NEW DTW-BASED METHOD TO SIMILARITY SEARCH IN SUGAR CANE REGIONS REPRESENTED BY CLIMATE AND REMOTE SENSING TIME SERIES

L. A. S. Romani[1,2], *R. R. V. Goncalves*[3], *J. Zullo Jr.*[3], *C. Traina Jr.*[1], *A. J. M. Traina*[1]

[1]Computer Science Department, USP at Sao Carlos, PB 668 13560-970, Brazil
[2]Embrapa Agriculture Informatics, Campinas, Brazil
[3] CNPq and Cepagri, University of Campinas, Campinas, Brazil
E-mail: {alvim, caetano, agma}@icmc.usp.br {jurandir, renata}@cpa.unicamp.br

1. INTRODUCTION

According to future scenarios assessed by specialists, extreme events may increase in frequency and intensity in the next years. These changes may cause natural disasters, food security problems and other effects on human environment. Studies indicate that global warming is due to natural and anthropogenic factors. One of the main causes of this warming is the increase in the emission of greenhouse gases. In this context, researchers at the twenty-first century have many urging challenges on finding alternatives for mitigation and adaptation. The replacement of fossil-fuel by fuel generated from renewable sources is a way of contributing to the decrease in the emission of greenhouse gases.

In Brazil, the main source of biofuel is sugar cane, which is a strategic agricultural crop for the Country. Sugar cane has an annual cycle and it is cultivated in large and contiguous fields, which allows the use of low-resolution satellites sensors, such as NOAA-AVHRR. AVHRR (Advanced Very High Resolution Radiometer) is a useful sensor on board the NOAA (National Oceanic and Atmospheric Administration) satellites. AVHRR images have been used to study land surface, such as crop area and yield estimation as well as climate applications.

NDVI (Normalized Difference Vegetation Index) is one of the vegetation indexes most widely used and can be obtained by the combination of visible and near-infrared channels of AVHHR. NDVI is correlated with green biomass [1] and leaf area [2]. Many works have analyzed the correlation among variables obtained through remote sensing data, such as NDVI and indexes that summarize the agroclimate conditions, such as WRSI (Water Requirement Satisfaction Index). WRSI represents a fraction of the amount of water consumed by the plant and the amount of water that would be used by it to ensure maximum productivity. WRSI is generated from water balance simulation. These two indexes can be used to characterize regions that produce sugar cane, since the NDVI indicates the state of vegetation and WRSI represents the climate conditions.

However, the task of finding similar regions by analyzing the time series of NDVI and WRSI is not simple. In [3] was presented a method to find NDVI time series similar to other NDVI series from different regions. This approach has combined a distance function and an algorithm for similarity search. Although it appears effective to find similar series, this method cannot detect similarity when two distinct series are combined.

To deal with this limitation, we propose a new method to similarity search considering two-dimensional objects, i.e. objects represented by two different series representing both indexes. This method takes advantage of the well-known Dynamic Time Warping (DTW) distance function [4, 5] weighted by the correlation between series and the variance of each one. This approach allows the specialists to make comparisons between regions considering distinct series that represent them, as well as combining attributes of different types of sensors. Thus, specialists can use an automatic method to analyze a huge volume of time

Thanks to Embrapa, FAPESP, CNPq and CAPES for funding.

series finding similarities and clustering among them. Detection of similar regions aims at understanding the distribution of certain crops and facilitates on monitoring of these crops. This paper is organized as follows. Section 2 shows study area and methodology. Section 3 describes experiments and discusses results. Section 4 concludes the paper.

2. MATERIAL AND METHODS

NOAA-AVHRR images used in this paper have been stored and managed by Cepagri (www.cpa.unicamp.br), a research center of the University of Campinas, Brazil. NOAA-16 and NOAA-17 images gathered from April, 2001 to March, 2008 were used in the experiments. The study site is located in an important region of sugar cane production in the state of Sao Paulo, Brazil. This region is located between the geographic coordinates $54^o00'$ and $43^o30'$ west longitude and $25^o30'$ and $19^o30'$ south latitude. Regions in the same Landsat scene, belonging to orbit/point 220/75 were also selected to perform the experiments.

