Active learning e sua aplicação no monitoramento da cana-de-açúcar utilizando o algoritmo *SVM*

João Paulo da Silva[1], *Jurandir Zullo Júnior*[2], *Luciana Alvim Santos Romani*[3]

[1] Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, São Paulo, Brasil, joao.silva@feagri.unicamp.br

[2] Centro de Pesquisas Meteorológicas e Climáticas Aplicadas a Agricultura, Universidade Estadual de Campinas, Campinas, São Paulo, Brasil, jurandir@cpa.unicamp.br

[3] Embrapa Informática Agropecuária, Empresa Brasileira de Pesquisa Agropecuária, Campinas, São Paulo, Brasil, luciana.romani@embrapa.br

RESUMO

A cana-de-açúcar é um dos pilares do agronegócio brasileiro e, por apresentar intensa dinâmica expansionista, demanda metodologias que subsidiem a criação de estratégias políticas e econômicas que promovam a sustentabilidade da produção. Este artigo propõe uma nova abordagem de monitoramento de áreas canavieiras baseada na classificação de séries temporais de imagens de satélite associada à técnica de *Active Learning*. A interação do usuário especialista no aprendizado do algoritmo de classificação através desta técnica utilizando parâmetros sazonais das séries temporais gerou um conjunto de treino otimizado que promoveu a redução do custo operacional de monitoramento da ocupação da cana-de-açúcar. A correlação de cerca de 90% observada entre as análises conduzidas neste trabalho com dados oficiais indica que a metodologia proposta pode ser utilizada no monitoramento agrícola devido à similaridade entre os resultados associada ao baixo custo operacional envolvido.

PALAVRAS-CHAVE: Mineração de dados, Análise sazonal, Classificação, Cana-de-açúcar.

ABSTRACT

Sugarcane is one of the pillars of Brazilian agribusiness and, because of its intense expansionary dynamics, demands methodologies that subsidize the creation of political and economic strategies that promote the sustainability of production. This paper proposes a new approach

for sugarcane areas monitoring based on the classification of satellite images time series associated to the *Active Learning* technique. The interaction of the expert user in the classification algorithm's learning through this technique using time series seasonal parameters generated an optimized training set that promoted operational cost reduction for sugarcane occupation monitoring. The correlation of approximately 90% observed between this study's analysis with official data indicates that the proposed methodology can be used in agricultural monitoring due to the similarity between its results associated with low operating cost involved.
KEYWORDS: Data mining, Seasonal analysis, Classification, Sugarcane.

INTRODUÇÃO

A partir da primeira década de 2000, período de intensificação da produção e comercialização das *commodities* no mundo, a agroindústria canavieira do estado de São Paulo, principal produtor do país, expandiu sua área de ocupação passando de 2,5 milhões para 5 milhões de hectares, segundo a União da Indústria de Cana-de-açúcar (UNICA, 2016). A análise dessa dinâmica expansionista evidencia a necessidade de adotar técnicas de monitoramento da área destinada à cana-de-açúcar para subsidiar a criação de estratégias políticas e técnicas que reduzam seus impactos sobre os recursos naturais existentes (BUCKERIDGE et al., 2012; DAMATTA et al., 2010).

O sensoriamento remoto tem evoluído e se mostrado uma importante ferramenta de monitoramento ambiental. No caso da cana-de-açúcar, destaca-se o programa Canasat[1] (RUDORFF et al., 2010), desenvolvido pelo Instituto Nacional de Pesquisas Espaciais (Inpe) e que, entre 2003 e 2013, monitorou a dinâmica da ocupação canavieira da região Centro-Sul do Brasil. Os levantamentos realizados por esse projeto, entretanto, eram feitos com um grande esforço de interpretação visual, o que poderia limitar sua condução e expansão à disponibilidade de capital humano, especialmente em períodos de desaceleração econômica. Assim, é importante também que as técnicas de monitoramento ambiental que utilizam dados remotos apresentem escalabilidade, para que a intervenção humana e, consequentemente, o seu custo operacional, sejam os menores possíveis.

