A methodology for determining optimal feature subset for classification in object-based image analysis

Damien Arvor [1]
Alexandre Wiefels[1]
Nathalie Saint-Geours [2]
Claudio Almeida [3]
Kenji Ose [2]
Laurent Durieux [1]

[1] IRD - UMR ESPACE DEV 228
500, rue Jean-François Breton, 34093 Montpellier Cedex, France
{damien.arvor, laurent.durieux}@ird.fr

[2] IRSTEA - UMR TETIS
500, rue Jean-François Breton, 34093 Montpellier Cedex, France
{nathalie.saint-geours, kenji.ose}@teledetection.fr

[3] Instituto Nacional de Pesquisas Espaciais - INPE
INPE/CRA, Parque de Ciência e Tecnologia do Guamá, Av. Perimetral 2651, Belém, CEP 66077-830- PA, Brazil
claudio@dsr.inpe.br

Abstract. In GEOBIA, remote sensing experts benefit from a large spectrum of characteristics to interpret images (spectral information, texture, geometry, spatial relations, etc). However, the quality of a classification is not always increased by inserting a higher number of features. The experts are then used to define classification rules based on a laborious "trial-and-error" process. In this paper, we propose a methodology to automatically determine an optimal subset of features for discriminating features. This process assumes that a reference land cover map is available. The method consists in ranking the features according to their potential for discriminating two classes. This task was performed thanks to the Support Vector Machine-Ranking Feature Extraction (SVM-RFE) algorithm. Then, it consists in training and validating a classification algorithm (SVM), with an increasing number of features: first only the best-ranked feature is included in the classifier, then the two best-ranked features, etc., until all the N features are included. The objective is to analyze how the quality of the classification evolves according to the numbers of features used. The optimal subset of features is finally determined through the analysis of the Akaike information criterion. The methodology was tested on two classes of pastures in a study area located in the Amazon. Two features were considered as sufficient to discriminate both classes.

Keywords: Remote Sensing, Image Processing, GEOBIA, SVM.

1. Introduction

Since the early 2000s, geographic object-based image analysis (GEOBIA) has appeared as a new paradigm shift in remote sensing image processing. GEOBIA relies on automated methods to partition remote sensing imagery into meaningful image-objects and the assessment of their characteristics to generate new geographic information. In GEOBIA, remote sensing experts benefit from a large spectrum of characteristics to interpret images (spectral information, texture, geometry, spatial relations, etc). However, as mentioned by Van Niel et al. (2005) and Bruzzone et al. (2000), the quality of a classification is not always increased by inserting a higher number of features. This phenomenon, called peaking, is known in remote sensing as the "Hughes effect". Consequently, a good classification must be based on a subset of relevant features identified due to their ability to discriminate the classes of interest. In GEOBIA, this process is usually performed visually by an expert through a

“trial-and-error” process. On the one hand, this makes GEOBIA really efficient for interpreting high resolution images since it allows the user to integrate his expert knowledge in the classification process, but, on the other hand, the final accuracy of the classification depends too much on the remote sensing expert knowledge, i.e. two experts will define different rules for a same class and thus produce different (but potentially correct) maps. In order to achieve more robust results, it is then necessary to assist the remote sensing expert in its interpretation process by advising him which features to select. Image Information Mining (IIM) techniques can thus appear useful for rapidly acquiring knowledge on better rules to map land cover classes. But as mentioned by Durbha et al. (2010), whereas *"earlier efforts were focused mainly on the reduction of features using clustering approaches [...] little was reported on the selection of the best feature subset”*. For these authors, such a task should be led by combining predictive-models with feature selection and feature-generation approaches. In this paper, we tested such a combined approach to automatically define features of interest and the corresponding classifying method for discriminating two classes of pasture in a study area located in the state of Para, Brazilian Amazon.

2. Study area

The study area is located in the surrounding of the city of Altamira, in the Brazilian Amazon, state of Para. This city is located along the transamazonian road and was thus affected by large land use changes during the last 40 years. Amongst main land use changes that occurred, the conversion of primary forest into pastures is a major issue in the region as in the entire Amazon. Such conversions have major impacts on biodiversity and regional climate. However such impacts are differentiated depending on the kinds of pastures considered, clean pasture, dirty pastures or regeneration of pastures. It thus becomes essential to monitor this diversity of pastures.

