INTEGRATION OF WORLDVIEW-2 AND LIDAR DATA TO MAP A SUBTROPICAL FOREST AREA: COMPARISON OF MACHINE LEARNING ALGORITHMS

C. Sothe[1], C. M. de Almeida[1], M. B. Schimalski[2], V. Liesenberg[2]

[1]Division of Remote Sensing, National Institute for Space Research, Brazil
[2]Department of Forest Engineering, Santa Catarina State University, Brazil

ABSTRACT

This work is committed to explore the integration of airborne LiDAR data and WorldView-2 (WV-2) images to classify land cover and land use in a rural area with the presence of a subtropical forest. Different methods were used for this purpose: two artificial neural networks (ANN) and three decision trees forests. The results demonstrated that the inclusion of LiDAR data significantly improved the classifications in all methods. Excluding the Convolutional Neural Network, the classification algorithms had a nearly similar performance, and none of them achieved the best accuracy for all adopted classes. Forest by Penalizing Attributes (FPA) attained the best general result, with a Kappa index of 0.92, while Rotation Forest obtained the best result in the classification of the two vegetation classes.

Index Terms— Forest succession stages, Artificial Neural Network, Decision Forest, Data fusion.

1. INTRODUCTION

The Mixed Ombrophilous Forest (MOF) or Araucaria Moist Forest is one of the main phytophysiognomies of the Atlantic Rain Forest in Southern Brazil [1]. It is characterized by the presence of *Araucaria angustifólia* (Parana´s pine tree) as an emerging species over the canopy, actuating as an indicator of this forest phytophysiognomy. It is regarded as one of the country´s most threatened forest category, for only 7% of its original cover is still standing [2]. The greatest part of its remnants is composed of small isolated patches (<50 ha) of secondary forests in initial (SS_1) and intermediate succession stages (SS_2) [3]. Nevertheless, although *Araucaria angustifólia* is on top of MOF main species ranking, it is declared as "vulnerable" in Brazil´s Official List of Plants Extinction Endangered Species [4]. Considering this, a precise mapping of its few remaining forest remnants becomes crucial, so as to ease the implementation of management, surveillance, administration, and conservation strategies.

In this respect, field surveys are efficient; however, they tend to be time-consuming and costly to be executed at a large scale and in a systematic way. These restraints have fostered initiatives based on remote sensing as an effective means to map the vegetation succession stages [5-7] and tree species, as those at risk of extinction [8].

Data fusion, i.e. the integration of data coming from manifold sources, is still a challenging topic in Remote Sensing and Digital Image Processing [9]. A promising approach for the land cover and land use classification, especially in the case of different forest classes, is the combined use of passive multispectral data with active sensors data, like Light Detection and Ranging (LiDAR), since they provide complementary information on the targets of interest (spectral response and vertical structure, respectively) [10]. Besides these data, the choice of the classifier is also decisive for a reliable land cover and land use mapping. In this sense, machine learning is a relatively new scientific field, which is yet constantly evolving in a fast way and likely to yield ever better results [11]. The increasing use of such methods in latest years is due to several factors: their capacity to learn complex patterns, which makes it possible to apply them to faulty or noisy data; the possibility of incorporating *a priori* information in the analysis; and its independence in relation to the data statistical distribution. This latter advantage renders available the addition of data from multiple sensors, auxiliary variables and even categorical variables [12].

In face of what has been previously exposed, the goal of this work is to analyze the integration of WorldView-2 (WV-2) multispectral data with airborne laser scanning (ALS) data for the semiautomatic land cover and land use mapping in a subtropical forest area of the Atlantic Rain Forest. The specific goals are: i) analyze the performance attained in classifications relying on the use of multispectral data alone and compare them with classifications also using LiDAR data and; ii) test and compare different machine learning methods, like those more often used, such as Random Forest (RF) [13] and Multilayer Perceptron (MLP), as well as barely explored algorithms in forest applications, like Rotation Forest (RotF) [14], Forest by Penalizing Attributes (FPA) [15], and Convolutional Neural Network (CNN).

2. MATERIAL AND METHODS

The study area is located in the municipality of Painel, Santa Catarina state, south of Brazil. The area belongs to the Atlantic Rain Forest biome and shelters the Araucaria Moist Forest phytophysiognomy, characterized by the presence of the *Araucaria angustifólia* species.

