SENTINEL-2 MULTIDIMENSIONAL DATA CUBES FOR CROP MONITORING TIME SERIES CLASSIFICATION

Priscilla Azevedo dos Santos[1], *Maria Antônia Falcão de Oliveira*[1], *Thales Sehn Korting*[1], *Marcos Adami*[1] *and Ieda Del'Arco Sanches*[1]

[1]Postgraduation Program in Remote Sensing - PGSER, Division of Earth Observation and Geoinformation - DIOTG, National Institute for Space Research - INPE, Av. dos Astronautas, 1758 - Jardim da Granja, São José dos Campos - SP, Brazil
{priscilla.santos; maria.oliveira; thales.korting; marcos.adami; ieda.sanches}@inpe.br

ABSTRACT

Geoprocessing and remote sensing play an important role when it comes to monitoring land use and land cover using large volumes of data (Big data). In this context, Satellite Time Series Image (Data Cubes) emerge as an alternative to manage Big data mining and classification. Combining information and describing data using time series analysis methods, like Time-Weighted Dynamic Time Warping (TWDTW), for pattern recognition and classification in diverse areas, becomes possible to observe and understand land use and land cover changes as agricultural expansion and crop monitoring. Thus, this work aims to classify crops dynamics in the western portion of Bahia - Brazil, using machine learning and data cubes. Our results showed consistency and feasibility in mapping agricultural targets on a monthly base, with a reasonable classification accuracy over 70% for the produced maps.

Keywords – *Satellite time series, MSI, crop maps, land use, monitoring.*

1. INTRODUCTION

Monitoring agricultural crops through remote sensing is challenging task, since many crops have similar spectral characteristics, hindering to recognize and separating agricultural uses [1, 2]. In this way, time series data analysis tools emerge as an alternative that allows distinguishing agricultural crops types by evaluating, for example, phenological metrics extracted from temporal data and associating them with intrinsic crops characteristics such as leaf geometry and texture, also verifying their spatial distribution, spectral behavior, planting and harvesting time, among others [2–4].

Remote sensing images data cubes allow the analyst to classify land use and land cover. This procedure can be performed by analyzing and distinguishing agricultural crops characteristics using first by the **temporal approach** (scale daily, monthly, annual or historical time series) and, secondly, **spatial** (geographical distribution) and **spectral** (pixel-pixel multiband analysis) approaches [5–7].

This study purpose is to produce crops maps (land use level) through the use of Sentinel-2A/B image data cubes and agricultural database (field references). It brings a classifying approach of multidimensional data cubes using supervised machine learning algorithms to obtain maps in monthly level of the main crops existing in the study area.

2. MATERIAL AND METHODS

2.1. Study Area

The study area corresponds to LEM+ dataset [8] in a region of interest (ROI) represented by the interval coordinates: $46°23'53''$W – $45°29'39''$W and $14°0'33''$S – $11°45'52''$S, in the western portion of the Bahia state, Brazil. The dataset includes monthly land use information about 1854 fields from October 2019 to September 2020 in the area. The majority of the 16 land uses classes are crops, as shown in Figure 1 [9], distributed along Luís Eduardo Magalhães (LEM) and other Bahia state municipalities.

