USO DA MINERAÇÃO DE DADOS NA CLASSIFICAÇÃO DO ALGODÃO UTILIZANDO SÉRIES-TEMPORAIS DE IMAGENS MODIS

João Paulo Sampaio Werner [1], Júlio César Dalla Mora Esquerdo [2], Stanley Robson de Medeiros Oliveira [3]

[1]Mestre em Engenharia Agrícola pela Faculdade de Engenharia Agrícola – Unicamp, Av. Cândido Rondon, n°501, Campinas – SP, wernerjoaopaulo@gmail.com; [2]Doutor em Engenharia Agrícola e pesquisador da Embrapa Informática Agropecuária, Av. André Tosello, nº 209, Campinas – SP, julio.esquerdo@embrapa.br; [3]Doutor em Ciência da Computação e pesquisador da Embrapa Informática Agropecuária, Av. André Tosello, nº 209, Campinas – SP, stanley.oliveira@embrapa.br

RESUMO

O objetivo do trabalho foi avaliar o uso de técnicas de mineração de dados extraídos de séries temporais de índices vegetativos do sensor MODIS para a classificação de padrões temporais do cultivo do algodão herbáceo. A partir da série temporal de imagens, foram gerados perfis espectro-temporais e extraídas 11 métricas fenológicas na forma de imagens de decomposição. Com as informações das métricas fenológicas e dados de referência terrestre, técnicas de mineração de dados foram aplicadas para gerar regras de classificação que, posteriormente, foram utilizadas para separar os padrões com cultivo de algodão de outras coberturas vegetais. Os resultados encontrados demonstraram a capacidade dos modelos para discriminar padrões de algodão de outras coberturas.

Palavras-chave — índice de vegetação, métricas fenológicas, árvore de decisão, *TIMESAT*.

ABSTRACT

The objective of this work was to evaluate the use of data mining techniques extracted from time series of vegetation indexes from MODIS sensor in order to classify temporal patterns extracted from herbaceous cotton crops. From the set of time series images, spectral-temporal profiles were generated and 11 phenological metrics were extracted as decomposition images. Using the information from the phenological metrics and land reference data, data mining techniques were applied to generate classification rules that were later used to separate the the cotton patterns from other vegetation covers. The results showed the ability of the models to discriminate cotton patterns from other vegetation covers, such as pasturelands, forests and other types of crops .

Key words — *vegetation index, phenological metrics, decision tree, TIMESAT.*

1. INTRODUÇÃO

Após a consolidação do cultivo do algodoeiro no cerrado brasileiro na década de 90, o algodão se apresenta como uma das principais *commodities* do Brasil. O estado de Mato Grosso desenvolveu uma cotonicultura sólida nos últimos anos e ocupa a liderança na produção do algodão no Brasil, sendo responsável por 66% da produção nacional na safra 2016/17, com 627 mil hectares de área colhida [1].

Nexto contexto, a identificação e quantificação das áreas de cultivo do algodão são estratégicas para o setor produtivo, uma vez que permitem o monitoramento da cultura no campo e auxiliam na tomada de decisões por parte de vários setores da economia, além de fornecer apoio ao planejamento das instituições governamentais.

Embora muitos estudos utilizem as variáveis espectrais para a identificação de culturas agrícolas por meio de técnicas de sensoriamento remoto e classificação digital de imagem com relativo êxito, ocorre que alguns objetos na superfície terrestre podem ter um comportamento espectral similar ao da classe de interesse, difultando o processo de identificação do alvo. Além disso, uma mesma cultura pode encontrar-se em diferentes estádios de desenvolvimento, e se apresentar com diferentes padrões espectrais, embora pertença a uma mesma classe de uso da terra.

Alguns trabalhos evidenciaram que quando se inclui o domínio temporal na abordagem, sobretudo com o uso de imagens de índices vegetativos, estes problemas são minimizados, uma vez que cada cultura agrícola segue um padrão espectro-temporal específico, tornando o processo mais eficiente [2][3][4][5].

No entanto, ainda persiste o desafio de desenvolver métodos sistemáticos que aprendam com o padrão espectro-temporal de culturas agrícolas baseado nas informações extraídas em séries temporais de imagens, de forma a automatizar o processo de identificação das principais culturas praticadas.

Assim, as técnicas de mineração de dados têm demonstrado forte potencial na caracterização das coberturas vegetais e culturas agrícolas, uma vez que permitem transformar as informações contidas em grandes volumes de dados em conhecimento e, portanto, são de grande aplicabilidade em estudos que utilizam séries temporais [6][7][8].

