Combining Time Series Features and Data Mining to Detect Land Cover patterns: a Case Study in Northern Mato Grosso State, Brazil

Alana K. Neves[1], Hugo do N. Bendini[1], Thales S. Körting[1], Leila M. G. Fonseca[1]

[1]Instituto Nacional de Pesquisas Espaciais – INPE

Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil

alana.neves@inpe.br,{hnbendini,tkorting,leila}@dpi.inpe.br

Abstract. *One product of the MODIS sensor (Moderate Resolution Imaging Spectroradiometer) is the EVI2 (Enhanced Vegetation Index). It generates images of around 23 observations each year, that combined can be interpreted as time series. This work presents the results of using two types of features obtained from EVI2 time series: basic and polar features. Such features were employed in automatic classification for land cover mapping, and we compared the influence of using single pixel versus object-based observations. The features were used to generate classification models using the Random Forest algorithm. Classes of interest included Agricultural Area, Pasture and Forest. Results achieved accuracies up to 91,70% for the northern region of Mato Grosso state, Brazil.*

1. Introduction

Since the 50s, Amazon's occupation was characterized by expanding the agriculture frontier, which resulted in extensive and accelerated transformations. This period was marked by high and continuous deforestation rates, especially in the areas located in the so-called “arc of deforestation” (BECKER, 1990, 2009). Currently, in the Amazon, large areas of pasture, agriculture, reforestation and secondary vegetation can be found and much of the primary forest is limited to protected areas (BECKER, 2009).

Due to its complexity, there is still no complete understanding of the dynamic of landscape evolution in the Amazon region. This is because of the great heterogeneity of land use and occupation since the implementation of old governmental colonization projects and new federal infrastructure projects. To help in understanding the Amazon landscape, INPE (Brazil's National Institute for Space Research), in partnership with EMBRAPA (Brazilian Agricultural Research Corporation), produces land cover data about Legal Amazon, in a project known as TerraClass – mapping of land use and land cover change in legal Amazon deforested areas (COUTINHO et al., 2013). TerraClass presents to the society information related to which are the current main activities (spatially and numerically) in deforested areas in a specific year. TerraClass information is currently available for years 2008, 2010 and 2012.

To achieve the proposed goal, most of the TerraClass interpretation and classification is done visually and manually, which is a very time consuming task. The annual agriculture mapping of TerraClass is based on an automatic method, which used minimum and maximum values of NDVI (Normalized Difference Vegetation Index) time series. In agricultural areas, vegetation indices present low values in the beginning of agricultural cycle and high values in vegetation peaks. The difference between these two moments above a certain limit corresponds to agriculture pattern (ADAMI et al., 2015). Although there are some efforts to automate its methodology, there is still space to study more adequate data and methods to improve automatic classification results.

Since the 70s, acquisition data through remote sensing is a practice of increasingly importance and, more and more, is becoming fundamental in the knowledge of Earth's phenomena. Interpreting these phenomena only by *in situ* observations would require such an amount of resources (human, time and money).

Remote sensors, like MODIS (Moderate Resolution Imaging Spectroradiometer), have been responsible for systematically collect images of Earth, which can be converted into image time series (VUOLO, 2012). MODIS products include vegetation indices, capable of providing spatial and temporal comparisons of global vegetation conditions. The well-known vegetation indices available are the NDVI and the EVI2 (Enhanced Vegetation Index). NDVI is more sensitive to the presence of pigments such as chlorophyll, while EVI2 is related to changes in canopy structure, such as Leaf Area Index (LAI), vegetation type and vegetation physiognomy (HUETE et al., 2002). For this reason, the study of EVI2 time series also allows to obtain information about soil cover.

One of the techniques used to manipulate large amount of observations present in a time series is data mining. Data mining consists of a supporting tool, through the discovering of correlations, patterns and trends in data, combining technologies of pattern recognition, mathematics and statistics (LAROSE, 2014). Such techniques have already been employed in remote sensing, combining data mining techniques and vegetation indices time series.