A utilização de técnicas de mineração de dados em sensoriamento remoto tem demonstrado ser uma alternativa eficiente aos métodos tradicionais para este tipo de aplicação. Diferentes trabalhos envolvendo esta abordagem para monitoramento agrícola vêm sendo propostos para diversas finalidades, dentre as quais está a classificação de áreas agrícolas (XIE

[1] Disponível em: http://www.dsr.inpe.br/laf/canasat/

et al., 2014; ZHANG; FENG; YAO, 2012). Um dos principais entraves neste tipo de tarefa, porém, é a criação de um conjunto de treinamento que expresse as principais características de cada classe envolvida. Para contornar este problema, Crawford, Tuia e Yang (2013) apresentam uma extensa revisão sobre *Active Learning*, uma abordagem para a criação de um conjunto de treinamento baseada na interação do usuário especialista ou em critérios para otimização da criação do conjunto de treino, capaz de reduzir seu tamanho em até 90% sem que haja perda na acurácia dos resultados (TUIA et al., 2009).

Desse modo, esse trabalho teve como objetivo principal apresentar uma metodologia para monitoramento de áreas canavieiras baseada na classificação de séries temporais de imagens de satélite que possa ser adotada como alternativa a métodos tradicionais baseados na interpretação visual de imagens remotas, reduzindo o custo e aumentando a eficiência deste tipo de aplicação.

MATERIAL E MÉTODOS

A área e o período de estudo empregados na condução deste trabalho representam todas as safras agrícolas de cana-de-açúcar ocorridas no estado de São Paulo, principal produtor nacional, entre 2006 e 2012, período de grande expansão da área ocupada com a cultura.

As séries temporais foram construídas a partir de imagens do Índice de Vegetação da Diferença Normalizada (NDVI, em inglês) do sensor MODIS-Terra, com resolução temporal de 16 dias e espacial de 0,25km, sendo cada série representativa do período compreendido entre abril e março do ano seguinte que correspondem, respectivamente, ao início e fim de cada safra.

Preparação e pré-processamento das séries temporais

A preparação e pré-processamento é uma etapa crítica de projetos de mineração de dados que pode influenciar diretamente o sucesso das análises e a robustez dos resultados obtidos. Neste trabalho, esta etapa se deu pela (1) filtragem e extração de parâmetros sazonais das séries temporais; e (2) redução da dimensionalidade do conjunto de dados.

Com o auxílio da ferramenta Timesat (EKLUNDH; JÖNSSON, 2012) foram extraídos parâmetros sazonais das séries temporais relacionados ao tempo cronológico do ciclo vegetativo, como o (1) início; (2) o fim; (3) a duração; e (4) o tempo para máximo vigor vegetativo, e também aos valores assumidos pelo NDVI no ciclo, como (5) o valor máximo; (6) o valor base; (7) a amplitude; (8) a taxa de crescimento; e (9) a taxa de declínio após o máximo vigor vegetativo. Além das séries sazonais (SS) resultantes da extração sazonal, foram mantidas na condução do processamento dos dados séries temporais sem extração da sazonalidade (SES) para avaliação da contribuição deste processo no resultado final.