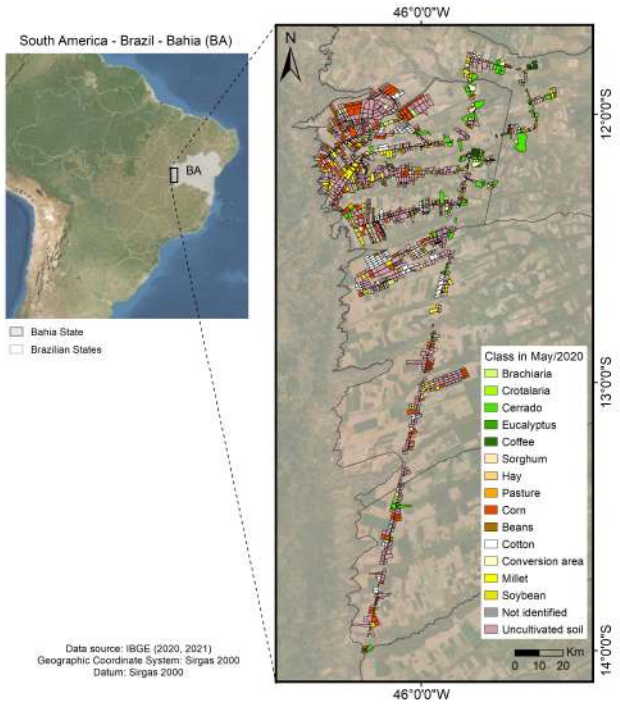

Figure 1: Study area location map: field's data collection (class distribution for May/2020) [9].

2.2. Methodology

The Analysis-Ready Data (ARD) cubes and image collections availability were verified through the Brazil Data Cube (BDC) SpatioTemporal Asset Catalog (STAC) service. From the BDC catalog we retrieved: Sentinel-2A/B MSI collection ARD data cubes (identified as "*S2-SEN2COR_10_16D_STK-1*") related to the 16-day stack (a function identifies the most appropriate pixel in a 16-day interval) [5] with 10m spatial resolution and obtained according to a bounding box (bbox)

defined around the study area samples. Point values were extracted from each created cube and, therefore, randomly separated in portions of training (70%) and test (30%), considering a homogeneous distribution in landscape.

In an attempt to improve the train samples quality and remove possible noises, Self-Organizing Maps (SOM) [7] were created and validated (original samples versus new samples). Supervised classification was applied, where three machine learning algorithms were tested, based on decision trees and kernel trick (Random Forest - RF, Support Vector Machines - SVM and eXtreme Gradient Boosting Machine - XGBoost or XGB).

Within the best model estimated, Probability cubes (Pc) were build and, then, smoothed using smoothness assumption (bayesian smoothing, reduction of “salt-and-pepper" effect). Pc's were classified by each cube (based on tiles) and then a mosaic from the output images were created. A labelled image mosaic was built as output for each monthly cube.

Finally, produced crop monitoring classified mosaics maps (12 in total) were validated using the previously separated test samples and statistical metrics were extracted to evaluate their accuracy (Overall accuracy - OA, kappa - K, Acurracy intervals related to classification confidence levels) [10]. All processing procedures were developed in RStudio script [11]. Figure 2 shows the processing flow adopted in the present study.

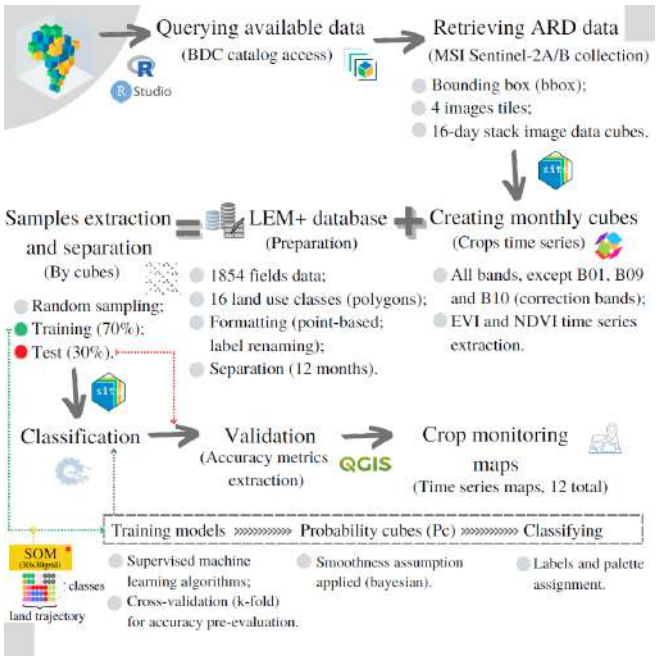

Figure 2: Methodology flowchart.

