

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. Goncalves³, J. Zullo Jr.³, C. Traina Jr.¹, A. J. M. Traina¹

¹Computer Science Department, USP at Sao Carlos, PB 668 13560-970, Brazil

²Embrapa Agriculture Informatics, Campinas, Brazil

³CNPq and Cepagri, University of Campinas, Campinas, Brazil

E-mail: {alvim, caetano, agma}@icmc.usp.br {jurandir, renata}@cpa.unicamp.br

ABSTRACT

Brazil is an important sugar cane producer, which is the main resource for ethanol production, a renewable source of energy. This agricultural commodity is important to the country economy, becoming fundamental to improve models that assist the crops monitoring process. Vegetation indexes originated from remote sensing images and agrometeorological indexes can be combined to represent sugar cane fields in a regional scale. However, finding different regions with similar patterns to classify or analyze their characteristics is a non-trivial task. Accordingly, this paper presents a method to find similar sugar cane fields represented by series of vegetation and agrometeorological indexes. The proposed method combines a weighted distance function with an algorithm to find similar objects. Results were coincident in the most cases with the classification done by experts, finding regions with similar characteristics of climate and productivity. Consequently, this approach can help in decision making processes by agricultural entrepreneurs.

Index Terms— Vegetation index, ethanol, agrometeorological index, climate change, data mining

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 [1, 2, 3]. 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 AVHRR. NDVI is correlated with green biomass [4] and leaf area [5]. 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 [6] 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 [7, 8] weighted by the correlation between series and the variance of each one. This approach allows the specialists to make comparisons between regions considering distinct

Thanks to Embrapa, FAPESP, CNPq, CAPES, SticAmsud and Microsoft Research for funding.

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 series finding similarities and clustering among them. Detection of similar regions aims at understanding the distribution of certain crops as well as aiding the monitoring of these crops. This paper is organized as follows. Section 2 shows the background and the methods developed. 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°00' and 43°30' west longitude and 25°30' and 19°30' south latitude. Ten regions in the same Landsat scene, belonging to orbit/point 220/75 were selected to perform the experiments as it can be seen in Figure 1.

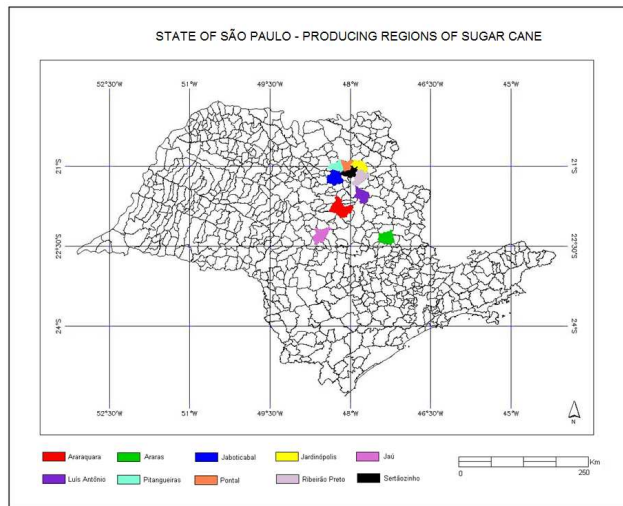


Fig. 1. Test area is located in Sao Paulo, an important state of northeastern Brazil, which is responsible for the major sugar cane production in the country.

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 the Maximum Value Composite of NDVI images [9]. These processing methods were performed by the NavPro system [10]. 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.

We have calculated the WRSI using meteorological data in order to consider also agrometeorological information to the producing areas of sugar cane. We have calculated a water balance, maximum and real evapotranspiration, on a 10-day, biweekly and monthly periods to describe the WRSI. An one-year season of sugar cane was defined to calculate the water balance. In this case, planting begins in August and harvest begins in April. This pattern was established because NOAA-AVHRR image has a low spectral resolution and each pixel of the image shows a combination of spectral responses of several sugar cane areas, which are in different stages of development.

We have used the water balance calculus proposed by Thornthwaite and Mather [11] in this experiment. In the water balance, some variables such as real, potential and maximum evapotranspiration were calculated. Evapotranspiration is the sum of evaporation and plant transpiration. WRSI is calculated as the ratio between real evapotranspiration and maximum evapotranspiration. This index varies from zero to one and represents a fraction of the amount of water consumed by the plant and the amount of water that would be used by the plant to ensure maximum productivity.

Once defined the test area, we describe below the computational method that was proposed in this work to identify similar sugar cane areas. 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 2.

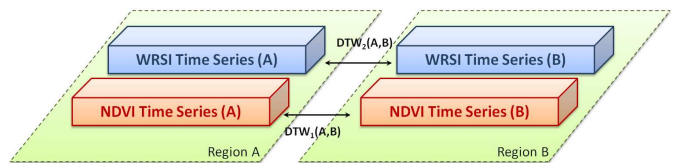


Fig. 2. 1st step: Calculation of DTW of two series of the same variable

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:

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

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

Equation 1 shows how to calculate the Euclidean distance,

only if n is equal to m .