The WV-2 scenes used in this work were acquired in May of 2012. Initially, they were converted to radiance images, using the

command Radiometric Calibration of ENVI 5.0. Next, the images were once again converted to surface reflectance using the tool Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes (FLAASH). After atmospheric correction, the WV-2 multispectral bands with 2 m of spatial resolution were pansharpened with the WV-2 panchromatic band, with 0.5 m, by means of the Gram-Schmidt method. Finally, the WV-2 scene was orthorectified based on the rational polynomial coefficients, using parameters provided by the images metadata and the LiDAR-derived Digital Terrain Model (DTM). For geocoding the data, an orthoimage with 0.39 m of spatial resolution acquired by the Airborne System for Acquisition and Post-processing Images (SAAPI) was used as reference together with 15 ground control points.

The LiDAR data, on their turn, were acquired by the Leica ALS-60 sensor onboard of a flight undertaken in January of 2011, with an average density of 7 points/m². Firstly, outliers were removed and then the DTM and the Digital Surface Model (DSM) were generated using the *Adaptive* TIN filter. Next, the Canopy Height Model (CHM) was extracted by the subtraction between the DSM and DTM.

The following stage consisted in the feature extraction from both datasets. From the WV-2 images, texture metrics were calculated based on the four multispectral bands (red, green, blue and near infrared), with a 5x5 window and southwest direction, using the grey-level co-occurrence matrix (GLCM) proposed by Haralick et al. [16]. Four vegetation indices (VI) were computed (IV): the Normalized Difference Vegetation Index (NDVI), the Simple Ratio (SR), the Red-edge Normalized Difference Vegetation Index (NDVI-RedEdge) and the Optimized Soil Adjusted Vegetation Index (OSAVI).

As to the LiDAR data, besides the CHM, seven elevation and intensity-derived metrics were extracted with the LidR package of R: percentage of intensity returned below the 90th percentile of height (ipcumzq90), intensity mean (imean), maximum intensity (imax), 10th percentile of height distribution (zq10), kurtosis of height distribution (zkurt), skewness of intensity distribution (iskew), mean height (zmean). The spatial resolution of all LiDAR-derived attributes was 1 m, so as to make their resolution compatible with that of the WV-2 data and to assure that there were no missing values in the resulting images. Two datasets were then created: one solely composed by the original bands of WV-2 and their attributes (texture & VI), named as *WV2*, and the other one composed by these data together with the ALS-derived attributes, named as *full*.

The multiresolution segmentation available at eCognition software was applied to the second dataset. A weight value of 1 was assigned to the original bands of WV-2 and to the LiDAR-derived CHM band, while a weight value of 0 was applied to the remaining bands, in order to avoid an oversegmentation. The segmentation parameters were heuristically defined and the scale factor was set to 15, the shape factor to 0.2 and the smoothness factor to 0.5. After this stage, training samples were collected for the eight classes based on the SAAPI orthoimage and field observations. The classes and the number of respective segments used as training samples were: natural vegetation in a secondary intermediate stage of recovery (SS_2) (61), natural vegetation in a secondary early stage of recovery (SS_1) (59), reforestation (corresponding to areas with the exotic species *Pinus* sp.) (63), araucaria (dominant tree species in the region) (40), agriculture (crops) (68), bare soil (21), fields (109) and water bodies (31). The samples as well as all remaining segments were exported as a shapefile containing all the attributes previously mentioned and the layer mean of all bands, totalizing 52 attributes for the *full* dataset and 44 attributes for the *WV2* dataset.

In the sequence, two databases were created in Attribute-Relation File Format (ARFF) in order to drive the software Waikato Environment Knowledge Analysis (WEKA), version 3.7. The next stage concerned the supervised classification of both datasets. For this task, five machine learning classifiers were selected, two ANNs (MLP and CNN) and three based on decision trees forests (RF, RotF and FPA). Initially, preliminary tests were carried out using a cross-validation process, in which different parameter settings were evaluated in relation to the default values of WEKA. It was observed that in the great majority of cases, the default values had a slightly higher performance and they have been kept in the final classifications. Only for the FPA algorithm, 50 trees were employed instead of the default value of 10, and for the CNN, a convolutional layer of values 100-3-3-2-2 was assigned, respectively corresponding to the number of feature maps, patch-width, patch-height, pool-width, and pool-height.