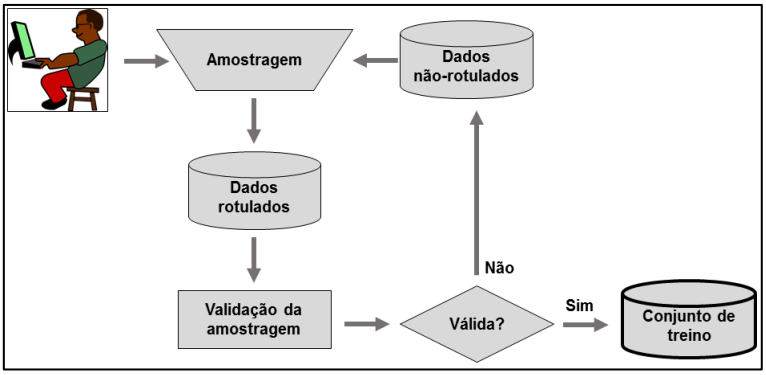
Para a redução de dimensionalidade dos dados, foi aplicada, tanto nas séries SS quanto nas séries temporais SES, a Análise de Componentes Principais (ACP) (RICHARDS, 2012), que por meio da transformação dos atributos originais em autovetores, permite a redução do conjunto de dados para que este contenha apenas os principais autovetores do conjunto inicial. Neste trabalho, tanto para as séries sazonais quanto para as séries temporais SES, foram retidos apenas os autovetores que apresentaram autovalores superiores a 0,7, os quais, segundo Jolliffe (1972) e Jolliffe (1973), detêm cerca de 90% da variância do conjunto de dados original.

Classificação das séries temporais

A classificação das séries sazonais e SES foi feita utilizando-se o algoritmo *Support Vector Machine* (SVM) (CORTES; VAPNIK, 1995), considerado como um dos algoritmos mais sofisticados atualmente para análise de dados remotos (MOUNTRAKIS; IM; OGOLE, 2011) em função do seu desempenho em situações onde o tamanho do conjunto de treino é reduzido (MANTERO; MOSER; SERPICO, 2005) ou quando o número de atributos é elevado, como no caso de séries temporais (WANG et al., 2010). Seu funcionamento baseia-se na escolha de registros essenciais (ou vetores suporte) do conjunto de treino para a determinação de um hiperplano otimizado (e suas margens) de separação dos dados em duas classes distintas.

Para aplicação da abordagem *Active Learning* (Figura 1), foram estabelecidos critérios de seleção por parte do usuário especialista que expressassem, confiavelmente, características espectrais exclusivas de cada uma das classes amostradas. Com o auxílio da ferramenta ENVI, versão 5.2, foram amostradas regiões representativas de seis diferentes classes de cobertura da terra – *área urbana*, *hidrografia*, *área de mata*, *cultura anual*, *cana-de-açúcar* e *pastagem* – a partir de imagens formadas por composições das séries sazonais para serem utilizadas na construção de um conjunto de treino, sendo que o critério adotado para avaliação da distinção entre as amostras foram as medidas de distância Jeffries-Matusita (JM) e a Divergência Transformada (DT) (RICHARDS, 2012). Tomando como referência o estudo de classificação de uso e cobertura da terra realizado por Cai e Zhang (2010), foi definido que a amostragem de uma classe deveria ser refeita até que o grau de separabilidade medido entre esta e as outras classes fosse igual ou superior a 1,9. Para comparação dos resultados da classificação, as regiões amostradas sobre as composições das séries sazonais foram replicadas, sem alterações, na classificação das séries temporais SES.

Figura 1. Fluxograma de aplicação da *Active Learning* para construção do conjunto de treino otimizado.



Fonte: Autor

As estimativas anuais de área plantada geradas a partir da classificação das séries temporais foram cruzadas com dados do projeto Canasat/INPE e de outras metodologias de estimativa de área plantada utilizadas por órgãos oficiais para comparação e validação dos resultados.

RESULTADOS E DISCUSSÃO

Preparação e pré-processamento dos dados

Com a extração dos parâmetros sazonais das séries temporais, além da atenuação primária de ruído, verificou-se também que parâmetros sazonais relacionados ao vigor vegetativo se mostraram mais relevantes que aqueles ligados ao tempo cronológico na distinção de diferentes coberturas vegetais na área estudada. Contudo, os períodos de início, meio e fim do ciclo fenológico da vegetação também foram coincidentes com o clima da região, caracterizado por um período quente e úmido no início e fim do ano, quando o desenvolvimento vegetativo é mais intenso, seguido por outro mais frio e seco no intervalo, quando ocorre o repouso da vegetação, segundo dados do Instituto Nacional de Meteorologia (INMET) (2017).