3. RESULTS

3.1. LEM+ dataset preparation and defining monthly regular image data cubes

Due to the input formats required by sits and BDC, and also in order to work with monthly time series, the data set had to be adjusted before creating the cubes, involving procedures such as: point sample extraction, attributes table columns renaming and tables separation (total of 12), sequentially described. Thus, monthly ARD cubes could be generated having as input the informed tables and Sentinel-2A/B spectral bands information (B2 - Blue, B3 - Green, B4 - Red, B5 to B7 - Vegetation Red-Edge, B8 - Near Infrared, B8A - Vegetation Red Edge, B11 to B12 - Short Infrared and B12) in 10m spatial resolution (provided by BDC ARD Data) [12].

Then, the main dataset was separated into training and testing portions as a pre-classification (pixel-by-pixel) sampling process and then extracted from the cubes. Figure 3 shows the dataset spatial distribution separated into training and testing portions to be used in classification and accuracy verification processes, using a homogeneously random sampling in portion of 70/30%, respectively.

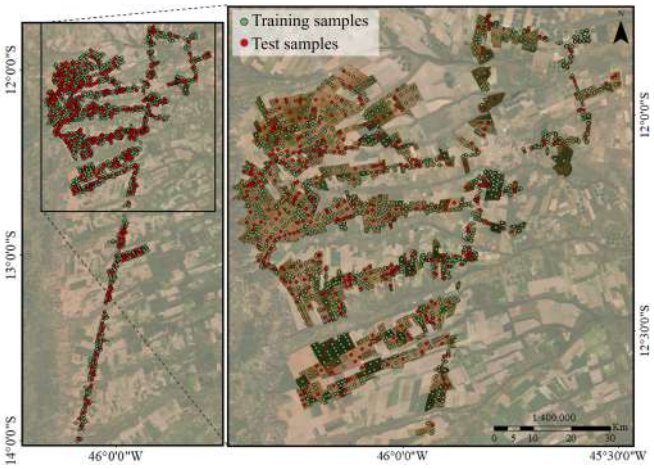

Figure 3: Training and test crop samples spatial distribution.

3.2. Data cubes classification: validation and accuracy measurements

The next step was to classify land use (agricultural use, crop level) by performing a supervised pixel-by-pixel classification with machine learning algorithms. The models (RF, SVM and XGB) were trained using reference samples indicated in Figure 3 (train samples), containing information on land use classes related to the month of analysis. The aforementioned models were implemented following the standard sits function parameters [7] and using an initial randomness (seed settings) equal to 127.

Concerning choosing the best fitted classification algorithm, an k-fold cross-validation (where k=5) was performed and the results can be seen in Table 1. Cross-validation uses part of the available samples to fit the classification model and a different part to test it [13], but is important to acknowledge that this technique should be taken as a measure of training data model‘s performance and not as an overall map accuracy estimation.

Results show that the model which best classified the crops data was with XGBoost in almost all data cubes, followed by RF and, finally, SVM (Table 1). Table 1 also shows kappa values and confidence levels (Lower and Upper) quantified in classification process.

Since the XGB model obtained the best accuracy (higher OA value) compared to the other models (Table 1), expressed

Month	MM	OA	K	A_L	A_U
M_1	XGB	0.8882	0.6819	0.8698	0.9048
M_2	RF	0.8674	0.5703	0.8477	0.8854
M_3	RF	0.5359	0.2806	0.5083	0.5633
M_4	XGB	0.7123	0.5492	0.6867	0.7369
M_5	XGB	0.8170	0.6758	0.7947	0.8379
M_6	XGB	0.5952	0.4796	0.5679	0.6221
M_7	XGB	0.6538	0.5018	0.6272	0.6797
M_8	RF	0.6669	0.5474	0.6405	0.6926
M_9	RF	0.6515	0.5948	0.6249	0.6775
M_{10}	XGB	0.6523	0.5859	0.6256	0.6782
M_{11}	RF	0.6847	0.5975	0.6586	0.7099
M_{12}	XGB	0.7502	0.5300	0.7257	0.7735

OA= Overall Accuracy; K = kappa; A_L = Accuracy Lower; A_U = Accuracy Upper; RF= Random Forest and XGB = Gradient Boosted Machine; M_1 = October/2019; M_2 = November/2019; M_3 = December/2019; M_4 = January/2020; M_5 = February/2020; M_6 = March/2020; M_7 = April/2020; M_8 = May/2020; M_9 = June/2020; M_{10} = July/2020; M_{11} = August/2020 and M_{12} = September/2020.

