COMPARING PHENOMETRICS EXTRACTED FROM DENSE LANDSAT-LIKE IMAGE TIME SERIES FOR CROP CLASSIFICATION

Hugo do Nascimento Bendini[1], *Leila Maria Garcia Fonseca*[1], *Marcel Schwieder*[2], *Thales Sehn Körting*[1], *Philippe Rufin* [2,3], *Ieda Del Arco Sanches*[1], *Pedro J. Leitão* [2,4], *Patrick Hostert* [2,3]

[1]National Institute for Space Research (INPE, Brazil)
[2] Geography Department, Humboldt-Universität zu Berlin
[3] Integrative Research Institute on Transformations of Human-Environment Systems (IRI THESys)
[4] Department Landscape Ecology and Environmental Systems Analysis, Technische Universität Braunschweig

ABSTRACT

In this research, we compared two different sets of land surface phenological metrics (phenometrics) derived from dense satellite image time series to classify agricultural land in the Cerrado biome. We derived phenometrics from a dense Enhanced Vegetation Index (EVI) data cube with an 8-day temporal resolution and subjected them to classification using the Random Forest (RF) algorithm. We used a hierarchical classification with four levels, from land cover to crop rotation classes. We then evaluated the classification results comparing the use of phenometrics extracted using TIMESAT software [1], those obtained by polar representation, proposed by Körting et al. (2013) and the combination of both. We concluded that the accuracies of semi-perennial and winter crop classes increase substantially when using TIMESAT metrics combined with Polar features, and the misclassifications between single crops with non commercial crops are reduced.

***Index Terms*—** Big data, random forest, land use and land cover mapping, multi-sensor

1. INTRODUCTION

Brazils position as one of the most relevant agribusiness frontiers in the world, as well as having large areas of native vegetation, providing environmental services with global impact, reinforces the need to ensure sustainable agricultural development. Accurately mapping methods of the distribution of agricultural areas, and its evolution over time, are therefore essential. TerraClass is the official LULC mapping program in Brazil, starting from Legal Amazon deforested areas [2] moving all the way to the Cerrado [3]. Recently, MapBiomas has proposed to carry out the annual automatic mapping of all biomes of the country in a Landsat-like resolution. However, in addition to the limitations of each initiative [4], agriculture was mapped with little thematic detail. Remotely sensed image time series are valuable for agricultural mapping, as they allow for the monitoring of the highly dynamic crops development. Several studies highlighted the benefits of time series for agricultural mapping in Brazil [5]. The use of crop phenological parameters extracted from image time series can be an important strategy for the development of agriculture mapping methods. Different phenological metrics were explored for this purpose, mainly using MODIS time series [5, 6] and recently also with Landsat-like images [7, 8]. Nonetheless, the full potential of these metrics derived from Landsat-like images has not yet been fully explored, considering the range of possible metrics. TIMESAT software [1] is already a well known tool to generate phenometrics and it is valuable for crop mapping. However, some studies showed limitations regarding the detection of seasons within a specified period, as it depends on smoothing methods and regular time series. Recently, Körting et al. (2013) proposed phenometrics obtained from time series in polar representations, which are based on fixed intervals of time, defined by the quadrants [9]. These metrics were already applied in crop mapping [9], but there is still no analysis of its benefits for this task. Furthermore, in Brazil there is a lack of the development of methods capable of classifying agriculture, semi-automatically, in Landsat scale and with such level of detail. The main objective of this research is to compare different sets of phenometrics derived from dense satellite image time series to classify agricultural land in the Cerrado biome.

2. METHODOLOGY

For the development of this study we carried on a field experiment applying the method in two areas within the Cerrado (Figure 4), in the west of Bahia State (A) and southeastern of Mato Grosso (B). The first area features mostly single-cropping regimes of soybean, but there are also double cropping systems of soy and cotton, and minor winter crops in irrigated areas. It is considered to be the newest agriculture

Thanks to the World Bank for funding through the FIP (Forest Investment Program). This study was financed in part by the Coordenao de Aperfeioamento de Pessoal de Nvel Superior (CAPES) Finance Code 001.

frontier in Brazil. The study area in southeastern Mato Grosso is characterized by intensive double-cropping rotations of soy / maize and soy / cotton. We collected training points based on the representativeness of cropping systems of the Cerrado. We conducted the survey during the growing peak and collected information from farmers in the study area. The team also collected training points from other sites in the Cerrado biome. For the thematic class definition we used a hierarchical approach with 4 levels: the cropland classes (Level 1 - L1), consisting of a crop class and a non-cropped class. Both are divided in a second level (Land cover class - Level 2), where the crop class is divided into another level, grouping annual crops and semi-perennial crops. The non-cropped class consists of natural vegetation classes, divided into three main Cerrado physiognomies: forest, savanna and natural grasslands [10]. And also Perennial crops, Planted Forest and Pasture. From the annual crop class a Crop Group (Level 3 - L3) is defined by the main agricultural practices in the Cerrado region. A Crop Rotation class (Level 4 - L4) is the most detailed level of thematic detail and it consists on crop rotation types definitions. We used the TerraClass maps and photo interpretation of Google Earth imagery to collect additional samples for the non-crop classes. Field boundaries were digitized over Google Earth imagery to obtain a ground polygon database. Finally, 841 polygons were generated, where randomly pixels were sampled from each polygon.