Desse modo, este trabalho visa promover ganhos na automatização nos processos de identificação das áreas algodoeiras no estado de Mato Grosso. O objetivo deste trabalho foi classificar padrões espectro-temporais característicos da cultura do algodão herbáceo utilizando técnicas de mineração de dados e séries temporais de índices vegetativos do sensor MODIS.

2. MATERIAL E MÉTODOS

2.1. Área de estudo

A região de estudo, o estado de Mato Grosso, está localizada na região centro-oeste do Brasil, com uma área de 903 mil km². O estado é responsável por 66% da área plantada de algodão no Brasil [1]. A Figura 1 apresenta o mapa dos principais produtores de algodão no estado de Mato Grosso.

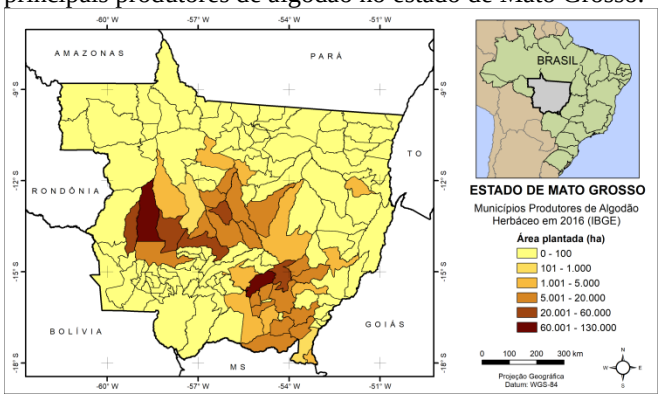

Figura 1. Área de estudo e os municípios produtores de algodão herbáceo na safra 2015/16, segundo IBGE.

2.2. Dados de entrada

2.2.1. Série temporal de imagens MODIS

Como o trabalho foi fundamentado pela abordagem espectro-temporal, foram utilizados índices vegetativos das imagens do sensor MODIS, a bordo dos satélites Terra (MOD13Q1) e Aqua (MYD13Q1). As imagens correspondem à área de estudo e englobam quatro safras agrícolas (2012/2013 a 2015/2016), com a finalidade de identificar os padrões espectro-temporais do algodão para o mesmo período.

Os produtos MOD13Q1 e MYD13Q1 fornecem dados dos índices vegetativos *Normalized Difference Vegetation Index* (NDVI) e *Enhanced Vegetation Index* (EVI) em composições de máximo valor de 16 dias, com resolução espacial de 250m. Dessa forma, cada safra foi composta por 46 composições de 16 dias (23 de cada satélite), que totalizaram 184 imagens para cada índice vegetativo (NDVI e EVI). As imagens foram disponibilizadas já processadas em seu formato final de trabalho pela Embrapa Informática Agropecuária, por meio do Banco de Produtos MODIS [9].

Além disso, para minimizar os efeitos causados pela presença de nuvens nas composições MODIS, que podem interferir nos valores digitais, foi aplicado o método de filtragem *Flat Botton Smoother*. Trata-se de um filtro simples e conservador, adaptado do método proposto por [10]. Sua aplicação se deu a partir de rotinas desenvolvidas em linguagem *Interactive Data Language (IDL*).

A definição dos períodos das séries temporais de cada safra respeitou o ciclo de produção do algodoeiro no campo, sendo consultado o calendário agrícola na região em estudo. Assim, definiu-se que a semeadura ocorreu entre 1° de dezembro a 28 de fevereiro de cada ano-safra.

2.2.2. Dados de campo

Como referência terrestre para treinamento do classificador foram utilizados dados amostrais de localização geográfica das áreas de cultivo de algodão obtidas por trabalhos de campo, cedidos pela Embrapa Informática Agropecuária. Os dados também incluiram a localização de outras culturas comerciais e de cobertura (soja, milho, milheto, sorgo, crotalária, cana-de-açúcar), pastagem e vegetação nativa presentes na área de estudo, que serviram como base para a definição do padrão espectro-temporal de cada superfície.

2.3. Geração e extração das métricas fenológicas

A análise dos perfis temporais das áreas de referência foi realizada por meio do software *TIMESAT* [11], a partir do qual foram geradas métricas da fenologia do algodão extraídas das séries temporais de índices de vegetação e da parametrização dos perfis temporais utilizando funções matemáticas de ajuste.

A análise das métricas fenológicas pode permitir a discriminação de culturas agrícolas por meio da investigação dos diferentes efeitos causados pela sazonalidade, que se relacionam com o desenvolvimento vegetativo ao longo do tempo. A Figura 2 ilustra um exemplo de um perfil temporal do NDVI e as 11 métricas fenológicas extraídas por meio do *TIMESAT*. A linha em vermelho representa o perfil ajustado por meio de uma função de filtragem aplicada aos dados brutos (em azul).

