Brazil is an important sugar cane producer, which is the main resource for ethanol production, a renewable source of energy. Due to the strategical importance of this agricultural commodity, it is necessary to improve models that assist the crops monitoring process. Recently, remote sensing images have also been used to crops monitoring. Vegetation index images obtained by operations between satellite channels, for instance, can be taken over a season, showing the development of crops. Specialists in agrometeorology need methods which aim at understanding and mining these datasets to discover interesting patterns and knowledge. Accordingly, this paper presents a methodology to analyze NDVI time series using a distance function based on dynamic time warping distance (DTW) to perform similarity search. The experiments were done for NDVI multi-temporal images from seven harvests regarding the period from April/2001 to March/2008. NDVI time series was generated from NOAA-AVHRR images of a relevant sugar cane producer region in Brazil. Two different distance functions were compared and DTW reached better results than Euclidean distance. The proposed method allowed comparing harvests in different regions and in the same time series. Results of similarity search on NDVI time series demonstrate the efficacy of the use of distance function to similarity search in remote sensing data. This approach is appropriate to assess patterns in a long time series of multi-temporal images and can assist in the process of decision making by agricultural entrepreneurs.

INTRODUCTION

According to official data from Brazilian government, the agribusiness contributed with 23.3% of national GDP and 36.4% of exports in 2007. One of the most important crops is the sugar cane, which is used to produce sugar, ethanol, and energy. Brazil occupies the top position in the world ranking of the sugar cane production and has an important role to attend the world’s sugar and ethanol needs. Then, this agricultural commodity is strategic to the economy of the country. The development of new models to assist in the monitoring, forecasting or analyzing process is a priority to researchers and government. There are many sugar cane producing areas and Sao Paulo state is the main producer of ethanol in Brazil. This region has a strategic importance for the country economy and to
guarantee the Brazilian self-sufficiency in this important, renewable source of energy.

Remote sensing data can be used to improve traditional agrometeorological methods for harvest monitoring or forecasting. Nowadays, these data are more accessible and there are appropriate technology (software and hardware) to receive, distribute, manipulate and process long time series of satellite images. In this scenario, several satellites can be used to assist in monitoring and forecasting, especially the satellites of the National Oceanic and Atmospheric Administration (NOAA). A useful sensor over the past few years has been the Advanced Very High Resolution Radiometer (AVHRR) on board the NOAA satellites. AVHRR images have been used to study land surface and climatic applications (Los et al., 1994). Many works have described the use of NOAA-AVHRR images in land surfaces studies, such as drought investigation (Bayarjargal et al., 2006; Bajgiran et al., 2008), crop area and yield estimation (Dalezios et al., 2001; Liu and Kogan, 2002).

The use of remote sensing in the Brazilian research has been also increased due to the benefits for a country that has a territory with continental dimensions. The sugar cane crops are cultivated in large and contiguous fields, which contribute to the use of low resolution satellite images such as NOAA-AVHRR. AVHRR sensor is widely used for studies of ecosystems due to the availability of long time series of its image data. Additionally, the global coverage and free data access are also advantages that make AVHRR data attractive for many studies.

The AVHRR sensor has channels in the visible and near-infrared spectrum and, combinations of these channels can be used to indicate the amount and state of vegetation (Tucker, 1979), usually known as vegetation indices. Normalized Difference Vegetation Index (NDVI), proposed by Rouse et al. (1973) is the most used index. NDVI is correlated with green biomass (Anyamba and Tucker, 2005) and leaf area (Xavier and Vettorazzi, 2004; Wang et al., 2005).

Seasonal profiles can be obtained when sequential NDVI observations are taken over a season. These profiles show the progression of crop canopy emergence, maturation and senescence (Esquerdo et al., 2006). The NDVI values over the profile indicate the crop performance and are related to crop yields. In the literature, there are several works that analyze NDVI time series to improve the agriculture monitoring (Lucas and Schuler, 2007). In this context, this paper presents a model to analyze harvest through similarity search using a metric for time series similarity. Two distance function (Euclidean distance and DTW) were used to similarity search in NDVI time series. Experimental results of nearest-neighbor queries over NDVI time series are discussed.