Costa et al. (2015) used EVI data to differentiate pasture and native grassland in the Brazilian biome named Cerrado, comparing Support Vector Machine, Multilayer Perceptron and Autoencoder algorithms. Others efforts to classify land cover in Amazon include the use of Naïve Bayes, Nearest Neighbor and Optimum Path Forest algorithms (NOMA et al., 2013; BARBOSA ET AL., 2015). Random Forest algorithm is not so usual in remote sensing applications, but it is a powerful machine learning and it as expanding its applicability in land studies by remote sensing (RODRIGUEZ-GALIANO et al., 2012), even using vegetation indices data (NITZE et al., 2015) for image acquisition optimization for land cover classification.

Time series can be related to land patterns using feature extraction (GALFORD et al., 2008). There are several types of features, such as basic and polar features (KÖRTING, 2012), that can be combined to assist in classification models. Thus, this work aims to generate classification models to detect land cover testing different time series features, in a test area in northern Mato Grosso state, Brazil, which belongs to the arc of deforestation.

2. Methodology

In Figure 1 we present a flowchart of the employed methodology, which will be better explained as follows.

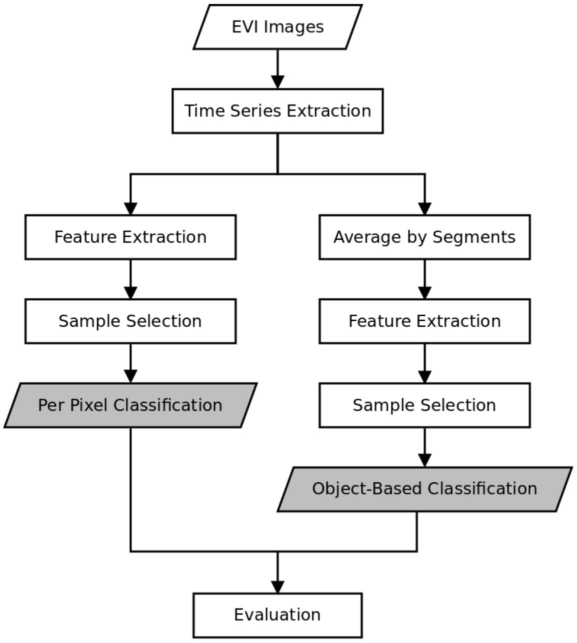

Figure 1. Methodology flowchart

2.1. Study Area

The study area (Figure 2) chosen for this work is the path-row 227-068 of TM sensor Landsat 5 satellite. The scene is located in the northern Mato Grosso (MT), Brazil and covers part of eight municipalities: Juara, Nova Canaã do Norte, Itaúba, Tabaporã, Porto dos Gauchos, Itanhangá, Ipiranga do Norte and Sinop. The scene belongs to the agriculture frontier in the arc of deforestation. MT is one of the three Brazilian states with the largest deforested area in the Amazon (MARGULIS, 2003).

