engenharia agrícola
Journal of the Brazilian Association of Agricultural Engineering
ISSN: 1809-4430 (on-line)

DATA MINING TECHNIQUES FOR SEPARATION OF SUMMER CROP BASED ON SATELLITE IMAGES[1]

Doi:http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v37n4p750-759/2017

WILLYAN R. BECKER[2], JERRY A. JOHANN[3*], JONATHAN RICHETTI[2],

LAÍZA C. DE A. SILVA[2]

[3*]Corresponding author. Western Paraná State University (UNIOESTE)/ Cascavel - PR, Brazil.
E-mail: jerry.johann@hotmail.com

ABSTRACT: Due to the difficulty in discriminating soybean and corn in mappings obtained by the time series of satellite images, this study aimed to apply the data mining techniques to separate soybean and corn. Pure pixels selection from Landsat-8 were extracted and used to build a standard spectro-temporal EVI profile for both crops. These profiles were obtained with the Timesat software and, further incorporated in the Weka software. Five out of eleven variables of the standard spectro-temporal EVI profile for each crop were found through the decision tree, a data mining technique. These five variables were sufficient to achieve the separation of soybean and corn crops with an accuracy of 96.3% and a kappa index of 0.92.

KEYWORDS: corn, EVI, J48, soybean, Weka.

INTRODUCTION

Brazilian agricultural production stands out as a significant part of the country's economy, accounting for 21% of Gross Domestic Product (GDP) in 2015. This contributes to the development of regions, increasing the growth of agroindustry, and grain processing units. According to CONAB (2015), Brazil is the second largest producer of soybeans and corn exporter in the world.

The methodology of crop forecasting made by official organizations, as IBGE (Brazilian Institute of Geography and Statistics) (2002) and SEAB (Paraná Department of Agriculture) (2016), is widely used, however, optimizing the forecasting with new methods is important increse the objetivity of the process (Johann et al., 2012).

In this sense, many researches have used series of spectral-temporal vegetation indexes, obtained from satellite images (Esquerdo et al., 2011; Souza et al., 2015; Grzegozewski et al., 2016; Johann et al., 2016), to create agricultural mappings. However, soybean and corn crops have some similarities as the spectral signatures and the vegetative cycle, resulting in difficulties to map these cultures. To Johann et al. (2016), this confusion in mappings leads to errors in the quantification of cultivated areas. Grzegozewski et al. (2016) and Souza et al. (2015) have made important advances using vegetation indexes to map and separate summer crops in Brazil, but in crop years that sowing dates of both cultures are close the separation difficulty persists.

The automatic identification of cultivated areas is one of the most important steps in the crop forecasting process (Nonato & Oliveira, 2013). The improvement in the estimate of cultivated area of each crop directly influences the agricultural production estimates of the respective crop year (Assad et al., 2007) making decisions regarding the commodity harder to do.

One line of study that has recently been addressed to overcome the challenges in improving mappings is the use of data mining techniques (Souza et al., 2010; Nonato & Oliveira, 2013). Data mining is the main step in the process of Knowledge Discovery in Databases (KDD) and aims to find relationships or hidden patterns in databases (Fayyad et al., 1996). For Fayyad et al. (1996), the knowledge discovery is a sequence of an iterative processes that ensues the following steps: 1 -

[1] Research developed in the discipline of "Data Mining and Knowledge Discovery" studied in the Graduate Program in Agricultural Engineering - PGEAGRI, from Western Paraná State University (UNIOESTE) – Cascavel Campus - PR/ BR, in 2015.

[2] Western Paraná State University (UNIOESTE)/ Cascavel - PR, Brazil.

Received in: 7-1-2016
Accepted in: 3-1-2017

Preprocessing (cleaning, integration, selection), 2 - Data transformation (extraction, normalization), 3 – Data mining (application of algorithms in search of implicit and useful knowledge), 4 - Assessment and interpretation (evaluation of the standards obtained and presentation of knowledge).

The decision tree is a data mining technique used for classifying and predict through machine learning (Tan et al., 2009). For this, the construction of standards with the training data is made and from the tree obtained it is possible to classify new unknown samples. Such trees consist of an internal hierarchy and external nodes that are connected by branches. Through a logical test, each node represents a decision on a variable that branches to the next descendant node or to the final result (Crivelenti et al., 2009; Nonato & Oliveira, 2013).

Thus, the aim of this research was to apply the data mining technique called decision tree on variables (attributes) obtained from spectral-temporal profiles of the EVI from the MODIS' sensor in order to do a separation at corn and soybean areas in the state of Paraná.

MATERIAL AND METHODS

The methodology followed the steps established by Fayyad et al. (1996): preprocessing, transformation, data mining, evaluation and interpretation (FIGURE 1).

Data

The study area comprises the State of Paraná (FIGURE 2), located in the southern region of Brazil, which has 399 municipalities, subdivided into 10 mesoregions. To select pure pixels, it was used satellite images from Landsat-8. The 221/77, 222/76, 222/77, 222/77, 222/78 and 223/77 path/row from the OLI sensor (Operational Land Imager) with spatial resolution of 30 meters were used.