Os dados de NDVI e EVI, presentes nos produtos MODIS, variaram entre -2.000 a 10.000, sendo que os dados inválidos foram marcados com o valor -3.000. É importante ressaltar que o *TIMESAT* não extrai métricas a partir da série temporal bruta, mas sim de uma série ajustada por meio de uma de suas três funções matemáticas de ajuste (Logística Dupla, Gaussiano Assimétrico e *Savitzky-Golay*). Neste trabalho, o filtro *Savitzky-Golay* foi o que melhor se ajustou às características de sazonalidade das áreas algodoeiras. Este filtro baseia-se em uma janela móvel (tamanho 4), que utiliza ajuste linear de mínimos quadrados por meio de sucessivas equações polinomiais.

O *TIMESAT* requer a definição de um conjunto de parâmetros de entrada para o cálculo do início e final da safra, os quais são definidos pelo usuário a partir de uma interface gráfica disponível pelo *software*, considerando os tipos de cobertura vegetal de interesse. Uma vez definidos os parâmetros de entrada, outro módulo do *TIMESAT* foi utilizado para calcular as métricas fenológicas no conjunto de imagens de cada safra analisada, pixel a pixel.

Uma das saídas do *TIMESAT* são imagens de decomposição, que representam cada uma das onze métricas fenológicas disponíveis, considerando os dois índices de vegetação. De posse dessas imagens, uma rotina escrita em

linguagem *IDL* foi escrita, as mesmas foram compiladas e executadas, com a finalidade de extração de seus valores e conversão para formato de planilha, legível pelo *software Waikato Environment for Knowledge Analysis (WEKA)* [13], onde foi aplicada a etapa de mineração. Na extração usou-se a localização geográfica presente nos dados de campo.

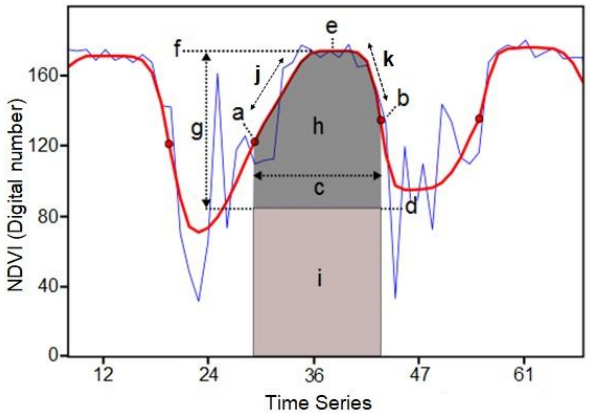

Figura 2. Representação gráfica do perfil temporal do NDVI, registrados pelo MODIS, e sua derivação em métricas fenológicas no *TIMESAT*. Adaptado: manual do *software* [12].

a) Início do ciclo da cultura: quando a função ajustada à série temporal é aumentada para 20% da amplitude, tendo como referência o valor mínimo esquerdo.
b) Final do ciclo da cultura: quando a função ajustada à série temporal é decrescida para 20% da amplitude, tendo como referência o valor mínimo direito.
c) Comprimento do ciclo da cultura: tempo entre o início e final do ciclo da cultura.
d) Valor base: média dos valores mínimos encontrados nos lados esquerdo e direito.
e) Posição da metade do ciclo: média do tempo em que a função ajustada à esquerda aumentou 80% e à direita diminuiu em 80%.
f) Pico de máximo da cultura: valor máximo do índice de vegetação para a função ajustada à série temporal.
g) Amplitude sazonal: diferença entre o valor máximo da cultura e o valor base.
h) Pequena integral: área de baixo da curva da função ajustada à série temporal entre o início e final do ciclo da cultura, a partir do valor base.
i) Grande integral: área total sob a curva da função ajustada à série temporal entre o início e o final do ciclo da cultura.
j) Taxa de crescimento: razão da diferença entre os níveis 20% e 80% do lado esquerdo e sua correspondente diferença de tempo.
k) Taxa de senescência: valor absoluto da razão da diferença entre os níveis 20% e 80% do lado direito e sua correspondente diferença de tempo.