3. Data and methods

3.1. Data : TerraClass

In order to define the best subset of features, we need a preliminary land cover map for training and validating our approach. Although we should ideally use maps performed through photo-interpretation and/or field campaigns, we here used an already existing land cover map of the study area produced under the frame of the TerraClass project.

The TerraClass project aims at qualifying the land cover in cleared areas in the Legal Amazon thanks to remote sensing images. This project resulted in the production of a land cover map of the Amazon for the year 2008, and updates for 2010 should be soon released. This project is led by the CRA/INPE (Regional Center in the Amazon-Belem), the Embrapa Oriental Amazon (Belem-PA) and the Embrapa Informatics (Campinas-SP). The land cover maps are freely available at: http://www.inpe.br/cra/projetos_pesquisas/terraclass.php.

The land cover maps introduce 11 classes, namely: primary forest, secondary forest, annual crop, clean pasture, dirty pasture, regeneration with pasture, urban area, extraction site, non observed area, hydrography, land cover mosaic.

3.2. Methodology

The method proposed for extracting classification rules is based on three main steps. The first step consists in pre-processing the data in order to: i) retrieve polygonal objects from the TerraClass map, ii) extract a number N of features (regarding spectral, textural, and geometrical properties of the objects) for each object based on the corresponding Landsat

image, and iii) build a training sample and a validation sample of objects with their associated features, for the two classes of interest. Once data are ready, the second step consists in ranking the N selected features based on their ability to discriminate the two classes of interest. Finally, the third step consists in training and validating a classification algorithm, with an increasing number of features: first only the best-ranked feature is included in the classifier, then the two best-ranked features, etc., until all the N features are included. The objective is to analyze how the quality of the classification evolves according to the numbers of features used.

1) Data preparation: Feature extraction, Object selection, Outlier detection

We first acquired the TerraClass map for the study area from INPE in shapefile format. The corresponding Landsat image (with only 3 spectral bands) was also acquired from the same source. We integrated both data (vector and raster) in the TerraView software proposed by INPE (www.dpi.inpe.br/**terraview**/). We then used the GeoDMA plug-in to extract 41 features of interest for each polygon. GeoDMA is an image data-mining tool proposed by INPE (Korting et al., 2008). The extracted features proposed in GeoDMA refer to polygon properties (shape, area, etc) and to raster-polygon properties (mean values for each band, textural features, etc). It is noteworthy that features on topological relations are not included. Finally, we obtained two tables (for the two classes of interest) where each row refers to an object (e.g. polygon) and each column refers to the associated features. At this step, the tables were imported to the R software for further processing. As the classes represented very large numbers of polygons, we randomly selected 1000 objects per class.

Our objective is to find the optimal feature subset to discriminating two pasture classes on Landsat images based on a vectorial land cover map. However, this vectorial mal (i.e. the TerraClass map) was produced based on multi-source data including Landsat, MODIS and PRODES deforestation maps. Thus, in cases where classes are derived from non Landsat sources, we might find a large heterogeneity of Landsat-based features for one class (e.g. annual crops are detected with MODIS time series and might appear as non vegetated areas on a Landsat image acquired during the intercropping period, also clouds detected on Landsat can be classified as forests if PRODES data mentioned that the corresponding area had never been cleared). The issue is that, in order to perform an efficient learning procedure, we need to work with objects as homogeneous as possible. For this purpose, we filtered the tables through eliminating outliers. The filtering processing first consisted in eliminating objects that contained inconsistent values. For example, textural features measured for very small features were not computed so that we deleted them. Secondly, we normalized the tables and computed the Mahalanobis distance for each object in order to detect outliers (Arvor et al., 2008). We then discarded the 20% less confident objects of the sample.

Finally we divided both datasets in two subsets of same size (n = 400). One table was designed for training the feature ranking and classification algorithms whereas the other table was designed to validate the classification algorithms. The size of training sample = 400 is consistent with the requirements fixed by Van Niel et al. (2005), i.e. the training sample should be 10 to 30 times the number of features (number of features = 41 in our case).

2) Feature ranking

Once the datasets were prepared, we applied a procedure to rank features according to their potential for discriminating the two classes of interest. To achieve this objective, we applied the SVM-RFE (Support Vector Machine Recursive Feature Extraction) method. This algorithm proposed by Guyon et al. (2002) returns a ranking of the features of a classification problem by training a Support Vector Machine (SVM) with a linear kernel and removing the

feature with smallest ranking criterion. The *svmrfeFeatureRanking* function was used in R from package (e1071).