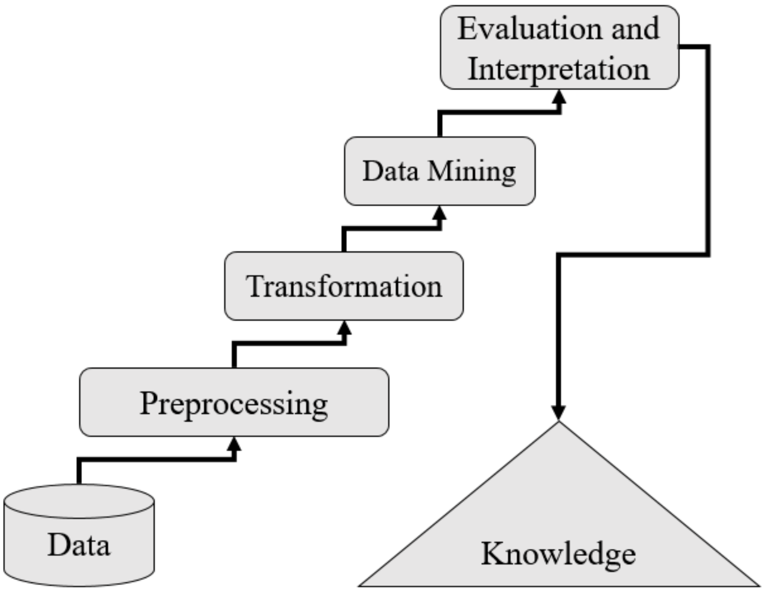

FIGURE 1. Knowledge Discovery in Databases process. Source: Adapted from Fayyad et al. (1996).

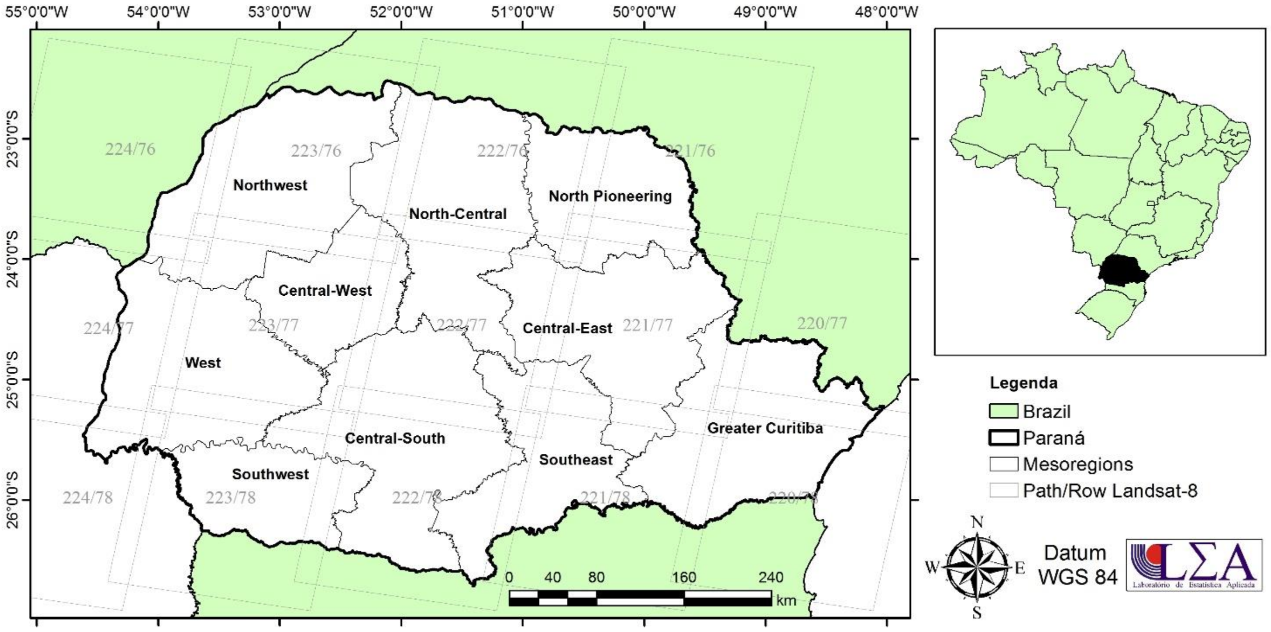

FIGURE 2. Study area location map and path/row of Landsat-8 satellite.

The EVI (*Enhanced Vegetation Index*) (Hunte et al., 2002) from MODIS sensor products MOD13Q1 and MYD13Q1, tile h13v11, with 250 meters of spatial resolution were used. The MOD13Q1 from Terra platform and MYD13Q1 from Aqua platform used together have a temporal resolution of 8 days (NASA, 2014). These combined products were used to elaborate a time series covering all the phenological cycle of both cultures (i.e., from pre-sowing to harvesting). Therefore, a spectral-temporal series was elaborated among 08/13/2014 to 05/01/2015, totaling 34 images of the MODIS sensor in the crop year 2014/2015.