A partir da aplicação da ACP nas séries sazonais, foram retidos dois dos nove autovetores originais em todas as safras analisadas. Os autovetores retidos corresponderam, em média, a 87% da variância do conjunto original de dados. No caso das séries temporais SES, com exceção das safras 2006-2007 e 2009-2010, para as quais foi retido apenas um, todas as outras safras resultaram também na retenção de dois autovetores, e a variância média contida nestes autovetores resultantes foi de cerca de 93%. Utilizando a abordagem da aplicação da ACP em dados remotos para modelagem ecossistêmica do Pantanal, Almeida et al. (2015) conseguiram descrever 99% da variância contida no conjunto original de dados nos três primeiros autovetores resultantes da aplicação da ACP. Réjichi e Chaabane (2015), por sua vez,

em trabalho de classificação de séries temporais, conseguiram atingir 90% de acerto com o modelo utilizando apenas o primeiro autovetor resultante da aplicação da ACP.

Classificação das séries temporais

A construção do conjunto de treino do modelo de classificação a partir da abordagem *Active Learning* satisfez os critérios estabelecidos em 100% dos casos para as séries sazonais, atingindo, na maioria dos casos, o grau máximo separabilidade (2,0) entre as classes para ambas as medidas de distância, JM e DT. No trabalho realizado por Cai e Zhang (2010), a distância JM obtida pelos autores variou entre 1,86 e 1,98, que na média foram piores que as distâncias obtidas no presente trabalho. Na replicação da medida de distância entre as classes para os mesmos pontos amostrais utilizando as séries SES, contudo, os resultados obtidos foram divergentes em 19% das comparações válidas, principalmente naquelas envolvendo as classes *cultura anual* e *cana-de-açúcar*, para as quais as distâncias JM e DT médias foram, respectivamente, 0,78 e 0,66. Nas safras 2006-2007 e 2009-2010, para as quais a aplicação da ACP resultou na retenção de apenas 1 autovetor, não foi possível realizar o cálculo da distância entre as classes devido a restrições na aplicação das metodologias.

Na validação do mapeamento da área plantada com cana-de-açúcar, foram amostrados pontos aleatórios na área de estudo a partir do mapeamento do projeto Canasat/INPE para comparação com os resultados gerados pela classificação das séries sazonais e séries temporais SES. Os resultados da validação do modelo de classificação para a classe minoritária (*cana-de-açúcar*) e para todas as classes estão descritos, respectivamente, nas Tabelas 1 e 2.

Tabela 1. Valores médios das métricas de avaliação da classificação das séries sazonais e séries temporais sem extração sazonal (SES) referentes à classe minoritária.

Métrica	Séries sazonais	Séries temporais SES
Falso positivo	57%	61%
Falso negativo	60%	59%
Recall	40%	41%
Especificidade	85%	80%
Precisão	43%	39%

Tabela 2. Valores médios das métricas de avaliação da classificação das séries sazonais e séries temporais sem extração sazonal (SES) referentes a classificação geral.

Métrica	Séries sazonais	Séries temporais SES
Kappa	25%	21%
Acurácia Geral	75%	72%
Recall	63%	61%
Especificidade	63%	61%
Precisão	30%	28%

Apesar de ser amplamente utilizado em trabalhos sobre classificação de dados remotos, o uso do SVM engloba, em sua maioria, a classificação de dados com maior resolução espacial

e espectral (MOUNTRAKIS; IM; OGOLE, 2011). Sua aplicação em séries temporais, no entanto, ainda representa um desafio. Em alguns exemplos encontrados sobre o tema (XIE et al., 2014; XUE; DU; FENG, 2014), nota-se a agregação de outros parâmetros, como o uso de imagens de maior resolução espacial e/ou outras bandas espectrais, para subsidiar o algoritmo e melhorar a precisão da classificação final. Neste trabalho, foram utilizadas apenas séries temporais de NDVI e os resultados da sua análise sazonal.