2.1. Remotely Sensed Image Time Series

We used all available ETM+ and OLI data for all the sites (Path/Row 226/070, 225/070, 226/071, 225/071, 220/068, 220/069, 220/070, 219/069 and 219/075), acquired between April 2013 and April 2017. Assuming an 8-day temporal resolution, this 4-year period contains 186 potential observations. The images were obtained from the Center Science Processing Architecture (ESPA) of the US Geological Survey (USGS). These data are provided with level 1 geometric correction (L1TP). We used the Enhanced Vegetation Index (EVI). The limiting factors of a dense time series are sensor errors and cloud cover. To overcome these constraints, Schwieder et al. (2016) used a weighted ensemble of Radial (Gaussian) Basis Function (RBF) convolution filters to approximate the missing data in a Landsat time series. To approximate the given EVI observations into dense 8-day time series without data gaps, we used the RBF approach [11] with some adaptations.

2.2. Phenometrics

We obtained the phenological parameters using TIMESAT V3.2 software [1], where seasonal data are extracted from the time series for each growing season of the focal year (between August 4, 2015, to October 1, 2016). We fitted the time series using the Savitzky-Golay filter [12, 1]. A set of 13 phenometrics were derived for each season. Parameters included day-of-the-year (DOY) of start, mid, end, and length of season and phenological proxies like peak and base value, seasonal amplitude or rate of increase and decrease. Detailed information on the calculation of TIMESAT parameters are found in Jönsson and Eklundh [1]. Besides the TIMESAT phenometrics, we also used the phenometrics proposed by Körting et al. (2013) [9], which are also called polar features, since the purpose is to represent the time series by projecting the values onto angles in the interval [0,2π]. Let a cycle be the function f(x)=(x,y,T), where (x,y) is the spatial position of a point, and T is a time interval $t_1, ...t_N$, and N is the number of observations in such a cycle. The cycle can be visualized as a set of values $v_i \in V$, where v_i is a possible value of $f(x, y)$ in time t_i. Let its polar representation be defined by the function $g(V) \longrightarrow \{A, O\}$ (A corresponds to the abscissa axis in the Cartesian coordinates, and O to the ordinate axis). where: $a_1 = v_1 cos\frac{2\pi}{N} \in A, i = 1, ...N$ and $o_1 = v_1 sin\frac{2\pi}{N} \in O, i = 1, ...N$. Considering a_N+1=a_1 and o_N+1=o_1, we can obtain the coordinates of a closed shape. We then calculate the area of the resulting shape for each of the quadrants $[0, \frac{\pi}{2}]$, $[\frac{\pi}{2}, \pi]$, $[\pi, \frac{3\pi}{2}]$ and $[\frac{3\pi}{2}, 2\pi]$, and are supposed to represent the seasons.

2.3. Random Forest Classification

After the feature extraction, we used our complete field database to train RF [13] and obtained a classifier for each level, considering 3 set of phenometrics: 1) TIMESAT phenometrics (TIMESAT); 2) Polar phenometrics (Polar) and 3) TIMESAT combined with Polar phenometrics (TIMESAT+Polar). The models were trained with a set of 40,385 samples. We applied a hierarchical classification approach by which L1 classes domains are isolated, and land cover is classified by correspondence at each domain for the subsequent nomenclature levels. RF is a classification technique in which the dataset is randomly divided into smaller subsets, and a decision tree is built from each subset. RF needs two parameters to be tuned including the number of trees (ntree), and the number of variables. The ntree parameter values of each RF classification model (L1, 2, 3 and 4) were respectively 50, 50, 70 and 90.

3. RESULTS AND DISCUSSIONS

The validation was done using exhaustive method based on the Monte Carlo simulation [14], where 1000 simulations were carried out by randomly selecting 70% of the samples to train a RF classification model for each hierarchical level. The remaining 30% were used for validation. For each subdivision, a confusion matrix was calculated, and the average confusion matrix was used to derive the overall accuracy and the class f1-scores. Table 1 shows the overall accuracy for each model and each set of phenometrics.

Table 1. Overall accuracy for each model and each set of phenometrics.