Preprocessing

Samples of corn and soybean cultures were collected by a Landsat-8 false-color composition (RGB-564) (FIGURE 3), that has the shape of a MODIS sensor pixel (250mx250m). The sample selection was made in the five path/row mentioned using ArcGis 10.0 software totalizing 139 corn fields and 210 soybean fields which include all the mesoregions of study (FIGURE 2).

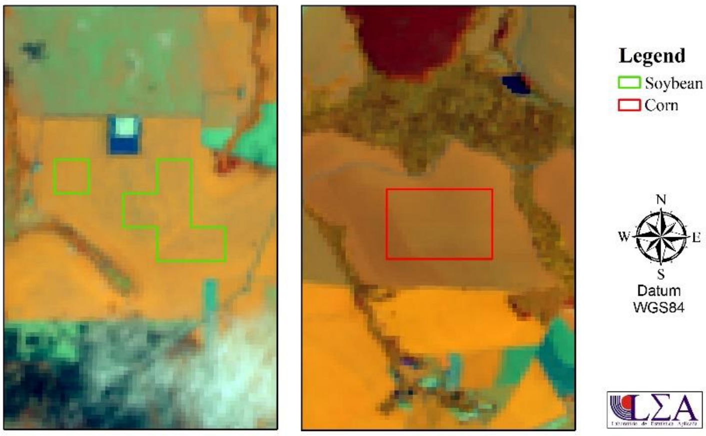

FIGURE 3. Soybean and corn areas with pure pixels delimitations in Landsat-8 images, RGB-564 composition.

It was applied the Flat filter (Esquerdo et al., 2011) in the MODIS time series images, with the purpose to minimize noise effects that might exist. This filter replaces inconsistent vegetation index values with the lowest adjacent value.

Transformation

After each crop sample selected the extraction of spectral-temporal profile values for each culture was made in software ArcGIS 10.0. After this, the extractions were imported into the software Timesat (Eklundh & Jönsson, 2015) which applied the filter Savitzky-Golay (Savitzky & Golay, 1964) to obtain 11 attributes. These attributes being: Seeding Date (SD), Harvest Date (HD), Cycle of Culture (CC), Base Value of EVI ($\mathrm{BASE_{EVI}}$), Date of Maximum Vegetative Development (DMVD), Maximum Value of EVI ($\mathrm{MAX_{EVI}}$), Range ($\mathrm{RANGE_{EVI}}$), Small Integral (S-INT), Large Integral (L-INT), illustrated in FIGURE 4, Left Derivative (L-DER) and Right Derivative (R-DER) (Johann et al., 2016). All attributes were used in data mining process, resulting in the decision tree.

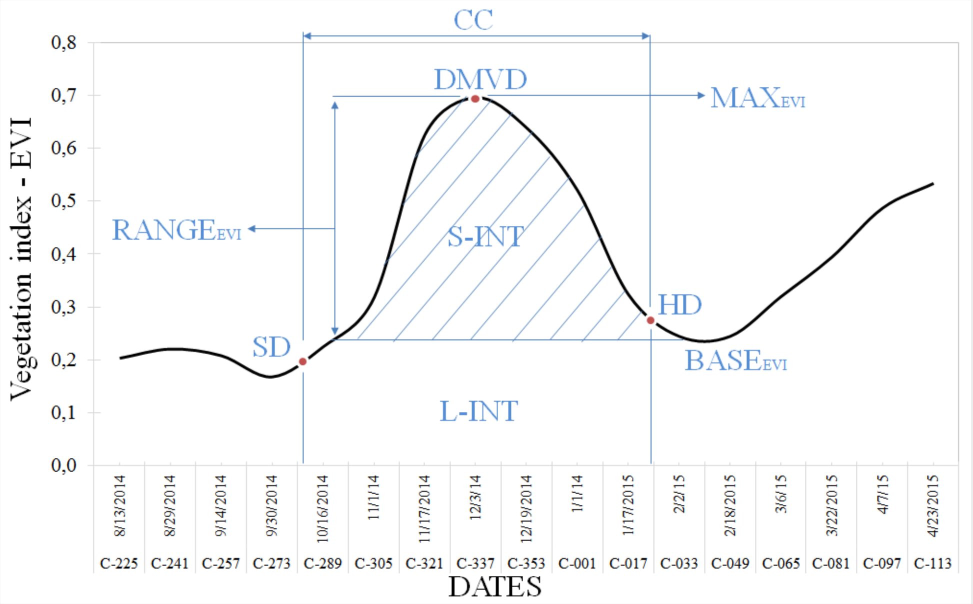

FIGURE 4. Standard of a spectro-temporal EVI profile for a summer crop.