For the accuracy assessment, polygons meant for analyses were delimited in the image for each class, with the aid of the orthoimage and field observations. Based on these polygons, 50 pixels were randomly selected in each class. This procedure was repeated for each classification, so as to keep the independence among the validation samples. The number of samples was defined according to Congalton and Green [17], who recommend a minimum of 50 samples for maps of less than one million acres in size and fewer than 12 classes. After the samples collection, error matrices were elaborated based on a cross-checking between the classified maps and their respective validation samples. The following indices were extracted from such matrices: a) overall accuracy (OA); (b) producer´s accuracy (PA), (c) user´s accuracy (UA); (d) Kappa index and conditional Kappa per class (Kc) [17]. The z tests for the Kappa indices were executed with a level of significance of 5%, i.e., with a confidence interval of 95%. The value of the normal distribution z is obtained by the ratio of the difference between two distinct Kappa indices to the difference of their respective variances. When $z > 1.96$, the test is significant, the null hypothesis is rejected, and hence, there is significantly statistical difference between the two values.

3. RESULTS AND DISCUSSION

As expected, the *full* dataset obtained a significant increase in accuracies for all classifiers. In the case of the *WV2* dataset, the Kappa values ranged from 0.59 to 0.81 for CNN and MLP, respectively, while for the *full* dataset, the Kappa values oscillated from 0.68 to 0.92 for CNN and FPA, respectively (Figure 1b). It was observed that LiDAR data was mainly important for the

discrimination between the two succession stages classes (SS_1 and SS_2) and also between SS_2 and *araucaria*. When located amidst the forest, araucaria tends to be a dominant species in the canopy, and the use of LiDAR data, which provide information on the forest vertical structure, increased the accuracy of this class. The same thing happened in the discrimination of the two succession stages, since SS_1, besides differences in the species composition, has a shorter height and a smaller number of vertical strata than SS_2. Other authors also observed that the inclusion of LiDAR data in the classification process was advantageous, chiefly in the case of spectrally similar classes, like forest species and typologies [10, 18].

Regarding the classifiers, in the case of the *full* dataset, FPA attained the best general result (Figure 1a), with a Kappa index of 0.92 (Figure 1b), although it did not significantly differ from MLP, RF, and RotF, which also reached good results: Kappa of 0.88, 0.88 and 0.91, respectively. Adnan and Islam [15], idealizers of FPA, compared it with other methods, such as RF and RotF. The authors clarified that FPA, contrary to RF, uses the whole feature space, and hence, generates trees with high individual precision. When the algorithms performances were compared for the *WV2* dataset, it was observed that MLP (Kappa of 0.81) was significantly superior to CNN, FPA and RF, not differing much from RotF (Kappa of 0.77). CNN, on its turn, was significantly inferior to further classifiers, both for the *full* and *WV2* datasets. According to Pasupa and Sunhem [19], CNN needs a great number of training samples, in order to avoid overfit, what may have been one of the reasons for the poor performance of this algorithm. Moreover, its parameters have not been widely explored in this study.

None of the algorithms obtained the best results for all classes (Figure 1c). The RotF classifier with the *full* dataset (RotF_*full*) attained the best results for two vegetation classes: SS_2, with Kc of 0.98, and *araucaria*, with Kc of 0.96. For SS_1, the best result was reached with FPA_*full*, Kc of 0.81. The *reforestation* class had a Kc value of 1.0 with FPA_*full*, MLP_*full* and RF_*full*. Only the class *field* had a better performance in an experiment without LiDAR data, what can be explained by the fact that this class shows a reduced elevation range. In this case, the best result was attained with MLP_*WV2*, Kc of 0.93. As to the processing time, the decision trees forests showed to be faster than ANN (MLP and CNN), what represents an advantage of such methods, especially when dealing with datasets of bulky dimension.

4. CONCLUSION

This study showed that the integration of LiDAR data with WV-2 multispectral data led to a significant increase in classifications accuracies for all tested algorithms. This increase was mainly observed in the case of spectrally similar classes owing marked elevation differences, like the vegetation succession stages and the tree species araucaria.

