DATA MINING TECHNIQUES FOR IDENTIFICATION OF SPECTRALLY HOMOGENEOUS AREAS USING NDVI TEMPORAL PROFILES OF SOYBEAN CROP

JERRY A. JOHANN[2], JANSLE V. ROCHA[3], STANLEY R. DE M. OLIVEIRA[4], LUIZ H. A. RODRIGUES[5], RUBENS A. C. LAMPARELLI[6]

ABSTRACT: The aim of this study was to group temporal profiles of 10-day composites NDVI product by similarity, which was obtained by the SPOT Vegetation sensor, for municipalities with high soybean production in the state of Paraná, Brazil, in the 2005/2006 cropping season. Data mining is a valuable tool that allows extracting knowledge from a database, identifying valid, new, potentially useful and understandable patterns. Therefore, it was used the methods for clusters generation by means of the algorithms K-Means, MAXVER and DBSCAN, implemented in the WEKA software package. Clusters were created based on the average temporal profiles of NDVI of the 277 municipalities with high soybean production in the state and the best results were found with the K-Means algorithm, grouping the municipalities into six clusters, considering the period from the beginning of October until the end of March, which is equivalent to the crop vegetative cycle. Half of the generated clusters presented spectro-temporal pattern, a characteristic of soybeans and were mostly under the soybean belt in the state of Paraná, which shows good results that were obtained with the proposed methodology as for identification of homogeneous areas. These results will be useful for the creation of regional soybean "masks" to estimate the planted area for this crop.

KEYWORDS: annual crop, MODIS, SPOT Vegetation, MAXVER, K-Means.

TÉCNICAS DE MINERAÇÃO DE DADOS PARA IDENTIFICAÇÃO DE ÁREAS ESPECTRALMENTE HOMOGÊNEAS, UTILIZANDO PERFIS TEMPORAIS DE NDVI DA CULTURA DA SOJA NO ESTADO DO PARANÁ

RESUMO: O objetivo deste trabalho foi agrupar, por semelhança, perfis temporais do produto NDVI decendial, obtido pelo sensor *SPOT Vegetation*, para os municípios produtores de soja no Estado do Paraná, na safra agrícola de 2005/2006. A Mineração de Dados é uma ferramenta valiosa que permite extrair conhecimento de uma base de dados, identificando padrões válidos, novos, potencialmente úteis e compreensíveis. Neste sentido, adotou-se a abordagem de geração dos *clusters* pelos algoritmos *K-Means*, *MAXVER* e *DBSCAN* no *software* WEKA. Foram gerados *clusters* com base no perfil temporal médio de NDVI dos 277 municípios produtores de soja do Estado, e os melhores resultados foram encontrados com o algoritmo *K-Means*, agrupando os municípios em seis *clusters*, utilizando o período do início de outubro ao final de março, equivalente ao ciclo vegetativo da cultura. Metade dos *clusters* gerados apresentou padrão espectro-temporal característico de soja e esteve, em sua grande maioria, sob o cinturão da soja do Estado do Paraná, demonstrando os bons resultados encontrados com a metodologia proposta, em termos de identificação de áreas homogêneas. Estes resultados serão úteis na geração de "máscaras" de soja regionalizadas para estimativa de área plantada desta cultura.

PALAVRAS-CHAVE: cultura anual, MODIS, *SPOT Vegetation*, *MAXVER, K-Means.*

---

[1] Extraído do trabalho final das disciplinas "Preparação de Dados para Mineração de Dados" e "Mineração de Dados e Descoberta de Conhecimento" cursadas no programa de pós-graduação da Faculdade de Engenharia Agrícola, da Universidade Estadual de Campinas - UNICAMP, em 2008.

[2] Dr. em Engª Agrícola, Professor Adjunto da UNIOESTE, Cascavel-PR, jerry.johann@hotmail.com ou jerry.johann@unioeste.br.

[3] Dr. em Sensoriamento Remoto, Prof. Associado, Programa de Pós-Graduação da Feagri/UNICAMP, Campinas-SP, jansle.rocha@feagri.unicamp.br.

[4] Dr. em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Professor do Programa de Pós-Graduação da Feagri/UNICAMP, stanley.oliveira@embrapa.br.

[5] Dr. em Engª Agrícola, Prof. Livre Docente do Programa de Pós-Graduação da Feagri/UNICAMP,lique@feagri.unicamp.br.

[6] Doutor em Engª de Transportes, Pesquisador do Núcleo Interdisciplinar de Planejamento Energético-NIPE,Professor do Programa de Pós-Graduação da Feagri/UNICAMP, rubens.lamparelli@gmail.com.

Recebido pelo Conselho Editorial em: 29-10-2010

Aprovado pelo Conselho Editorial em: 7-5-2012

INTRODUCTION

The agricultural production plays a crucial and strategic role in the economy of Brazil. According to FAOSTAT (2009), the harvested area of soybean in the world, in the crop year of 2005/2006, was 94.93 million hectares with a production of 214.24 million tons, in which Brazil was responsible for respectively 23.23% and 24.49% of this production and area (IBGE, 2008). In Brazil, the state of Paraná was responsible for 48.3% of the planted area and 52.8% of soybean production in the crop year of 2005/2006, indicating the importance of this state in the soybean complex (IBGE, 2008).

Soybean has two important characteristics: short cycle and crops in large areas, requiring care in monitoring and tracking. Remote sensing has proved to be a valuable tool in agricultural monitoring due to a synoptic view and the periodicity for obtaining information concerning large areas of the land surface (LABUS et al., 2002). REES (1990) also pointed out that the application of this tool is related to the monitoring of the extention, vigor and type of vegetation covering. However, it is necessary the knowledge of the spectral pattern of these surfaces, since different targets have different spectral signatures (SMITH, 2001; JENSEN et al., 2002).

In this regard, JIANYA et al. (2008), JENSEN et al. (2002) and FERREIRA et al. (2008) suggested the use of multi-temporal satellite images to study the changes in the Earth's surface. Moreover, one crop presents a high dynamic spectro-temporal feature and requires the monitoring with vegetation indices in multiple dates (HOLBEN, 1986), which has allowed to well describing this characteristic, reflecting the vegetation conditions along its phenological cycle, as shown by various studies (FONTANA et al., 1998; LABUS et al., 2002; RUDORFF et al., 2005; ESQUERDO, 2007; RIZZI & RUDORFF, 2007). One of the most used vegetation indices for this purpose has been the Normalized Difference Vegetation Index (NDVI), proposed by ROUSE et al. (1973), according to the studies of LUNETTA et al. (2006), YI, et al. (2007), WARDLOW & EGBERT (2008), MERCANTE et al. (2009), FERNANDES et al. (2011) and ARAÚJO et al. (2011).

Soybean is an example of this dynamic spectro-temporal feature, making its monitoring more complex when considering all phenological phases. Thus, the evaluation of the NDVI temporal profile of soybean, per municipality, generates a large amount of data which may become a difficult task, since the spectro-temporal pattern may vary according to the location. In this context, the data mining (DM) is a valuable tool because it allows for analyzing large volumes of data, aiming to extract from them useful information (knowledge). According to FAYYAD et al. (1996), DM is the nontrivial process of identifying valid, novel, potentially useful and understandable patterns in data.

