DATA MINING TECHNIQUES APPLIED TO ALOS-2/PALSAR-2 SATELLITE IMAGERY FOR LAND USE AND LAND COVER CLASSIFICATION

Flávio Fortes Camargo[1,2], Edson Eyji Sano[2], Cláudia Maria de Almeida[3], José Cláudio Mura[3]

[1]National Department for Transportation Infrastructure (DNIT), CEP: 02167-000, São Paulo, SP, Brazil, flavio.fortes.camargo@gmail.com; [2]University of Brasília (UnB), Institute of Geosciences, Brasília, DF, Brazil, edson.sano@gmail.com; [3]National Institute for Space Research (INPE), CEP: 12227-010, São José dos Campos, SP, Brazil, almeida@dsr.inpe.br, jose.mura@inpe.br

ABSTRACT

This paper proposes a workflow for the classification of synthetic aperture radar (SAR) images obtained by the ALOS-2/PALSAR-2 satellite, aiming at the land use and land cover mapping. The study area is located in the western portion of Federal District of Brazil. The presented approach combines multiresolution segmentation, object attributes, and iterative machine learning procedures. A set of 397 attributes was generated based on the amplitude images, HH and HV polarizations. These attributes were processed in the WEKA 3.8 software using the J48 decision tree, Random Forest and Multilayer Perceptron Artificial Neural Network classifiers. Classification results attained Kappa indices higher than 0.70, especially the Multilayer Perceptron Artificial Neural Network algorithm (Kappa = 0.87). This workflow demands low time processing and has potential to be reproduced for other study sites or SAR images obtained at different wavelengths.

Key words — Machine learning, Weka, decision tree, random forest, multilayer perceptron.

1. INTRODUCTION

Remote sensing has consolidated itself as an efficient set of technologies for land use and land cover (LULC) mapping. The ability to systematically image large areas of terrain at different spatial resolutions and in different regions of the electromagnetic spectrum makes remote sensing an important tool for identifying, characterizing and quantifying different LULC classes [1].

In addition to the advances in remote sensing, data mining has emerged as an interesting technique for processing and analysis of different databases [2] [3] [4]. The data mining corresponds to an interdisciplinary field that combines artificial intelligence, data management and visualization, machine learning, mathematical algorithms, and statistics. It offers different methodologies for decision making, problem solving, analysis, planning, diagnostics, pattern recognition, integration, prevention, learning, and innovation [3] [4].

Several LULC mapping activities have used different data mining techniques successfully, such as the decision tree (DT) [5], the Random Forest (RF) [6], and the Artificial Neural Network [7].

In this context, this paper presents a new procedure for the classification of SAR images obtained by the ALOS-2/PALSAR-2 L-band to discriminate LULC classes. The procedure involves the combined use of image segmentation techniques and DT, RF and Multilayer Perceptron Artificial Neural Network (MLP ANN) classifiers available in the WEKA 3.8 software. The proposed approach is iterative and reproducible, allowing the obtainment of quick results with high levels of accuracy.

2. MATERIAL AND METHODS

2.1. Study area

The study area, with approximately 356 km², is located in the western portion of the Federal District of Brazil, more specifically between 15º 39' 02" and 15º 53' 26" of south latitude and between 47º 54' 02" and 48° 01' 41" of west longitude (Figure 1).

This area was selected due to the relatively large diversity of LULC types, with emphasis on the presence of Cerrado native vegetation (shrublands) in the central part of the area, consolidated and under-consolidation urban areas in the southern part, and a reservoir (Santa Maria dam) for domestic consumption, water catchment in the northern part.

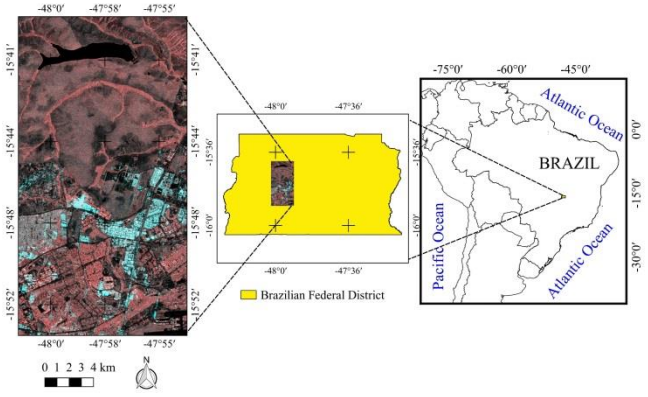

Figure 1. Location of the study area in the Federal District. The image corresponds to the RGB color composite of ALOS-2/PALSAR-2 satellite, HH and HV polarizations (RGB/HV-HH-HH) and overpass on April 29, 2015.

