DETECTING ATLANTIC FOREST PATCHES APPLYING GEOBIA AND DATA MINING TECHNIQUES

C. D. Girolamo Neto [a], A. C. M. Pessôa [b], T. S. Körting [a], L. M. G. Fonseca [a]

[a] Image Processing Division – National Institute for Space Research – INPE – cesare@dsr.inpe.br; (thales; leila)@dpi.inpe.br
[b] Remote Sensing Division – National Institute for Space Research – INPE – ana.pessoa@inpe.br
Av. dos Astronautas, 1758, São José dos Campos, SP, Brazil.

KEY WORDS: Land cover, Classification, Landsat-8, Random Forest, Artificial Neural Networks, Feature selection

ABSTRACT:

Brazilian Atlantic Forest is one of the most devastated tropical forests in the world. Considering that approximately only 12% of its original extent still exists, studies in this area are highly relevant. In this context, this study maps the land cover of Atlantic Forest within the Protected Area of 'Macaé de Cima', in Rio de Janeiro State, Brazil, combining GEOBIA and data mining techniques on an OLI/Landsat-8 image. The methodology proposed in this work includes the following steps: (a) image pan-sharpening; (b) image segmentation; (c) feature selection; (d) classification and (e) model evaluation. A total of 15 features, including spectral information, vegetation indices and principal components were used to distinguish five patterns, including *Water*, *Natural forest*, *Urban area*, *Bare soil/pasture* and *Rocky mountains*. Features were selected considering well-known algorithms, such as Wrapper, the Correlation Feature Selection and GainRatio. Following, Artificial Neural Networks, Decision Trees and Random Forests classification algorithms were applied to the dataset. The best results were achieved by Artificial Neural Networks, when features were selected through the Wrapper algorithm. The global classification accuracy obtained was of 98.3%. All the algorithms presented great recall and precision values for the Natural forest, however the patterns of Urban area and Bare soil/pastures presented higher confusion.

1. INTRODUCTION

Land use and cover (LUC) analysis can be used to determinate how a specific area is being used, highlighting the anthropogenic interactions with the environment. In order to access patterns of LUC changes, it is vital to use data from remote sensing imagery (Brannstrom *et al.*, 2008). This technology allows the generation of LUC maps, showing areas being occupied by pastures, crops, natural vegetation, river courses and other features. They can also indicate areas of risk or those heavily degraded.

One of the most devastated Brazilian biomes is the Atlantic Forest. The second largest Brazilian forest has only 12% of its initial extent preserved (Ribeiro *et al.*, 2009). A large part of the occupation of this biome has occurred due to the expansion of urban centers and agricultural areas. Among Rio de Janeiro and São Paulo states, most of Atlantic Forest patches are usually in Protected Areas (PA's). LUC analysis on these territories is even more important when considering the possibility of degrading preserved natural vegetation areas (Figueroa & Sánchez-Cordero, 2008).

A procedure used to perform the LUC classification is the Geographic Object-Based Image Analysis (GEOBIA), which aims to classify an image based on similar characteristics of its objects. In addition to the spectral properties, GEOBIA can evaluate features associated with the shape, texture, contextual and semantic relationships of objects, increasing the chances of a more reliable classification (Camargo *et al.*, 2009).

Another methodology that has been constantly used on image classification is the Data Mining (DM). DM helps GEOBIA in the process of identifying patterns on objects. In this context, some DM classification algorithms have been used on LUC analysis, for instance: Decision Trees were used for vegetation mapping (Colstoun *et al.*, 2003), temporal analysis of agricultural crops (Körting, 2012) and urban LUC (Pinho *et al.*, 2012). Random Forests were applied to classify LUC on various locations (Smith, 2010; Müller *et al.*, 2015). Artificial Neural Networks were used assessing Natural vegetation LUC (Moreira *et al.*, 2013) and Agricultural areas (Andrade *et al.*, 2013). The choice to use each algorithm requires the analysis of the problem. The results of Decision Trees are easy to visualize (Witten *et al.*, 2011), Random Forests can avoid overfitting, and are also not very sensitive to noisy data (Breiman, 2001). The main advantage of Artificial Neural Networks is to solve complex problems (Haykin, 2009) and may outperform other classifiers on LUC classification (Song *et al.*, 2012).