According to REZENDE (2005) and LAXMAN & SASTRY (2006), the DM process involves domain knowledge, problem identification, pre-processing, pattern extraction, post-processing and the use of the gained knowledge. During the pre-processing phase, the domain knowledge and the problem identification help in selecting the data set. In the pattern extraction phase, it is defined the DM task, i.e., it should be defined a descriptive activity (association rules, summarization, grouping or clusters) or a predictive activity (classification, regression), according to the desired goals and then the algorithm which will be used for this task. Finally, in post-processing phase, after the selection of the most important or relevant patterns, the knowledge obtained should be used to solve the identified problem. The prediction activities in DM aim decision making process. The generation of clusters is a descriptive task that aims to segment a data set into a number of homogeneous subgroups, which are at the same time, distinctly heterogeneous between each other.

RIE & OSAMU (2001) identified cloudiness information in long temporal series of images through clusters, by using meteorological satellite images. The information about such clusters was inserted in a relational database, which enabled users to make queries. ZHANG et al. (2008) reported almost the same procedures for analyzing time series of meteorological satellite images using DM techniques to improve weather forecast.

Thus, the objective of this study was to group the temporal profiles of 10-day composites NDVI data obtained by the SPOT Vegetation sensor, for main soybean producing municipalities in the state of Paraná, during the 2005/2006 cropping season and to identify homogeneous areas of soybean production regarding sowing dates and vegetative peak.

MATERIAL AND METHODS

The study area was the state of Paraná, in the southern region of Brazil, located between the parallels 22°29'S and 26°43'S and the meridians 48°2'W and 54°38'W (Figure 1). The state has 399 municipalities and, according to IBGE (2008), 363 of them presented some production and/or soybean cultivated area in the 2005/2006 cropping season. A pre-selection was carried in order to exclude some municipalities presenting small planted areas and/or very low production. Thus, a total of 86 municipalities were excluded from the study, considering those with production of less than 1,000tons, and less than 1,000ha of cultivated area or less than 5% of cultivated area, in relation to the municipality area, leaving 277 municipalities with the highest expression regarding soybean production for analysis.

To gather the average NDVI temporal profile of each municipality for the 2005/2006 cropping season, as exemplified in Figure 2, it was necessary to map the crop areas, generating a crop mask(Figure 1). This mask was created from 16-day composites NDVI images of the MODIS sensor (Moderate Resolution Imaging Spectroradiometer) with spatial resolution of 250m (MODIS, 2008). However, from this mask, it was used only the geographical coordinates of pixels to locate the areas with soybean production in the municipalities, since the average NDVI profile was generated from dekadal images of the SPOT Vegetation sensor (VITO, 2008).

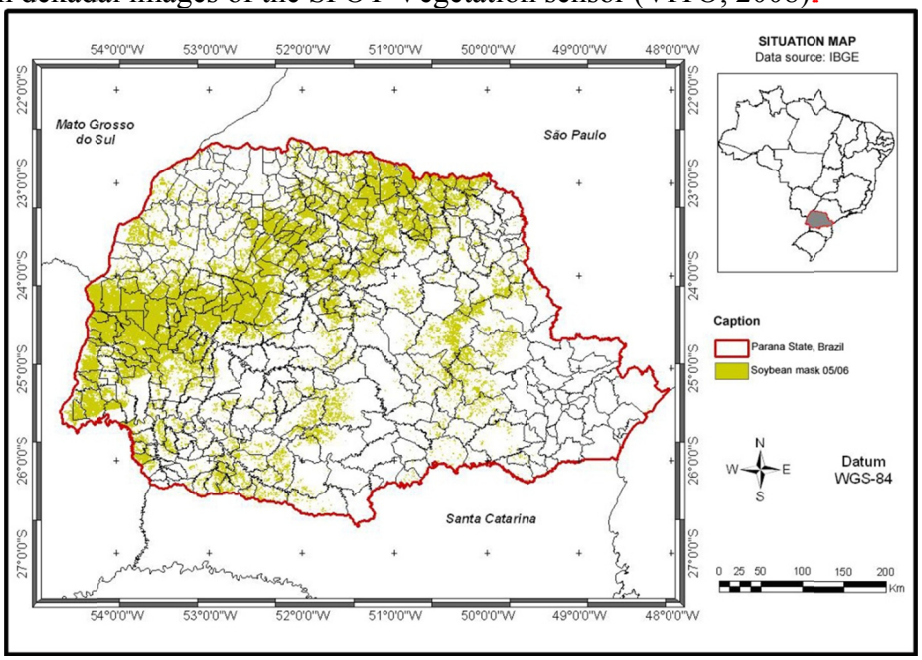
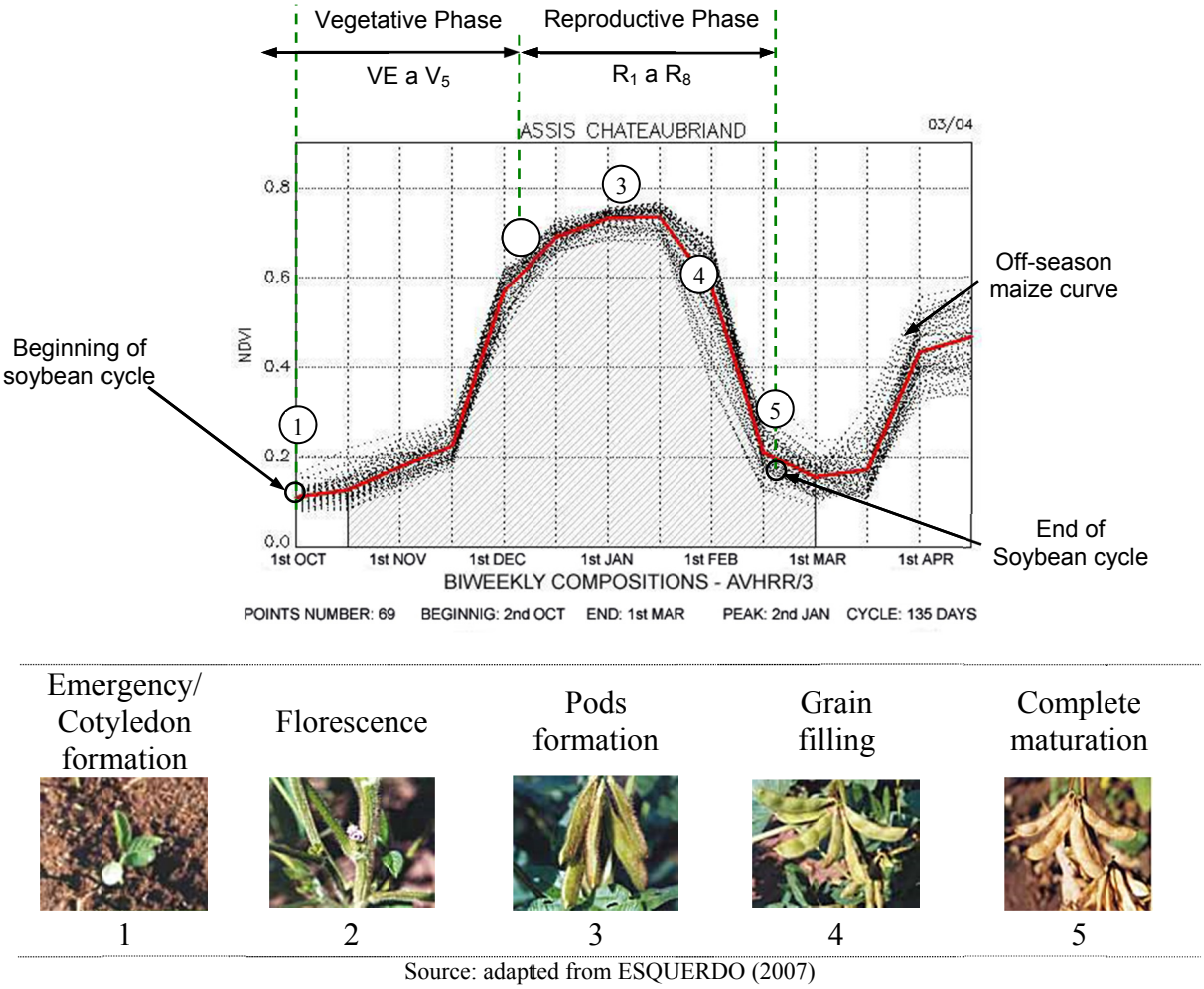