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 [6]. This system guarantees that each image has less than 30% of pixels covered by clouds, without noise, and high elevation passes. Masks were generated to guarantee that only pixels classified as sugar cane fields were processed, eliminating urban areas, soil, and other kinds of vegetation.

The new similarity measure proposed can be described as a weighting of a distance function DTW using correlation and variance factors. In the first step, DTW values were calculated between time series of the same variable, as it can be seen in Figure 1.

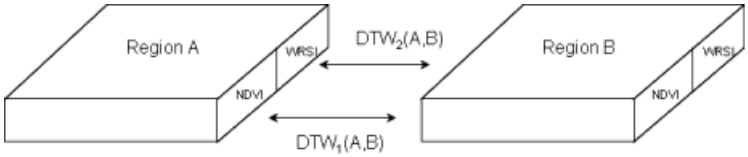


Fig. 1. 1^{st} step: Calculation of DTW of two series of the same variable

DTW is a well-known efficient and effective distance function to compare time series, thus it was chosen in this work. Let be two time series Q and C, of lengths n and m respectively, where:

$Q = q_1, q_2, \ldots, q_n$

$C = c_1, c_2, \ldots, c_m$

Equation 1 shows how to calculate the Euclidean distance, only if n is equal to m.

$$d(q_i, c_i) = \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2} \tag{1}$$

To align two sequences using DTW, an n-by-m matrix was built where the (i_{th}, j_{th}) element of the matrix contains the Euclidean distance $d(q_i, c_j)$ between two points q_i and c_j. A warping path W is a contiguous set of matrix elements that defines a mapping between Q and C. There are many warping paths, but DTW is a sum of w_k elements in the path that minimizes the warping cost. The DTW calculation is given by Equation 2.

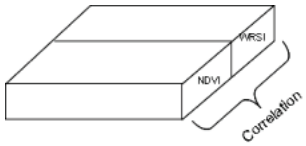
$$DTW(Q, C) = min\left\{\sqrt{\sum_{k=1}^{K}(w_k)}/K\right\} \tag{2}$$

The second step of our proposed method is based on the calculation of Pearson correlation between the two series and the variance of each series. Suppose that the region is a 2D region given by the two series (NDVI and WRSI), as illustrated in

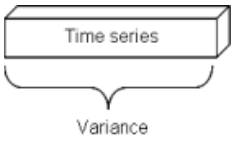
Figure 2a. The correlation factor indicates how these two series are related. The correlation calculation factor is given by Equation 2.

$$f_C(A, B) = C_A * C_B \tag{3}$$

where $f_C(A, B)$ is the correlation factor, C_A is the Pearson correlation value between two time series for region *A* and C_B is the Pearson correlation value between two time series for region *B*.



(a) 2^{nd} step: calculation of correlation factor (f_C)



(b) 3^{rd} step: calculation of variance factor (f_V)

Two time series (NDVI and WRSI) used to represent the features of a region may have a different distribution. In this case, it was included the variance factor to mitigate this possible variation (Figure 2b). The variance factor is given by Equation 2.

$$f_{V_1}(A, B) = (1 - V_{A_1}) * (1 - V_{B_1}) \tag{4}$$

where $f_{V_1}(A, B)$ is the variance factor for a given time series 1 in the regions A and B, V_{A_1} is the variance for time series 1 of region A and V_{B_1} is the variance for time series 1 of region B.

The last step corresponds to the weighting of DTW using variance and correlation factors. The new distance (*MultiDist*) between regions A and B is given by Equation 2.

$$MultiDist(A, B) = ((DTW_1(A, B) * f_{V_1}(A, B)) + (DTW_2(A, B) * f_{V_2}(A, B))) * f_C(A, B) \tag{5}$$

It was used nearest-neighbor query to perform similarity queries to find the closest region to the query center. That is, given a region A of interest - the center of the query - which are the regions with smaller distances (higher similarities) to this region A? Then, given a query object q_q and a dataset of objects (regions) T, the nearest neighbor is the object such that $NNQuery(q_q) = \{q_n \in T | \forall q_i \in T, d(q_q, q_n) <= d(q_q, q_i)\}$. An example of a nearest neighbor query in sugar cane regions database is: "find the regions in T that are the most similar to region A".