The remainder of this paper is structured as follows. Next section presents the material and methods employed to the development of this work. The following section discusses the experiments and results achieved so far. Finally, the last section concludes this work and suggests further developments.

MATERIAL AND METHODS

NOAA-AVHRR images used in this work have been stored and managed by the Center of Meteorological and Climatic Research applied to Agriculture of the State University of Campinas. The database has more than 6 terabytes of images and store data since April 1995. NOAA-16 and NOAA-17 images gathered from April/2001 to March/2008 were used in the experiments.

The test area is located in an important region of sugar cane crops in the state of Sao Paulo, Brazil. Ten regions (Araraquara, Araras, Jaboticabal, Jardinopolis, Jau, Luis Antonio, Pitangueiras, Pontal, Ribeirao Preto and Sertaozinho) in the same Landsat scene, that belonging to orbit/point 220/75 were selected to perform the experiments.

The raw image transmitted by NOAA satellite can contain problems and distortions. AVHRR images have geometric distortions caused by the Earth curvature, rotation and satellite clock errors, attitude errors and imprecise orbital (Rosborough and Baldwin, 1994). Then, for land applications, these distortions must be corrected to avoid errors where high geometric precision is required. Thus, AVHRR images must be submitted to a pre-processing system that comprehends phases as follows:

- format conversion from raw images to intermediate format;
- radiometric calibration;
- geometric correction;
- identification of pixels classified as cloud and
- generation of maximum value of NDVI images.

These processing methods were performed by the NAVPRO system (Esquerdo et al, 2006), which was used to
do the necessary geometric corrections. NAVPRO is an automatic set of C-shell scripts that call the subroutines of NAV (NAVigation) (Emery et al., 1989), developed by the Colorado Center for Astrodynamics Research (CCAR), Aerospace Engineering Sciences, with the University of Colorado, Boulder, USA.

This system guarantees that each image has less than 30% of pixels covered by clouds, without noise, and high elevation passes. The last phase of NAVPRO system was the generation of several products. One of them combined channels 1 (red) and 2 (near infrared) to calculate NDVI images. The effects of shadow, aerosols and water vapor were minimized by the generation of Maximum Value Composite (MVC) of NDVI images, as described by (Holben, 1986). The MVC were made using only images from the same satellite. Masks were generated to guarantee that only the pixels classified as sugar cane field were processed, eliminating urban areas, soil, and other kinds of vegetation. Figure 1 presents a flowchart describing the methodology to correct NDVI images.

![Flowchart describing the methodology used to geometric correction of NDVI images.](image)

The proposed methodology uses a distance function and an algorithm to perform similarity search to find the series most similar to a time series (query center) that is being analyzed. A distance function or metric can be defined as a similarity measure between two datasets, in this case two time series. A distance function must comply the four axioms presented below to be considered a metric, rendering together to the dataset a metric space:

1. \( d(s_1, s_2) \geq 0 \)
2. \( d(s_1, s_2) = 0 \) when \( s_1 = s_2 \), otherwise \( d(s_1, s_2) > 0 \)
3. \( d(s_1, s_2) = d(s_2, s_1) \) (symmetric)
4. \( d(s_1, s_3) = d(s_1, s_2) + d(s_2, s_3) \) (triangle inequality) for any \( s_i \) pertaining to the data domain.

There are some distance functions that do not satisfy the fourth axioms. These distance functions are called, in specific situations, as pseudo metrics. There are many distance functions for particular applications. The most widely used distance functions are those from the Minkowski family (or \( L_p \) norm). The Euclidean distance corresponds to \( L_2 \) which is commonly used to calculate the distance between multi-dimensional arrays and vectors.

For the following definitions will be considered two time series \( Q \) and \( C \), of length \( n \) and \( m \) respectively, where:

\[
Q = q_1, q_2, ..., q_n
\]
\[
C = c_1, c_2, ..., c_m
\]
Equation 1 shows how to calculate the Euclidean distance.

\[ d(q_i, c_j) = \sum_{i=1}^{n} (q_i - c_j)^2 \]  

(1)

Dynamic Time Warping is a very efficient distance function to compare time series (Berndt and Clifford, 1994). Its main objective is to keep close time series that have similar behavior, but are delayed or distorted along the time axis. Thus, this technique has a good sensibility to warping because the comparisons between corresponding points are not rigid. In this case, points of a series can be compared to adjacent ones in other series, as illustrated in Figure 2.