Considering the importance of LUC information on Atlantic Forest areas and the potential of GEOBIA and DM techniques on image classification, this study aims to map the land use and cover of Atlantic Forest within the PA of 'Macaé de Cima', in Rio de Janeiro State, Brazil, using GEOBIA and DM techniques on an OLI/Landsat-8 image.

2. METHODOLOGY

2.1 Study site and Data

The study site is the PA of Macaé de Cima – Rio de Janeiro State (Figure 1). It is located between the coordinates of 22°17´S-22°27´S and 42°35´W-42°12´W, on the municipalities of Macaé and Nova Friburgo. It has an area of 350.000 square meters on steep region with rocky mountains and small valleys, with 72% of Atlantic Forest cover (INEA, 2007).

Data was obtained from the Operational Line Imager (OLI) sensor from Landsat-8 satellite, path/row 216/76 acquired on 10/14/2014.

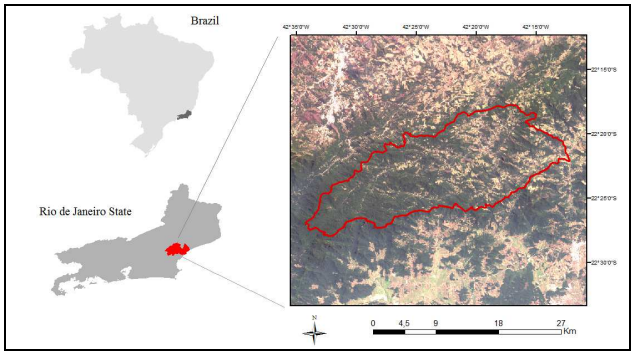

Figure 1. Study site on a true color composition (R4, G3, B2) from the OLI sensor.

2.2 Image Pan-sharpening

The original image was pan-sharpened using the Gram-Schmidt method (Laben *et al.*, 2000). Bands 2 to 7 (0.450 to 2.300 µm) from OLI sensor were used, since this type of pan-sharpening method has generally good results when it is carried out with the same sensor bands (Klonus & Ehlers, 2009). In addition, this method presented good results when compared to other pan-sharpening methods when applying LUC classification (Bendini *et al.*, 2015).

2.3 Image Segmentation

In order to generate the objects for the classification process, the pan-sharpened images were segmented on eCognition software (version 8.64), using the Multi-resolution Segmentation algorithm (eCognition, 2011). Bands 2 to 7 were used in this process and the internal parameters in the algorithm (Scale, Shape and Compactness) were calibrated in order to optimize the segmentation results.

2.4 Dataset generation and feature selection

The features obtained from the processed image are on Table 1. All of them presented mean values from each object (excluding max. diff.).

Feature	Meaning
Band_1	Surface reflectance from band 1
Band_2	Surface reflectance from band 2
Band_3	Surface reflectance from band 3
Band_4	Surface reflectance from band 4
Band_5	Surface reflectance from band 5
Band_6	Surface reflectance from band 6
Band_7	Surface reflectance from band 7
PCA1	Principal Component 1
PCA2	Principal Component 2
PCA3	Principal Component 3
NDVI	Normalized Difference Vegetation Index
NBR	Normalized Burn Ratio
B6/B5	Relation Band6/Band5
Brightness	Average of means from bands 1 to 7
Max. Diff.	Maximum difference between bands
Class	Name of the class

Table 1. Features of the dataset.

The objects could be classified to distinguish 5 patterns, including *Water*, *Natural forest*, *Urban area*, *Bare soil/pasture* and *Rocky mountains*. A visual interpretation of the image was done in order to select approximately 120 samples of each class. A high resolution image, acquired on 30/05/2014 by Google Earth software, was used as a support for the interpretation. The Water class had no more than 80 samples, so it was excluded from the data set, since there was no more Water areas left to be identified in the image.