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

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.

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

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 3. The correlation factor indicates how these two series are related. The calculus of the correlation factor is given by Equation 2.

$$f_C(A, B) = C_A * C_B \quad (3)$$

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 .



Fig. 3. 2nd step: calculation of correlation factor (f_C)

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 4). The variance factor is given by Equation 2.

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

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 (*Multi-Dist*) between regions A and B is given by Equation 5.

$$\begin{aligned} MultiDist(A, B) = & ((DTW_1(A, B) * f_{V_1}(A, B)) \\ & + (DTW_2(A, B) * f_{V_2}(A, B))) \\ & * f_C(A, B) \end{aligned} \quad (5)$$



Fig. 4. 3rd step: calculation of variance factor (f_V)

We employed similarity queries (nearest-neighbor) to find the closest region to the query center. That is, “given a region of interest A - 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 q_n such that $NNQuery(q_q) = \{q_n \in T | \forall q_i \in T, d(q_q, q_n) \Rightarrow d(q_q, q_i)\}$. 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

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 use:

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

The two approaches were used and generated a rank with the most similar regions to the query center. Table 1 shows the results for the region of Jaboticabal as a query center. The methods *sumDTW* and *multiDist* presented different ranks for the same query, as is shown in Table 1. The rank proposed by the experts also appears in the same table.

In this experiments, *multiDist* presented results more similar to the rank given by the specialists than the other method. Dividing the list of regions in two groups, it can be seen that the regions geographically closer to Jaboticabal appear in the top five ranking. The regions appearing in the latest ranking positions are more geographically distant and probably have small differences in climate that have been captured by the WRSI. The proposed method is closer to the results provided by specialists with one position of lag in the ranking. When we just sum DTW values calculated for each series, the results do not follow the specialists.

Table 1. Comparative ranking for similarity search in different regions

Results for Jaboticabal as query center					
Regions	Experts ranking	sumDTW		multiDist	
		ranking	values	ranking	values
Araraquara	8	7	0.08680	8	0.02359
Araras	6	6	0.07875	6	0.02053
Jardinopolis	5	3	0.06788	4	0.01611
Jau	7	8	0.88589	7	0.02079
Luis Antonio	9	9	0.08899	9	0.02388
Pitangueiras	1	5	0.07071	5	0.01862
Pontal	2	1	0.01982	1	0.00609
Rib. Preto	4	2	0.06639	3	0.01484
Sertaozinho	3	4	0.06854	2	0.01370

According to the experts, 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. Consequently, these regions have the same methods of planting and possibly the same cut-off date of sugar cane to the corresponding development stage of Jaboticabal (query center). This explains the similarity among series. Thus, a method that approximates the ranking done by specialists can help identify similar regions in a given region with greater extents, as it can be difficult to do manually.

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 reached better results.

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] L.A. Vincent, T.C. Peterson, V.R. Barros, M.B. Marino, M. Rusticucci, Carrasco G., E. Ramirez, L.M. Alves, T. Ambrizzi, M.A. Berlatto, A.M. Grimm, J.A. Marengo, L. Molion, D.F. Moncunill, E. Rebello, Y.M.T. Anunciao, J. Quintana, J.L. Santos, J. Baez, G. Coronel, J. Garcia, M. Trebejo, I. Bidegain, M.R. Haylock, and D. Karoly, "Observed trends in indices of daily temperature extremes in south america 1960-2000," *Journal of Climate*, vol. 18, pp. 5011-5023, 2005.
- [2] L.V. Alexander, X. Zhang, T.C. Peterson, J. Caesar, B. Gleason, A.M.G.K. Tank, M. Haylock, D. Collins, B. Trewin, F. Rahimzadeh, A. Tagipour, K. R. Kumar, J. Revadekar, G. Griffiths, L. Vincent, D.B. Stephenson, J. Burn, E. Aguilar, M. Brunet, M. Taylor, M. New, P. Zhai, M. Rusticucci, and J.L. Vasquez-Aguirre, "Global observed changes in daily climate extremes of temperature and precipitation," *Journal of Geophysical Research*, vol. 111, pp. 1-22, 2006.
- [3] A.R. Ganguly and K. Steinhäuser, "Data mining for climate change and impacts," in *IEEE International Conference on Data Mining Workshops*, IEEE, Ed. 2008, pp. 385-394, IEEE.
- [4] 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.
- [5] Q. Wang, S. Adiku, J. Tenhunen, and A. 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.
- [6] 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.
- [7] 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.
- [8] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358-386, 2005.
- [9] B. N. Holben, "Characteristics of maximum value composite images from temporal avhrr data," *International Journal of Remote Sensing*, vol. 7, pp. 1417-1435, 1986.
- [10] 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.
- [11] C. W. Thornthwaite and J. R. Mather, "The water balance," *Climatology*, vol. 8, no. 1, pp. 104, 1955.