IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-2010-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IGARSS52108.2023.10282190

INVESTIGATING SAR-OPTICAL DEEP LEARNING DATA FUSION TO MAP THE BRAZILIAN CERRADO VEGETATION WITH SENTINEL DATA

Paulo Silva Filho[1,2], *Claudio Persello*[2], *Raian V. Maretto*[2], *Renato Machado*[1]

Aeronautics Institute of Technology, São José dos Campos-SP, Brazil[1]
Faculty of Geo-Information Science and Earth Observation, University of Twente,
Enschede, The Netherlands[2]

ABSTRACT

Despite its environmental and societal importance, accurately mapping the Brazilian Cerrado's vegetation is still an open challenge. Its diverse but spectrally similar physiognomies are difficult to be identified and mapped by state-of-the-art methods from only medium- to high-resolution optical images. This work investigates the fusion of Synthetic Aperture Radar (SAR) and optical data in convolutional neural network architectures to map the Cerrado according to a 2-level class hierarchy. Additionally, the proposed model is designed to deal with uncertainties that are brought by the difference in resolution between the input images (at 10m) and the reference data (at 30m). We tested four data fusion strategies and showed that the position for the data combination is important for the network to learn better features.

Index Terms— Cerrado, deep learning, SAR-optical data fusion, semantic segmentation, remote sensing

1. INTRODUCTION

The Cerrado is the second largest biome in South America with an approximate area of 2 million km^2. It is considered one of the most diverse ecosystems in the world, with more than 11.000 different plant species and a unique vast fauna [1]. The biome's natural vegetation can be divided in three main formation levels: forest, savannah, and grassland. These formations can be subdivided into lower levels up to 25 phytophysiognomies [2]. Despite its environmental importance, the biome has undergone a rapid transformation process due to human activities in the last 50 years. It is estimated that almost 50$\%$ of the biome's natural vegetation has been suppressed, and several other areas are still under pressure from agriculture, livestock, and coal production. This rapid transformation is causing many negative consequences such as biodiversity loss, growth of invasive species, soil erosion, and water pollution. In fact, less than 3$\%$ of the area is under strict protection. In that sense, mapping and monitoring the remaining vegetation is essential to guide public policies for the Cerrado preservation and the sustainable development of the anthropic areas [3].

Nevertheless, producing accurate high-resolution land use and land cover (LULC) maps of the Cerrado is a big challenge. In addition to its large extension, other characteristics of the region can pose problems for the mapping, such as the seasonality of natural changes, high dynamic land use pressure, high cloud and smoke coverage, spectral similarities and high heterogeneity of the natural vegetation formations. These difficulties are present even in mapping approaches based on visual interpretation [4]. Recent researches have indicated that higher spatial resolution is essential to discriminate multiple plant physiognomies. On the other hand, a higher resolution means more data, which makes visual interpretation extremely expensive and impractical. Therefore, researches on machine learning are needed to analyze large volumes of Earth observation data, and obtain a highly accurate classification. Recent works in the literature have explored these techniques using optical data with classic machine learning algorithms [5, 6], such as random forest, support vector machines, and others.

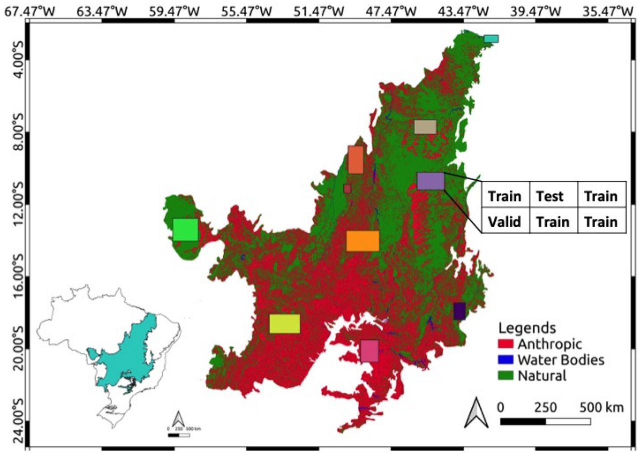

Fig. 1. Cerrado areas of the produced dataset.

The resolutions of the optical sensors used in the works vary from 30m (Landsat) to 2m (WorldView-2). Despite some promising results using deep learning methods at 2m resolution images [7], there is still some spectral confusion between classes that the optical data cannot deal with.

SAR data is a good complementary source of information. Because of SAR canopy penetration capability, it is more effective in capturing features that can describe structural elements of the vegetation [1, 8]. Therefore, the combination of SAR and optical data is expected to improve the classification of the different vegetation types, even in lower-resolution images. Few works have attempted to fuse information from both sensors to map Cerrado's vegetation using classic machine-learning algorithms [9, 1, 10]. To the best of our knowledge, deep learning has not yet been explored for this matter.