On the prepared dataset a feature selection process was carried out. This process tends to reduce the computational cost and raise the classification accuracy, by eliminating irrelevant and redundant features of the dataset (Guyon & Elisseeff, 2003). Tree feature selection methods were used: Wrapper (John & Kohavi, 1997), Correlation Feature Selection (Hall, 1999) and the GainRatio (Witten *et al.*, 2011).

2.5 Classification and evaluation

The classification was performed on the WEKA software, version 3.7.9. (Hall *et al.*, 2009). The algorithms used for classification were Decision Trees (5 and 10 instances per leaf), Artificial Neural Networks (back propagation with one hidden layer with 10 neurons) and Random Forest (100 trees). As a first task, these classifiers were evaluated on a 10-fold cross validation. Later on, an external data set was used for the evaluation. It contained 30 visually classified samples of each class that were not on the training set. The confusion matrix was used to determine measures such as accuracy, error, precision and recall.

3. RESULTS AND DISCUSSION

The GainRatio algorithm ranked the features from the dataset. From the 16 features we decided to reduce the number to 10, according to the values of information gain ratio. The best merit found on the new dataset was from the NDVI feature, so we decided to test how this feature would influence the classification. We tested the dataset with and without the NDVI feature for the algorithms presented before. The accuracies of the classifiers are presented on Table 2.

	With NDVI		Without NDVI	
Algorithm	Cross validation	Test set	Cross validation	Test set
Decision Tree (5)	96,6	96,7	96,6	96,7
Decision Tree (10)	96,8	95,8	96,8	95,8
Random Forest	96,6	96,7	97,0	97,5
Artificial Neural Networks	97,9	97,5	98,1	97,5

Table 2. Overall accuracy (%) of GainRatio feature selection with and without NDVI.

The influence of using the NDVI feature was very low. There was no alteration on Decision Trees results and it was noticed a slight improvement on the 10-fold cross validation without using the NDVI on Artificial Neural Networks and Random Forest. Usually, NDVI tends to apply better results on the classification when it is used with others sensors bands (DeFries & Townshend, 1994). However, when we analyse the remaining features selected we noticed that band 4 was always selected on all modelling situations. When band 4 is removed instead of the NDVI, the same results are obtained. When both were removed, all results were the same or lower than the previous

classification, so we assumed that band 4 and NDVI might have correlated information for the classification (since NDVI uses band 4 on its formula). The results of the other algorithms can be found on Table 3.

	CFS		Wrapper	
Algorithm	Cross validation	Test set	Cross validation	Test set
Decision Tree (5)	96,6	96,7	96,2	95,0
Decision Tree (10)	96,8	95,8	96,4	95,0
Random Forest	97,2	96,7	97,2	96,7
Artificial Neural Networks	97,2	97,5	98,1	98,3

Table 3. Overall accuracy (%) of Correlation Feature Selection and Wrapper for the three algorithms.

The best results were obtained by Artificial Neural Networks with the Wrapper feature selection. The algorithm which presented the second best results was the Random Forest with the GainRatio feature selection. Both scored 98.1% accuracy on 10-fold cross validation. We show the confusion matrix of both of these classifiers in order to enhance this discussion (Table 4 and 5). We selected the matrix from the cross-validation because it contained more instances and the errors became more visible, but similar results were spotted on the test set Matrix.

		Classified as			
		Bare soil /pasture	Natural forest	Urban area	Rocky mountains
Actual class	Bare soil /pasture	119	0	2	0
	Natural forest	0	121	0	0
	Urban area	2	1	107	3
	Rocky mountains	0	0	1	112

Table 4. Confusion matrix for Artificial Neural Networks with Wrapper feature selection.