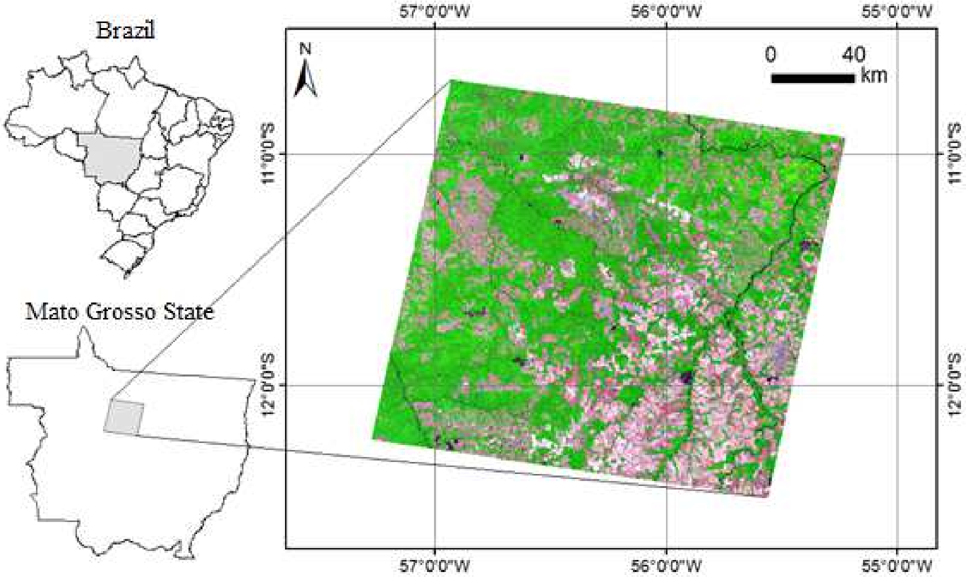

Figure *2*. Study area: TM sensor (Landsat 5) path-row 227-068.

2.2 Time Series Extraction

The data used in our experiments was obtained from the modified 2-band EVI2 (Equation 1) from the product MOD13Q1 of MODIS sensor, with spatial resolution of 250 meters and temporal resolution of 16 days (SOLANO et al., 2010).

$$EVI2 = 2.5\frac{\rho_{NIR}-\rho_{\text{Red}}}{1+\rho_{NIR}+\rho_{\text{Red}}} \qquad (1)$$

where ρ_{NIR} is the Near Infra-red reflectance and ρ_{Red} is the reflectance of the red band.

Because of its temporal resolution, EVI2 generates cycles of around 23 observations each year, that combined can be interpreted as time series. The 46 (23 for 2008 and 23 for 2010) EVI2 images were downloaded from http://earthexplorer.usgs.gov/. For each year, images were ordered by time (Figure 3) and time series were composed for each pixel.

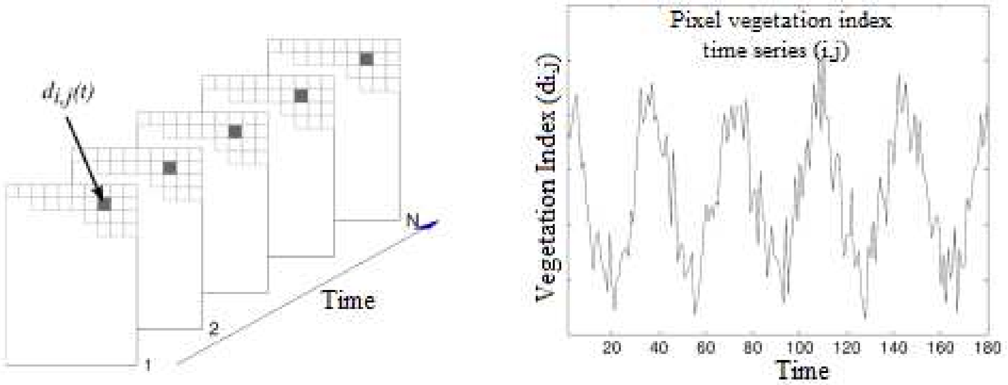

Figure 3. Per-pixel feature extraction and time series composition (Adapted from Eklundh & Jönsson, 2012).

2.3. Feature Extraction

Two approaches were used: per pixel, where each pixel has its respective time series, and object-based. Using objects means that the imagery was partitioned into homogeneous regions, so spatial, spectral and temporal characteristics can be included in the analysis (HAY & CASTILLA, 2006). In this work, objects from TerraClass were used to group pixels with similar behavior and their time series were represented by the average of all-time series of pixels present in each object. Since TerraClass information is available for years 2008 and 2010 (also 2012), we used time series from the years 2008 and 2010 in our analysis.

With temporal resolution of 16 days and spatial resolution of 250 m, EVI2 data from MODIS generates cycles of around 23 observations each year, that combined can be interpreted as time series. Several features can be extracted from each time series. In this work, two groups of features were extracted, according to the methodology proposed by Körting (2012): the so called basic and polar features. Basic features includes statistical measures such as mean, standard deviation, minimum and maximum values of the curve (Figure 4).