3. EXPERIMENTS AND RESULTS

A preprocessing step was employed on the NOAA-AVHRR images for the period 2001-2008, which were corrected using NAVPRO system. Thus, it was generated monthly NDVI images using the Maximum Value Composite technique [7]. The NDVI values were extracted from images of 10 sugar cane producer regions. Additionally, it was generated WRSI values for the same 10 regions.

Experiments were performed with 10 regions composed of two time series (NDVI and WRSI) each one. Three agrometeorologists classified the regions and ranked them considering one specific region (as a query center). The average of their classification is shown in Table 1. This ranking made by specialists was used as (ground truth) reference to access the fidelity provided by the automated result.

In order to validate the proposed method, we performed experiments employing two approaches that uses:

1. *sumDTW*: sum of the DTW distances calculated for each series in different regions,
2. *multDist*: weighting the DTW distance using correlation and variance factors.

The two approaches were used and generated a rank with the most similar regions to query center. Table 1 shows the results

for the region of Jaboticabal as a query center. The methods *sumDTW* and *multDist* presented different ranks for the same query, as in shown in Table 1. The rank proposed by the specialists also appears in the same table.

Table 1. Comparative ranking for similarity search in different regions

Results for Jaboticabal as query center					
Regions	Specialists ranking	sumDTW		multiDist	
		ranking	values	ranking	values
Araraquara	8	7	0.086806	8	0.023596
Araras	6	6	0.078751	6	0.020538
Jardinopolis	5	3	0.067882	4	0.016110
Jau	7	8	0.885896	7	0.020796
Luis Antonio	9	9	0.088991	9	0.023888
Pitangueiras	1	5	0.070715	5	0.018625
Pontal	2	1	0.019823	1	0.006091
Ribeirao Preto	4	2	0.066390	3	0.014848
Sertaozinho	3	4	0.068546	2	0.013708

In this experiments, *multiDist* presented results more similar to the rank given by the specialists than the other method. According to the specialists, regions that appear in the top positions in the ranking are geographically closer and have a climate more similar to Jaboticabal, which was used as the query center.

4. CONCLUSIONS

This paper presented a new method to analyze regions with sugar cane fields using remote sensing and climate data. The *MultiDist* method weighs the DTW distance function and provides an algorithm to accomplish similarity searching. Two different approaches - sumDTW and multiDist - were compared. Experiments indicate the method that considers correlation between two time series and their variance.

The proposed method provides a valuable tool to help the specialists on automatically analyzing different regions. The method allows experts to study areas aggregating information on biomass and climate data, as it supports similarity search of two-dimensional objects.

As a further research direction, the proposed method can be extended to work with multidimensional objects. In addition, other correlation calculation formulas should also be considered.

5. REFERENCES

[1] A. Anyamba and C. J. Tucker, "Analysis of sahelian vegetation dynamics using noaa-avhrr ndvi data from 1981-2003," *Journal of Arid Environments*, vol. 63, no. 3, pp. 596–614, 2005.

[2] Quan Wang, Samuel Adiku, John Tenhunen, and Andr Granier, "On the relationship of ndvi with leaf area index in a deciduous forest site," *Remote Sensing of Environment*, vol. 94, no. 2, pp. 244–255, 2005.

[3] L. A. S. Romani, J. Zullo Jr, C. R. Nascimento, R. R. V. Goncalves, C. Traina Jr., and A. J. M. Traina, "Monitoring sugar cane crops through dtw-based method for similarity search in ndvi time series," in *Fifth International Workshop on the Analysis of Multi-temporal Remote Sensing Images*, Groton, Connecticut, 2009, pp. 171–178.

[4] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD Workshop*, Seattle, WA, 1994, pp. 359–370.

[5] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358386, 2005.

[6] J. C. D. M. Esquerdo, J. F. G. Antunes, D. G. Baldwin, W. J. Emery, and Jurandir Zullo Jr, "An automatic system for avhrr land surface product generation," *International Journal of Remote Sensing*, vol. 27, no. 18, pp. 3925–3942, 2006.

[7] B. N. Holben, "Characteristics of maximum value composite images from temporal avhrr data.," *International Journal of Remote Sensing*, vol. 7, pp. 1417–1435, 1986.