FIGURE 1. Vectors of municipalities in the state of Parana and soybean mask for the 2005/2006 cropping season, generated from MODIS/NDVI images.

To represent this average NDVI profile 1,000 samples (pixels) were randomly selected from each dekadal temporal series for each of the 277 municipalities, since some of them had less than 1,000 pixels of soybean. The system for extracting image data was developed by ESQUERDO et al. (2006) and ESQUERDO (2007), using IDL language within the IDL/ENVI software. The definition of the number of samples per municipality was carried out by simulation and, in tests, it was found that reducing the number of samples hardly changed the dekadal average of its NDVI. Then, these data were exported in spreadsheet format. The software's used in the above procedure were the ENVI 4.5, IDL 6.2, ArcMap 9.2 (ArcGIS) and IDRISI KILIMANJARO 14.2.

The WEKA software (Waikato Environment for Knowledge Analysis) (WITTEN & FRANK, 2005) was used to generate the clusters, based on the municipal average of dekadal NDVI temporal profile. This software aggregates algorithms from different methods/paradigms, to proceed the statistical and computational analysis of the data provided, using DM techniques, which allows to achieve new knowledge, either inductive or deductive.



Source: adapted from ESQUERDO (2007)

FIGURE 2. NDVI temporal profile for Assis Chateaubriand municipality, during the 2003/2004 cropping season.

Among the clustering methods, the partition, density and probabilistic are the main methods (GUIDINI & RIBEIRO, 2006). In the partitioning method, the most used algorithm is the K-Means, proposed by MACQUEEN (1967), which identify classes of objects with similar characteristics, and those closest to a determined centroid are in general determined by Euclidean distance or Manhattan distance. However, the number of clusters must be a priori defined by the analyst, who chooses later the best set of clusters. This is a disadvantage of this method; moreover, this method is sensitive to noise and discrepant values in the data set.

The DBSCAN (Density Based Spatial Clustering of Applications with Noise), proposed by ESTER et al. (1996), is a method to group objects based on density, enabling the discovery of the number of clusters in an arbitrary manner without requiring the user to define it. This algorithm requires that the user define two parameters, the maximum radius of the surrounding area (epsilon = Eps) and the minimum number of points (MinPts) within this radius. Thus, the clusters are dense regions of objects in a data space, which are separated by regions of low density. The objects that lie in that region of low density are generally characterized as outliers or noise, which is the great advantage of this algorithm over the other one.

Among the probabilistic methods, the algorithm Maxver (Expectation-Maximization), proposed by DEMPSTER et al. (1977), also known as Gaussian Mixture, has been most widely

used to group data. It is based on statistics of maximum likelihood to estimate the parameters of the normal distribution. The data is a mixture of ***n*** univariate normal distributions of the same σ^2 variance and the averages of each normal distribution are estimated, i.e, the hypothesis that maximizes the likelihood of such means and, through an iterative process, the clusters are formed.

The purpose was to characterize homogeneous areas of soybean production of the 277 municipalities in the state of Paraná, so an average NDVI temporal profile was generated for each one. Thus, to group these 277 municipalities regarding the NDVI temporal profile, three simulations with different periods were performed to generate the clusters. In the first Simulation (S1) it was considered the entire analysis period (from 1st Sep 2005 to 3rd_May 2006). In the second Simulation (S2) the dekads between September 2005 and May 2006 were removed and the third Simulation (S3) considered only the period between the first dekad of October 2005 (01_Oct 05) and the third dekad of March 2006 (03_Mar 06) to generation the clusters. The main purpose of reducing the amount of dekads was just to adjust the analysis for the period that included the soybean crop cycle in the state.

To generate the clusters, it was used DBSCAN, K-Means and Maxver algorithms with in the WEKA software on mode "use training set". The K-Means and Maxver methods require that the user define the desired number of clusters; however, the Maxver also allows the algorithm to find the number of clusters automatically. Thus, several tests were conducted to find the best number of clusters to group the 277 municipalities. The DBSCAN method, which determines the number of clusters automatically, the MinPts was set in six for a single cluster and in each group (default) and the Eps ranged from 0.1 to 2.0.

RESULTS AND DISCUSSION

In Figure 3, it is shown, as an example, the NDVI temporal profile average and the NDVI coefficient of variation (CV_NDVI) of the 2005/2006 cropping season of the municipality of Marechal C. Rondon - Paraná (PR), Brazil.

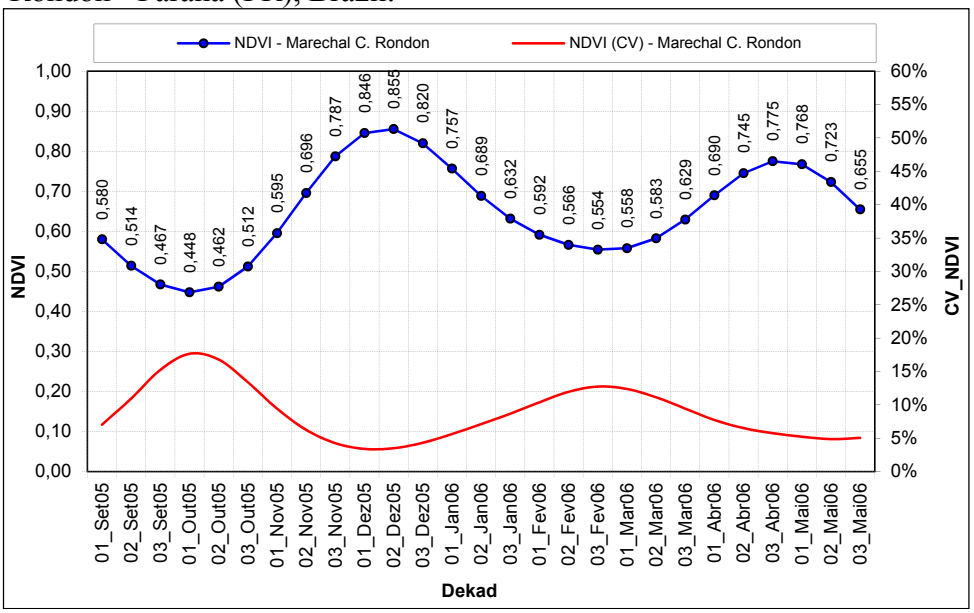


FIGURE 3.NDVI temporal profile and its corresponding coefficient of variation (CV_NDVI) of Marechal C. Rondon, during the 2005/2006 cropping season.

