# FEATURE EXTRACTION FOR NDVI AVHRR/NOAA TIME SERIES CLASSIFICATION

*W. L. da Silva*[1], *R. R. V. Gonçalves*[2], *A. S. Siqueira*[1], *J. Zullo Jr*[2], *F.A.M. Gomes Neto*[1]

[1]Department of Applied Mathematics
State University of Campinas

[2]Center of Meteorological and Climatic Research Applied to Agriculture
State University of Campinas

## ABSTRACT

One of the biggest problems of agribusiness in Brazil is related to estimation and forecasting of agricultural crops. In this problem, time series classification enters as a way to help production estimation. In this paper, we are concerned with the development of an automatic classifier that identifies the areas covered with the sugarcane culture by using Normalized Difference Vegetation Index (NDVI) time series, from the AVHRR/NOAA data warehouse of Center of Meteorological and Climatic Research Applied to Agriculture (CEPAGRI). We assumed that a multidimensional space generated by information obtained in the harmonics is a appropriate space to study the similarity between time series. Here we used the word *features* of a series to refer the coefficients extracted by time series in Fourier decomposition. The proposed methodology has shown to be efficient with a high success rate for the classification of the culture of sugarcane in images from Jaboticabal city, in Brazil, 2004/2005.

***Index Terms*—** Classification criteria, harmonic analysis, mean feature curve.

## 1. INTRODUCTION

In Brazil, agribusiness turnover in 2009 was about 23% of Gross Domestic Product (GDP), equivalent to R$ 735 billions[1], which gives a sense of dimension and potential of national agriculture. Otherwise, one of the problems faced in Brazil's agribusiness is related to estimation and forecasting of agricultural crops. Without a reliable number, economic agents become oriented by information generated from different sources, which are not always concordant. An efficient crop estimation system is essential to any country that depends directly on agriculture. The usual way to obtain data to estimate crop in São Paulo state (Brazil) consists on field surveys by census method, in which all the sample units are considered, or by sampling, in which only a portion of the sample units are considered. One of the disadvantage of these methods is their expensive and complex structure.

The area classification favor the use of rescheduling land, to avoid overproduction of one product, shortage of others, supply domestic and foreign markets, to promote some products considered essential to the national economy, to estimate the losses from pests, diseases and natural phenomena such as drought and flood, which are common in tropical countries like Brazil.

The main goal is to develop an automatic classifier that identifies the areas covered with the sugar cane culture by using Normalized Difference Vegetation Index (NDVI) time series, from the AVHRR/NOAA data warehouse of Center of Meteorological and Climatic Research Applied to Agriculture (CEPAGRI) [2]. We used that an NDVI annual series is sufficient to identify the sugarcane [3], i.e. its peaks, valleys and information from its curve are sufficient to classify the cane. We assume that a multidimensional space generated by the information obtained in the harmonics (amplitude, additive term and phase angle) is an appropriate space to study the similarity between NDVI series.

## 2. MATERIAL AND METHODS

The study area is the Jaboticabal city in state of São Paulo. That city is a leading producer of sugarcane in Brazil, located in Southeast (21$^o$ 15' 18" S and 48$^o$ 19' 19" W). The NDVI multi-temporal images were obtained from NOAA satellites from crop season 2004/2005. For each crop season, the NDVI series varied from the beginning of planting in April, passing through the period of greatest vigor and going up to end of the harvest in March. The raw image transmitted by the NOAA satellite can contain problems and distortions. Therefore, all images were processed according to the following steps: format conversion from raw images to intermediate format; radiometric calibration; geometric correction; masking of clouds and generation of Maximum Value Composite of NDVI images. These processing methods were performed by the NavPro system [2]. This system guarantees that each

Thanks to Capes and Fapesp agency for funding.

image has no noise and less than 30% of pixels covered by clouds.