Data mining, evaluation and interpretation

To execute the data mining process the software Weka was used (Hall et al., 2009). For the classification was used the decision tree J48, a modification of algorithm C4.5. The database used contained 11 attributes and 349 instances, containing representative samples of soybean and corn for the state of Paraná.

Cross validation, the percentage of correctly classified instances, and the Kappa index were the evaluation methods for the classification (Tan et al., 2009). The potential of the decision tree model was evaluated in relation to the percentage of correctly classified instances and the Kappa index in the training data of the database.

RESULTS AND DISCUSSION

Initially the descriptive analysis of the obtained results obtained will be presented, and secondly the classification obtained with the decision tree.

Descriptive analysis

In the 11 attributes extracted by Timesat software (FIGURE 5 to FIGURE 7), the studied cultures have different characteristics, mainly in the attributes related to dates (SD, DMVD and HD). The corn crop was anticipated in relation to soybeans, especially when observing SD (FIGURE 5a) and DMVD (FIGURE 5c). As for the period that the crop remains in the field, corn has a larger cycle than soybeans (FIGURE 5d and FIGURE 6). The same analyses also might be observed in SD (FIGURE 5a) and HD (FIGURE 5b), since corn is sown earlier but has the date of harvest close to soybeans.

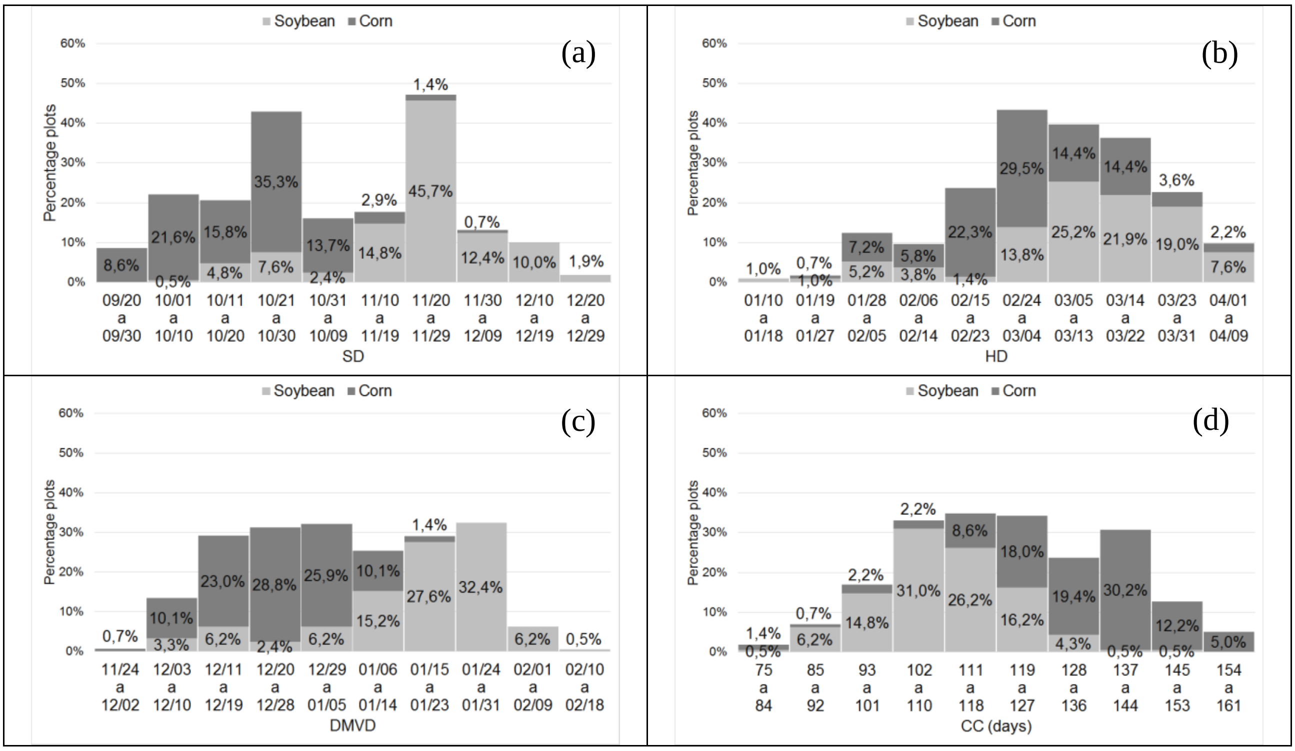

FIGURE 5. Percentage plots of the variables extracted by Timesat software: Sowing date - SD (a); harvesting date - HD (b); date of maximum vegetative development - DMVD (c) and cycle of culture - CC (d).

The value of EVI in DMVD (MAX_{EVI} in FIGURE 7a) is higher for soybean crop than for corn (FIGURE 6). Although both crops have similar phenological cycles, soybean has typically shorter cycle and higher EVI, while corn has a longer cycle and lower EVI. However, the opposite occurs to the EVI values in $BASE_{EVI}$ (Figure 7b); soybean presents a slightly larger value than corn.