It can be seen that in late September, and more specifically, in the first dekad of October 2005 (01_Oct 05), the lowest value of NDVI (0.446) occurred, since it is the period just after winter, when vegetation is dry and soil is uncovered, leading to a low reflectance, hence justifying the low value of NDVI. This period represents the first phenological phase of the soybean crop in the municipality that involves sowing, seed germination and initial development, corroborating what ADAMI (2010) defined. It was found that low NDVI values represent high coefficient of variation

values (CV_NDVI), since at this phase there is a great diversity of crops: some are being prepared for sowing, others are in seed germination phase and others are in the initial development phase, justifying the high value of CV_NDVI. Following, NDVI values increase due to the second phenological phase (green cover dominance, when the crop is in the vegetative, flowering, pod formation and/or grain filling phase). In this phase, peak vegetative or Maximum Vegetative Development (MVD) is reached, with maximum values of NDVI (0.856) in the second dekad of December 2005 (02_Dec05); in the same period, the lowest CV_NDVI value occurs, since the majority of farms in the municipality are in the vegetative phase, hence justifying the low variability of NDVI at this time. The MVD usually occurs between the phenological phases R1 (beginning of flowering) and R3 (beginning of pod formation) of soybean development (ADAMI, 2010), which is one of the most important periods to define the final crop yield. This demonstrates the importance of knowing it all over the state.

The last crop phenological phase, maturation, senescence and desiccation of leaves, can be identified by the reduction of NDVI values. This is due to the effect of exposure of dry vegetation and soil. Further increase in CV_NDVI reinforces this phase identification. For this municipality and year studied, the lowest NDVI (0.553) occurred in the third dekad of February 2006 (03_Feb06). This indicates the end of summer crops cycle and the beginning of winter crops sowing. In the municipality evaluated, it is characteristic farmers to sow off-season maize (maize crop during winter) soon after summer crops harvest. This justifies, in Figure 3, the identification of a new crop cycle, with vegetative peak in the third dekad of April 2006 (03_Apr06) and subsequent decrease of NDVI values after this phase, similar to what occurred with the soybean crop.

To achieve the clusters, 277 NDVI temporal profiles (one for each county) were considered, similar to those discussed in Figure 3. They were grouped by the DBSCAN, K-Means and Maxver methods for the three simulations periods (S1: 01_Sep05 to 03_May06; S2: 01_Oct05 to 03_Mar06; S3: 01_Oct05 to 03_Mar06). Results show that regardless of the clustering method, different configurations (in terms of clusters) were found for simulation periods S1, S2 and S3.

Results found by using the DBSCAN method varying Eps from 0.1 to 2.0 are presented in Table 1. For Eps≥0.7, all the 277 municipalities were grouped into a single cluster. When Eps=0.1, municipalities were considered outliers for all simulations periods. For Eps values between 0.2 and 0.5, municipalities groups varied from one cluster to multiple clusters, depending on the period to be considered. However, it is worth observing that certain municipalities were grouped as outliers in most simulations. This suggests that these municipalities have NDVI temporal profile remarkably different from the other municipalities. In general, this algorithm did not have good performance with this database.

Concerning simulation, a procedure similar to DBSCAN was performed for the K-Means and Maxver algorithms. The main difference between both was to determine a desired number of clusters for the latter. Different numbers of clusters were tested, aiming to find the best grouping of municipalities. In order to validate these results, for each number of clusters generated for both K-Means and Maxver, graphs of average NDVI profile of municipalities grouped in each cluster were generated for comparative analysis. For example, when two clusters have been defined (for K-Means or Maxver), all the 277 municipalities have been represented in two graphs (cluster0 and cluster1). Similar graphs were generated for different number of clusters for the three simulation periods. Other method used to analyze the results was to create a classification method in WEKA with four algorithms: J48 (decision tree), SMO (support vector machine), Multilayer Perceptron (neural networks) and Naive Bayes (probabilistic model). For the classification task, the NDVI data set of defined periods in each simulation were used as predictive attributes and a target attribute was created, whose values were labels of the clusters generated in the previous phase. For example, for the first simulation (S1), 28 attributes (NDVI from 01_Sep05 to 01_May06 and cluster) were considered. Similar procedures were used for the other two simulations (S2 and S3).

TABLE 1. Number of clusters generated by DBSCAN method for the three defined periods of simulation.

Epsilon (Eps)	S1 (01_Sep05 a 03_May06)	S2 (01_Oct_05 a 03_May06)	S3 (01_Oct05 a 03_May06)
0.1	Outliers=277	Outliers=277	Outliers=277
0.2	Outliers=277	Outliers=277	C0=10; C1=5; C2=6; C3=6; Outliers=250
0.3	C0=7; C1=9; C2=9; Outliers=252	C0=24; C1=36; C2=17; C3=16; C4=8; C5=7; Outliers=169	C0=181; C1=7; C2=7; Outliers=82
0.4	C0=58; C1=58; C2=65; C3=6; Outliers=90	C0=232; Outliers=45	C0=244; C1=5; C2=6; Outliers=12
0.5	C0=238; Outliers=39;	C0=260; Outliers=17	C0=273; Outliers=4
0.6	C0=272; Outliers=5	C0=275; Outliers=2	C0=277
0.7 to 2.0	C0=277	C0=277	C0=277

Legend: C0 = *cluster*0; C1 = *cluster*1; C2 = *cluster*2; C3 = *cluster*3; C4 = *cluster*4; C5 = *cluster*5; *outlier* = noise.

In a general way, different clusters were found for these two methods (K-Means and Maxver) and the three period simulations performed. Among the analysis, the best results were found for the third simulation (S3: 01_Oct05 to 03_Mar06). In Tables 2 and 3 the results of classification methods are summarized, respectively, for the grouping method Maxver and K-Means. It can be observed that most classification algorithms presented a good performance regardless of the number of clusters used. This generated a difficulty to determine the ideal number of clusters for the 277 municipalities. In order to determine the best clustering algorithm and the best number of clusters, the results of the classification methods and graph analysis of the average NDVI behavior profile. Thus, six was found to be the best result regarding clustering and K-Means considered the best algorithm.