Table 1: Cross-validation accuracy metrics results for Best Monthly-Models (MM) applied machine learning supervised classifiers. Metrics are associated with each month best results.

mainly by monthly analysis (total number of months), it was used as the classifier in classification process. So, XGB was applied in classification workflow, generating probability cubes (PC).

In post-classification stage, a bayesian smoothing (BS) was applied under the generated probability cubes, in order to use the class probability to estimate if there were a classification error due to a spatial autocorrelation effect existent between a pixel and its neighbors, adjusting the probabilities for the pixel based on our prior beliefs (assumption of correlation between pixels in neighborhood). Also, BS reduces “salt-and-pepper” effect in classified PC's.

Finally, values were assigned to each classes existed in generated probability cubes. Classified images cubes accuracy measures followed the best practices proposed by [10], which uses an area-weighted technique in order “to eliminate bias attributable to map classification error, where the error-adjusted area estimated has confidence intervals to quantify sampling variability in the mentioned area".

Therefore, cross-validation performed was only made to access an accuracy preview because of the biases inherent in that resampling method in training data [7, 14]. So, here we used test samples as an independent validation data set, so that classification can be validated and expressed by model‘s accuracy.

To perform validation and retrieve accuracy metrics, the sits accuracy function was applied. Table 2 shows the post-classification model’s accuracy metrics by validation method, where almost all months reached more than 70% in overall accuracy (OA), except tenth and eleventh months, which reached 68.30% and 68.45% values in OA, respectively.

3.3. Monthly based crop maps

Figure 4 shows the 12 maps produced by performing a machine learning classification, for monitoring monthly based time series of Sentinel-2A/B images cubes. Classified images cubes (labeled Pc) were merged, composing a mosaic with the four classified tiles. These maps produce a sense of stationary type analysis at a monthly level, but serially in the entire evaluated time interval of one agricultural year.

MM	OA	K	A_L	A_U
M_1	0.8792	0.5660	0.8484	0.9058
M_2	0.8619	0.4666	0.8298	0.8900
M_3	0.8608	0.6009	0.8277	0.8896
M_4	0.7458	0.4654	0.7040	0.7845
M_5	0.8060	0.5170	0.7671	0.8410
M_6	0.7122	0.5974	0.6662	0.7552
M_7	0.7923	0.5841	0.7526	0.8282
M_8	0.7763	0.6070	0.7357	0.8134
M_9	0.7091	0.6435	0.6609	0.7540
M_{10}	0.6830	0.5840	0.6354	0.7280
M_{11}	0.6845	0.5609	0.6383	0.7281
M_{12}	0.7670	0.4522	0.7268	0.8039

OA= Overall Accuracy; K = kappa; A_L = Accuracy Lower; A_U = Accuracy Upper; M_1 = October/2019; M_2 = November/2019; M_3 = December/2019; M_4 = January/2020; M_5 = February/2020; M_6 = March/2020; M_7 = April/2020; M_8 = May/2020; M_9 = June/2020; M_{10} = July/2020; M_{11} = August/2020 and M_{12} = September/2020.

Table 2: Validation accuracy metrics results for Gradient Boosted Machine (XGB) Monthly-Model (MM) applied in machine learning supervised classification.

4. DISCUSSION

In the agricultural use analysis, the output pixels that composed the cubes may not be sensitive to the point of faithfully representing the change of use class in the region, especially on the scale of longer interval (1 agricultural year), by accessing only the cube (requiring a careful and sensitive point analysis at land change trajectory throughout all time interval - 1 agricultural year - by using Web Land Trajectory Service - WLTS, not approached here). It's due to the BDC approach that evaluates time first and then space [4, 5].