2.2. Material

ALOS-2/PALSAR-2 amplitude images (16 bits) obtained on April 29, 2015 at the StripMap High Sensitive mode, that is, 6-meter spatial resolution and dual polarization (HH and HV), provided by the Japan Aerospace Exploration Agency (JAXA) under the Kyoto & Carbon initiative, were selected. The processing level was 3.1, namely, georeferenced to the Universal Transverse Mercator (UTM) projection and WGS84 datum. The images were obtained with an incidence angle of 32.9º and descending orbit.

In order to identify the LULC types in the study area, this study considered a set of aerial orthophotos mosaics from 2013. These mosaics were published on the internet by the State Secretariat for Land Management and Housing of the Federal District [8]. The semi-detailed vegetation map of the Brasília National Park [9] was also used as a reference map. Based on the visual inspections of these materials, potential areas for collecting samples of digital numbers (regions of interest) in the ALOS-2/PALSAR-2 images were defined for training and validation of the selected classifiers.

The following thematic classes were found in the study area: gallery forest, Cerrado shrubland, wooded Cerrado, pasture and degraded Cerrado, bare soil, urban area with corner reflection, urban area without corner reflection, water reservoir, and paved highway. Cerrado shrubland corresponds to a mosaic of trees, shrubs and grasses with varying proportions and predominance of shrubs [10]. In the wooded Cerrado, the occurrence of trees is smaller in relation to the Cerrado shrubland.

2.3. Image segmentation

For the segmentation of the ALOS-2/PALSAR-2 images (HH and HV polarizations), the multiresolution segmentation algorithm available in the eCognition 8.7 software [11] was used. Only one level of segmentation was generated employing equal weights to the HH and HV amplitude images. The segmentation parameters were empirically defined by trial and error. The scale parameter 350 was selected and higher weights were assigned to the criteria of homogeneity, shape and smoothness, rather than color and compactness.

After segmentation, the following attributes were calculated: layer values, texture, pixel-based, to-neighbors (for each segmented input image), geometric and scene. Thus, for each available image (HH amplitude, HV amplitude, HH intensity and HV intensity) attributes were extracted from the six categories mentioned above. This process resulted in a set of 397 layers of segments with different attributes that were exported for processing in the WEKA 3.8 software. The texture attributes were obtained based on the updated Haralick method, which employs both the Gray Level Co-Occurrence Matrix (GLCM) and the Gray Level Difference Vector (GLDV) [11].

2.4. Image classification and validation

At least 30 samples of segmentation objects from the HH and HV polarizations were collected for each representative LULC class, except for the reservoir and pasture and degraded Cerrado classes, because of their relatively small areas of occurrence in the study area. Samples were randomly divided in two groups, one for training (minimum of 10 samples per class) and another for validation (minimum of 20 samples per class).

According to [12], the recommended minimum number of validation samples per class would be 50, considering study areas with size smaller than 4,047 km² (or 1,000,000 acres) and number of classes fewer than 12. However, in this study, objects generated by segmentation were used for validation. These objects have variable sizes and aggregate many pixels. Given the size of the study area (356 km²), it was not feasible to collect 50 objects per class. Thus, for validation purposes, 20 samples (segments) per class were considered reasonable, considering a minimum thematic accuracy of 85%, according to the study conducted by [13].

In this context, two files were generated in the shapefile format, one for the training of classifiers (with 79 objects) and another for validation of classifications (with 182 objects). A third shapefile was also generated with all samples (with 4,868 objects), containing all attribute fields generated in the eCognition 8.7 software.

For the validation of the classification results, confusion matrices were generated. Kappa indices from different classifiers were compared to each other to verify which one presented the best performance. In this case, hypothesis tests were analyzed based on the standard normal distribution (Z test) [12].