TABLE 2. Proportion of correctly classified instances from the Maxver algorithm achieved in the third simulation (NDVI between 01_Oct05 to 03_Mar06) for the 277 municipalities in the State of Paraná, Brazil.

Number of Clusters	Algorithm I J48 (C4.5)	Algorithm II SMO	Algorithm III MultilayerPerceptron	Algorithm IV NaiveBayes
14 (default)	88.45%	94.59%	99.64%	100.00%
2	99.28%	99.28%	100.00%	100.00%
3	94.95%	98.56%	100.00%	100.00%
4	93.50%	97.47%	100.00%	100.00%
5	93.50%	96.39%	100.00%	100.00%
6	94.59%	96.75%	100.00%	100.00%
7	94.95%	97.11%	100.00%	100.00%
8	93.14%	97.11%	100.00%	100.00%

TABLE 3. Proportion of correctly classified instances from the algorithm K-Means achieved in the third simulation (NDVI between 01_Oct05 to 03_Mar06) for the 277 municipalities in the state of Paraná.

Number of Clusters	Algorithm I J48 (C4.5)	Algorithm II SMO	Algorithm III MultilayerPerceptron	Algorithm IV NaiveBayes
2	96.03%	100.00%	100.00%	98.20%
3	94.59%	99.28%	100.00%	96.03%
4	92.42%	99.64%	100.00%	95.31%
5	89.89%	96.03%	100.00%	95.67%
6	94.22%	94.59%	100.00%	96.39%
7	89.53%	94.95%	100.00%	98.20 %
8	90.25%	95.31%	100.00%	98.20 %

Table 4 presents a contingency table where can be seen the overall accuracy (EG) between K-Means and Maxver, both with six clusters. EG is defined as the sum of the main diagonal divided by the total number of registers in the data (n=277). In this case, the EG was 84.48%, showing that although while the two methods have different heuristics grouping, there was a high degree of similarity between them. Among the simulations, except for the grouping with two clusters (EG=93.50%), the other results were worse than with six clusters, defined as the best.

TABLE 4. Contingency table with six clusters between the K-Means and Maxver algorithms for the third simulation (NDVI between 01_Oct05 to 03_Mar06) for the 277 municipalities in the state of Paraná.

K-Means (6)	Maxver(6)						
	cluster4	cluster0	cluster3	cluster1	cluster5	cluster2	General Total
cluster0	19	-	-	-	-	-	19
cluster1	-	56	9	-	12	-	77
cluster2	-	-	38	-	1	3	42
cluster3	3	-	-	55	-	2	60
cluster4	-	-	-	11	33	1	45
cluster5	-	-	1	-	-	33	34
General Total	22	56	48	66	46	39	277

Figures 4-9 show the graphs of the average NDVI profile between the first dekad of October 2005 (01_Oct05) and the third dekad of March 2006 (03_Mar06) for the K-Means method with six clusters. Main differences between the graphs for each of the six clusters, were for NDVI values at the beginning of the crop cycle (01_Oct05: highlighted with red brackets on the Y axis) and vegetative peak (high NDVI values: blue rectangle highlighted in the graphs). In general, the clusters 0; 3 and 4 (Figures 4; 5; 6) showed temporal profiles that are more similar to soybean crop. For the clusters 1; 2 and 5 (Figures 7; 8; 9) NDVI profiles presented less variation throughout the crop cycle. As can be seen in Figure 11, the clusters 0; 3 and 4 coincide, mostly, with the soybean belt mask showed in Figure 1, i.e., municipalities most representatives of soybean production and planted area in 2005/2006 crop season.

An agricultural zoning with the regionalization of cultivars and soybean seeding periods was proposed by KASTER & FARIAS (2011). Thus, the state of Paraná was subdivided into two macro regions (1 and 2) and five micro regions of soybean crop (MRS 103, MRS 104, MRS 201, MRS 202 and MRS 203), as illustrated in Figure 11. In this new zoning, the classification of cultivars was organized by Relative Maturity Groups (GMR) proposed by ALLIPRANDINI (2005) and the Number of Days to Maturity (NDM). Because of this, soybean cultivars were grouped into three groups (G1-G3) for macro regions 1 and 2. For macro region 1 (MRS 103 and MRS 104) the following characteristics were established: G1 - short cycle, with NDM $\leq$ 130 days and GRM $\leq$ 6.3; G2 - average cycle, with 131 $\leq$ NDM $\leq$ 145 and 6.4 $\leq$ GMR $\leq$ 7.4; G3 - long cycle with NDM $\geq$146 and GRM $\geq$ 7.5. For macro region 2 (MRS 201, MRS 202, MRS 203) these characteristics were established as follows: G1 (NDM $\leq$ 125 days and GRM $\leq$ 6.7), G2 (126 $\leq$ NDM $\leq$ 135 and 6.8 $\leq$ GMR $\leq$ 7.6) and G3 (NDM $\geq$ 136 and GRM $\geq$ 7.7). The following soybean seeding periods were established : Oct/21 to Nov/30 (MRS 103), Oct/21 to Dec/10 (MRS 104), Oct/01 to Nov/30 (MRS 201) and Nov/10 to Nov/30 (MRS 202 and 203).The cluster0 had 19 municipalities with a vegetative peak between 01_Dec05 and 03_Dec05 and with NDVI values at the beginning of the crop, ranging from 0.25 to 0.45 (Figure 4). It is possible to observe in Figure 11 that they are grouped in the western region of the state (MRS 201), more specifically, in the Lakes Region (Lake Itaipu), municipalities which has the characteristic of plant off-season maize, which explains the early planting and, consequently, the fact that the vegetative peak (or MVD related to the phases R1 to R3 of soybean) occur before in comparison to other regions, supporting the recommendations of planting dates from 10/01 given by KASTER & FARIAS (2011) and with the results found by ARAÚJO et al. (2011) who mapped the areas with summer crops in state of Paraná, using images

from SPOT Vegetation and rainfall data of the municipalities. This information was also confirmed by technicians of the Agricultural Research Center Cooperative (COODETEC) based in Cascavel-PR, one of the largest producers of soybean seeds in Brazil.

For cluster3, shown in Figure 5, with 60 other municipalities, the difference was due to the date of the vegetative peak, which was between 03_Dec05 and 02_Jan06, showing a delay in the development of soybean in these municipalities comparing with cluster0. It is possible to observe in Figure 11, that part of these municipalities are located in the MRS 201 and 202, however, most in MRS 203 which has recommendation of sowing from 10/11 (KASTER & FARIAS, 2011) justifying this delay in vegetative peak culture.