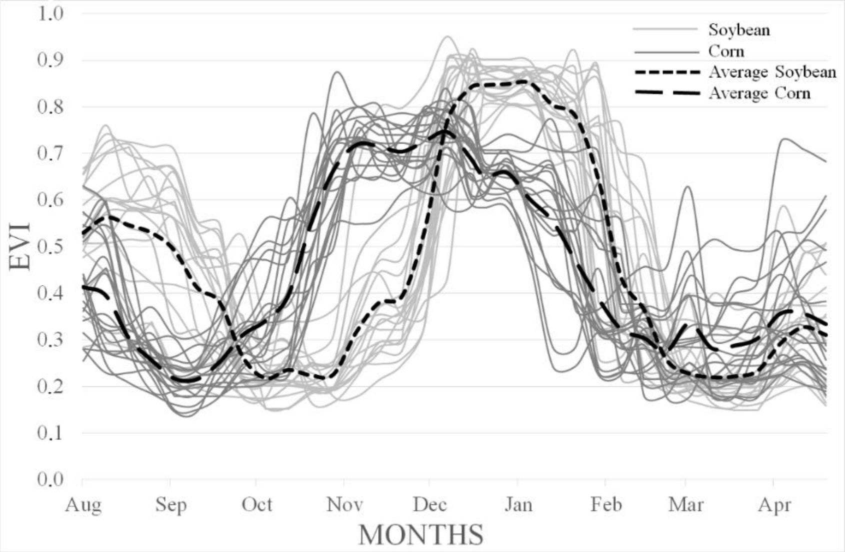

FIGURE 6. Spectro-temporal EVI profiles from all soybean and corn areas.

The left and right derivatives, L-DER (Figure 7c) and R-DER (Figure 7d) respectively, highlights that soybean has higher values than corn. The slope of the soybean vegetation index being higher (FIGURE 6) during the vegetative growth and senescence periods is explained by L-DER and R-DER. In addition, the soybean crop has $BASE_{EVI}$ value (Figure 7b) lower than corn and MAX_{EVI} value (Figure 7a) higher than corn.

Using the small integral - S-INT (FIGURE 7e), which considers the area on the EVI spectrum-temporal profile until the $\mathrm{BASE}_{\mathrm{EVI}}$ value; it was not possible to observe differences between cultures. However, observing the large integral - L-INT (FIGURE 7f), that considers the area on the EVI spectrum-temporal profile to the X axis of each crop; the values were higher for corn. Lastly, the $\mathrm{RANGE}_{\mathrm{EVI}}$ (FIGURE 7g) corroborates what was observed for $\mathrm{BASE}_{\mathrm{EVI}}$ and $\mathrm{MAX}_{\mathrm{EVI}}$ values, being higher for soybean and lower for corn.