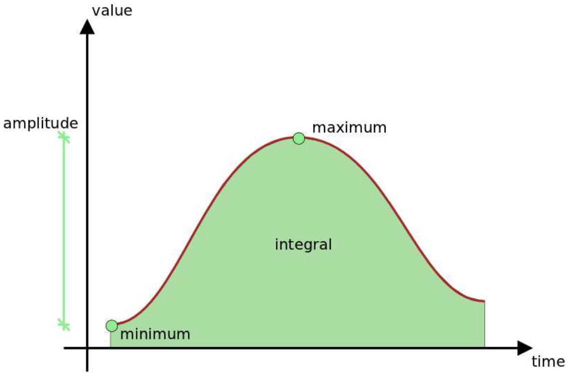

Figure 4. Example of basic features representation.

Many natural phenomena can be represented by cyclical patterns, such as agriculture. Cycles can be characterized by rise and fall oscillations in series. To support the cycles visualization, a way of plotting was proposed, adapted from Edsall et al. (1997), where each cycle value is projected in angles in the interval [0,2π] (Figure 3) (KÖRTING, 2012). This projection generates an object with a closed contour, whose properties can represent some specific behavior of the original time series.

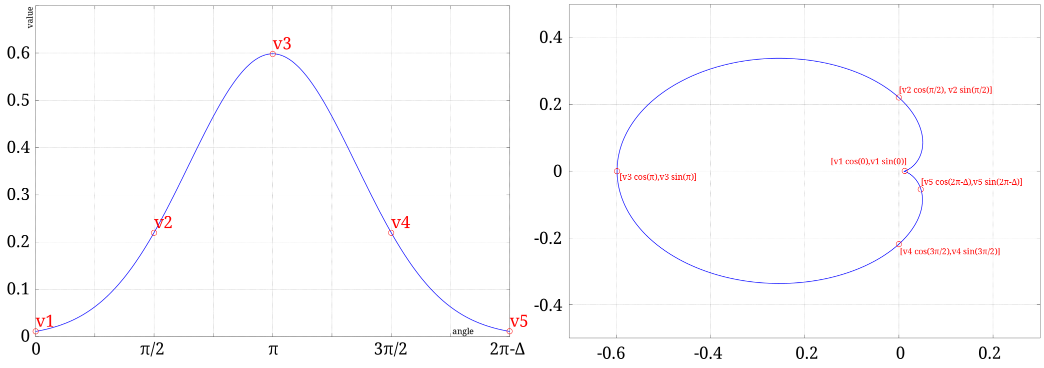

Figure 5. Example of a cycle in a time series. On left, one time series example. On the right, its polar representation according to Körting (2012).

From the polar visualization of these objects, several features can be generated such as eccentricity, angle of orientation, area per season and others. Both basic and polar features are described in the Table below.

Table 1. Description of basic and polar features from time series.

Name	Description	Type	Range
Amplitude	The difference between the cycle's maximum and minimum values. A small amplitude means a stable cycle.	Basic	[0, 1]
Area	Area of the closed shape. A higher value indicates a cycle with high EVI values.	Polar	≥0
Area per Season	Partial area of the closed shape, proportional to a specific quadrant of the polar representation. High value in the summer season can be related to the	Polar	≥0

	phenological development of a cropland.		
Circle	Returns values close to 1 when the shape is more similar to a circle. In the polar visualization, a circle means a constant feature.	Polar	[0, 1]
Cycle’s maximum	Relates the overall productivity and biomass, but it is sensitive to false highs and noise.	Basic	[0, 1]
Cycle’s mean	Average value of the curve along one cycle.	Basic	[0, 1]
Cycle’s minimum	Minimum value of the curve along one cycle.	Basic	[0, 1]
Cycle’s std	Standard deviation of the cycle’s values.	Basic	≥0
Cycle’s sum	When using vegetation indices, the sum of values over a cycle means the annual production of vegetation.	Basic	≥0
Eccentricity	Return values close to 0 if the shape is a circle and 1 if the shape is similar to a line.	Basic	[0,1]
First slope maximum	It indicates when the cycle presents some abrupt change in the curve. The slope between two values relates the fastness of the greening up or the senescence phases.	Basic	[-1, 1]
Gyration radius	Equals the average distance between each point inside the shape and the shape’s centroid. Smaller values stand for shapes similar to a circle.	Polar	≥0
Polar balance	The standard deviation of the areas per season, considering the 4 seasons. Small value point to constant cycles, e.g. the EVI of water (with a small Area), or forest (with a medium Area).	Polar	≥0