2.4. Modelagem para classificação das áreas de algodão

Foram gerados dois bancos de dados constituídos pelos valores das métricas fenológicas e sazonais derivadas das séries temporais de NDVI e EVI, respectivamente, extraídas no *TIMESAT*. As 11 métricas representam as variáveis independentes (atributos numéricos) e o atributo-meta contém duas classes: a) algodão (cultura de interesse); b) não algodão (todos os demais usos e coberturas da terra). Assim, os conjuntos de dados ficaram constituídos por 3.211 instâncias, para o conjunto do NDVI, e 2.890 instâncias para o conjunto do EVI, com 12 atributos em cada. No primeiro, havia 1.159 instâncias classificadas como algodão, enquanto que no segundo, 1.167 instâncias. As métricas fenológicas foram calculadas pelo *TIMESAT* a partir da identificação das datas de início e final dos ciclos. Quando tais momentos não puderam ser identificados pelo *software*, nenhuma outra métrica foi calculada. Isso explica a diferença no número de instâncias encontradas pelo NDVI e EVI, uma vez que as curvas desses dois índices podem apresentar comportamentos distintos para um mesmo pixel, fazendo com que o resultado da detecção das datas seja diferente para cada índice.

Em seguida, foi utilizado o algoritmo de indução por árvore de decisão J48, conhecido como C4.5 [14]. Para tornar o modelo mais simples e genérico, melhorar a taxa de acerto do classificador e diminuir o sobreajuste, foram testados diferentes números de objetos por folha para podar a árvore de decisão. Além disso, quatro diferentes métodos para seleção dos atributos foram testados, com o propósito de retirar aqueles com baixa correlação em relação às classes: a) sem seleção de atributos, onde ocorreu a utilização de todos atributos, caracterizando-se pela ausência de seleção; b) seleção de atributos baseado em correlação (CFS), que pesquisa o conjunto de atributos correlacionados com a classe e não correlacionados entre si; c) o método *InfoGain*, que avalia o valor de um atributo medindo o ganho de informação em relação à classe; d) a abordagem *Wrapper*, que ocorre conjuntamente com o algoritmo básico de aprendizagem, em que a validação cruzada é utilizada para estimar a precisão do esquema de aprendizagem para um conjunto de atributos [15].

Os modelos de indução foram gerados por meio de validação cruzada (10 *folds*) e avaliados pelas medidas: acurácia, coeficiente Kappa e precisão por classe, calculadas pela matriz de confusão, cuja matriz é obtida pelo cruzamento entre as classes preditas pelos modelos e as classes observadas no campo (real) [15].

3. RESULTADOS E DISCUSSÃO

Foram gerados diversos modelos por meio do algoritmo J48, considerando diferentes métodos de seleção de atributos. As Tabelas 1 e 2 apresentam, de uma maneira geral, as avaliações dos modelos para os índices de vegetação NDVI e EVI, respectivamente. Observa-se que as variáveis independentes, representadas pelas métricas fenológicas, conseguiram predizer as classes com um ótimo desempenho para os dois índices de vegetação. Para os classificadores baseados no NDVI, nota-se a superioridade do método

Wrapper em relação ao demais, o qual conseguiu realizar a predição das classes, com alta acurácia, com apenas seis dos 11 atributos.

Tabela 1. Desempenho do classificador J48 para série temporal do NDVI.

Método de seleção	Acurácia	Kappa	Precisão por classe		N° de regras	N° obj por folha	Atributo Selecionado*
			Algodão	Não Algodão			
CFS	95,14%	0,89	0,95%	0,95%	12	6	i,h,e
Infogain	95,02%	0,89	0,94%	0,96%	12	8	h,i,c, f,g,b
Wrapper	96,08%	0,91	0,95%	0,97%	13	6	g,c,j, a,i,h

Tabela 2. Desempenho do classificador J48 para série temporal do EVI.

Método de seleção	Acurácia	Kappa	Precisão por classe		N° de regras	N° obj por folha	Atributo Selecionado*
			Algodão	Não Algodão			
CFS	95,26%	0,9	0,95%	0,96%	8	6	c,i,h,f,e
Infogain	94,80%	0,89	0,94%	0,95%	10	6	h,i,c, g,f,b
Wrapper	94,91%	0,89	0,94%	0,95%	11	6	j,i,h,e

*Atributos selecionados: a) início do ciclo; b) final do ciclo; c) comprimento do ciclo; d) valor base; e) posição da metade do ciclo; f) pico de máximo; g) amplitude sazonal; h) pequena integral (produtividade primária); I) grande integral (produtividade total); j) taxa de crescimento; k) taxa de senescência.

Para os classificadores baseados no EVI, o modelo gerado por meio do método CFS, com apenas cinco atributos, apresentou resultados melhores, possivelmente pelo fato deste método ser mais eficiente para lidar com atributos numéricos.