### 2.1. Feature extraction of time series by using the harmonic function

Although we deal with discrete data, we will support the theory of continuous functions to admit later that the data used can be understood as a sample of the continuous approach. To classify a pixel, we used the harmonic function of the NDVI series. The key idea behind the harmonic (Fourier) analysis is to decompose a signal into an infinite number of components (harmonics)[4]. Each component is made up of sine and cosine waves of same frequency. Let $f : [0, L] \to \mathbb{R}$ continuous, then:

$$\begin{aligned} f(x) &\sim \frac{1}{2}a_0 + \sum_{n=1}^{\infty}\left[a_n \cos\left(\frac{2\pi nx}{L}\right) + b_n \sin\left(\frac{2\pi nx}{L}\right)\right] \\ &= c_0 + \sum_{n=1}^{\infty}\left[c_j \cos\left(\frac{2\pi jx}{L} - \phi_j\right)\right]. \end{aligned} \quad (1)$$

The both representation of $f(x)$ on right hand side are, each one, forms of the harmonic function (Fourier series), so $a_j$ and $b_j$ are the Fourier coefficients, $c_j$ is the amplitude and $\phi_j$ is the phase angle of $j$-th term. Let us define the $j$-th harmonic as the $j$-th term of the Fourier series (for $j \geq 1$) given by:

$$a_j \cos\left(\frac{2\pi jx}{L}\right) + b_j \sin\left(\frac{2\pi jx}{L}\right)$$

or, by some convenient trigonometric transformations, we take $c_j \cos\left(\frac{2\pi jx}{L} - \phi_j\right)$, where: $c_0 = \frac{1}{2}a_0$, $c_j = \sqrt{a_j^2 + b_j^2}$ and $\phi = tg^{-1}\left(\frac{b_j}{a_j}\right)$ if $a_j \neq 0$ or $\phi = \frac{\pi}{2}$ otherwise. As we used the inverse tangent function, we have $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, so, whenever $a_j < 0$ we must use $\phi = tg^{-1}\left(\frac{b_j}{a_j}\right) + \pi$ to obtain the true phase angles $\phi \in [-\frac{\pi}{2}, \frac{3\pi}{2}]$.

For a finite set of data $y(t)$ with $t \in \{1, 2, \cdots, n\}$, we use these ideas to work with a adjusted data through the above series, where the coefficients were obtained by a least squares fit. Any data series, with $n$ points, can be well approximated by the following finite expansion using $m < \lfloor \frac{n}{2} \rfloor$ and exactly represented by $m = \lfloor \frac{n}{2} \rfloor$:

$$\begin{aligned} y_t &= \bar{y} + \sum_{k=1}^{m}\left[C_k \cos\left(\frac{2\pi kt}{n} - \phi_k\right)\right] \\ &= \bar{y} + \sum_{k=1}^{m}\left[A_k \cos\left(\frac{2\pi kt}{n}\right) + B_k \sin\left(\frac{2\pi kt}{n}\right)\right] \end{aligned} \quad (2)$$

where $\bar{y}$ is a is the arithmetic mean of the data and $A_k$, $B_k$, for a series of equally spaced data in time and without missing values, takes the form:

$$A_k = \frac{2}{n}\sum_{t=1}^{n} y_t \cos\left(\frac{2\pi kt}{n}\right), \quad (3)$$

$$B_k = \frac{2}{n}\sum_{t=1}^{n} y_t \sin\left(\frac{2\pi kt}{n}\right). \quad (4)$$

Notice that we are adopting a time series classifying feature based approach. It is also important to observe that approximation of a data series, can be done by use of any base, not necessarily the sines and cosines base. The coefficients found in this approach are the *features* of this series, i.e., *features of a series, related to a base, are the coefficients of the best possible adjustment of this series by a linear combination of elements of that base*. Roughly speaking, for each pixel, a time series of NDVI related to that pixel is adjusted by a trigonometric polynomial.

### 2.2. Classifier architecture

Our general purpose of classification is to identify structure in an unlabeled data set by objectively organizing data into a homogeneous group where the within-group-object distance from a center is less than a giving value.