As we randomly selected a predefined number of objects for performing the feature ranking, we are expected to get a certain variability in feature ranking (i.e. two runs give different rankings). To achieve more robustness, we ran the process 100 times. We then analyzed the mean rank for each feature in order to define the final feature ranking.

3) Determining the best combination of features for classification

Once the features have been ranked, the issue is to define if there exists an optimal combination of features to be used for discriminating two classes. As mentioned by Thomas *et al.* (1987): « *any final assessment of the accuracy of a classification rests upon the classification process itself and not directly upon the separability index selection of appropriate channels* ». Thus, searching for the most relevant combination of features must be based on classifications. For this purpose, we trained a classification algorithm (on the training samples) in order to classify the validation samples and thus validate the approach.

We tested the Support Vector Machine (SVM) from the *svm* function from the same R package (*e1071*). For training the classifiers, we firstly only used the best feature that was determined at the feature ranking step. Then, we performed new tests by including each feature, one-by-one according to its potential for discriminating two classes of interest, i.e. its rank. For each combination of features, the classifier was trained and then applied to classify the validation sample. We were then able to compare the successive classifications in order to determine the best one and thus define the best combination of features. In order to assess the quality of these successive classification, we computed traditional statistical indices, i.e. overall accuracy and Kappa index.

Here again, the process was run 100 times in order to achieve more robustness (because each classification was affected by variability linked to the training and validation samples randomly selected).

Once we had robust results on the classification accuracy of each successive classification (including one feature, two features,..., n features). We were able to find the peak at which best results are obtained. However, what we observe is not really a peak but a plateau. For which combination of features does the plateau starts then becomes the issue. To address this issue, we used the Akaike information criterion that measures the relative goodness of fit of a statistical model. It is used to assess the quality of a classification by considering both the classification accuracy (i.e. the Kappa index) and the number of features required. In other words, the best classification is the one that achieve the best accuracy with the lowest number of feature. Since we have always used a fix sample of 800 objects (400 objects for each class), we used the AICc formula defined for finite sample sizes (Equation 1) :

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \quad \text{(eq.1)}$$

$$AIC = 2k - 2\ln(L) \quad \text{(eq.2)}$$

where k it he number of features in the model, L is the mean value of the classification accuracy statistics (i.e. Kappa index), and n is the sample size. The lowest AICc value indicates the optimal compromise between the classification accuracy and the subset of features.

4. Results

1) Data preparation

The method was applied in order to discriminate two types of pastures, i.e. "clean" pasture and "regeneration with pasture" representing 4096 polygons and 2527 polygons in the TerraClass map, respectively. For each of these polygons, 41 features were extracted in order to build two databases (one per class) that were prepared as mentioned in section 3.2.1. Especially, the databases were filtered thanks to the Mahalanobis distance (Figure 2). Finally, we obtained two databases that were cut in two tables of equal size for training and validating the method.