3. RESULTS AND DISCUSSION

Figure 2 shows the classifications obtained by the J48 DT, RF, and MLP ANN algorithms, respectively. The visual analysis of Figure 2 shows that the classification obtained by the MLP ANN presented better spatial distribution of LULC classes. There is an appropriate delineation of both paved highway and urban areas. The reservoir in the northern part of the study area was correctly classified, with no confusion with bare soil. The Cerrado natural areas presented a spatial distribution similar to those observed in the orthophotos mosaics and also in the semi-detailed vegetation map of the Brasília National Park.

The confusion matrix of the J48 DT (Table 1) indicates that the overall accuracy was 0.76, while the Kappa index was 0.72. A high degree of confusion was found between bare soil and paved highway, which was somewhat expected because of the predominance of almost specular scattering in these areas due to the low surface roughness related to the relatively long wavelength of the ALOS-2/PALSAR-2 L-band. The bare soil presented 50% of correct classification and also presented commission errors with the paved highway and reservoir. The analysis of the pasture and

degraded Cerrado and reservoir was impaired given the small number of samples available for validation.

Classification \ Ground Truth	a	b	c	d	e	f	g	h	i	j	Total
a	**20**	0	0	0	0	0	0	0	0	0	20
b	0	**22**	1	0	0	0	0	0	9	0	32
c	0	2	**20**	0	0	0	0	0	0	0	22
d	0	0	0	**3**	3	0	0	0	0	0	6
e	0	0	0	1	**12**	6	0	1	0	0	20
f	0	0	0	0	9	**13**	0	2	0	0	24
g	0	0	1	0	0	0	**22**	0	0	0	23
h	3	0	0	0	0	0	0	**5**	0	0	8
i	1	2	0	0	0	2	1	0	**21**	0	27
j	0	0	0	0	0	0	0	0	0	**0**	0
	24	26	22	4	24	21	23	8	30	0	**182**

Overall accuracy: 0.76 Kappa index: 0.72

a = Wooded Cerrado; b = Gallery forest; c = Cerrado shrubland; d = Reservoir; e = Bare soil; f = Paved highway; g = Urban area with corner reflection; h = Pasture & degraded Cerrado; i = Urban area without corner reflection; and j = Unclassified.

Table 1. Confusion matrix of the classification obtained by the J48 DT algorithm.

The RF algorithm had a higher performance than the J48 DT: the overall accuracy was 0.79 and the Kappa index was 0.76 (Table 2). However, the results presented by the J48 DT and RF classifiers were not statistically different, at a significance level of 5%. This result differs from some studies that indicated that RF performs better than J48 DT [6].

The performance of the RF algorithm can be related to the size and distribution of the training samples used. RF is a bagging-type classifier, that is, decision trees are created from subsets of the same sample set with replacement. Thus, RF is sensitive to the characteristics of the sample set used [6]. The detailed analysis of the errors and accuracy shows that the RF classifier also presented high confusion between bare soil and paved highway.

The MLP ANN presented relatively high values of global accuracy (0.89), Kappa index (0.87) and low confusion between classes (Table 3). There was low confusion among classes and all obtained scores were ≥ 80%. Even classes with difficult separation presented low confusion (reservoir, bare soil, and paved highway). This classifier was able to better explore the set of attributes available, compared to the J48 DT and RF.

The test of hypothesis performed demonstrates that the results presented by the RF and MLP ANN classifiers are statistically different, at a significance level of 5%. Despite the low performance of RF, possibly due to problems in the training sampling, this result is consistent with other studies that also highlighted the high performance of the MLP ANN relative to other classifiers, including RF [14].

Classification \ Ground Truth	a	b	c	d	e	f	g	h	i	j	Total
a	**20**	0	0	0	0	0	0	0	0	0	20
b	1	**21**	2	0	0	0	0	0	8	0	32
c	0	1	**21**	0	0	0	0	0	0	0	22
d	0	0	0	**5**	1	0	0	0	0	0	6
e	0	0	0	0	**15**	5	0	0	0	0	20
f	0	0	1	1	7	**15**	0	1	0	0	25
g	0	0	0	0	0	0	**22**	0	0	0	22
h	3	0	0	0	0	0	0	**5**	0	0	8
i	1	4	0	0	0	1	1	0	**20**	0	27
j	0	0	0	0	0	0	0	0	0	**0**	0
	25	26	24	6	23	21	23	6	28	0	**182**

Overall accuracy: 0.79 Kappa index: 0.76

a = Wooded Cerrado; b = Gallery forest; c = Cerrado shrubland; d = Reservoir; e = Bare soil; f = Paved highway; g = Urban area with corner reflection; h = Pasture & degraded Cerrado; i = Urban area without corner reflection; and j = Unclassified.