To classify a pixel, we proposed a method to analyze the similarity between the features obtained from its NDVI series by the use of decomposition (2), and the average features of a control group. If the distance between them is less than a given tolerance, we say that pixel belongs to the sugarcane class, therefore, it is necessary a training phase before classification phase. In the training phase, we obtain a NDVI curve (*mean feature curve*) that well represent any curves of NDVI sugarcane annual series. In the training phase, we used data from Jaboticabal city for the period from April 2004 to March 2005. To build the mean feature curve, a mask were generated to guarantee that only pixels classified as sugar sugarcane fields were processed, eliminating urban areas, soil, and other kinds of vegetation.

We have an NDVI sugarcane annual series $\mathbf{S_i}$ with $n$ points. For each of these series, we have a vector of features with dimension size up to $2\lfloor\frac{n}{2}\rfloor + 1$, which we shall call **feature vector** $\mathbf{v^i}$. This vector is defined as follows: *the first coordinate $v_1^i$ refers to the additive term $\bar{y}$ of the harmonic decomposition. The coordinates $v_{2k}^i$ and $v_{2k+1}^i$ are respectively the values of amplitude ($C_k$) and phase angle ($\phi_k$) of the $k$-th harmonic*.

The number $m$ of harmonics can range from 1 to $\lfloor\frac{n}{2}\rfloor$ and the features vector dimension will always be equal to $2m+1$. Using then the features vectors, we can obtain the **mean feature vector** $\bar{\mathbf{v}}$, which is the result of the function minimization of $Q(\bar{\mathbf{v}}) = \sum[d(\mathbf{v^i}, \bar{\mathbf{v}})]^2$, where $d$ is the Euclidean dis-

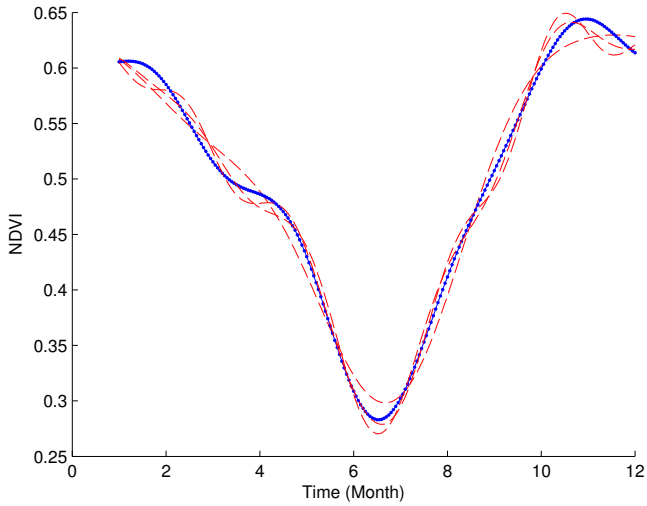


**Fig. 1**. Building mean feature curve from a sample of pixels.

tance. It is, therefore, a problem of quadratic objective function without constraints, where the solution is the arithmetic mean of the coordinates.

Thus, we have a curve defined by a trigonometric polynomial where the coefficients are the coordinates of the mean feature vector. That curve, denominated **mean feature curve**, is a good representation of the series. The basic concept behind this idea is that the behavior demonstrated by the variation of the NDVI has a strong seasonal component that can be well represented in the features space and thus well simulated by the mean feature curve. To avoid the risk of incorporating features of outliers, we choose to exclude these outliers from the test set by Chauvenet's criterion [5].

By choosing the appropriate number of harmonics, this classification may occur prior to the crop season end. For instance, if we consider $m = 3$, the feature vector will have dimension 7, which will allow employment of the method once we reach the sixth month. To say that a pixel belongs to sugarcane class, we have two criteria,

$$C_1 \quad : \quad \#S < \epsilon \text{ where } S = \{V_i \mid |V_i - \bar{v}_i| \geq \lambda\sigma_i\}, \quad (5)$$

$$C_2 \quad : \quad \sum_{i=1}^{m} \frac{|V_i - \bar{v}_i|}{\lambda\sigma_i} < \epsilon \quad (6)$$

where $1 \leq i \leq 2m + 1$ and $\sigma_i$ is the standard deviation of the data used to build the coordinate $i$ of $\bar{\mathbf{v}}$. In each criterion, $\lambda$ define the dispersion tolerance and $\epsilon$ is the acceptable tolerance value related to the difference between $\mathbf{V}$ and $\bar{\mathbf{v}}$, where $\mathbf{V}$ is the feature vector of the pixel to be classified.