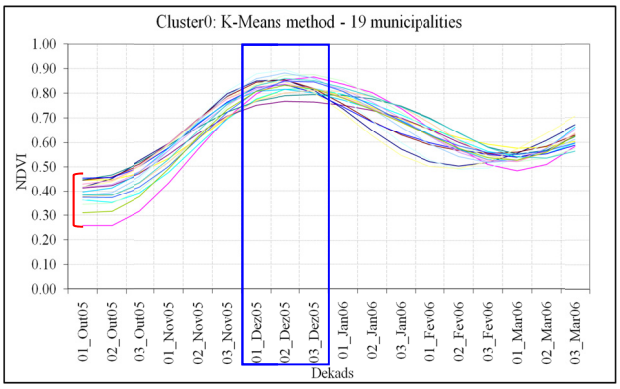


FIGURE 4. Average NDVI temporal profile behavior (01_Oct05 to 03_Mar06) of the 19 municipalities of cluster0 (K-Means).

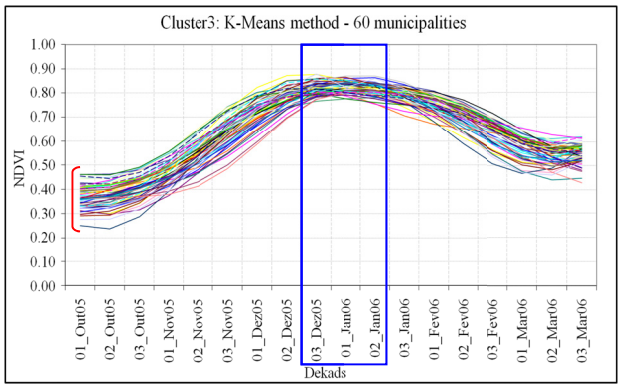


FIGURE 5. Average NDVI temporal profile behavior (01_Oct05 to 03_Mar06) of the 60 municipalities of cluster3 (K-Means).

For cluster4, with exception of the municipalities of Tibagi, Pirai do Sul and Carambei, located in MRS 104; the municipality of Roncador and Nova Cantu, located at MRS 103; and the municipality of Moreira Sales, São Manoel do Parana and Boa Esperança, located at MRS 202; the other 37 municipalities are located in MRS 203 (Figure 11). However, unlikely the municipalities grouped in cluster3, these had NDVI values higher at the beginning of the crop cycle (0.38 <NDVI <0.52) and a vegetative peak more late and long, ranging from 01_Jan06 and 02_Feb06 (Figure 6 and Figure 10). According to information obtained from the technicians of COODETEC, this MRS 203, especially the region closest to the border with the state of São Paulo, is characterized as a drier region, which makes farmers opt for medium cycle cultivars, which have longer flowering and grain filling, reducing the possibilities of reduced productivity for the eventual Indian summers, normal in this region.

Figure 7 shows the graph with the NDVI profiles of 77 municipalities, grouped in cluster1. The majority of these municipalities were located in the MRS 104 and in the eastern part of the MRS 103 (Figure 11). The beginning of the development of NDVI was high (between 0.45 and 0.73) with vegetative peak ranging 03_Jan06 and 03_Feb06, which is justified according KASTER & FARIAS (2011) recommendations of sowing until 12/10, the largest in the state, which consequently made the vegetative peak later, as demonstrated in Figures 7 and 10. Moreover, in the border between these two micro regions (103 and 104) are the highest altitudes of the state (above 1,000m) (VALERIANO & ABDON, 2007), which makes the month of October cold for sowing soybeans, according to the information gathered from the technicians of COODETEC, justifying the later sowing.

The 42 municipalities grouped in cluster2 (Figure 8) had NDVI values ranging from 0.70 to 0.80 at the vegetative peak, moreover, for a longer period (between the 02_Dec05 and 02_Feb06) when compared with the other clusters. These municipalities, in their great majority, are located on the border of the MRS 201 and 103 and west of MRS 103. And finally, the remaining 34 municipalities were grouped in cluster5 (Figure 9), which like the previous cluster, showed low NDVI values (<0.80) in the vegetative peak, but with a smaller amplitude (02_Dec05 to 01_Jan06)

that the cluster2. Practically all municipalities are located in MRS 202, northwest of the state (Figure 11), which has a characteristic of agriculture and intensive cultivation of sugar cane. Thus, this NDVI profile with small variation along the phenological cycle may be generated by pixels mixing problems in the generation of masks of culture, since the soybean large planting fields lies between the areas of sugar cane.

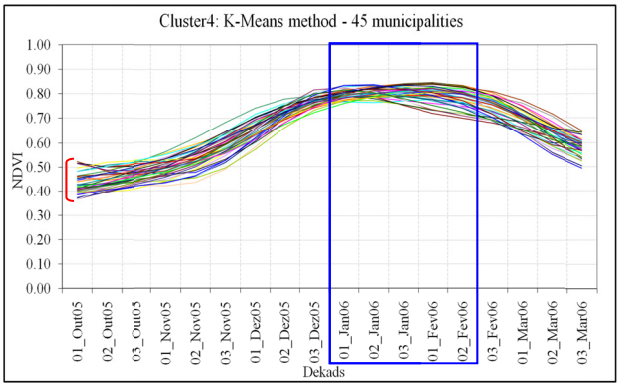


FIGURE 6. Average NDVI temporal profile behavior (01_Oct05 to 03_Mar06) of the 45 municipalities of cluster4 (K-Means).

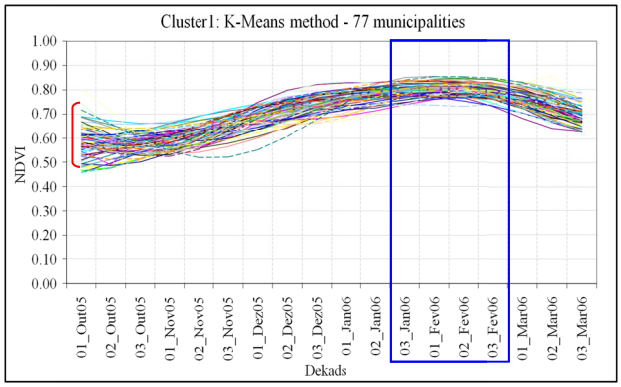


FIGURE 7. Average NDVI temporal profile behavior (01_Oct05 to 03_Mar06) of the 77 municipalities of cluster1 (K-Means).

Figure 10 shows a graph of the average NDVI temporal profile for each cluster, i.e., to cluster0 the average temporal profile of its 19 municipalities is shown, and so on for the other clusters. This graph shows the reason for the different clusters adjusted by K-Means algorithm. Considering as sowing date, according to ADAMI (2010), the point where the curve of the temporal series of NDVI begins to rise, with the Figure 10 it is possible to say that with the exception of cluster1, where the decennial of sowing occurred in 03_Oct_2005, for the other clusters it occurs in the first decennial of October 2005 (01_Oct_2005). Which may explain the large differences of dates of vegetative peak found are basically the use of cultivars with different cycles and also that each municipality profile, shown in Figure 4-9 and summarized by the average of each cluster in Figure 10, is the average of NDVI values of all soybean large planting fields, within each municipality.