The machine learning algorithms have a great potential for classifying datasets coming from manifold sources. With the exception of CNN, all other classifiers had a similar performance, and none of them attained the best accuracy for all land cover and land use classes. In spite of that, it is worth mentioning that the decision trees forests are advantageous in terms of processing time, particularly in the case of large datasets. In general terms, FPA had the highest Kappa for the *full* dataset (0.92), while MLP obtained the highest Kappa for the *WV2* dataset (0.81).

5. REFERENCES

[1] P. Higuchi, A. C. da Silva, T. de S. Ferreira, S. T. de Souza, J. P. Gomes, K. M. Silva, K. F. dos Santos, C. Linke, and P. da S. Paulino, "Influência de variáveis ambientais sobre o padrão estrutural e florístico do componente arbóreo, em um fragmento de Floresta Ombrófila Mista Montana em Lages, SC," *Ciência Florestal*, v. 22, n. 1, pp. 79- 90, 2012.

[2] A. C. Vibrans, A. L. de Gasper, and J. J. V. Müller, "Para que inventariar florestas? Reflexões sobre a finalidade do inventário florístico florestal de Santa Catarina," *Revista de Estudos Florestais*, vol. 14, no. 1, pp. 6-13, 2012.

[3] M. C. Ribeiro, J. P. Metzger, A. C. Martensen, F. J. Ponzoni, and M. M. Hirota, "The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation," *Biological Conservation*, vol. 142, pp. 1141-1153, 2009.

[4] Brasil, "Portaria MMA nº 443, de 17 de dezembro de 2014. Estabelece a Lista Nacional de Espécies da Flora Ameaçadas de Extinção," *Diário Oficial da União*, 18 de dezembro de 2014, seção 01, pp. 110 -121, 2014.

[5] D. Lu, G. Li, E. Moran, and W. Kuang, "A comparative analysis of approaches for successional vegetation classification in the Brazilian Amazon," *GIScience & Remote Sensing*, vol. 51, no. 6, pp. 695-709, 2014.

[6] A. G. Piazza, A. C. Vibrans, V. Liesenberg, and J. C. Refosco, "Object-oriented and pixel-based classification approaches to classify tropical successional stages using airborne high–spatial resolution images," *GIScience & Remote Sensing*, 2016.

[7] C. Sothe, C. M. de Almeida, V. Liesenberg, and M. B. Schimalski, "Evaluating Sentinel-2 and Landsat-8 data to map sucessional forest stages in a subtropical forest in Southern Brazil," *Remote Sensing*, vol. 9, no. 8, pp. 838, 2017.

[8] G. Omer, O. Mutanga, E. M. Abdel-Rahman, and E. Adam, "Performance of Support Vector Machines and Artificial Neural Network for Mapping Endangered Tree Species Using WorldView-2 Data in Dukuduku Forest, South Africa," *IEEE Journal of Selected Topics in applied Earth Observations and Remote Sensing*, 2015.

[9] M. Dalla Mura, S. Prasad, F. Pacifici, G. P. Fellow, J. Chanussot, and J. A. Benediktsson, "Challenges and Opportunities of Multimodality and Data Fusion in Remote Sensing," *Proceedings of the IEEE*, vol. 103, no. 9, 2015.