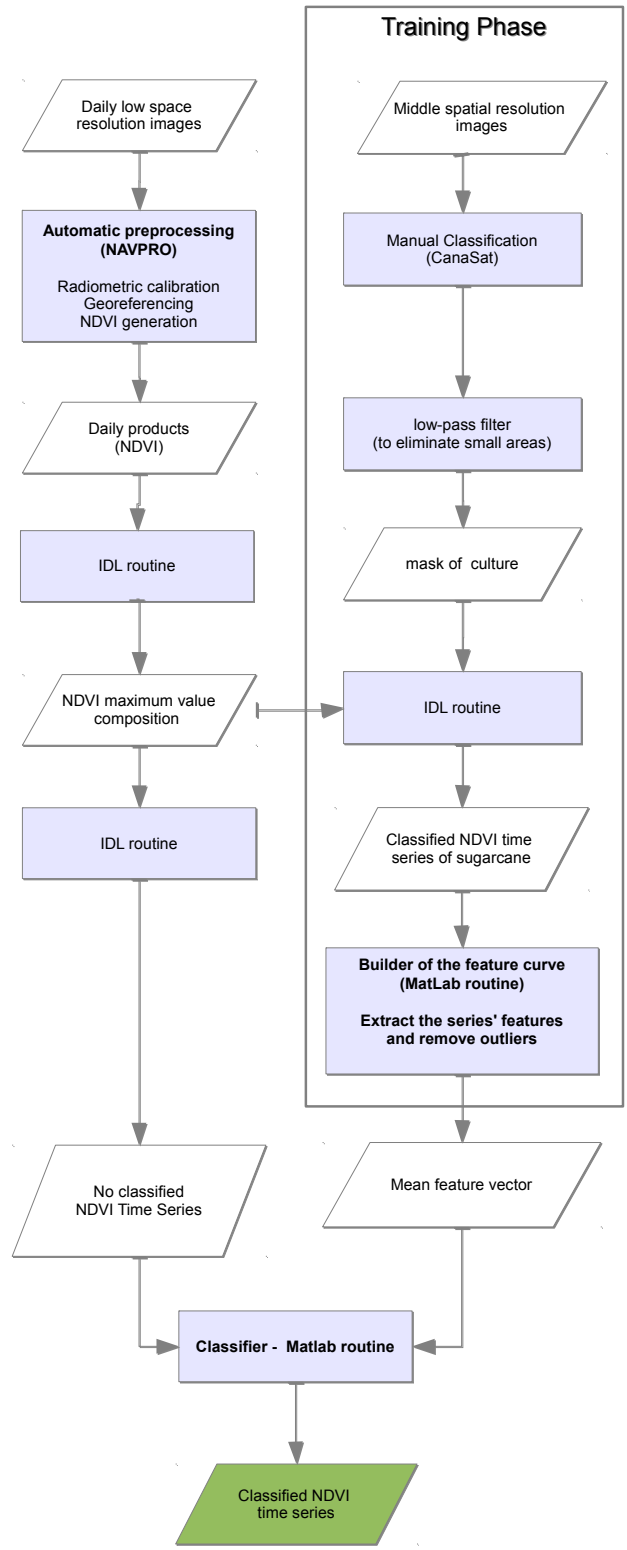


**Fig. 2**. All the processes involved in the classifier.

## 3. EXPERIMENTS AND RESULTS

We tested criteria (5) and (6) in the data from Jaboticabal in 2005. In each criterion, $\lambda$ and $m$ were varied to determine the best classifier performance based on the observation of the confusion matrices that were obtained.

To provide a better statistical analysis, the confusion matrices were obtained from the arithmetic mean of the confusion matrices obtained in the turns generated by cross-validation. Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against each other. We used the basic form of cross-validation in $k$-fold cross-validation where $k = 10$.