One of the main criticism of the machine learning works for the Cerrado is that most of them are restricted to a small study area that does not represent the diversity of the whole biome [3]. This work aims at investigating the data fusion of SAR and optical data for mapping the natural vegetation of the Cerrado at 10m resolution on a regional to national level, using deep learning semantic segmentation algorithms.

2. STUDY AREA

Ten areas of the Cerrado were selected to build a dataset using Sentinel-1 (S1) Single Look Complex (SLC) data and Sentinel-2 (S2) optical data from the year 2018, during the dry season (June to September). The total study area comprises an area of 142,801 km^2 and spreads across 55 S2 tiles. Each tile has an average area of 100×100km^2 with a 5km overlap between adjacent tiles. 36 of those tiles were used for training, 9 for validation, and 10 for testing (one tile per area). For the testing tiles, the overlapping areas were excluded to guarantee a proper accuracy assessment. Figure 1 presents the chosen areas and an example of how the tiles were separated.

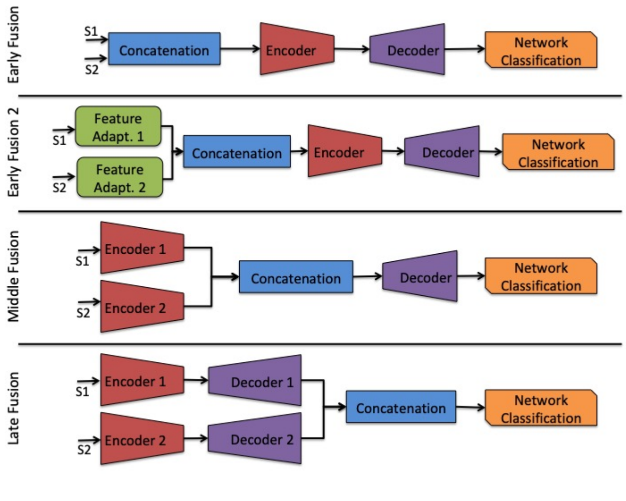

Fig. 2. Investigated SAR-optical fusion architectures.

The reference data was derived from the combination of the TerraClass Project 2018 LULC map [4] and the vegetation map produced by [6] in a hierarchical manner. TerraClass was used to produce the reference at a first hierarchical level, with three classes: 1) anthropic areas, 2) natural areas, and 3) water bodies. The natural areas were further subdivided into five second-level classes using both references: 1) natural non-vegetated areas (sandy and rocky outcrops), 2) grassland, 3) forest, 4) savannah and, 5) secondary vegetation areas.

The reference maps were downscaled from 30m to 10m resolution using nearest neighbor resampling. However, this interpolation introduces noise to the labeled dataset, affecting the performance of supervised learning methods. Moreover, another source of noise is the uncertainty on each reference map. The Terraclass map, for example, is reported to have a 90.7% overall accuracy, while the other reference's is 86%.

3. METHODOLOGY

Our methodology addresses three main aspects to achieve our goal: 1) learn from a noisy dataset due to different spatial resolution between the input data and the reference maps; 2) evaluate different deep learning data fusion architectures, and 3) apply a hierarchical classification approach according to the classes structure.

Table 1. Test set accuracies of the architectures with one and two outputs for the first level of the hierarchical classification

		S2 1-output	S2 2-outputs
Accuracy	OA	84.27%	**88.81%**
	Natural	84.03%	**93.30%**
	Anthropic	**85.75%**	80.81%
	Water	**39.50%**	20.26%
IoU	mIoU	**60.30%**	58.46%
	Natural	78.40%	**85.00%**
	Anthropic	63.91%	**70.64%**
	Water	**38.59%**	19.73%
F1-Score	mF1-Score	**73.86%**	69.21%
	Natural	87.89%	**91.89%**
	Anthropic	77.98%	**82.79%**
	Water	**55.69%**	32.96%

Our main approach to deal with the resolution noise was to design network architectures in an encoder-decoder architecture that infers the results at both resolutions (10m and 30m). While the 10m output was evaluated in the noisy resampled reference, the 30m output used the original reference. The back-propagation used the sum of the loss from each output to update the weights of the network. The optimization algorithm used was Adam, with an exponential learning rate decay. This novel two-output combination is expected to reduce the effects on the training of the noisy artifacts introduced by the nearest neighbor resampling of the reference data. During training, the accuracy of both outputs