Therefore, we used monthly time series analysis approach to try to overcome this limitation and better explore crops variability throughout the analyzed agricultural year. Results show that XGB classifier well-performed the representation of agricultural classes, with a general accuracy of more than 70%, presented in the validation (Table 1 and Table 2).

Some classes could not be mapped due to lack of samples in sufficient quantities in the dataset (unbalance). Although it was not addressed in the work, a balancing of the samples was analyzed in the workflow aiming at its applicability in the context of the research and to avoid such mentioned problems. However, as the uncultivated soil class had most of the total amount of dataset samples (more than 53.94% of samples distributed throughout the ROI), undersampling this class and oversampling the other 15 classes to reach a certain limiar could cause problems in the classification as overfitting or underfitting, and also, discrepant representation of the actual data distribution in the study area, respectively.

As a consequence, some classes like brachiaria and beans could not been represented due to the amount of representativa samples (Figure 4-a and 4-c). This is also a problem caused by dataset transformation (polygon to point) to suit sits processing requirement, in which was used the polygon‘s centroid to represent each sample (and class), suppressing essential information.

Seeking to overcome this problem, an approach using: (i) stratified sampling and balancing; (ii) random sampling based

on the proportion of distribution of these points per area of the polygons, also considering what was suggested by [15]; (iii) Latin hypercube sampling (simulation of Monte Carlo method) ; is encouraged and suggested for future work.

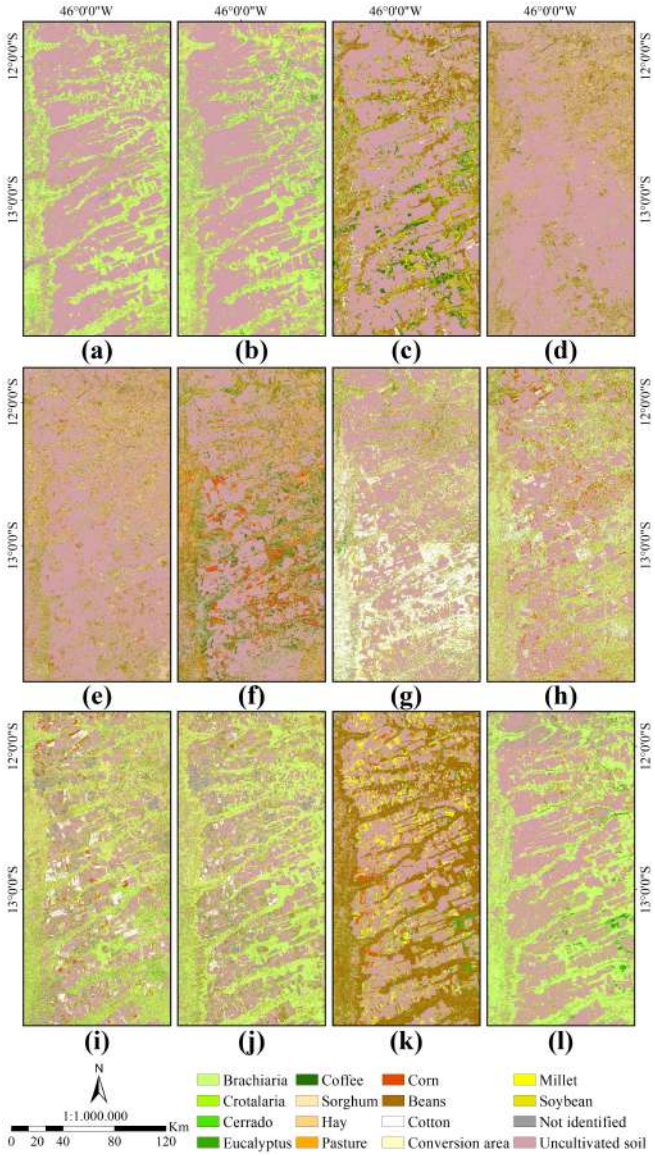

Figure 4: Crop maps based on monthly time series analysis, from October/2019 to September/2020 (a-l).

