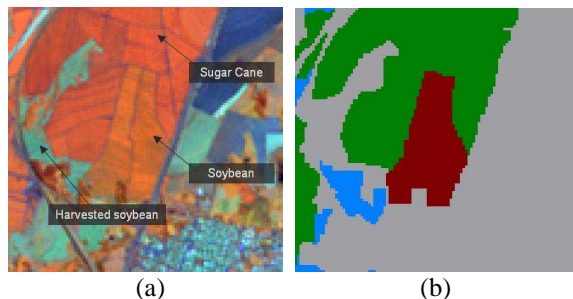


Figure 4: Separation of soybean areas not harvested in March: (a) detail of the the March image – composition R(4) G(5) B(3). (b) detail of the classified map showing the separation of the mentioned soybean plot.

The classification of sugarcane areas following the OBIA methodology has obtained a theme map whose Global Accuracy (GA) and kappa (k) indexes have been respectively 94% and 0.87 (Value-p=0,0000). The found value for Kappa coefficient actually represents the conformity of the classification with the reference map, and this has been confirmed by the Z test in which there is agreement between classifications and the reference image, for the adopted level of significance ($\alpha=5\%$).

In relation with the Global Accuracy index, according to Foody (2002) it is desirable for a classification to reach indexes above 85%, a condition that has also been confirmed, since our classification obtained GA of 94% (Table 1). The values found for Global Accuracy and Kappa indicate the quality of the classified map that has been generated.

		Reference map		
		Sugarcane	Others	Σ
OBIA	Classes			
	Cane	186	10	196
	Others	20	283	303
	Σ	206	293	500

Kappa coefficient = 0.87; Global Accuracy = 93.99%

Table 1 – Confusion matrix of the classification according to OBIA + DM methodology.

4. CONCLUSIONS

The present work has tried to investigate the viability of integrating two important approaches of Artificial Intelligence in sugarcane mapping based on orbital images of average spatial resolution. By exploring the potential of Data Mining and Object-Based Analysis, the present research brings relevant contribution for the search of automation as regards the process of classification of images from remote sensing, applied to agriculture mapping.

Results achieved with the experiments performed suggest that it is possible to automate the classification process with high levels of accuracy, which can be proven with the good Global Accuracy and Kappa indexes that could be reached, of 94% and 0.87 respectively.

One of the first aspects to consider when mapping sugarcane is the fact that the process is dynamic and progresses along the time. Therefore, the analysis of sugarcane patterns and the task of relating them to processes of soil use oblige the use of time series. If sugarcane patterns are analyzed based on data obtained in a single date, observation of the process shall be seriously damaged, since those patterns tend to be the result of a combination of processes of different periods, and different strategies of crop management. Considering these factors, the conducted experiments have used images that picture the field scene in several periods, respecting the timewise dynamics of sugarcane.

A relevant limitation in structural classification with the use of algorithm C4.5 is the high dependence of that algorithm in relation to quality of the training set, due to its high sensitivity to instances of training. In that sense, small disturbances in the training set may cause big alterations in the learned model. In training sets where the amount of noise is low, the algorithm generates structures that can obtain perfect adjustments to training instances. However, when much noise is present, that ability is damaged. From a certain depth onwards, decisions taken are based on small sets of instances. It provokes the fragmentation of concepts, leading to the re-application of several subtrees, and we may remind that smaller trees are easier to understand, and they have a better predictive performance.

ACKNOWLEDGMENTS

Our thanks to CNPq and Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp) for the scholarships and research funds granted for the realization of this research. Our thanks to INPE (Instituto Nacional de Pesquisas Espaciais, located in São José dos Campos, SP) for the data and for the infrastructure provided.

REFERENCES

Aguiar, D. A.; Silva, W. F.; Rudorff, B. F. T.; Sugawara, L. M.; Carvalho, M. A. Expansion of sugarcane in the State of São Paulo: crops 2003/2004 through 2008/2009. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14. (SBSR), 2009, Natal. Anais... São José dos Campos: INPE, 2009. p. 9-16. (Original in portuguese)

Blaschke, T., 2010. Object-based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 65, 2-16.

Baatz, M.; Schäpe, A. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. In: XII Angewandte Geographische Informationsverarbeitung, Wichmann-Verlag, Heidelberg, 2000.

- Bittencourt, G., 2006. *Artificial intelligence: tools and theories*. Florianópolis/SC: Editora da Universidade Federal de Santa Catarina (UFSC). 371 p. (Original in portuguese)
- Cohen, Y.; Shoshany, M., 2005. Analysis of convergent evidence in an evidential reasoning knowledge-based classification. *Remote Sensing of Environment*, v. 96, 518-528.
- Congalton, R. G.; Green, K., 1999. *Assessing the accuracy of remotely sensed data: Principles and practices*. Boca Raton: Lewis Publishers, pp. 11 – 70.
- Definiens, 2006. *Definiens professional 5: reference book*. Munich: Definiens (The Imaging Intelligence Company), 122 p
- DeFries, R. S., & Chan, J., 2000. Multiple criteria for evaluating machine learning algorithms for land-cover classification from satellite data. *Remote Sensing of Environment*, 74, 503–515.
- Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R., 1996. *Advances in knowledge discovery and data mining*. Cambridge: MIT Press, 560 p.
- Foody, G. M., 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, v. 80 (1), p. 185-201.
- Good, P. I., 2001. *Resampling methods: a practical guide to data analysis*. Berlin: Birkhauser Verlag AG. 256 p.
- Hsu, W.; Lee, M; Zhang, J., 2002. Image mining: trends and developments. *Journal of Intelligent Information*, v. 19 (1), p. 7-23.
- IBGE. Instituto Brasileiro de Geografia e Estatística. Systematic survey of agricultural production - December 2008. Rio de Janeiro-RJ: IBGE, 2008. 80 p. (Original in portuguese)
- Ponzoni, F. J.; Shimabukuro, Y. E. Remote sensing in the study of vegetation. 1st ed. São José dos Campos, SP: Parêntese Editora, 2007. 144 p. (Original in portuguese)
- Quinlan, R., 1993. *C4.5: programs for machine learning*. San Francisco: Morgan Kaufmann, 316p.
- Rouse, J.W.; Haas, R. H.; Schell, J. A.; Deering, D. W., 1973. Monitoring vegetation systems in the great plains with ERTS. In: Earth Resources Technology Satellite-1 Symposium, 3., Washington. *Proceedings...* Washington: NASA, 1974, v.1, p.309-317.
- Rudorff, B. F. T.; Sugawara, L. M.; Aguiar, D. A.; Silva, W. F.; Goltz, E.; Aulicino, T. L. I. N.; Carvalho, M. A.; Arenas-Toledo, J. M.; Brandão, D. *Determination of the area planted with sugarcane in the South-Center region through images from sensing satellite - crops 2005/06 through 2008/09*. São José dos Campos: INPE, 2009. 151 p. (Original in portuguese)
- Schroeder, T. A.; Cohen, W. B.; Song, C.; Canty, M. J.; Yang, Z., 2006. Radiometric correction of multi-temporal Landsat data for characterization of early successional forest patterns in western Oregon. *Remote Sensing of Environment*, v. 103 (1), p. 16-26.
- Weka (2010), “*Weka 3 - Data Mining with Open Source Machine Learning Software in Java*”, <http://www.cs.waikato.ac.nz/ml/weka/> (Accessed 12 march 2010).
- Witten, I.; Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2. Ed. San Francisco, CA: Morgan Kaufmann Publishers, 524 p.