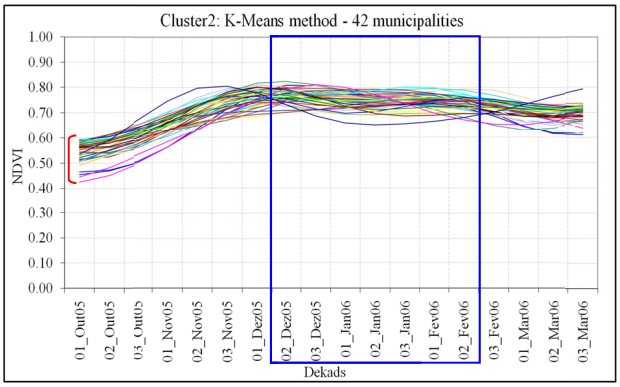


FIGURE 8. Average NDVI temporal profile behavior (01_Oct05 to 03_Mar06) of the 42 municipalities of cluster2 (K-Means).

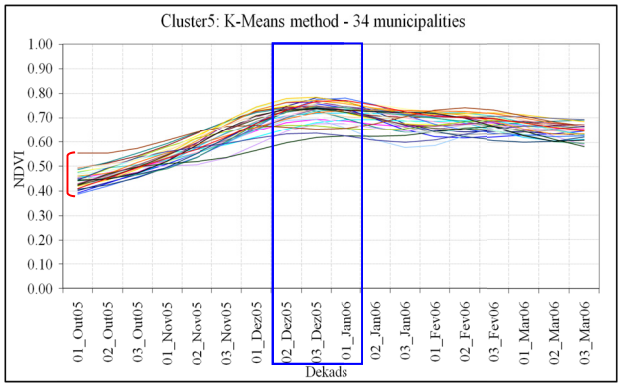


FIGURE 9. Average NDVI temporal profile behavior (01_Oct05 to 03_Mar06) of the 34 municipalities of cluster5 (K-Means).

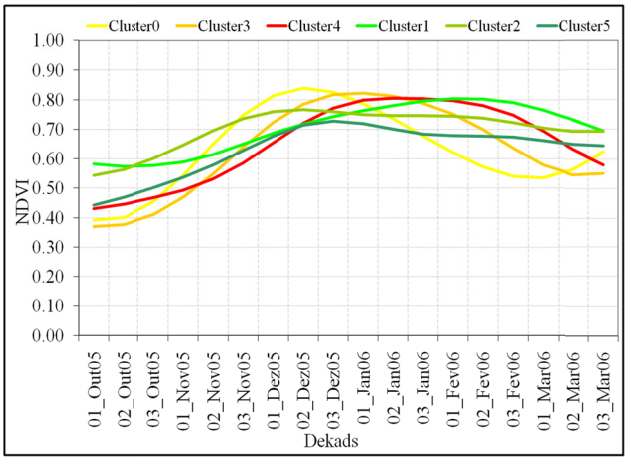


FIGURE 10. Average NDVI temporal profile behavior of municipalities in each cluster between 01_Oct05 to 03_Mar06 by K-Means method.

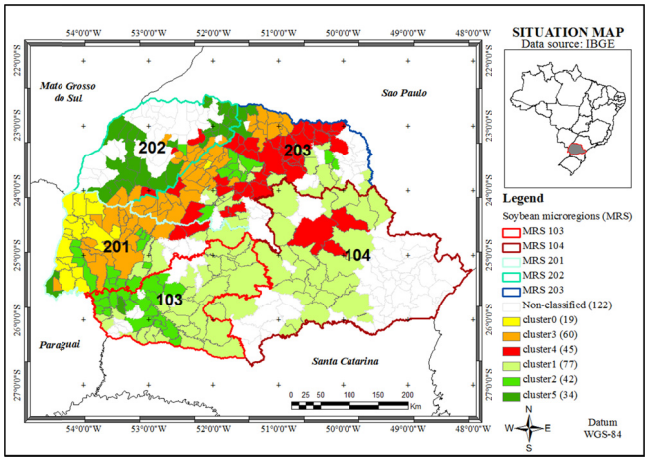


FIGURE 11. Map of spatial distribution of municipalities by six clusters generated by K-Means method.

CONCLUSIONS

The average temporal profile of NDVI by municipality showed different behavior patterns when compared to the vegetation index in the state of Paraná.

It was found that the coefficient of variation of NDVI (CV_NDVI) can be used to indicate the beginning of the sowing of crops, as well as to emphasize their vegetative peak. This information is important because the occurrence of hydric deficiency in the phases of sowing, germination of crops, flowering, pod formation and/or grain filling may indicate possible reductions in productivity, important diagnostic for crop forecast by municipality and/or state.

Regarding clusters, the DBSCAN algorithm was not effective in any simulation performed. However, the clustering algorithms K-Means and Maxver showed high rates of overall accuracy, indicating that although they have different heuristics for generating clusters, the results for this case study are convergent.

The municipalities grouped in clusters 0;3 and 4 presented a NDVI temporal profile which is characteristic of soybean and in general they were located in the soybean belt, or on the mesoregions 201 and 203 that produces soybean in the state of Paraná, in the crop year of 2005/2006, reaveling the promising results found by the proposed methodology. Moreover, it was possible to demonstrate that the data mining techniques were effective in the identification of homogeneous areas of soybean production in the state of Paraná.

In general, the results obtained from the identification of homogeneous areas, in terms of NDVI, may be useful for generating masks of summer crops more suited to sowing calendar, by homogeneous areas, contributing to improve the estimated planted area of such crops in the state of Paraná.

REFERENCES

ADAMI, M. *Estimativa da data de plantio da soja por meio de séries temporais de imagens MODIS*. 2010. 163f. Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2010.

ALLIPRANDINI, L.F. Proposta de nova classificação das cultivares de soja segundo grupos de maturação. In: REUNIÃO DE PESQUISA DE SOJA DA REGIÃO CENTRAL DO BRASIL,27., 2005.*Resumos*...Londrina: Embrapa Soja, 2005. p.116-123.

ARAÚJO, G. K. D.; ROCHA, J. V.; LAMPARELLI, R. A. C.; ROCHA, A. M. Mapping of summer crops in the state of Paraná, Brazil, through the 10-day spot vegetation NDVI composites. *Engenharia Agrícola*, Jaboticabal, v.31, p.760-770, 2011.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, v.39, n.1, p.1-38, 1977.

ESQUERDO, J. C. D. M. *Utilização de perfis multi-temporais do NDVI/AVHRR no acompanhamento da safra de soja no oeste do Paraná*. 186f. Tese (Doutorado em Engenharia Agrícola) - Universidade Estadual de Campinas, Campinas, 2007.

ESQUERDO, J. C. D. M.; ANTUNES, J. F. G.; BALDWIN, D. G.; EMERY, W. J.; ZULLO JÚNIOR, J. An automatic system for AVHRR land surface product generation. *International Journal of Remote Sensing,* Basingstoke, v.27, p.3925-3942, 2006.

ESTER, M.; KRIEGEL, H. P.; SANDER, J.; XUI, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2., 1996. Portland.

FAOSTAT - FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. *ProdSTAT – Crops*. 2009. Disponível em: <http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor>. Acesso em: 5 jun. 2009.

FAYYAD, U.; SHAPIRO, G. P.; SMYTH, P. Knowledge discovery and data mining: towards a unifying framework. In: PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING,2.,1996. Portland.