Em trabalho relativo à classificação de séries temporais de NDVI para monitoramento da vegetação, Carrão, Gonçalves e Caetano (2008) concluíram que a variabilidade espectral foi mais determinante que a temporal na classificação da cobertura vegetal, destacando que coberturas que apresentem características semelhantes podem ter sua distinção prejudicada na classificação final. No estado de São Paulo, o comportamento espectral da cana-de-açúcar assemelha-se ao das pastagens (XAVIER et al., 2006), também amplamente cultivada no estado. No trabalho desenvolvido por Cai e Zhang (2010), que apresenta fundamentos similares aos que foram aplicados neste trabalho, os autores obtiveram acurácia geral e índice Kappa acima de 90%, porém, enfatizaram que a sua amostragem para validação da classificação foi feita sobre áreas estrategicamente amostradas, ao passo que neste trabalho a amostragem para validação espacial do modelo foi feita inteiramente ao acaso para evitar o enviesamento dos resultados.

Na comparação da área plantada estimada pelo modelo com as estimativas do Canasat/INPE, Instituto Nacional de Geografia e Estatística (IBGE) (IBGE, 2017) Companhia Nacional do Abastecimento (Conab) (CONAB, 2017) (Tabela 3), foram observadas semelhanças que elucidam o potencial da metodologia apresentada, especialmente em relação à classificação feita utilizando as séries sazonais (SS).

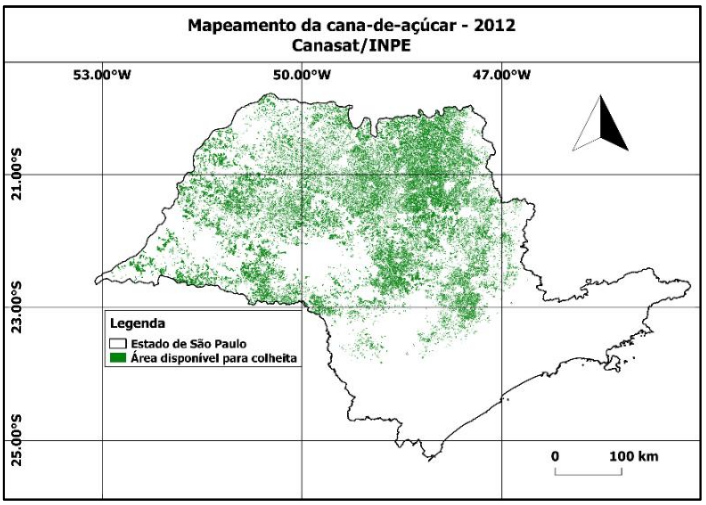
Tabela 3. Estimativa de área plantada com cana-de-açúcar pelo modelo de classificação das séries temporais e por estimativas oficiais.

Safra	Séries sazonais (SS)	Séries temporais (SES)	Canasat/INPE	IBGE	Conab
	Área plantada (hectare)				
2006-2007	4.195.048	2.128.026	3.016.262	3.498.265	3.288.200
2007-2008	2.890.060	4.193.774	3.289.761	3.890.414	3.824.200
2008-2009	4.048.791	2.194.391	3.746.039	4.541.509	3.882.100
2009-2010	2.485.076	6.060.411	4.562.179	4.977.077	4.129.900
2010-2011	4.511.049	2.850.983	4.810.327	5.071.205	4.357.000
2011-2012	4.298.490	6.590.394	4.664.610	5.216.491	4.370.100
2012-2013	4.675.054	8.232.584	4.601.335	5.172.611	4.419.460

Os dados observados na Tabela 3 mostram que as estimativas geradas pela classificação das séries sazonais, com exceção das safras 2006-2007 e 2009-2010, tiveram uma boa correlação com as estimativas oficiais utilizadas para levantamento da área cultivada com cana-de-açúcar. O índice de correlação de Pearson (r), gerado sem considerar as duas safras