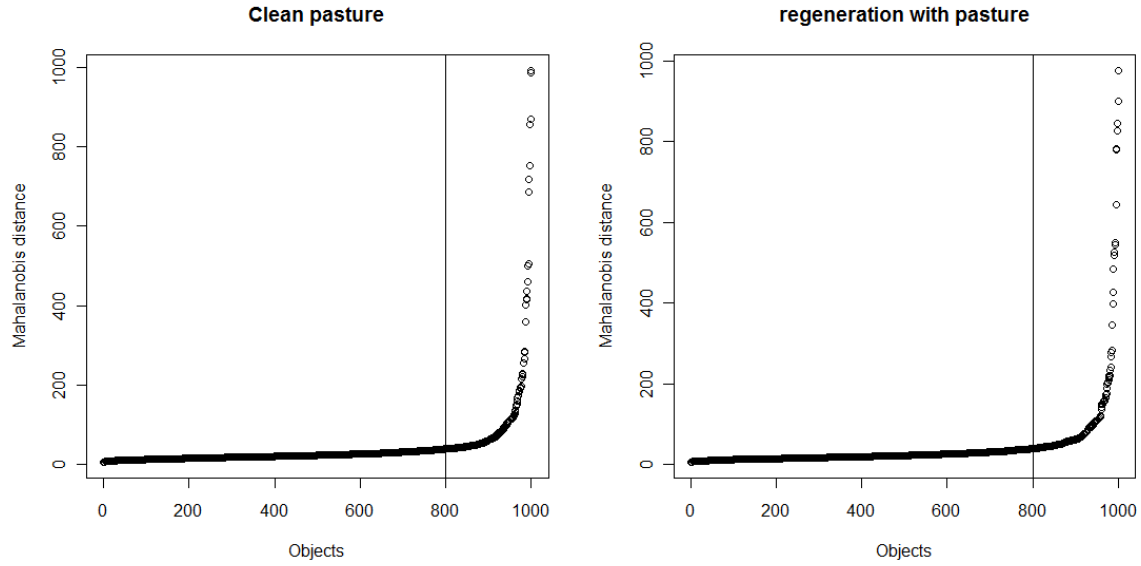

Figure 1. Sorted Mahalanobis distance of the 1000 objects randomly selected for each class. 20% of objects (points located at the right of the vertical line) were considered as outliers and thus discarded.

2) Feature ranking

The SVM-RFE algorithm was applied 100 times to rank the features. We then defined the mean rank of each feature. The results are partly introduced in table 1. They highlight the importance of entropy indexes for discriminating both classes.

Table 1. Ten best features ranked by the SVM-RFE method. The mean rank is given after the method has been run 100 times. The numbers 0,1 and 2 in the feature's name refer to different Landsat bands (Blue = 0, Green=1, Red=2)

Feature Names	Mean Rank
rp_entropy2_0	3.66
rp_sum_2	4.06
rp_entropy2_1	4.09
p_area	5.19
rp_entropy2_2	6.39
p_box_area	8.46
rp_homogenity_2	9.15
p_peimeter	9.74
rp_sum_0	9.8
rp_sum_1	11.68

3*) Determining the best combination of features for classification*

Once the features have been ranked, we trained the classifier for different subsets of features, including features one-by-one according to their ranking, from the highest rank to the lowest rank. The process was run 100 times and the classification statistics were computed (figure 3). The SVM classifier led to good results (mean Overall Accuracy > 0.90, Kappa > 0.80). Visually, a plateau appears on the Mean Kappa curve after the inclusion of the 13th best ranked feature. However, the analysis of the Akaike information criterion (figure 4) reaches it minimal value after the inclusion of the 4th feature. This indicates that the optimal subset of feature is only composed of 4 features, i.e. rp_entropy2_0, rp_sum_2, rp_entropy2_1 and p_area.