We compare the performances of $C1$ and $C2$ criteria with the parameters $m$, number of harmonics ranged from 3 to 6 and $\lambda, \epsilon \in \{1, \frac{3}{2}, 2, \frac{5}{2}, 3\}$. Here, $\epsilon$ is a number related to infractions that the vector can commit to still be considered sugarcane. In $C_1$ criterion, $\epsilon$ is a integer, but in $C_2$ that is not necessary. Notice that $\epsilon$ must be a positive number less or equal than $2m + 1$.

The parameter values that generated percentage of accuracy for true-positive (TP) and true-false (TF) of less than $65\%$ were discarded. Thus, we obtained a clear perception that the criterion $C1$ was more effective for TP detection, but has a lower success rate for TF detection. On the other hand, criterion $C2$ seems more appropriate as a dissimilarity measure and less efficient for TF detection.

The best results were obtained by $\epsilon = 2$ and one of the best performance was reached for $C_1$ with $m = 5$ and $\lambda = 2$ with the following confusion matrix:

**Prediction outcome**

| | | |
|---|---|---|
| **actual value** | 0.88 | 0.12 |
| | 0.13 | 0.77 |

Good results were also observed for criterion $C1$ with $m = 3$, $\lambda = 1.5$; $m = 6$, $\lambda = 2$ and for $C2$ criterion with $m = 3$ and $\lambda = 3$. These results are respectively described by the following confusion matrices:

| | |
|---|---|
| 0.76 | 0.14 |
| 0.12 | 0.78 |

| | |
|---|---|
| 0.86 | 0.14 |
| 0.30 | 0.70 |

| | |
|---|---|
| 0.70 | 0.30 |
| 0.10 | 0.90 |

## 4. CONCLUSION AND FUTURE WORKS

The proposed methodology was efficient with a high success rate for the classification of the culture of sugarcane in AVHRR/NOAA images, with advantage over classification based on mask, because it is able to detect sugarcane expansion.

The $C1$ criterion, based on integer infractions, has been more appropriate for the sugarcane pixels recognition than the $C2$ criterion. However, for identification of no sugarcane pixels, the $C2$ criterion, based on the infraction strength, was more appropriate. Overall $C_1$ has shown a better performer than $C_2$, but maybe is possible build a better criterion from mixing C1 and C2. Thus we would have two turns for classifier iteration, where after the classification made by one of the criteria, we extract pixels of the class that criterion best recognizes. In the remaining database we reapply the classifier based on other criteria. The idea is, in some sense, improves the classifier performance in a balanced way, i.e. improving TP and TF at the same time.

Regarding future work, we must emphasize that the trigonometric basis is not the only base, so they can be replaced by Wavelets. Besides that, Euclidean distance can be replaced by a metric that allows the pixel to be partially pertinent to the sugarcane class, this can be done by using *fuzzy sets* at the Fourier coefficients. That way you can deal with the hassle of manipulate low-resolution images using a crisp approach.

## 5. REFERENCES


[1] CEPEA, "Pib do agronegócio," http://www.cepea.(...) esalq.usp.br/pib/other/Pib_Cepea_1994_2009.xls, accessed on February 21, 2011.

[2] J. C. D. M. Esquerdo, J. F. G. Antunes, D. G. Baldwin, W. Emery, and J. Zullo, "An automatic system for avhrr product surface generation," *Internation Journal o Remote Sensing*, vol. 27, pp. 3925–3942, 2006.

[3] R. R. V. Gonçalves, C. R. Nascimento, J. Zullo Jr., and L. A. S. Romani, "Relationship between the spectral response of sugar cane, based on avhrr/noaa satellite images, and the climate condition, in the state of so paulo (brazil), from 2001 to 2008," in *5th International Workshop on the Analysis of Multi-temporal Remote Sensing images (Multitemp 2009)*, Groton Connecticut Daniel L. Civco, Ed., Ed., USA, 2009, pp. 315–322.

[4] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, Elsevier Academic Press, 2st, edition, 2006.

[5] J. R. Taylor, *An Introduction to Error Analysis*, California: University Science Books, 2st, edition, 1997.