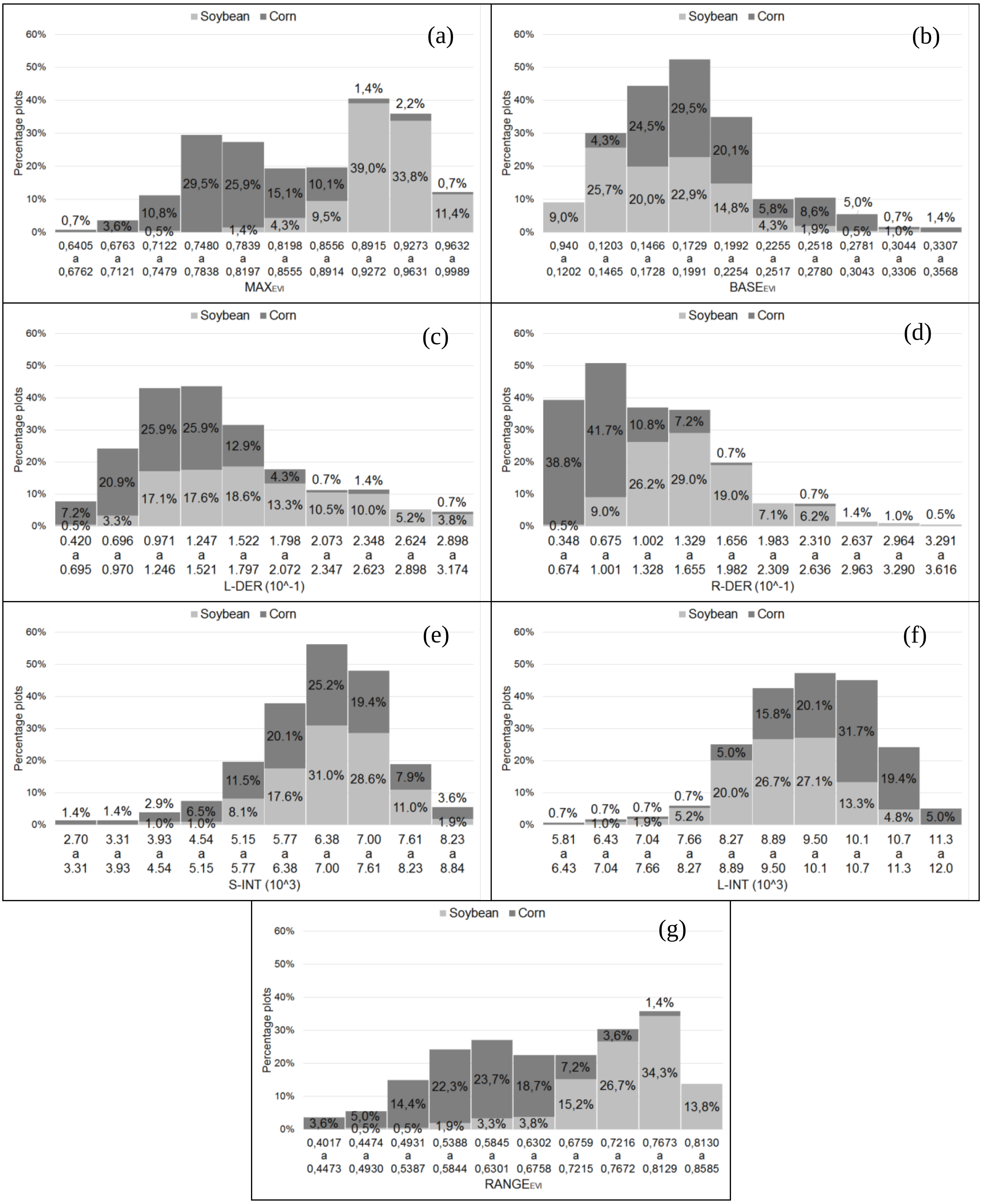

FIGURE 7. Percentage plots of the variables extracted by Timesat software: $\mathrm{MAX}_{\mathrm{EVI}}$ (a); $\mathrm{BASE}_{\mathrm{EVI}}$ (b); left derivative - L-DER (c); right derivative - R-DER (d); small integral – S-INT (e); large integral - L-INT (f) and range - $\mathrm{RANGE}_{\mathrm{EVI}}$ (g).

Decision tree

The decision tree (FIGURE 8) identified five attributes as necessary predictors for the differentiation of corn and soybean crops: Maximum EVI (MAX_{EVI}), Date of Sowing (SD), Cycle of culture (CC), Date of Maximum Vegetative Development (DMVD) and large integral (L-INT). Thus, seven classification rules were generated (Table 1), three (R1 to R3) for the corn crop and four (R4 to R7) for the soybean.

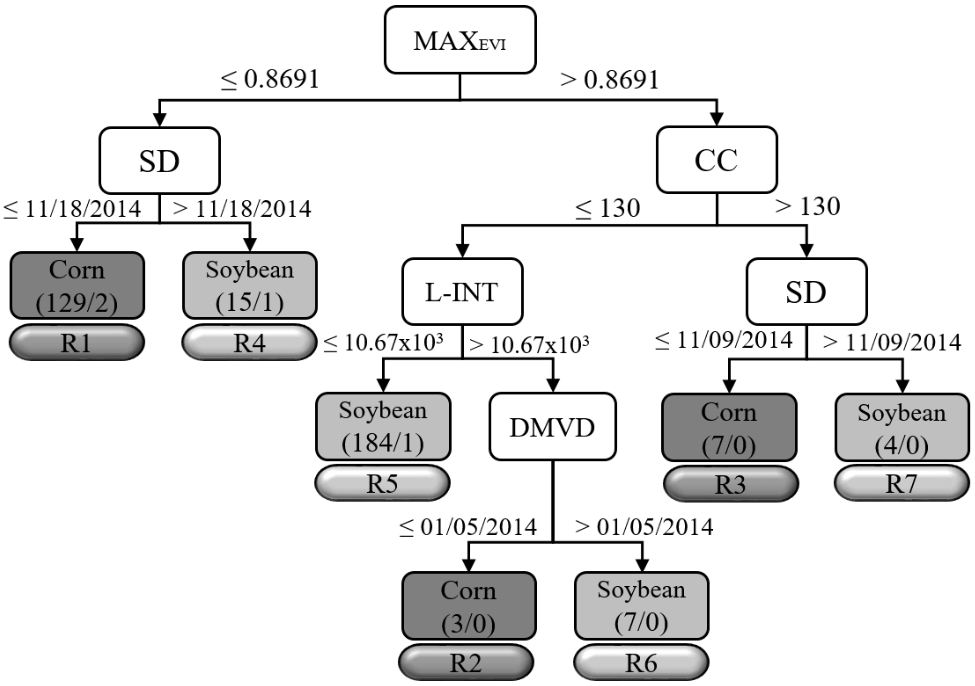

FIGURE 8. Decision tree for soybean and corn map separation. Maximum EVI - MAX_{EVI}; sowing date - SD; cycle of culture - CC; large integral - L-INT and date of maximum vegetative development - DMVD.