		Classified as			
		Bare soil /pasture	Natural forest	Urban area	Rocky mountains
Actual class	Bare soil /pasture	117	0	3	1
	Natural forest	0	120	0	1
	Urban area	6	0	107	1
	Rocky mountains	1	0	1	111

Table 5. Confusion matrix for Random Forest with GainRatio feature selection (without NDVI).

Our main objective within this paper is to evaluate the land cover of Atlantic forest. For this purpose, the recall values for this class were 100% and 99,2% for the respective classifiers. This result shows that almost all areas of Natural forest were classified correctly. There was almost no confusion of other classes being classified as Natural forest, since the precision values of the models were 99,2% and 100%. Both of these models worked very well on classifying the forest patches and can be used to map these tiles. The only confusion was a misclassified Forest area with a Rocky mountain area. This error may have occurred because of the presence of vegetation in some mountains, which behave similarly as a vegetation patch (Figure 2).

Figure 2. Rocky mountains with some vegetation at the peak of Pedra Riscada – Macaé/RJ.

When considering the other classes, some deficiencies can be spotted. There was a clear confusion between Bare soil/pasture and Urban area. On both classifiers, samples of Bare Soil/pasture were misclassified as Urban area and vice-versa. The Random Forest model was more problematic with this confusion, since the 9 samples were misclassified. The recall values for the Urban area class was considered low on both models (94,7% and 93,8%) when compared to the other classes. The confusion between these classes is a problem noticed before (Kux & Araujo, 2008; Novack & Kux, 2010). Considering this is a protected area and it has small agricultural actives, these classes might be encountered together and some objects may contain a mix of both classes.

4. CONCLUSIONS

Brazilian Atlantic Forest suffers from devastation and must be protected. Thus, the mapping of its remaining patches is crucial. An automated procedure was adopted involving GEOBIA and DM techniques. Algorithms as Artificial Neural Networks and Random Forest produced encouraging results on identifying these areas (up to 98.3% accuracy). The only misclassification occurred with some Rocky mountains that have vegetation on it. When considering other investigated patterns, the confusion between Bare soil/pasture and Urban area was more notable. Overall, using GEOBIA with classification algorithms as Artificial Neural Networks or Random Forest is a viable tool for mapping remaining areas of Atlantic Forest on protected areas. However, further adaptations might be required when on sites that have a different class distribution.

5. REFERENCES

Andrade, L. N.; Vieira, T. G. C.; Lacerda, W. S.; Volpato, M. M. L. & Davis Jr, C. A. 2013. Application of artificial neural networks in the classification of coffee areas in Machado, Minas Gerais State. *Coffee Science*, 8(1), pp. 78-90.

Bendini, H. N.; Girolamo Neto, C. D.; Körting, T. S.; Marujo, R. F. B.; Tranbaquini, K.; Eberhardt, I. D. R.; Sanches, I. D.; Fonseca, L. M. G. 2015. Effects of Image Fusion Methods on Sugarcane Classification with Landsat-8 Imagery. In*: XVII Brazilian simposion of remote sensing*, pp. 2498-2505.

Brannstrom, C.; Jepson, W.; Filippi, A. M.; Redo, D.; Xu, Z.; & Ganesh, S. 2008. Land change in the Brazilian Savanna (Cerrado), 1986–2002: comparative analysis and implications for land-use policy. *Land Use Policy*, 25(4), pp. 579-595.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1), pp. 5-32.

Camargo, F. F.; Florenzano, T. G.; Almeida, C. M.; Oliveira, C. G. & Feitosa, R. Q. 2009. Object-Based Analysis and ASTER/Terra Data for Classifying Relief Units. *Bol. Ciênc. Geod.*, 15(1), pp. 81-102.

Colstoun, E. C. B.; Story, M. H.; Thompson, C.; Commisso, K.; Smith, T. G. & Irons, J. R. 2003. National Park vegetation mapping using multitemporal Landsat 7 data and a decision tree classifier. *Remote Sensing of Environment*, 85(3), pp. 316-327.