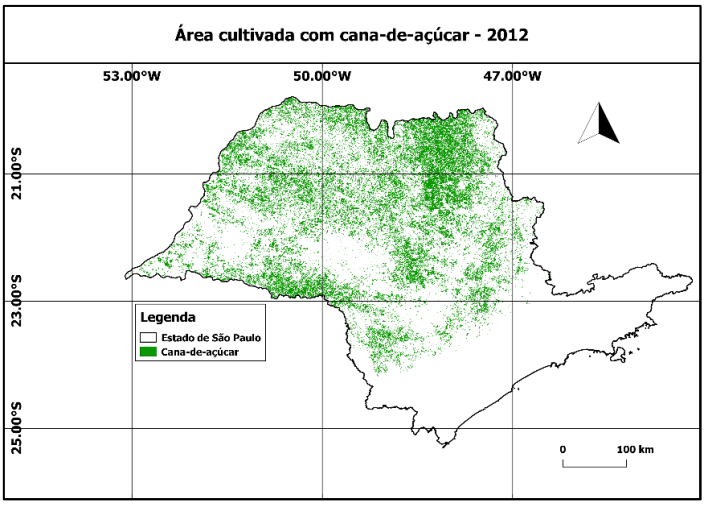
mencionadas acima, indicou correlação de 0,9 entre a estimativa gerada utilizando as séries sazonais com o mapeamento do Canasat/INPE; 0,84 com as estimativas da Conab (2017) e 0,95 com as estimativas do IBGE (2017). No caso das séries temporais SES, este índice foi de, no máximo, 0,61, mesmo após desconsiderar as safras 2006-2007 e 2009-2010. As Figuras 2 e 3 ilustram a distribuição da área plantada com cana-de-açúcar pela metodologia utilizada no projeto Canasat/INPE e pela classificação das séries sazonais.

Figura 2. Mapeamento da área cultivada com cana-de-açúcar pelo projeto Canasat/INPE.



Fonte: Adaptado de RUDORFF et al. (2010)

Figura 3. Mapeamento da área cultivada com cana-de-açúcar pela classificação das séries sazonais.



Fonte: Autor

Conclusão

Modelos de classificação de séries temporais de NDVI pelo algoritmo SVM para identificação de áreas canavieiras tem sua taxa de acerto dependente da cobertura do terreno por outros tipos de vegetação, mesmo quando aplicadas técnicas que promovam melhor diferenciação da cobertura vegetal, como a análise sazonal.

A abordagem *Active Learning* adotada na construção do conjunto de treino do modelo se mostrou eficiente na construção de um conjunto de treino compacto, mas que expressasse as principais características espectrais dos alvos a partir dos critérios de amostragem estabelecidos. Metodologias tradicionais de monitoramento agrícola baseado em imagens de satélite são muito dependentes do esforço humano demandado para a obtenção dos resultados, principalmente aqueles baseados em interpretação visual de imagens.

Excluindo-se as safras ocorridas em 2006-2007 e 2009-2010, a estimativa de área plantada obtida pela classificação das séries sazonais apresentou correlação variando entre 84% e 95% com as estimativas oficiais, porém com custo operacional sensivelmente menor. Com isso, a metodologia proposta pode ser adotada em conjunto com outras fontes de dados e outros mapeamentos afim de aumentar a precisão no mapeamento da cana-de-açúcar no estado de São Paulo.

Referências

ALMEIDA, T. I. R. et al. Principal component analysis applied to a time series of MODIS images: the spatio-temporal variability of the Pantanal wetland, Brazil. Wetlands ecology and management, v. 23, n. 4, p. 737-748, 2015.

BUCKERIDGE, M. S. et al. Ethanol from sugarcane in Brazil: a 'midway'strategy for increasing ethanol production while maximizing environmental benefits. GCB Bioenergy, v. 4, n. 2, p. 119-126, 2012.

CAI, H.; ZHANG, S. Regional land cover classification from MODIS time-series and geographical data using support vector machine. In: Information Computing and Telecommunications, 17610., 2010, Beijing. Anais… Beijing: IEEE, 2010. p. 102-105.