As Tabelas 1 e 2 também apresentam os resultados para os diferentes valores de objetos por folha, variando de dois a 12. Para os modelos baseados no NDVI (Tabela 1), os melhores valores variaram entre seis e oito, enquanto que para o EVI, o número seis apresentou os melhores resultados. Para ambos os conjuntos de dados, as métricas fenológicas produtividade primária, produtividade total e comprimento do ciclo foram os atributos que apresentaram o maior ganho de informação.

4. CONCLUSÕES

A aplicação das técnicas de mineração de dados às séries temporais de imagens de índices de vegetação NDVI e EVI, disponíveis pelos produtos do sensor MODIS, possibilitou a construção de modelos capazes de identificar áreas de cultivo de algodão.

A partir das regras dos modelos gerados, estas poderão ser testadas para mapear e estimar as áreas algodoeiras, considerando qualquer uma das quatro safras abordadas (2012/2013 a 2015/2016) de maneira rápida e objetiva, já que os resultados demonstraram um ótimo desempenho.

5. REFERÊNCIAS

[1] CONAB, Acompanhamento da safra brasileira [de] grãos: safra 2016/17: décimo segundo levantamento, Brasília, DF, v. 4, n. 10, jul. 2016. 170 p. Disponível em: <http://www.conab.gov.br/OlalaCMS/uploads/arquivos/17_09_12_10_14_36_boletim_graos_setembro_2017.pdf>. Acesso em: 16 nov. 2017.

[2] Arvor, D.; Jonathan, M.; Meirelles, M. S. P.; Dubreuil, V.; Durieux, L. Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil, International Journal of Remote Sensing, v. 32, n. 22, p. 7847–7871, 2011.

[3] Brown, J. C.; Kastens, J. H.; Coutinho, A. C.; Victoria, D, C.; Bishop, C. R. Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data. Remote Sensing of Environment, v. 130, p.39–50, 2013.

[4] Kastens, J.; Brown, J.; Coutinho, A. ; Bishop, C. R.; Esquerdo, J. Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso, Brazil. PLoS ONE 12 (4). 2017.

[5] Picoli, M. C. A.; Camara, G.; Sanches, I.; Simões, R.; Carvalho, A.; Maciel, A.; Coutinho, A.; Esquerdo, J.; Antunes, J.; Begotti, A. Arvor, D.; Almeida, C. Big Earth observation time series analysis for monitoring Brazilian agriculture, ISPRS Journal of Phtogrammetry and Remote Sensing, 2018. No prelo.

[6] Johann, J. A.; Rocha, J. V; Oliveira, S. R. M; Rodrigues. L. H. A; Lamparelli, R. A. C. Data mining techniques for identification of spectrally homogeneous areas using NDVI temporal profiles of soybean crop. Engenharia Agrícola Jaboticabal, v. 33, n. 3, p.511-524. 2013.

[7] Vieira, M.A; Formaggio, A.R; Rennó, C.D; Atzberger, C; Aguiar, D.A; Mello, M.P. Object Based Image Analysis and Data Mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. Remote Sensing of Environment. 2012.

[8] Zhou, F.; Zhang, A.; Townley-Smith, L. A data mining approach for evaluation of optimal time-series of MODIS data for land cover mapping at a regional level, ISPRS Journal of Photogarmmetry and Remote Sensing, v. 84, p. 114-129, 2013.

[9] Esquerdo, J.C.D.M.; Antunes, J.F.G.; Andrade, J.C. de. Desenvolvimento do banco de produtos MODIS na base estadual brasileira. Simpósio Brasileiro de Sensoriamento Remoto, 15, Curitiba. Anais... São José dos Campos: Inpe, p. 7596-7602, 2011.

[10] Wardlow, B.D.; Kastens, J.H.; Egbert, S.L. Using USDA Crop Progress Data for the Evaluation of Greenup Onset Date Calculated from MODIS 250-meter Data. Photogrammetric Engineering and Remote Sensing, vol.72, n.11, p.1225-1234. 2006.

[11] Jönsson, P.; Eklundh, L. Timesat - a program for analyzing time-series of satellite sensor data, Computers and Geosciences, v. 30, p. 833 – 845. 2004.

[12] Eklundh, L. e Jönsson, P. Timesat 3.2 Software Manual, Lund and Malmö University, Sweden, 88p. 2015.

[13] Witten, I. H.; Frank, E. Data mining: pratical machine learning tools and techniques. 2nd ed., 525 p., San Francisco: Morgan Kaufmann, 2005.

[14] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA,1993.

[15] Han, J.; Kamber, M; Pei, J. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2011. 770p.