DeFries, R. S. & Townshend, J. R. G. 1994. NDVI-derived land cover classifications at a global scale. *International Journal of Remote Sensing*, 15(17), pp. 3567-3586.

eCognition, 2011. *eCognition Developer 8.64.1 users guide*. Trimble Germany.

Figueroa, F. & Sánchez-Cordero, V. 2008. Effectiveness of natural protected areas to prevent land use and land cover change in Mexico. *Biodiversity and Conservation*, 17(13), pp. 3223-3240.

Guyon, I.; Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research.* 3, pp. 1157-1182.

Hall, M. A. 1999. Correlation-based feature selection for machine learning. 178p.

Hall, M. A.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. 2009. The WEKA Data Mining Software: An Update; *SIGKDD Explorations.* 11(1) pp. 10-18.

Haykin, S. 2009. *Neural Networks and Learning Machines.* Prentice-Hall.

INEA - Instituto Estadual do Ambiente/RJ (*Rio de Janeiro state environmental institute).* 2007. http://www.inea.rj.gov.br/Portal/Agendas/BIODIVERSIDADEEAREASPROTEGIDAS/UnidadesdeConservacao/INEA_008619#/ConselhoConsultivo. (06 may 2016).

John, G. H.; Kohavi, R. 1997. Wrappers for feature subset selection. *Artificial Inteligence.* 97(1-2), pp. 273-324.

Klonus, S. & Ehlers, M. 2009. Performance of evaluation methods in image fusion. In: *12th International Conference on Information Fusion, IEEE.* pp. 1409-1416.

Körting, T. S. 2012. GeoDMA: a toolbox integrating data mining with object-based and multi-temporal analysis of satellite remotely sensed imagery. 119 p. http://urlib.net/8JMKD3MGP7W/3CCH86S. (03 apr. 2016).

Kux, H. J. H. & Araujo, E. H. G. 2008. Object-based image analysis using QuickBird satellite image and GIS data, case study Belo Horizonte (Brazil). In: Blaschke, T.; Lang, S.; Hay, G. J. *Object-Based Image Analysis– Spatial Concepts for Knowledge-Driven Remote Sensing Applications*. pp. 531-553.

Laben, C. A.; Bernard V. & Brower W. 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening, US Patent 6,011,875. http://www.google.com/patents/US6011875. (18 apr. 2016).

Moreira, G. F.; Fernandes, R. B. A.; Fernandes Filho, E. I.; Vieira, C. A. O. & Santos, K. A. 2013. Automated Classification of Use and Land Cover from Landsat Images. *Revista Brasileira de Geografia Física*, 6(1), pp. 58-65.

Müller, H.; Rufin, P.; Griffiths, P.; Siqueira, A. J. B. & Hostert, P. 2015. Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian savanna landscape. *Remote Sensing of Environment*, 156, pp. 490-499.

Novack, T. & Kux, H. J. H. 2010. Urban land cover and land use classification of an informal settlement area using the open-source knowledge-based system InterIMAGE. *Journal of Spatial Science*, 55(1), pp. 23-41.

Pinho, C. M. D.; Fonseca, L. M. G.; Körting, T. S.; Almeida, C. M.; Kux, H. J. H. 2012. Land-cover classification of an intra-urban environment using high-resolution images and object-based image analysis. *International Journal of Remote Sensing*, 33(19), pp. 5973-5995.

Ribeiro, M. C.; Metzger, J. P.; Martensen, A. C.; Ponzoni, F. J.; Hirota, M. M. 2009. The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. *Biological Conservation*, 142(6), pp. 1141–1153.

Smith, A. (2010). Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm. *Journal of Spatial Science*, 55(1), pp. 69-79.

Song, X.; Duan, Z. & Jiang, X. 2012. Comparison of artificial neural networks and support vector machine classifiers for land cover classification in Northern China using a SPOT-5 HRG image. *International Journal of Remote Sensing*, 33(10), pp. 3301-3320.

Witten, I. H.; Frank, E.; Hall, M. A. 2011. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.