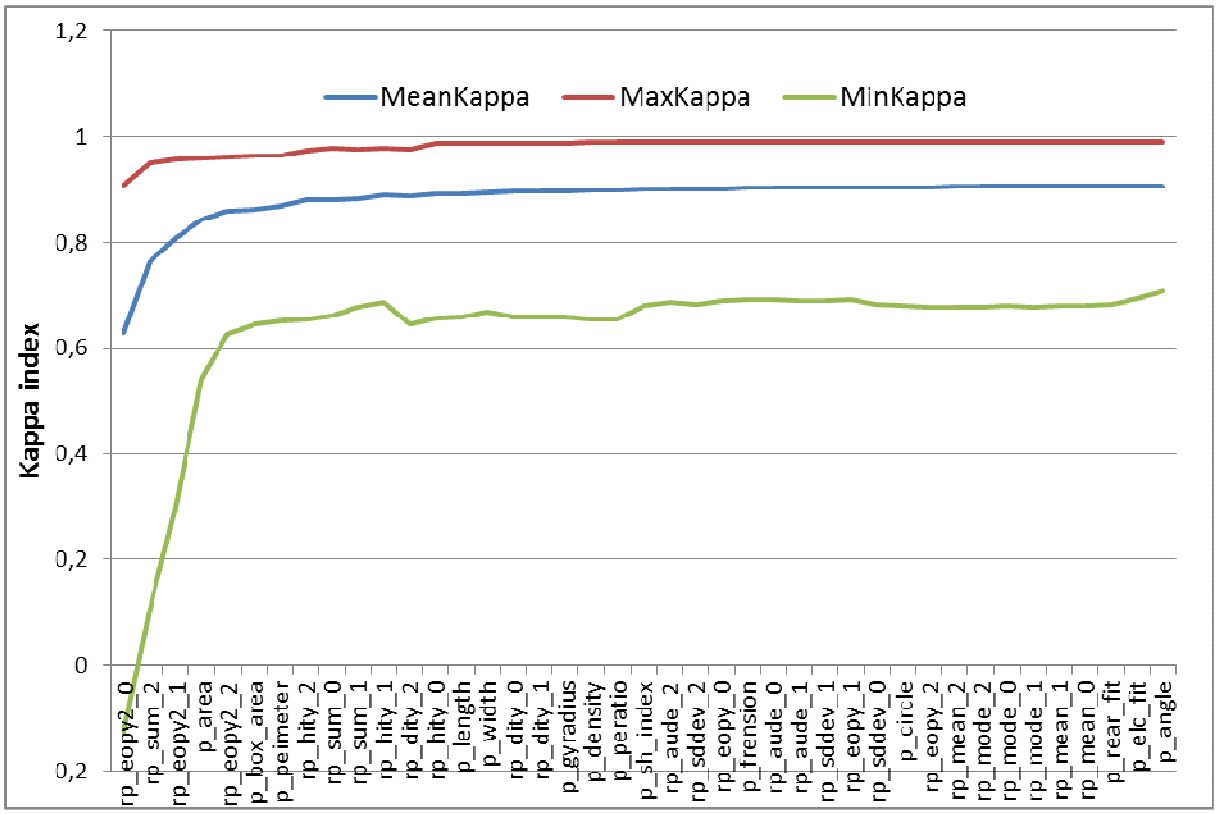

Figure 2. Mean, Max and Min Kappa indices computed based on SVM classifications obtained for different combinations of features (from 1 to 41 features).

5. Discussion and conclusion

This paper introduced preliminary results illustrating the global approach used for identifying an optimal subset of features to discriminate classes in GEOBIA. However, the method is not complete at this time and further improvements are required to achieve validating the approach.

First, considering the current paper, various points need to be discussed:

- More tests should be carried out with other classes to check for the robustness of the results obtained.
- The input data used in this study only include three spectral bands. More tests are required with all the Landsat spectral bands and with the inclusion of spectral indices (NDVI, WBI, etc). Moreover, spatial relations were not considered. Finally, it would be necessary to eliminate redundant features (many features are correlated between each other).

- The mapping has not been performed. This task is an issue since it would depend on the quality of the segmentation process that will not produce same objects as we had in the TerraClass maps.