To understand the output of classification process made by the software Weka, each branch of the decision tree ends in what is called a leaf (which can be soybean or corn), that is understood as a classification rule (Table 1). Thus, for each leaf or rule the number of hits /number of errors are presented between parenthesis (FIGURE 8). Thus, for R1, which is a corn classification branch, 129 instances (areas) classified as corn were obtained when the maximum value of EVI (MAX_{EVI}) ≤ 0.8691 and SD ≤ November 18th, 2014. However, two of these areas were, in fact, soybean plots, that is, (129/2), which corresponded to 98.45% of accuracy R1.

TABLE 1. Classification rules for the soybean and corn map separation decision tree.

Rules	Description	Result
R1	If $MAX_{EVI} \leq 0.8691$ and SD ≤ 11/18/2014	Corn
R2	If $MAX_{EVI} > 0.8691$ and CC ≤ 130 days and L-INT > 10.67×10^3 and DMVD ≤ 01/05/2015	Corn
R3	If $MAX_{EVI} > 0.8691$ and CC > 130 days and SD ≤ 11/09/2014	Corn
R4	If $MAX_{EVI} \leq 0.8691$ and SD > 11/18/2014	Soybean
R5	If $MAX_{EVI} > 0.8691$ and CC ≤ 130 days and L-INT ≤ 10.67×10^3	Soybean
R6	If $MAX_{EVI} > 0.8691$ and CC ≤ 130 days and L-INT > 10.67×10^3 and DMVD > 01/05/2015	Soybean
R7	If $MAX_{EVI} > 0.8691$ and CC > 130 days and SD > 11/09/2014	Soybean

Likewise, for R4, a soybean classification branch, 15 instances (areas) were classified as soybean, when the maximum value of EVI (MAX_{EVI}) ≤ 0.8691 and SD > at November 18th, 2014.

However, one of these areas was a sample of corn (15/1) with an accuracy of 93.33% in the R4. For the R5, soybean classification branch, 184 instances were classified as soybean when $MAX_{EVI} > 0.8691$ and $CC \leq 130$ days and L-INT $\leq 10.67x10^3$, but one of these plots was corn (184/1), corresponding to 99.46% accuracy for R5.

In summary, the most important rules for the decision tree were R1, R4 and R5. The rule R1 classified 92.8% of corn samples correctly and the rules R4 and R5 classified 94.8% of soybean samples correctly.

For the whole decision tree (FIGURE 8), 96.3% of instances were correctly classified with a Kappa index of 0.92, which emphasize the potential use of decision trees in the process of agricultural crops mapping. Similar results were found by Nonato & Oliveira (2013) in the identification of areas of sugarcane and Kumar et al. (2010) in soil cover.

In the first node of the decision tree (Figure 8) the MAX_{EVI} was the attribute that presented least entropy, corroborating with Megeto et al. (2014). Thereby, this represents the greatest information gain to make the separation between cultures. Thus, when the value of MAX_{EVI} was $\leq$ 0.8691 and the SD $\leq$ in November 18th, 2014 the area was corn (R1 of Table 1) otherwise the area was soybean (R4 of Table 1).

When the value of $MAX_{EVI} > 0.8869$, if the CC $\leq$ 130 days, and the L-INT $\leq 10.67x10^3$ it was soybean (R5 of Table 1) and if the L-INT was greater than $10.67x10^3$ and the DMVD < at January 5th, 2015 the area was corn (R2 from Table 1) otherwise the area was soybean (R6 from Table 1).

However, for $MAX_{EVI} > 0.8669$ and CC > 130 days and SD $\leq$ at November 9th, 2014 the areas were classified as corn (R3 of Table 1) and if SD > at November 9th, 2014 the areas were classified as soybean (R7 of Table 1).

Therefore, although both crops have similar phenological cycles corn sowing is anticipated in relation to soybeans and, therefore, its maximum vegetative development is reached before soybean (FIGURE 6). This situation is evidenced because SD and DMVD occur primarily for corn (FIGURE 8).

The attribute CC highlights that soybean has a smaller cycle and corn has the longer (FIGURE 5 and FIGURE 6). The results of MAX_{EVI} and SD corroborate with Zhong et al. (2016) results, that the corn crop has an early sowing in relation to soybean and that the maximum value reached of EVI is lower.

The L-INT attribute was also notorious for the decision tree (FIGURE 8), by representing the area under the curve to the X axis (FIGURE 4). In FIGURE 7f can be noted that corn has the highest values of L-INT.

CONCLUSIONS

The application of the decision tree allowed identifying that the maximum EVI, date of sowing, date of maximum vegetative development, cycle, and the major integral are the variables (or instances) that presents greater potential for soybean and corn crops separation in the state of Paraná.

The R1 and R5 rules were those that contained the most hits, with R1 being: 129 of the 139 corn areas (92.8%) and R5: 184 of the 210 soybean areas (87.6%), showing that the attributes involved in these rules are the most important to the soybean and corn separation in the state of Paraná. All other rules contributed to only 7.2% for corn and 12.4% for soybean, concluding that the attributes involved in them are more specific.