2.4. Samples Selection, Classification and Evaluation

After the feature extraction, the automatic classification was made on software WEKA 3.6 (HALL et al., 2009). We used the Random Forest algorithm, which creates a set of decision trees used to classify the full data set. The use of this algorithm in remote sensing applications is relatively new, but it has proven to be powerful in land-cover classification (RODRIGUEZ-GALIANO et al., 2012). The number of decision trees to be used is defined by the domain's expert. In our experiments we defined this parameter empirically, based on the accuracy of results and the time needed (computational cost) to classify all data. Models were built using training samples from the year 2008 and reevaluated in 2010. To evaluate the classification accuracy in 2008, we divided the samples in two subsets. 66% of the data was used for training and 34% was used for testing. Three interest classes were discriminated: Forest, Pasture and Agriculture.

In the Random Forest algorithm, data are partitioned randomly in many subsets by the Bootstrap technique (resampling with replacement), in which some records may appear several times in the same subset while others do not appear even once. Each subset generates a decision tree and all the decision trees have a vote with a certain weight to contribute in the decision about the class that will be assigned to the object (HAN et al., 2011).

We also tested different combination of features:

- Time Series, Basic Features and Polar Features;
- Basic and Polar Features;
- Time Series and Basic Features;
- Time Series and Polar Features;
- Only Time Series;
- Only Basic Features;
- Only Polar Features.

The classification generated by TerraClass is based on the interpretation of Landsat TM scenes, therefore objects from TerraClass were produced at the scale of 30m, differently from our input data from MODIS, whose spatial resolution is 250m. Land cover patterns from TerraClass include Annual Agriculture, Clean Pasture, Dirty Pasture, Forest, Urban Area, Mining, Occupation Mosaic, Regeneration with Pasture, Reforestation, Non Forest, Hydrography and Secondary Vegetation. Since there are different types of pasture, and also other classes which are unable to be recognized in MODIS spatial resolution, it was necessary to made some masking and a reclassification, where "Clean Pasture" and "Dirty Pasture" became a single class named Pasture. At Table 2, the reclassification made in TerraClass data to facilitate the comparison with the automatic classification. Those classes included in "Others" were not analyzed in the automatic classification, therefore their pixels were masked. Then it was considered that the image is composed only by the three targets of interest.

Table 2. Reclassification of TerraClass data for validation

TerraClass	Reclassification
Annual Agriculture	Agriculture
Clean Pasture Dirty Pasture	Pasture
Forest	Forest
Urban Area Mining Occupation Mosaic Regeneration with Pasture Reforestation Non Forest Hydrography Secondary Vegetation	Others

To test the model accuracy, we used evaluation and performance measures. As an evaluation measure of classification, one Error Matrix per year was generated for each approach.

3. Results and Discussion

The behavior of EVI2 time series in the study area can be seen in Figure 6, by pixel and object-based approach. These curves were generated with the mean of all the pixels or object time series for each class. High EVI2 values are observed on the periods between January and April, as well between October and December. This behavior reflects what is expected for the vegetation on this region, according to the annual seasonality, decreasing in greenness through the dry season and increasing during the rainy season, with an annual mean of 0.46 EVI2 and a maximum around 0.6 EVI2. The agriculture system had a more complex behavior, showing peaks next to 0.7 EVI2 between December – January, and March – April, and higher standard deviation (0,2). Similar values were found by Galford et al. (2008) in a study for detect croplands in Mato Grosso using time series wavelet analysis. In the Figure 6a we can see a more constant behavior in forest, around 0.5 EVI2, while in Figure 6b we observe that the forest mean temporal behavior was similar to pasture.