5. CONCLUSIONS

Accurate land cover and land use change (LULCc) maps are important to subsidize public policies the implementation related to climate change, agriculture dynamics, food security, climate change, land and soil management and environmental impacts prevention. These maps can help farmers in the correct land management and monitoring the development of each crop.

Our classifying approach of multidimensional data cubes using supervised machine learning algorithms and reference data proved to be a viable and consistent preliminary approach for crop mapping allowing the monitoring of agricultural areas in a monthly base, subsidizing several applications.

6. REFERENCES

[1] A. Bégué, D. Arvor, B. Bellon, J. Betbeder, D. De Abelleyra, R. P. D. Ferraz, V. Lebourgeois, C. Lelong, M. Simões, and S. R. Verón. Remote sensing and cropping practices: A review. *Remote Sensing*, 10(1):99, 2018.

[2] M. Weiss, F. Jacob, and G. Duveiller. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, 236:111402, 2020.

[3] M. Adami, B. F. T. Rudorff, R. M. Freitas, D. A. Aguiar, L. M. Sugawara, and M. P. Mello. Remote sensing time series to evaluate direct land use change of recent expanded sugarcane crop in brazil. *Sustainability*, 4(4):574–585, 2012.

[4] M.E. D. Chaves, M. de C. Alves, T. Sáfadi, M. S. de Oliveira, M. C. A. Picoli, R. E. O. Simoes, and G. A. V. Mataveli. Time-weighted dynamic time warping analysis for mapping interannual cropping practices changes in large-scale agro-industrial farms in brazilian cerrado. *Science of Remote Sensing*, 3:100021, 2021.

[5] K. R. Ferreira, G. R. Queiroz, L. Vinhas, R. F. B. Marujo, R. E. O. Simoes, M. C. A. Picoli, G. Camara, R. Cartaxo, V. C. F. Gomes, L. A Santos, et al. Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24):4033, 2020.

[6] M. E. D. Chaves, M. C. A. Picoli, and I. D. A. Sanches. Recent applications of landsat 8/oli and sentinel-2/msi for land use and land cover mapping: A systematic review. *Remote Sensing*, 12(18):3062, 2020.

[7] R. Simoes, G. Camara, G. Queiroz, F. Souza, P. R. Andrade, L. Santos, A. Carvalho, and K. Ferreira. Satellite image time series analysis for big earth observation data. *Remote Sensing*, 13(13):2428, 2021.

[8] L. V. Oldoni, I. D. A. Sanches, M. C. A. Picoli, R. M. Covre, and J. G. Fronza. "lem+ dataset: for agricultural remote sensing applications". *Mendeley Data*, v. 1, 2020. doi: 10.17632/vz6d7tw87f.1. `https://data.mendeley.com/datasets/vz6d7tw87f/1` (Accessed on 08/20/2022).

[9] L. V. Oldoni, I. D. A. Sanches, M. C. A. Picoli, R. M. Covre, and J. G. Fronza. Lem+ dataset: for agricultural remote sensing applications. *Data in Brief*, v. 33:pp. 106553, 2020.

[10] P. Olofsson, G. M. Foody, S. V. Stehman, and C. E. Woodcock. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129:122–131, 2013. doi: 10.1016/j.rse.2012.10.031.

[11] RStudio Team. *RStudio: Integrated Development Environment for R, version 2022.07.0.* RStudio, PBC., Boston, MA, 2020.

[12] Brazil Data Cube. Brazil data cube products - data cube collections: S2-sen2cor_10_16d_stk-1. 2022. `https://brazil-data-cube.github.io/products/cube_col/S2-SEN2COR_10_16D_STK-1.html#`.

[13] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[14] T. S. Wiens, B. C. Dale, M. S. Boyce, and G. P. Kershaw. Three way k-fold cross-validation of resource selection functions. *Ecological Modelling*, 212(3-4):244–255, 2008.

[15] R. G. Congalton and K. Green. *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press, 2019.