CARRÃO, H.; GONÇALVES, P.; CAETANO, M. Contribution of multispectral and multitemporal information from MODIS images to land cover classification. Remote Sensing of Environment, v. 112, n. 3, p. 986-997, 2008

COMPANHIA NACIONAL DO ABASTECIMENTO. Séries históricas. Disponível em: <http://www.conab.gov.br/ >. Acesso em: 09/05/2017.

CORTES, C.; VAPNIK, V. Support-vector networks. Machine learning: Springer, v. 20, n. 3, p. 273-297, 1995.

CRAWFORD, M. M.; TUIA, D.; YANG, H. L. Active Learning: Any value for classification of remotely sensed data? Proceedings of the IEEE, v. 101, n. 3, p. 593-608, 2013.

DAMATTA, F. M. et al. Impacts of climate changes on crop physiology and food quality. Food Research International, v. 43, n. 7, p. 1814-1823, 2010.

EKLUNDH, L.; JÖNSSON, P. TIMESAT 3.2 with parallel processing-Software Manual. Relatório técnico, 2012. 88p.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Produção agrícola municipal. Disponível em: <http://www.ibge.gov. br/>. Acesso em: 20 mai. 2017.

INSTITUTO NACIONAL DE METEOROLOGIA. Normais climatológicas. Disponível em: < http://www.inmet.gov.br/portal/>. Acesso em: 21 mai. 2017.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. I: Artificial data. Applied statistics, p. 160-173, 1972.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. II: Real data. Applied statistics. p. 21-31, 1973.

MANTERO, P., MOSER, G., SERPICO, S. B. Partially supervised classification of remote sensing images through SVM-based probability density estimation. IEEE Transactions on Geoscience and Remote Sensing, v. 43, n. 3, p. 559–570, 2005.

MOUNTRAKIS, G.; IM, J.; OGOLE, C. Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, v. 66, n. 3, p. 247-259, 2011.

RÉJICHI, S.; CHAABANE, F. Feature extraction using PCA for VHR satellite image time series spatio-temporal classification. In: Geoscience and Remote Sensing Symposium, 2015, Milan. Anais… Milan: IEEE, 2015., p. 485-488.

RICHARDS, J. A. Remote sensing digital image analysis: An introduction. 3.ed. Berlin: Springer, 2012. 503p.

RUDORFF, B. F. T. et al. Studies on the rapid expansion of sugarcane for ethanol production in São Paulo State (Brazil) using Landsat data. Remote Sensing, v. 2, n. 4, p. 1057-1076, 2010.

TUIA, D. et al. Active Learning methods for remote sensing image classification. Geoscience and Remote Sensing, v. 47, n. 7, p. 2218-2232, 2009.

UNIÃO DAS INDÚSTRIAS DE CANA-DE-AÇÚCAR. Dados da produção canavieira. Disponível em: <http://www.unicadata.com.br/>. Data de acesso: 15/05/2017.

WANG, K. et al. Remote sensing of ecology, biodiversity and conservation: a review from the perspective of remote sensing specialists. Sensors, v. 10, n. 11, p. 9647-9667, 2010.

XAVIER, A. C. et al. Multi-temporal analysis of MODIS data to classify sugarcane crop. International Journal of Remote Sensing, v. 27, p. 755-768, 2006.

XIE, D. et al. Autumn crop identification using high-spatial-temporal resolution time series data generated by MODIS and Landsat remote sensing images. In: Geoscience and Remote Sensing Symposium, 2014, Québec, Anais… Québec: IEEE. 2014., p. 2118-2121.

XUE, Z.; DU, P.; FENG, L. Phenology-driven land cover classification and trend analysis based on long-term remote sensing image series. Journal of Selected Topics in Applied Earth Observations and Remote Sensing, v. 7, n. 4, p. 1142-1156, 2014.

ZHANG, J.; FENG, L.; YAO, F. Improved maize cultivated area estimation over a large scale combining MODIS–EVI time series data and crop phenological information. Journal of Photogrammetry and Remote Sensing, v. 50, n. 9, p. 102-113, 2012.