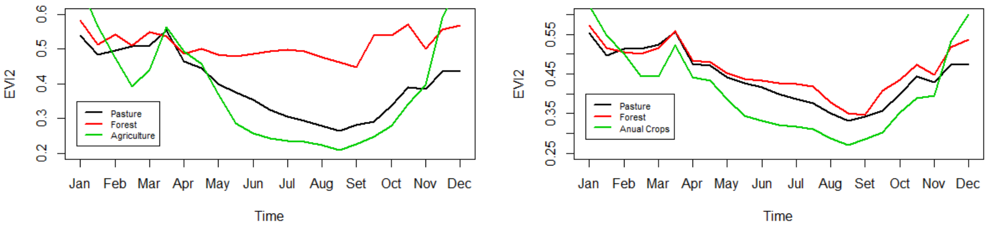

Figure 6. (a) EVI2 time series of both target by pixel based, with 65897, 219069 and 100335 for pasture, forest and agriculture respectively. (b) EVI2 time series of both target by object based approach, using 1685 segments for pasture, 1402 for forest and 492 for agriculture.

In our first experiment, resumed in Table 3, we tuned the Random Forest algorithm, by finding the best number of decision trees to be used. These performances were observed utilizing the full data (time series, polar and basic features) for the year 2008. The accuracy had little increase (less than 1%) while increasing the Number of Trees, and the Time to Build the Model almost doubled when compared to the previous. Thus, such little improvement in classification associated with the higher computational cost do not justify the use of more trees in the model. Therefore, it was chosen the number of 20 decision trees in all models used in the next results.

Table 3. Number of Trees and Performance Comparison

Number of Trees	10	20	50	100
Correctly Classified Instances (%)	90,97	91,3526	91,6056	91,6321
Time to Build Model (seconds)	257,73	482,14	1182,94	1941.07

Similarly to the obtained accuracies from Sato et al. (2013), the algorithm of Random Forest was satisfactory to distinguish patterns of land cover, although these authors have only used one Landsat image with four remote sensing products: MLME (Linear Spectral Mixture Model), NDVI (Normalized Difference Vegetation Index), NDWI (Normalized Water Index) and SAVI (Soil-Adjusted Vegetation Index).

The results of the second experiment are resumed in Table 4. The percentage of correctly classified instances for each approach shows that, in the situation studied in this work, both basic and polar features were efficient in distinguish Agriculture, Pasture and Forest, although using only the time series produced a better result (91,70%). All seven approaches obtained accuracies near 90%. Usually, the hit rate is higher in 2008, because the model was built in this year.

Table 4. Correctly Classified Instances (%) for each approach

	Per Pixel		Object-Based	
	2008	**2010**	**2008**	**2010**
Time Series, Basic Features and Polar Features	91,35	88,39	72,62	56,82
Basic and Polar Features	89,52	87,52	69,72	58,21
Time Series and Basic Features	91,43	88,34	72,62	57,37
Time Series and Polar Features	91,39	88,06	72,56	57,82
Only Time Series	91,70	88,09	72,31	54,22
Only Basic Features	89,38	87,00	69,65	57,14
Only Polar Features	84,84	83,33	64,96	53,77

Another important aspect is the fact that classification per pixel resulted in better accuracies (around 90%) than object based classification (around 60%). In Figure 7, the reference data from TerraClass and our result using automatic classification (only time series) can be compared.