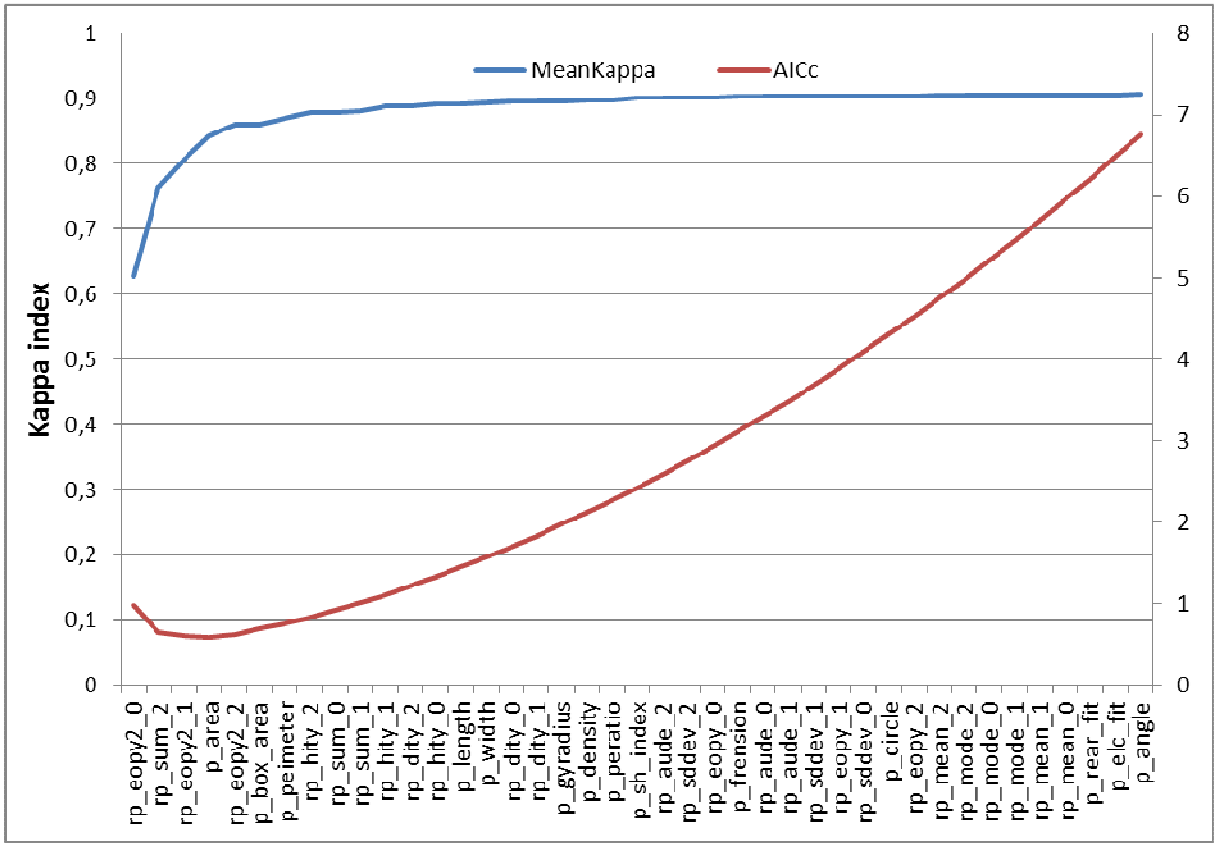

Figure 3. Akaike information criterion compared with the classification results obtained with the SVM classifier.

Second, various perspectives appear to enhance the approach:
- Other feature ranking and classification algorithms could be tested and compared with previous results.
- The Akaike criterion used to select the best subset of feature could evolve. For example, here we tested it with mean value of Kappa index although it could be tested with the maximized value.
- Other information criterion should be tested.
- The validation process should be performed on the datasets coming from another image to check for the ability of the approach to better take the spatial and temporal variability of data acquisition context into consideration.
- We need to assess if the knowledge extracted from such approach (i.e. the classification rules) can lead to the determination of "visual pattern" of land cover classes and thus improve the semantic description of land cover classes (e.g. a clean pasture is described by high or low entropy values and high or low "sum" values) so that they can be implemented in ontologies.
- Once the features are selected, we think in applying the Separability and Threshold method (Nussbaum et al., 2006; Marpu et al., 2008) to detect optimal thresholds to discriminate classes.
- The methodology will be tested in the BIO_SOS project. BIO_SOS is a EU-FP7 funded project *Biodiversity Multi-Source Monitoring System: From Space To Species* (BIO_SOS)

focusing on the development of tools and models for consistent multi-annual monitoring of protected areas exposed to human pressures and their surroundings in the Mediterranean and elsewhere (www.biosos.eu).

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Program FP7/2007-2013 under grant agreement n° 263435 for the BIOSOS project.

References

Arvor, D.; Jonathan, M.; Meirelles, M. S. P.; Dubreuil, V. Detecting outliers and asserting consistency in agriculture ground truth information by using temporal vi data from MODIS. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS). Beijing : China, 1031-1036. 2008.

Bruzzone, L.; Serpico, S. A technique for feature selection in multiclass problems. **International Journal of Remote Sensing**, v. 21, n. 3, p. 549-563, 2000.

Durbha, S.S.; King, R.L.; Younan, N.H. Wrapper-based feature subset selection for rapid image information mining. **IEEE Geoscience and Remote Sensing Letters**, vol. 7, n. 1, p. 43-47, 2010.

Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. **Machine Learning**, vol. 46, p. 389-422, 2002.

Korting, R.S.; Fonseca, L.M.G.; Escada, M. I. S.; Silva, F. C.; Silva, M. P. S. GeoDMA - A novel system for spatial data mining. in : IEEE Conference on Data Mining Workshops, 2008.

Marpu, P.R.; Niemeyer, I.; Nussbaum, S.; Gloaguen, R. A procedure for automatic object-based classification. In: Blaschke, T.; Lang, S.,; Hay, G. J. (Eds.). Object-Based Image Analysis. Spatial concepts for knowledge-driven remote sensing applications. Springer-Verlag Berlin Herdelberg, 2008. Cap. 2.4, p. 169-184.

Nussbaum, S.; Niemeyer, I.; Canty, M.J. SEATH - a new tool for autoomated feature extrcation in the context of object-based image analysis. In : 1st International Conference on Object-based Image Analysis (OBIA). Salzburg: Austria, 2006.

Thomas, I.; Ching, N.; Benning, V.; D'Aguanno, J. A review of multi-channel indices of class separability. **International Journal of Remote Sensing**, vol. 8, n.3, p. 331-350, 1987.

Van Niel, T.; McVicar, T.; Datt, B. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. **Remote sensing of environment, v.** 98**,** n. 4, p. 468-480, 2005.