![Figure 2. Comparisons between time series: a) conventional method; b) using DTW.](image)

To align two sequences using DTW, an \( nxm \) matrix is built where the \( (i^{th}, j^{th}) \) element of the matrix contains the Euclidean distance \( d(q_i, c_j) \) between the two points \( q_i \) and \( c_j \). Each element of the matrix corresponds to the distance between the points that it represents. A warping path \( W = (w_1, w_2, \ldots, w_k) \) is a contiguous set of matrix elements that defines a mapping between \( Q \) and \( C \). The adjustment route is defined by the following rules:

- it starts at \( w_1 = (1,1) \) and finishes at \( w_k = (m, n) \)
- the sequence of route must be to adjacent elements of the matrix (including diagonally adjacent cells)
- the points in \( W \) must be monotonically spaced in time, that is the sequence must not go back in the route.

There are many warping paths, but DTW is a sum of \( w_k \) elements in the path which minimizes the warping cost. DTW is calculated by the Equation 2.

\[ DTW(Q,C) = \min \left\{ \sum_{k=1}^{K} w_k / K \right\} \]  

(2)

where \( w_k \) is the \( k^{th} \) element of the adjustment route and \( K \) is the number of elements of the adjustment route.

In the equation 2, \( K \) in the denominator is used to compensate the size of the deviation between the two time series because the warping paths may have different lengths. Dynamic programming is an efficient way to found the path, which is employed in Equation 3.

\[ \gamma(i, j) = d(q_i, c_j) + \min \{ \gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1) \} \]  

(3)

Nearest-neighbor query was used to perform similarity queries to find the closest element to the query center. That is, given an element of interest – the center of the query – which are the elements of the dataset with smaller distances (higher similarities) to this element? Then, given a query object \( q_q \) and the set of data elements \( T \), the nearest neighbor is the element of \( T \) such that \( NNQuery(q_q) = \{ q_a \in T \mid \forall q_i \in T, d(q_q, q_a) \leq d(q_q, q_j) \} \). An example of a nearest neighbor query in a NDVI time series database is: “find the time series in \( T \) that is the most similar to time series \( A \”).

RESULTS AND DISCUSSIONS
In order to make the dataset uniform, NOAA-AVHRR images for a period from 2001 to 2008 were corrected using NAVPRO system. Thus, it was generated monthly NDVI images using the MVC technique. The NDVI values were extracted from images of 10 sugar cane producer regions.

Two datasets were generated to perform the experiments. One of them is composed of 70 time series with 12 values corresponding to one year of sugar cane harvest. The other dataset was generated with 10 complete time series, one for each region.

Experiments were performed with two datasets using Euclidean and DTW distance function. Similarity searches were computed for time series of the same region and distinct ones. In order to identify similar harvest in the dataset was accomplished queries using a specific harvest as a center. The algorithm was executed and returned a rank with the most similar series closest to the query center. Table 1 shows the results for the region of Jaboticabal, using the harvest of 2005-2006 as query center. DTW and Euclidean distance presented different ranks for the same query, as shown in Table 1.

**Table 1.** Rank and values of Euclidean distance and DTW for Jaboticabal.

<table>
<thead>
<tr>
<th>NDVI time series from Jaboticabal</th>
<th>DTW</th>
<th>Euclidean distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvest</td>
<td>rank</td>
<td>values</td>
</tr>
<tr>
<td>2001-2002</td>
<td>4</td>
<td>0.00239643</td>
</tr>
<tr>
<td>2002-2003</td>
<td>3</td>
<td><strong>0.00225431</strong></td>
</tr>
<tr>
<td>2003-2004</td>
<td>2</td>
<td>0.00219901</td>
</tr>
<tr>
<td>2004-2005</td>
<td>1</td>
<td>0.000976443</td>
</tr>
</tbody>
</table>

DTW and Euclidean distance have presented the same result until position equals 2. In the third position, the two distance functions diverge and present different results. Time series from 2002-2003 is more similar to the query center (2005-2006) than data from 2006-2007 when DTW was used. The result for the Euclidean distance is the opposite, as showed in bold at Table 1.