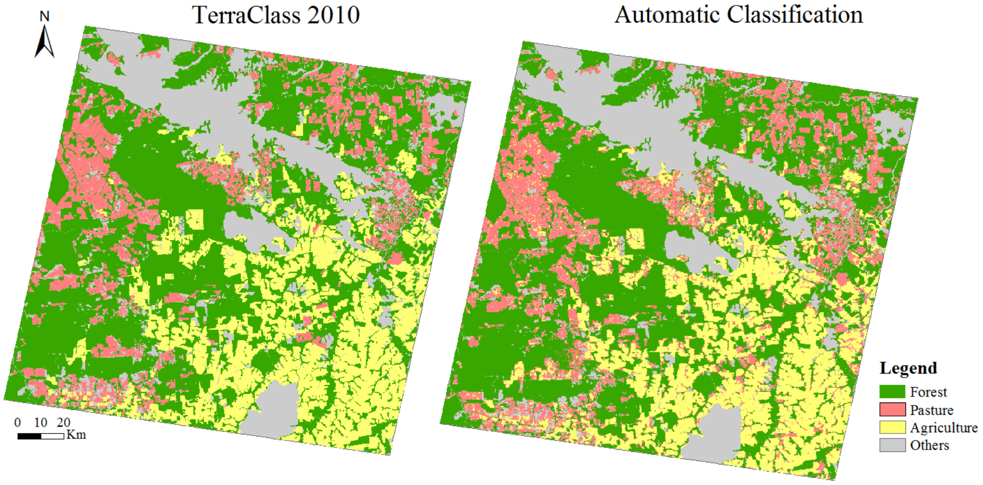

Figure 7. Comparation between Automatic Classification 2010 and its reference, TerraClass 2010. Elements of class 'Others' were not classified.

To have a more specific analysis for each class of interest, at Table 5 each class is represented and values are referring to the average of the seven approaches. As noticed before, in general the classification per pixel had a better performance, however when we observe the class Pasture, it shows that object based approaches increased correctly classified instances, which means that including spatial parameter in analyzes should improve its identification.

Table 5. Correctly classified instances (%) for each interest class.

	Per pixel		Object-based	
	2008	**2010**	**2008**	**2010**
Forest	95.82	92.81	37.65	20.75
Pasture	75.43	74.4	88.14	92.09
Agriculture	85.99	83.52	46.79	36.38

Despite the well-known good performance of object based classifications, in this case Forest and Agriculture had a better identification behavior when pixels time series were analyzed separately. According to Seyler, (2002), Pasture is a difficult class to be identified by only satellite sensor data. Because of the great quantity of mixed elements in its composition, like grass, trees, bush and others, it was harder to characterize it only by its behavior in time series.

4. Conclusions

Both basic and polar features from time series were satisfactory for the identification of the three interest classes. Forest and Agriculture classification had a great performance when using per pixel strategy, while Pasture was better differentiated when the object based approach were used. Random Forest algorithm showed to be robust enough to make a good separation between EVI2 patterns.

Although the automatic classification produced similar results to TerraClass data, it was inappropriate to make comparisons between mapped area for each approach because of the different spatial resolutions.

In future works, it is intended to analyze new interest classes and test results of segmentations that can include temporality of time series.

5. Bibliography

ADAMI, M; GOMES, A. R.; COUTINHO, A. C.; ESQUERDO, J. C. D. M.; VENTURIERI, A. Dinâmica de uso e cobertura da terra no estado do Pará entre os anos de 2008 a 2012. XVII Simpósio Brasileiro de Sensoriamento Remoto, 2015. **Anais...** João Pessoa, PB, 2015.

BARBOSA, D. P.; NOMA, A.; KÖRTING, T. S.; FONSECA, L. M. G. Um estudo experimental com classificadores baseados em regiões e perfis EVI. XVII Simpósio Brasileiro de Sensoriamento Remoto, 2015. **Anais...** João Pessoa, PB, 2015.

BECKER, B. K. Amazônia. **Série Princípios**. São Paulo: Ática, 1990. 92p.

BECKER, B. K. Amazônia: Geopolítica na virada do III milênio. Rio de Janeiro: Garamond, 2009. 172p.

COUTINHO, A. C.; ALMEIDA, C.; VENTURIERI, A.; ESQUERDO, J. C. D. M.; SILVA, M. **Projeto TerraClass: Uso e cobertura da terra nas áreas desflorestadas na Amazônia Legal**. Brasília, DF: Embrapa; Belém: INPE, 2013.