FERNANDES, J. L.; ROCHA, J. V.; LAMPARELLI, R. A. C. Sugarcane yield estimates using time series analysis of spot vegetation images. *Scientia Agrícola*, Piracicaba, v.68, p.139-146, 2011.

FERREIRA, L. G.; FERREIRA, N.C.; FERREIRA, M.E. Sensoriamento remoto da vegetação: evolução e estado-da-arte. *Acta Scientiarum*. *Biological Sciences*, Maringá, v.30, p.379-390, 2008.

FONTANA, D. C.; BERLATO, M. A.; BERGAMASCHI, H. Relação entre o índice de vegetação global e condições hídricas no Rio Grande do Sul. *Pesquisa Agropecuária Brasileira*, Brasília, v.33, n.8, 1399-1405, 1998.

GUIDINI, M.; RIBEIRO, C. Utilização da biblioteca TerraLib para algoritmos de agrupamento em Sistemas de Informações Geográficas. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA, 8., 2006, Campos do Jordão. *Anais*... São José dos Campos: INPE, 2006.

HOLBEN, B.N. Characteristics of maximum value composite images from temporal AVHRR data. *International Journal of Remote Sensing,* Basingstoke, v.7, n.11, p.1417-1435, 1986.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Banco de dados agregados - Sistema IBGE de Recuperação Automática – SIDRA*. 2008. Disponível em: <http://www.sidra.ibge.gov.br>. Acesso em: 9 jun. 2008.

JENSEN, J.R.; BOTCHWAY, K.; BRENNAM-GALVIN, E.; JOHANNSEN, C.J.; JUMA,C.; MABOGUNJE, A.; MILLER, R.; PRICE, K.; REINING, P.; SKOLE, D.; STANCIOFF, A.; TAYLOR, D.R.F. *Down to Earth:* Geographic information for sustainable development in Africa. Washington: National Academy Press, 2002. 155p.

JIANYA, G.; HAIGANG, S.; GUORUI, M.; QIMING, Z. A review of multi-temporal remote sensing data change detection algorithms. In: THE INTERNATIONAL ARCHIVES OF THE PHOTOGRAMMETRY, REMOTE SENSING AND SPATIAL INFORMATION SCIENCES, 37.,Beijing, 2008. p.757-762.

KASTER, M.; FARIAS, J.R.B. Regionalização dos testes de VCU - Valor de Cultivo e Uso de cultivares de soja - terceira aproximação. In: REUNIÃO DE PESQUISA DE SOJA DA REGIÃO CENTRAL DO BRASIL, 37., 2011, São Pedro. *Anais*...

LABUS, M. P.; NIELSEN, G. A.; LAWRENCE, R. L.; ENGEL, R.; LONG, D. S. Wheat yield estimates using multi-temporal NDVI satellite imagery. *International Journal of Remote Sensing,* Basingstoke, v.23, n.20, p.4169-4180, 2002.

LAXMAN, S.; SASTRY, P. S. A survey of temporal data mining. *Sadhana Academy Proceedings in Engineering Sciences*, Bangalore, v.31, n.2, p.173-198, 2006.

LUNETTA, R.S; KNIGHT, J.F.; EDIRIWICKREMA, J.; LYON, J.G.; WORTHY, D.L. Land-cover change detection using multi-temporal MODIS NDVI data. *Remote Sensing of Environment*, New York, v.105, p.142-154, 2006.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTIC AND PROBABILITY,5., 1967, Berkley. *Proceedings*... Berkley: University of California Press, Berkley, 1967. p.281-297.

MERCANTE, E.;LAMPARELLI, R.A.C.; URIBE-OPAZO, M.A.; ROCHA, J.V.. Características espectrais da soja ao longo do ciclo vegetativo com imagens landsat 5/TM em área agrícola no oeste do Paraná. *Revista Engenharia Agrícola*, Jaboticabal, v.29, p.328-338, 2009.

MODIS - *MODerate Resolution Imaging Spectroradiometer*. 2008. Disponível em: <http://modis.gsfc.nasa.gov>. Acesso em: 15 set. 2008.

MOREIRA, M.A. *Fundamentos do sensoriamento remoto e metodologias de aplicação*. São José dos Campos: INPE, 2001. 320p.

REES, W. G. *Physical principles of remote sensing.* Cambridge: Cambridge University Press, 1990. 247 p.

REZENDE, S. O. *Sistemas inteligentes:* fundamentos e aplicações. Barueri: Manole, 2005. 525p.

RIE, H,; OSAMU, K. Temporal rule discovery for time-series satellite images and integration with RDB. Congrès PKKD 2001: principles of data mining and knowledge discovery. In: EUROPEAN CONFERENCE ON PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY, 5., 2001, Freiburg. *Proceedings...* v.2168, p.204-215, 2001.

RIZZI, R.; RUDORFF, B. F. T. Imagens do sensor MODIS associadas a um modelo agronômico para estimar a produtividade de soja. *Pesquisa Agropecuária Brasileira*, Brasília, v.42, n.1, p.73-80, 2007.

ROUSE, J. W.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the Great Plains with ERTS. In: EARTH RESOURCES TECHNOLOGY SATELLITE - SYMPOSIUM, 3., 1973, Washington. *Proceedings*... Washington: NASA, 1974. v.1, p.309-317.

RUDORFF, B. F. T.; BERKA, L. M. S.; MOREIRA, M. A.; DUARTE, V.; XAVIER, A. C.; ROSA, V. G. C.; SHIMABUKURO, Y. E. Imagens de satélite no mapeamento e estimativa de área de cana-de-açúcar em São Paulo: ano safra 2003/04. *Agricultura em São Paulo,* São Paulo, v.52, n.1, p.21-39, 2005.

VALERIANO, M.de M.; ABDON, M.de M. Aplicação de Dados SRTM a estudos do Pantanal. *Revista Brasileira de Cartografia,* Rio de Janeiro,v.59, n.1, p.63-71, Abr. 2007.

VITO. *SPOT Vegetation* - Normalized Difference Vegetation Index (NDVI). 2008. Disponível em: <http://www.vgt.vito.be>. Acesso em: 15 set. 2008.

WARDLOW, B.D.; EGBERT, S.L. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. *Remote Sensing of Environment,*v.112, p.1096-1116, 2008.

WITTEN, I. H.; FRANK, E. *Data mining:* practical machine learning tools and techniques. 2.ed. São Francisco: Morgan Kaufmann, 2005. 525p.

YI, J. R.; SHIMABUKURO, Y. E.; QUINTANILHA, J. A. Identificação e mapeamento de áreas de milho na região sul do Brasil utilizando imagens MODIS. *Revista Engenharia Agrícola,* Jaboticabal, v.27, n.3, p.753-763, set/dez. 2007.

ZHANG, Z.; WU, W.; HUANG, Y. Effective spatio-temporal analysis of remote sensing data. In: PROGRESS research and development. Berlin: Springer, 2008. p.584-589. Disponível em: <http://springerlink.com/content/xm577273235k8661>. Acesso em: 16 mar. 2009.