Table 2. Confusion matrix of the classification obtained by the RF algorithm.

Classification \ Ground Truth	a	b	c	d	e	f	g	h	i	j	Total
a	**20**	0	0	0	0	0	0	0	0	0	20
b	1	**28**	0	0	0	0	0	0	3	0	32
c	0	0	**22**	0	0	0	0	0	0	0	22
d	0	0	0	**6**	0	0	0	0	0	0	6
e	0	0	0	0	**18**	1	0	1	0	0	20
f	0	0	0	1	3	**20**	0	0	0	0	24
g	0	0	0	0	0	1	**22**	0	0	0	23
h	2	0	1	0	1	0	0	**4**	0	0	8
i	0	2	0	0	0	2	1	0	**22**	0	27
j	0	0	0	0	0	0	0	0	0	**0**	0
	23	30	23	7	22	24	23	5	25	0	**182**

Overall accuracy: 0.89 Kappa index: 0.87

a = Wooded Cerrado; b = Gallery forest; c = Cerrado shrubland; d = Reservoir; e = Bare soil; f = Paved highway; g = Urban area with corner reflection; h = Pasture & degraded Cerrado; i = Urban area without corner reflection; and j = Unclassified.

Table 3. Confusion matrix of the classification obtained by the MLP ANN classifier.

5. CONCLUSIONS

All classifiers presented high performances. This can be attributed to the ability of these non-parametric classifiers to process various types and distributions of data. The performance of the RF classifier was lower than those reported in the literature, probably due to the size and distribution of the training samples used.

The classification procedures used in this study were effective in extracting information from the L-band ALOS-2/PALSAR-2 satellite images. This is a relevant finding, since some places in the world, especially in tropical countries, several applications rely only on SAR images due to persistent cloud coverage.