Figure 3 shows graph of the most similar time series (2004-2005) detected by both distance function. It indicates that the harvest of 2004-2005 had similar behavior when it is compared to 2005-2006 for Jaboticabal. Harvest of 2003-2004 also had similar trend to 2005-2006, which indicate a pattern of spectral response by NDVI from 2003 to 2006.

![Figure 3](image)

**Figure 3.** Graphs representing the results of similarity search for NDVI time series from Jaboticabal.

Figure 4 presents two graphs with different results for DTW and Euclidean distance to 2005-2006 as query center. In the Figure 4, graph a shows the second best result gotten by a similarity search using DTW and graph b shows the second best result for the query using the Euclidean distance. Looking at the graphs, one can see that the result provided by DTW is superior in 90%.

MultiTemp 2009 - The Fifth International Workshop on the Analysis of Multi-temporal Remote Sensing Images
July 28-30, 2009 • Groton, Connecticut
Figure 4. Graphs representing the results of similarity search for NDVI time series from Jaboticabal. (a) using DTW distance function. (b) using the Euclidian distance function.

Another experiment was performed to search time series similar to a query center (2005-2006) with the same period, but from different regions. Table 2 shows the results for the region of Jaboticabal, using the harvest of 2005-2006 as query center. DTW and Euclidean distance also presented different ranks, as shown in bold at Table 2.

Table 2. Rank and values of Euclidean distance and DTW for Jaboticabal.

<table>
<thead>
<tr>
<th>Regions</th>
<th>DTW</th>
<th>Euclidean distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rank</td>
<td>values</td>
</tr>
<tr>
<td>Araraquara</td>
<td>5</td>
<td>0.000977036</td>
</tr>
<tr>
<td>Araras</td>
<td>7</td>
<td>0.00102806</td>
</tr>
<tr>
<td>Jardinopolis</td>
<td>6</td>
<td>0.00100112</td>
</tr>
<tr>
<td>Jau</td>
<td>9</td>
<td>0.00173745</td>
</tr>
<tr>
<td>Luis Antonio</td>
<td>8</td>
<td>0.00140662</td>
</tr>
<tr>
<td>Pitangueiras</td>
<td>1</td>
<td>0.000182933</td>
</tr>
<tr>
<td>Pontal</td>
<td>3</td>
<td>0.000572769</td>
</tr>
<tr>
<td>Ribeirao Preto</td>
<td>4</td>
<td>0.000689431</td>
</tr>
</tbody>
</table>

In this experiment, DTW presented better results than Euclidean distance. Figure 5 illustrates two graphs with the time series from Pitangueiras and Sertaozinho.

Figure 5. Graphs representing the results of similarity search for NDVI time series from different regions. (a) the best result for similarity query using DTW and (b) the second result using DTW.
This result shows that Pitangueiras, Jaboticabal and Sertaozinho have similar spectral response for 2005-2006 harvest. This method is appropriate to analyze harvest from different periods and regions.

The last experiment was conducted with complete time series of all regions. The results of NNQuery with Jaboticabal as the query center showed that this NDVI time series is similar to Pitangueiras and Sertaozinho. Figure 6 illustrates these results.

![Figure 6](image)

**Figure 6.** Graphs representing the results of similarity search for complete NDVI time series from different regions. (a) and (b) using DTW.

**CONCLUSION**

This work presented a method to analyze harvests from different regions involving remote sensing data and an algorithm to accomplish similarity search. Two different distance functions – Euclidean distance and DTW - were compared and results indicated that DTW is more appropriate than Euclidean distance for the time series analyzed.

Accordingly, DTW-based method is an appropriate method to perform similarity search in NDVI time series. This approach makes the analysis of time series by the researchers easier because it finds similar series to a specific pattern presented to the automatic system. This technique has been used in time series mining in many areas, and this work demonstrated that this approach can also be successfully employed to multi-temporal images mining.

However, DTW has problems when the two time series differ in the Y-axis. Then, new methods should be proposed to improve the results for similarity search in multi-temporal images.

**REFERENCES**