The use of decision tree to crop classification shows a potential application in the agricultural mapping process especially in regions where the spectral separation of soybean and corn are difficult.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the research. They would also like to acknowledge the National Counsel of Technological and Scientific Development (CNPq) and the Technological Development of the Paraná State Research Foundation (Araucária Foundation) for the financial support, and the Western Paraná State University (UNIOESTE) for technical support.

REFERENCES

Assad ED, Marin RM, Evangelista SR, Pilau FG, Faria JRB, Pinto HS, Zullo Júnior J (2007) Sistema de previsão de safra de soja para o Brasil. Pesquisa Agropecuária Brasileira 42(5):615-625.

CONAB – Companhia Nacional de Abastecimento (2015) Perspectivas para a agropecuária. Safra 2015/2016. Brasília, 3:1-130.

Crivelenti RC, Coelho RM, Adami SF, Oliveira SRM (2009) Mineração de dados para inferência da relação solo-paisagem em mapeamentos digitais de solos. Pesquisa Agropecuária Brasileira 44:1707-1715.

Eklundh L, Jönsson P (2015) TIMESAT: A software package for time-series processing and assessment of vegetation dynamics. In: Kuenzer C, Deck S, Wagner W (eds). Remote Sensing Time Series - Revealing Land Surface Dynamics. Springer International Publishing. p141-158.

Esquerdo JCDM, Zullo Junior J, Antunes JFG (2011) Use of NDVI/AVHRR time series profiles for soybean crop monitoring in Brazil. International Journal of Remote Sensing 32:3711-3727.

Fayyad U, Shapiro GP, Smyth P (1996) Knowledge discovery and data mining: towards a unifying framework. In: International Conference on Knowledge Discovery and Data Mining. Portland, Proceedings…

Grzegozewski DM, Johann JA, Uribe-Opazo MA, Mercante E, Coutinho AC (2016) Mapping soya bean and corn crops in the State of Paraná, Brazil, using EVI images from the MODIS sensor. International Journal of Remote Sensing 37(6):1257-1275.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA Data Mining Software: An Update; SIGKDD Explorations 11(1):10-18.

Hunte ADL, Miura T, Rodriguez EP, Gao X, Ferreira LG (2002) Overview of the radiometric and biophysical performance of the *MODIS* vegetation indices. Remote Sensing of Environment 83(1-2):195-213.

IBGE – Instituto Brasileiro de Geografia e Estatística (2002) Pesquisas agropecuárias. IBGE, 2 ed. 92p.

Johann JA, Becker WR, Opazo MAU, Mercante E (2016) Uso de imagens do sensor orbital Modis na estimação de datas do ciclo de desenvolvimento da cultura da soja. Engenharia Agrícola 35:1-15.

Johann JA, Rocha JV, Duft DG, Lamparelli RAC (2012) Estimativa de áreas com culturas de verão no Paraná, por meio de imagens multitemporais EVI/Modis. Pesquisa Agropecuária Brasileira 47(9):1295-1306.

Kumar U, Kerle N, Punia M, Ramachandra TV (2010) Mining land cover information using multilayer perceptron and decision tree from MODIS data. Journal of the Indian Society of Remote Sensing 38(4):592-603.

Megeto GAS, Oliveira SRdeM, Del Ponte EM, Meira CAA (2014) Árvore de decisão para classificação de ocorrências de ferrugem asiática em lavouras comerciais com base em variáveis meteorológicas. Engenharia Agrícola 34(3):590-599.

NASA - National Aeronautics and Space Administration (2014) Technical specifications: moderate resolution imaging spectroradiometer (*MODIS*). Available: http://modis.gsfc.nasa.gov/about/design.php. Accessed: Jun, 2015.

Nonato RT, Oliveira SRdeM (2013) Técnicas de mineração de dados para identificação de áreas com cana-de-açúcar em imagens Landsat 5. Engenharia Agrícola 33(6):1268-1280.

Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36(8):1627-1639.

SEAB – Secretaria da Agricultura e do Abastecimento (2016) Divisão de estatísticas básicas – DEB. Available: http://www.agricultura.pr.gov.br/modules/conteudo/conteudo. php?conteudo=75. Accessed: Apr, 2016.

Souza CHdeW, Mercante E, Johann JA, Lamparelli RAC, Opazo MAU (2015) Mapping and discrimination of soya bean and corn crops using spectro-temporal profiles of vegetation indices. International Journal of Remote Sensing 36(7):1809-1824.

Souza ZM, Cerri DGP, Colet MJ, Rodrigues LHA, Magalhaes PSG, Mandoni RJA (2010) Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. Ciência Rural 40(4):840-847.

Tan PN, Steinbach M, Kumar V (2009) Introdução ao data mining: Mineração de dados. Rio de Janeiro, Ciência Moderna. 932p.