Table 2. Test set accuracies of the different data fusion architectures proposed for the first level of the hierarchical classification

		S1+S2 early 1	S1+S2 early 2	S1+S2 middle	S1+S2 late
Accuracy	OA	87.03%	**89.11%**	86.76%	81.57%
	Natural	85.93%	89.79%	**90.28%**	86.45%
	Anthropic	**89.55%**	88.06%	80.09%	72.27%
	Water	**78.39%**	71.91%	51.45%	33.98%
IoU	mIoU	61.43%	**68.64%**	65.60%	55.25%
	Natural	81.95%	**84.81%**	82.23%	76.21%
	Anthropic	70.35%	**72.96%**	66.36%	55.85%
	Water	31.99%	48.14%	**48.21%**	33.68%
F1-Score	mF1-Score	73.72%	**80.38%**	78.36%	69.52%
	Natural	90.08%	**91.78%**	90.25%	86.50%
	Anthropic	82.60%	**84.37%**	79.78%	71.67%
	Water	48.47%	64.99%	**65.05%**	50.39%

in the validation dataset was monitored. We selected the best model by the best accuracy on the 30m output, since this output is the most trustworthy. In order to test this hypotesis, we first tested two networks only with S2 data. One of the networks only had the output at 10m and the other one had the two outputs.

In terms of data fusion, we investigated four different deep learning architectures: two early fusion methods, one middle fusion, and one late fusion. Early fusion refers to the combination of the data before the encoder feature extraction of the network, while middle fusion happens before the decoder path of the network, and late fusion, before the fully-convolutional classification layer. Figure 2 shows a scheme with all four architectures. The main difference between the two tested early fusion methods is the addition of a convolution layer called feature adaptation before the encoder. The function of this layer is to prevent a raw combination of the inputs since they have different natures and statistical distributions.

The final classification was performed following the hierarchical structure of the LULC reference map. First, a classifier was built to segment the data into the 3 classes of the first level (i.e. anthropic, water and natural areas). The data fusion architectures were explored in this first level of classification. After the decision on the best data fusion method, another classifier with the same structure was trained to separate the natural areas in the other 5 categories of the second level of the LULC. At test time, a final classification map is produced by combining the outputs of each networks. All pixels classified as anthropic and water from the first network is directly imported to the final map, and the pixels with natural class receive the corresponding classification from the second network. The final map has 7 classes. For comparison, a non-hierarchical classifier with the 7 classes was also trained and evaluated, using the same data fusion network structure. All tests used the 10m bands of S2 and the intensity in both polarizations of S1 as input.

Table 3. Test set accuracies of the final classification

Level-2		Non-hierarchical	Hierarchical
Accuracy	OA	65.45%	**68.03%**
	Anthropic	86.41%	**88.06%**
	Water	60.93%	**71.91%**
	Nat. non-veg.	0.00%	0.00%
	Grassland	53.92%	**66.97%**
	Forest	**75.61%**	59.11%
	Savannah	53.24%	**60.25%**
	Sec. veg.	0.06%	**15.33%**
IoU	mIoU	35.75%	**38.18%**
	Anthropic	62.38%	**72.96%**
	Water	**55.86%**	48.14%
	Nat. non-veg.	0.00%	0.00%
	Grassland	42.42%	**47.56%**
	Forest	**46.92%**	42.81%
	Savannah	42.58%	**44.64%**
	Sec. veg.	0.06%	**11.16%**

4. RESULTS

Table 1 compares the results obtained when using only the 10m output and the two outputs. We can observe that there is an improvement of 4.54% on the overall accuracy (OA), and improvements on both anthropic and natural classes' IoU and F1-score. Despite these improvements, the network had a considerable drop in the detection of the water class. This drop little affected the OA because water is a minority class in the dataset, but it highly affected the mean intersection over union (mIoU) and mean F1-score.

Table 2 compare the different architectures for the fusion of S1 and S2. All architectures have 2 outputs, motivated by the results from the previous test. The early fusion architectures present better results than the other models. This behaviour happens because the combination in earlier stages makes the network learn better features with the data, than in each individual path. In terms of the OA, the amount of the improvement is just marginal, which could indicate that the SAR information is just marginally contributing to the network. If we observe the mIoU and mean F-1 score, the addition of the SAR information had a high contribution in differentiating the classes. The early fusion with the feature adaptation module presented the best results. Figure 3 shows the result in a small area for the level-1 classification.