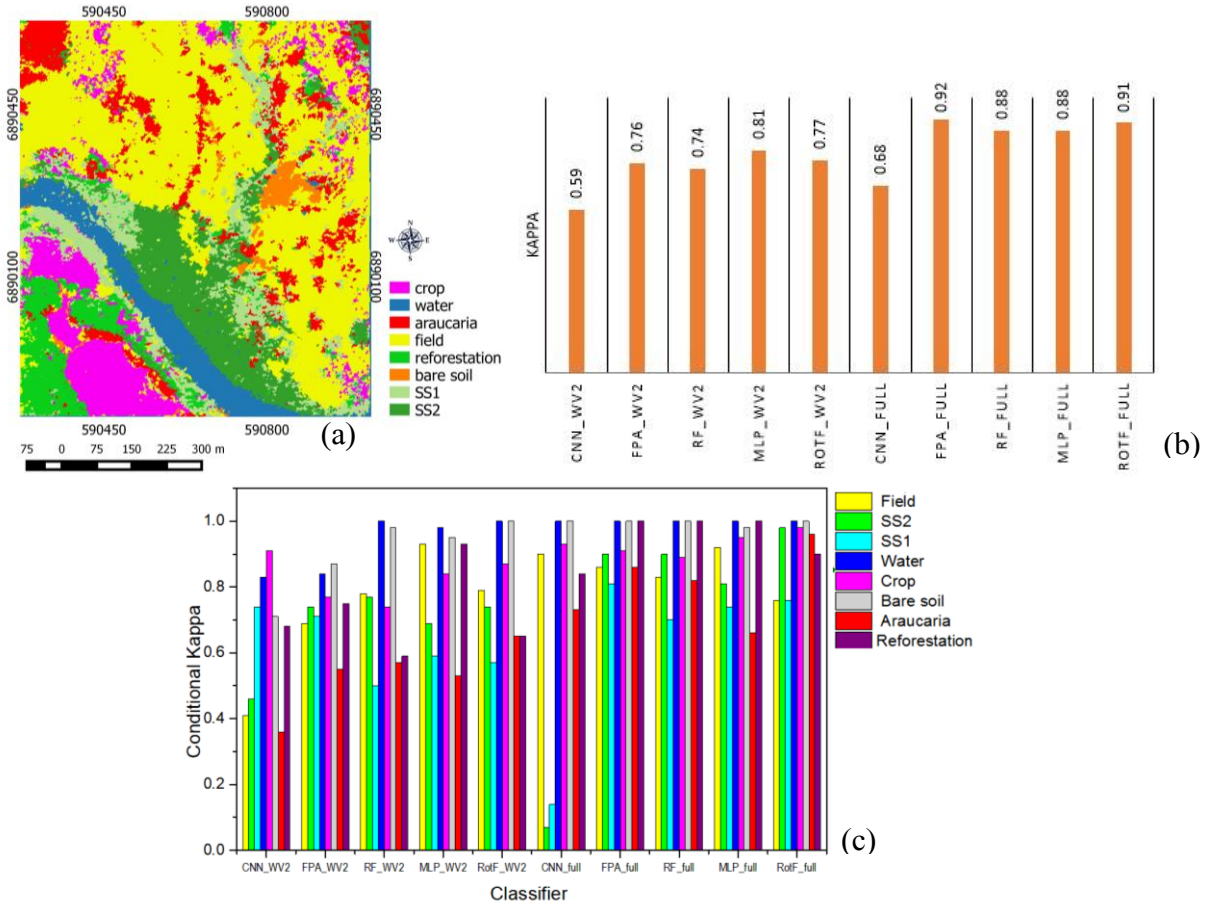

Figure 1. (a) Classification result using the FPA_*full* algorithm. (b) Kappa index of each classifier. (c) Conditional Kappa values per class for all classifiers.

[10] M. Dalponte, L. Bruzzone, and D. Gianelle, "Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data" *Remote Sensing of Environment,* vol. 123, pp. 258–270, 2012.

[11] V. F. Rodriguez-Galiano, and M. Chica-Rivas, "Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models," *Internat. Journal of Digital Earth*, 2012.

[12] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts, "Mapping Land-Cover Modifications Over Large Areas: A Comparison of Machine Learning Algorithms," *Remote Sensing of Environment*, vol. 112, no. 5, pp. 2272-2283, 2008.

[13] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation Forest: A New Classifier Ensemble Method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, 2006.

[15] N. Adnan, and Z. Islam, "Forest PA: Constructing a Decision Forest by Penalizing Attributes used in Previous Trees," *Expert Systems With Applications,* 2007.

[16] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, vol. 3, pp. 610–621, 1973.

[17] R. G. Congalton, and K. Green, "Assessing the accuracy of remotely sensed data: principles and practices," New York*: Lewis Publishers*, 1999.

[18] M. A. Cho, R. Mathieu, G. P. Asner, L. Naidoo, J. A. Van Aardt, A. Ramoelo, P. Debba, K. Wessels, R. Main, I. P. Smit and B.F Erasmus, "Mapping tree species composition in South African savannas using an integrated airborne spectral and LiDAR system," *Remote Sensing of Environment*, vol. 125, pp. 214–226, 2012.

[19] K. Pasupa, and W. A. Sunhem, "Comparison between Shallow and Deep Architecture Classifiers on Small Dataset," In: 8th International Conference on Information Technology and Electrical Engineering (ICITEE), *Proceedings*…Yogyakarta, Indonesia, 2016.