	Timesat	Polar	Timesat+Polar
L1	0.975	0.955	0.985
L2	0.972	0.971	0.995
L2 - non crop	0.955	0.935	0.968
L3	0.972	0.937	0.977
L4	0.936	0.871	0.956

A small increase in the overall accuracy when using Polar features can be observed. Neves et al. (2016) also found similar results combining basic features extracted from time series, with Polar features [15]. The L2 classification showed the higher increase. Figure 1 illustrates the increase in the class f1-score when using the combination of TIMESAT and the Polar phenometrics.

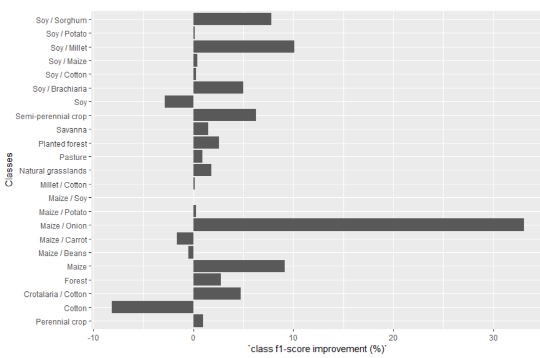

Fig. 1. Class f1-score improvement with the addition of the polar phenometrics.

There is a high increase in the "Maize / Onion" class when using the Polar features. By analyzing the confusion matrices, we observe that this class presented confusion with "Maize" class when using only TIMESAT metrics. An onion crop has a weak EVI response due to its canopy which exhibits an erectophile leaf angle distribution. So that, TIMESAT was not effective on the detection of this variation to capture the crop season. We selected phenometrics based on the Mean Decrease in Gini for all the models and present the boxplots in Figure 2.

Looking to Figure 3, where the boxplots corresponding to the class "Maize / Onion" shows a high range of distribution of the variable "Left Derivative of Season 2", the mean is zero, which means that for the most part of the samples of this class, TIMESAT did not detect the second season, We can also observe both inclusion and omission errors between the classes "Soy / Sorghum", "Soy / Maize", "Soy / Millet" and "Soy / Brachiaria" when using only TIMESAT metrics. This happened because the spectral and temporal response of grasses like "Maize" is similar; this was observed by other authors [5]. When using Polar features, we noticed a reduction on the confusion between "Soy" and "Maize", however, it increases misclassification with the other classes in general. Around 3% of the annual crop samples were wrongly classified being included as semi-perennial crop when using only TIMESAT features, so we can see a great increase on the accuracy of Semi-perennial crop class with the use of the Polar features. The 3D-scatterplot shows the 3 most important variables on the classification of Level 2 (Figure 3).

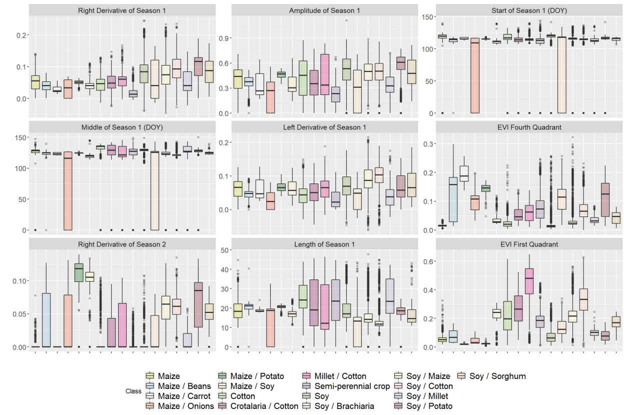

Fig. 2. Boxplots of selected phenometrics.

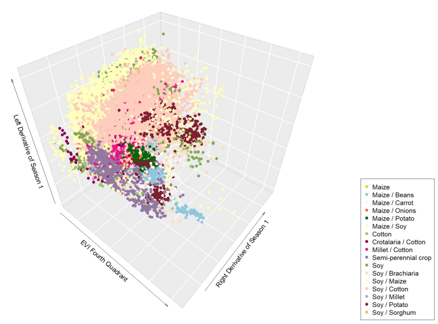

Fig. 3. 3D-scatterplot of important variables on the classification of Level 2.

Semi-perennial crops showed lower values of Left and Right derivatives and higher values of EVI in the fourth quadrant. The combination of these variables was important for separating semi-perennial crops from other crops. Finally, Figure 4 shows the maps of the study area, classified on L2 (for the non-cropped classes) and L4, using only TIMESAT phenometrics, in the right, and when combining to the Polar phenometrics, on the left.

When using the Polar features, there is much less salt and pepper noise, especially for the Semi-perennial class. We can also observe the inclusion errors of Semi-perennial class in the non-cropped class and other cropped classes.