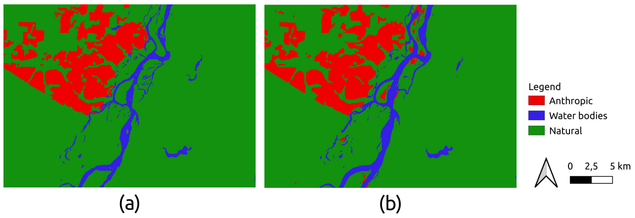

Fig. 3. Example of results obtained with the early fusion 2 architecture for the level-1 of the classification. (a) is the reference and (b) is the model prediction

The final hierarchical classification results are presented in table 3. The table also shows the non-hierarchical classification results. The hierarchical approach improved the accuracy in 2.58% and reduced the confusion for several classes that happened in the non-hierarchical classification. Neither models were able to identify the natural non-vegetated class, because of the high imbalance of this specific class. It represents less than 0.012% of the pixels in the dataset.

5. CONCLUSION

The results show that the location in which the fusion is done has a high impact on the results. It can bring more confusion to the classification instead of improving it. The addition of the second output at 30m and the combined loss function impacts the classification in the proposed dataset, but it only improved the learning of the majority classes. On one side, the hierarchical classification reduces the complexity of the classifier decision, but the final accuracy accumulates errors from previous levels. There is still yet room for improvement by exploring more bands of the S2 data, trying different combinations of the two output individual losses, and using different network structures that can better deal with SAR data.

6. REFERENCES

[1] F. de S. Mendes, D. Baron, G. Gerold, V. Liesenberg, and S. Erasmi, "Optical and SAR remote sensing synergism for mapping vegetation types in the endangered cerrado/amazon ecotone of Nova Mutum−Mato Grosso," *Remote Sensing*, vol. 11, no. 10, pp. 1161, 2019.

[2] F. Borghetti, E. Barbosa, L. Ribeiro, J. F. Ribeiro, and B. M. T. Walter, "South american savannas," *Savanna woody plants and large herbivores*, pp. 77–122, 2019.

[3] L. M. G. Fonseca, T. S. Körting, H. N. Bendini, C. Di G. Neto, A. K. Neves, A. R. Soares, E. C. Taquary, and R. V. Maretto, "Pattern recognition and remote sensing techniques applied to land use and land cover mapping in the brazilian savannah," *Pattern recognition letters*, vol. 148, pp. 54–60, 2021.

[4] L. F. F. G. Assis, K. R. Ferreira, L. Vinhas, L. Maurano, C. Almeida, A. Carvalho, J. Rodrigues, A. Maciel, and C. Camargo, "Terrabrasilis: a spatial data analytics infrastructure for large-scale thematic mapping," *ISPRS International Journal of Geo-Information*, vol. 8, no. 11, pp. 513, 2019.

[5] A. Alencar, J. Z. Shimbo, F. Lenti, C. B. Marques, B. Zimbres, M. Rosa, V. Arruda, I. Castro, J. P. F. M. Ribeiro, V. Varela, et al., "Mapping three decades of changes in the brazilian savanna native vegetation using landsat data processed in the google earth engine platform," *Remote Sensing*, vol. 12, no. 6, pp. 924, 2020.

[6] H. N. Bendini, L. M. G. Fonseca, M. Schwieder, P. Rufin, T. S. Korting, A. Koumrouyan, and P. Hostert, "Combining environmental and landsat analysis ready data for vegetation mapping: A case study in the brazilian savanna biome.," *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 43, 2020.

[7] A. K. Neves, T. S. Körting, L. M. G. Fonseca, A. R. Soares, C. Di G. Neto, and C. Heipke, "Hierarchical mapping of brazilian savanna (cerrado) physiognomies based on deep learning," *Journal of Applied Remote Sensing*, vol. 15, no. 4, pp. 044504, 2021.

[8] E. Fundisi, S. G. Tesfamichael, and F. Ahmed, "A combination of sentinel-1 radar and sentinel-2 multispectral data improves classification of morphologically similar savanna woody plants," *European Journal of Remote Sensing*, vol. 55, no. 1, pp. 372–387, 2022.

[9] F. F. Camargo, E. E. Sano, C. M. Almeida, J. C. Mura, and T. Almeida, "A comparative assessment of machine-learning techniques for land use and land cover classification of the brazilian tropical savanna using alos-2/palsar-2 polarimetric images," *Remote Sensing*, vol. 11, no. 13, pp. 1600, 2019.

[10] K. Lewis, Fernanda de V. Barros, M. B. Cure, C. A. Davies, M. N. Furtado, T. C. Hill, M. Hirota, D. L. Martins, G. G. Mazzochini, E. T. A. Mitchard, et al., "Mapping native and non-native vegetation in the brazilian cerrado using freely available satellite products," *Scientific reports*, vol. 12, no. 1, pp. 1–17, 2022.