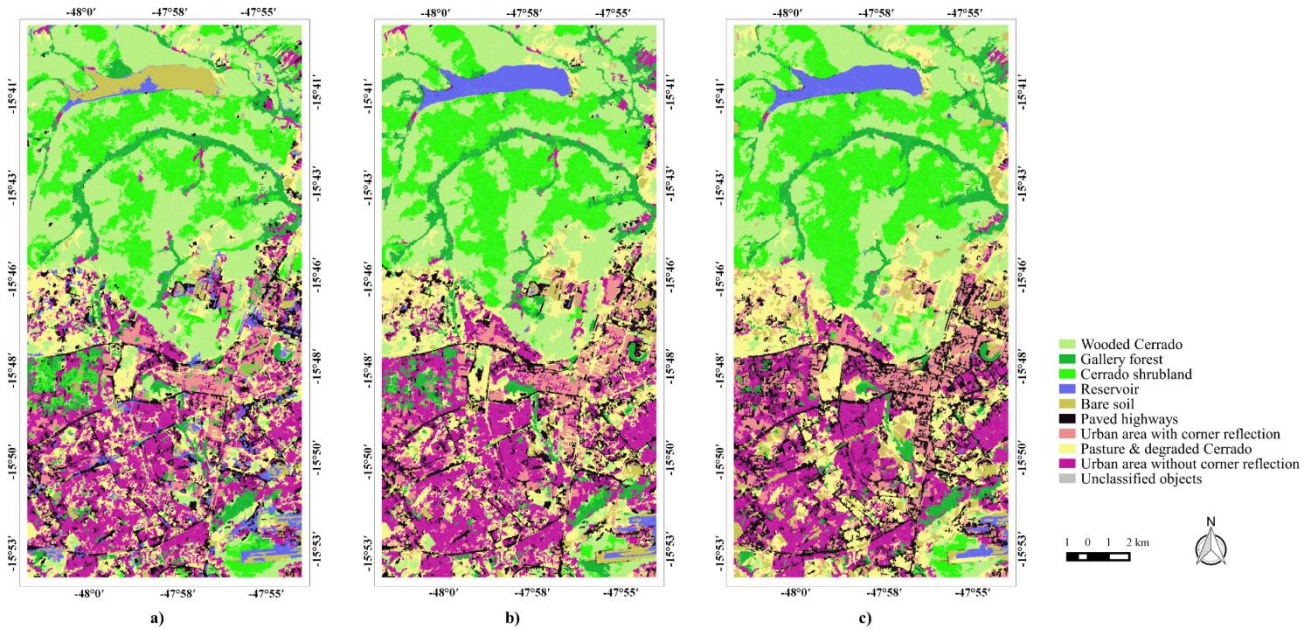

Figure 2. Classification results by the J48 DT (a), RF (b) and MLP ANN (c) algorithms.

Besides that, due to the intrinsic characteristics of the SAR images, classifications based on traditional parametric methods usually present poor performance, making the use of alternative classifiers, such as those explored in this work, even more important. Finally, it is important to emphasize that the intrinsic iterative character of the machine learning procedures provides higher objectivity in the classification task, making the entire process reproducible in time and for other LULC types.

6. REFERENCES

[1] Congalton, R. G.; Gu, J.; Yadav, K.; Thenkabail, P.; Ozdogan, M. "Global land cover mapping: a review and uncertainty analysis". *Remote Sensing*, v. 6, pp. 12070–12093, 2014.

[2] Piatetsky-Shapiro, G. and Fayyad, U. "An introduction to SIGKDD and a reflection on the term 'data mining'". *SIGKDD Explorations*, v. 2 (13), pp. 103–104, 2012.

[3] Tsai, H. "Global data mining: An empirical study of current trends, future forecasts and technology diffusions". *Expert Systems with Applications*, v. 39, pp. 8172–8181, 2012.

[4] Tsai, H. "Knowledge management vs. data mining: Research trend, forecast and citation approach". *Expert Systems with Applications*, v. 40, pp. 3160–3173, 2013.

[5] Körting, T. S.; Fonseca, L. M. G.; Câmara, G. "GeoDMA – geographic data mining analyst". *Computers & Geosciences*, v. 57, pp. 133–145, 2013.

[6] Belgiu, M. and Drăguț, L. "Random forest in remote sensing: A review of applications and future directions". *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 114, pp. 24–31, 2016.

[7] Foody, G. M. "Impacts of sample design for validation data on the accuracy of feedforward neural network classification". *Applied Sciences*, v. 7 (888), pp. 1-15, 2017.

[8] SEGETH/DF. Geoportal SEGETH DF. http://geoportal.segeth.df.gov.br/mapa/. Accessed in 17 September 2016.

[9] Ferreira, M. E.; Ferreira, L. G.; Sano, E. E.; Shimabukuro, Y. E. "Spectral linear mixture modelling approaches for land cover mapping of tropical savanna areas in Brazil". *International Journal of Remote Sensing*. v. 28 (2), pp. 413–429.

[10] Ribeiro, J. F. and Walter, B. M. T. "As principais fito-fisionomias do Cerrado". In: Sano, S. M.; Almeida, S. P.; Ribeiro, J. F. (eds.). *Cerrado: Ecologia e Flora*, Planaltina: Embrapa Cerrados, pp. 151-199, 1998.

[11] Trimble. "eCognition Developer 8.7: Reference Book". Munich: *Trimble*, 2011.

[12] Congalton, R. G. and Green, K. "Assessing the accuracy of remotely sensed data: principles and practices". Boca Raton: *CRC Press*, 200 p, 2009.

[13] Genderen, J. L. and Lock, B. F. "Testing land-use map accuracy". *Photogrammetric Engineering & Remote Sensing*, v. 43 (9), pp. 1135–1137, 1977.

[14] Shiraishi, T.; Motohka, T.; Thapa, R. B.; Watanabe, M.; Shimada, M. "Comparative assessment of supervised classifiers for land use–land cover classification in a tropical region using time-series PALSAR mosaic data". *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 7 (4), pp. 1186–1199, 2014.