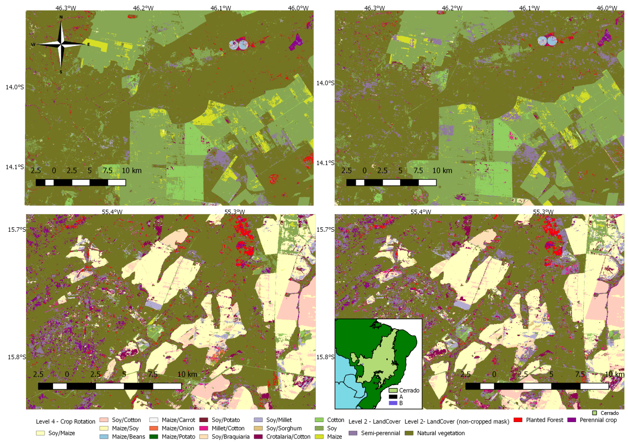

Fig. 4. Maps of the study areas (L2 non-cropped classes and L4), using only TIMESAT phenometrics in the right, and when combining to the Polar phenometrics, in the left.

4. FINAL CONSIDERATIONS

The main goal of this research was to compare different sets of phenometrics derived from dense satellite image time series to classify agricultural land in the Cerrado biome. Our tests showed that, when using TIMESAT metrics combined with Polar features, the accuracies of semi-perennial classes, winter crops increase substantially and misclassifications between single crops with non commercial crops are reduced.

5. REFERENCES

[1] P. Jönsson and L. Eklundh, "TIMESAT - a program for analyzing time-series of satellite sensor data," *Computers & Geosciences*, vol. 30, no. 8, pp. 833–845, 2004.

[2] C. A. Almeida, A. C. Coutinho, J. C. D. M. Esquerdo, M. Adami, A. Venturieri, C. G. Diniz, N. Dessay, L. Durieux, and A. R. Gomes, "High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data," *Acta Amazonica*, vol. 46, no. 3, pp. 291–302, 2016.

[3] "INPE: Mapeamento de uso e cobertura vegetal do cerrado," http://www.dpi.inpe.br/tccerrado/, Accessed: 2017-03-01.

[4] A.K. Neves, T.S. Körting, L.M. Fonseca, G.R. Queiroz, L. Vinhas, K.R. Ferreira, and M.I.S. Escada, "Terraclass x Mapbiomas: Comparative assessment of legend and mapping agreement analysis," in *Proceedings of XVIII GEOINFO*, 2017, pp. 295–300.

[5] D. Arvor, M. Jonathan, M. S. Penello Meirelles, V. Dubreuil, and L. Durieux, "Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil," *International Journal of Remote Sensing*, vol. 32, no. 22, pp. 7847–7871, 2011.

[6] J. C. Brown, J. H. Kastens, A. C. Coutinho, D. C. Victoria, and C. R. Bishop, "Classifying multiyear agricultural land use data from mato grosso using time–series MODIS vegetation index data," *Remote Sensing of Environment*, vol. 130, pp. 39–50, 2013.

[7] Z. Pan, J. Huang, L. Zhou, Q.; Wang, Y. Cheng, H. Zhang, G. A. Blackburn, J. Yan, and J. Liu, "Mapping crop phenology using NDVI time–series derived from HJ-1 A/B data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 34, pp. 188–197, 2015.

[8] H. N. Bendini, L. M. G. Fonseca, M. Schwieder, T. S. Körting, P. Rufin, I. D. A. Sanches, Leit ao, and P. Hostert, "Detailed Agricultural Land Classification in the Brazilian Cerrado based on Phenological Information from Dense Satellite Image Iime Series," *International Journal of Applied Earth Observation and Geoinformation (in press)*, 2019.

[9] T. S. Körting, L. M. G. Fonseca, and G. Câmara, "GeoDMA-Geographic data mining analyst," *Computers & Geosciences*, vol. 57, pp. 133–145, 2013.

[10] J. F. Ribeiro and B. M. T. Walter, "Fitofisionomias do bioma cerrado.," *Embrapa Cerrados-Capítulo em livro científico (ALICE)*, 1998.

[11] M.. Schwieder, P. J. Leito, M. M. C. Bustamante, L. G. Ferreira, A. Rabea, and P. Hostert, "Mapping Brazilian savanna vegetation gradients with Landsat time series," *International Journal of Applied Earth Observation and Geoinformation*, vol. 52, pp. 361–370, 2016.

[12] X. Zhang, M. A. Friedl, C. B. Schaaf, A. H. Strahler, J. C. F. Hodges, F. Gao, B. C. Reed, and A. Huete, "Monitoring vegetation phenology using MODIS," *Remote sensing of environment*, vol. 84, no. 3, pp. 471–475, 2003.

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

[14] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo method*, vol. 10, John Wiley & Sons, 2016.

[15] A. K. Neves, H. N. Bendini, T. S. Korting, and L. M. G. Fonseca, "Combining time series features and data mining to detect land cover patterns: a case study in northern Mato Grosso state, Brazil," *Revista Brasileira de Cartografia*, vol. 68, no. 6, 2016.