COSTA, W.; FONSECA, L.; KÖRTING, T. Classifying grasslands and cultivated pastures in the Brazilian cerrado using Support Vector Machines, Multilayer Perceptrons and Autoencoders. **Lecture Notes in Computer Science**. 1ed.: Springer International Publishing, 2015, v. 9166, p. 187-198.

EDSALL, R.; KRAAK, M.; MACEACHREN, A.; PEUQUET, D. Assessing the efectiveness of temporal legends in environmental visualization. **Proceedings of GIS/LIS**, Citeseer, p. 677{85, 1997. Available in: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.8955&amp;rep=rep1&amp;type=pdf>.

Eklundh, L., and Jönsson, P., 2012, TIMESAT 3.2 with parallel processing - Software Manual. Lund University, 88 pp.

GALFORD, G. L.; MUSTARD, J. F.; MELILLO, J.; GENDRIN, A.; CERRI, C. C.; CERRI, C. E. P. Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil. **Remote Sensing of Environment**. v. 112, p. 576-587. 2008.

HALL M.; FRANK, E., HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **The WEKA Data Mining Software: An Update**. SIGKDD Explorations, Volume 11, Issue 1. 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

HAY, G. J.; CASTILLA, G. Object-based image analysis: strengths, weaknesses, opportunities and threats (swot). **The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**. OBIA, 2006.

HUETE, A.; DIDAN, K.; MIURA, T.; RODRIGUEZ, E. P.; GAO, X.; FERREIRA, L. G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. **Remote Sensing of Environment**, v. 83, n. 1, p. 195-213, 2002.

KORTING, T. S.; FONSECA, L. M.; ESCADA, M. I. S.; SILVA, F. C.; SILVA, M. P. S. GeoDMA: a novel system for spatial data mining. **IEEE International Conference on Data Mining Workshops, Pisa, Italia**, 2008. **Anais…** Pisa, Italia, 2008.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. John Wiley and Sons, Inc, 2 ed. 2014.

MARGULIS, S. **Causas do Desmatamento da Amazônia Brasileira**. 1ed. Brasília: Banco Mundial, 2003.

Nitze, Ingmar; Barrett, Brian; Cawkwell, Fiona (2015). Temporal optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series. **International Journal of Applied Earth Observation and Geoinformation**, 34, 136-146.

NOMA, A.; KORTING, T. S.; FONSECA, L. M. G. Uma comparação entre classificadores usando regiões e perfis evi para agricultura. XVI Simpósio Brasileiro de Sensoriamento Remoto. **Anais...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2013. p. 2250–2257.

Rodriguez-Galiano, V. F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J.P. An assessment of the effectiveness of a Random Forest classifier for land-cover classification. **ISPRS Journal of Photogrammetry and Remote Sensing**. v.67, p.93-104, 2012.

SATO, L. Y.; SHIMABUKURO, Y. E.; KUPLICH, T. M.; GOMES, V. C. F. Análise comparativa de algoritmos de árvore de decisão do sistema WEKA para classificação de uso e cobertura da terra. XVI Simpósio Brasileiro de Sensoriamento Remoto, 2013. **Anais...** Foz do Iguaçu, PR, 2013.

SOLANO, R.; DIDAN, K.; JACOBSON, A.; HUETE, A. MODIS Vegetation Index User's Guide (MOD13 Series). 2010.

SEYLER, F.; CHAPLOT, V.; MULLER, F.; CERRI, C. E. P.; BERNOUX, M.; BALLESTER, V.; FELLER, C.; CERRI, C. C. C. Pasture mapping by classification of Landsat TM images. Analysis of the spectral behavior of the pasture class in a real medium-scale environment: the case of the Piracicaba Catchment (12 400 km², Brazil). **International Journal of Remote Sensing.** v. 23. n. 23. p. 4985-5004. 2002.

VUOLO, F.; MATTIUZZI, M.; KLISCH, A.; ATZBERGER, C. Data service platform for MODIS Vegetation Indices time series processing at BOKU Vienna: current status and future perspectives . Proc. SPIE 8538, **Earth Resources and Environmental Remote Sensing/GIS Applications III**, 85380A (October 25, 2012); doi:10.1117/12.974857.