Diferenciação de áreas cana de açúcar e pastagem através de técnicas de mineração de dados

Victor Danilo Manabe[1]
Jansle Vieira Rocha[1]
Rubens Augusto Camargo Lamparelli[1]

[1]Universidade Estadual de Campinas – UNICAMP/FEAGRI
Av. Candido Rondon, 501 – 13083-874 – Campinas – SP, Brasil
{victor.manabe, jansle.rocha}@feagri.unicamp.br
rubens.lamparelli@gmail.com

Abstract. The study of the sugarcane dynamics has a direct influence on the composition of agricultural production, the direct and indirect impacts on biodiversity, social and human development and the definition of public policies, among others. Therefore it becomes important to map areas of cultivation of sugarcane on a regional scale using remote sensing. This study aimed to evaluate data mining techniques to differentiate areas of sugarcane and pasture using NDVI data from Terra/MODIS sensor. Attribute selection and balancing classes contributed to the improved performance of classification models. The best result was using the neural network classifier (Multilayer Perceptron) with a 72.49% of accuracy and 0.45 Kappa index. Thus, it was noticed the potential in the application of data mining techniques for classification of crops, using time series of vegetation index.

Palavras-chave: classificação, séries temporais, índice de vegetação.

1. Introdução

A cana-de-açúcar vem tomando lugar de destaque no agronegócio brasileiro devido a grande demanda para produção de etanol causada pelo aumento da comercialização de carros bicombustível. (Rudorff et al., 2010). O estudo da dinâmica da cana de açúcar tem influência direta em questões tais como a composição da produção agrícola, os impactos diretos e indiretos sobre a biodiversidade, o desenvolvimento social e humano e a definição de políticas públicas, entre outros.

A utilização de séries temporais de índices de vegetação para mapeamento e monitoramento de culturas agrícolas apresenta potencial para produzir resultados com maior antecedência, precisão e, ainda, com menor custo operacional do que as técnicas convencionais (FAO, 1998). Além disto, essa alternativa possibilita a execução de análises compatíveis com a escala regional e nacional, com uma freqüência mais elevada, compatível com a dinâmica da agricultura e com as demandas específicas do mercado. As imagens de NDVI (Índice de Vegetação da Diferença Normalizada) do sensor Terra/MODIS apresentam larga utilização, tendo o produto MOD13Q1 imagens a partir de composições de 16 dias com resolução de 250 m (Latorre et al., 2007). Moraes (2012) e Fernandes et al. (2011), utilizaram informações extraídas de séries temporais de índice de vegetação para a identificação de áreas de cana-de-açúcar em grandes áreas no estado de São Paulo.

No sensoriamento remoto podem ser aplicadas técnicas de mineração de dados para melhorar a analise dos dados (Hansen et al., 2000). A extração de padrões relevantes e novos que estão implícitos nos dados é a base da mineração de dados. A tarefa de classificação consiste em predizer uma variável dependente em função de um conjunto de dados relacionados (Han e Kamber, 2006).

Desta forma, diversos trabalhos foram feitos com aplicações de técnicas de mineração de dados para classificação de diferentes alvos em imagens de satélite. Foody e Mathur (2006) utilizam o classificador Support Vector Machine (SVM) para o mapeamento de áreas agrícolas, tendo como amostras para geração do modelo, áreas com mistura espectral. A diferenciação de culturas anuais utilizando series temporais de índices de vegetação é feita por Peña-Barragán et al. (2011) com a tarefa de classificação de Redes Neurais. A árvore de decisão é um dos classificadores mais utilizados em sensoriamento remoto, tendo sido aplicada para o mapeamento de cana de açúcar por Vieira et al. (2012) utilizando índices de vegetação e bandas espectrais.

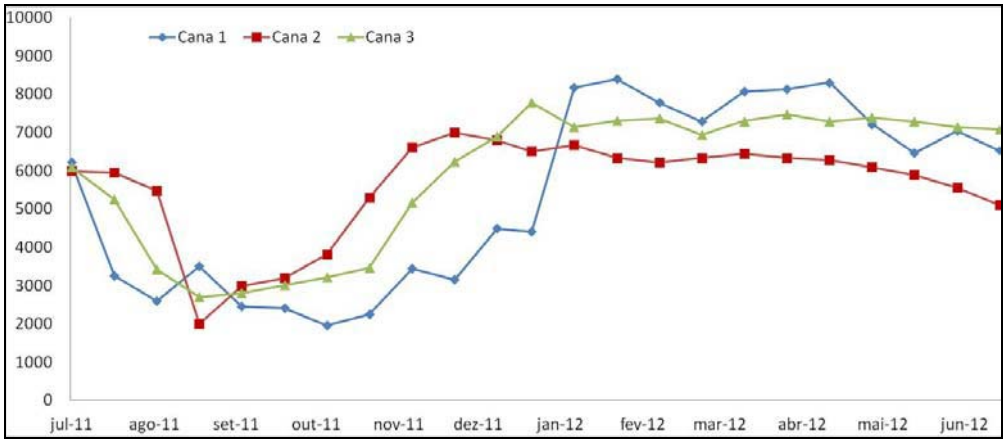
Assim, este trabalho tem como objetivo avaliar técnicas de mineração de dados para diferenciação de áreas de cana de açúcar e pastagem utilizando dados NDVI do sensor Terra/MODIS.

2. Material e Métodos

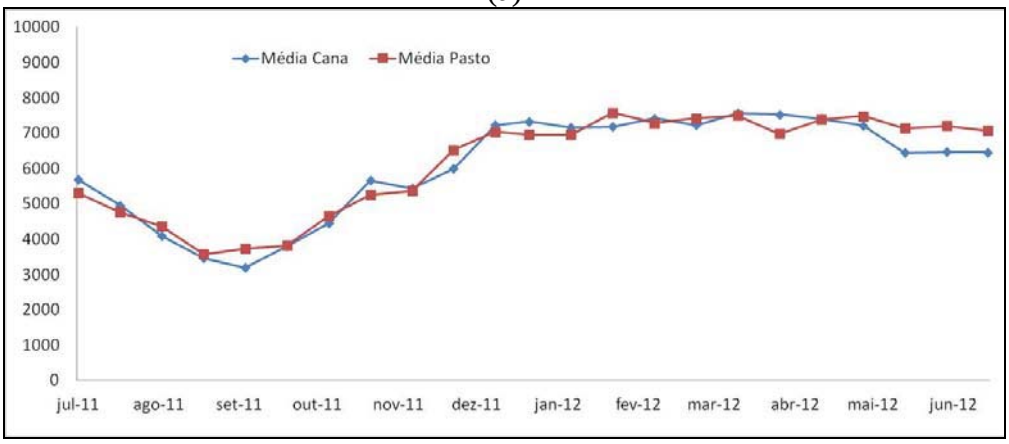
A área de estudo se localiza na região de Iturama – MG divisa com o estado São Paulo. Esta área se caracteriza pela expansão recente da cana-de-açúcar sobre áreas de pastagem. Foram coletados somente dados de áreas de cana de açúcar e pastagem tomando o cuidado de adquirir informações de pixels puros para os alvos.

O mapeamento de cana-de-açúcar pode ser realizado a partir da análise do comportamento temporal dos índices de vegetação NDVI em cada pixel ao longo da safra, partindo da premissa de que cada alvo possui um comportamento característico. A cana-de-açúcar, teoricamente, apresenta um comportamento temporal típico, apresentando valores baixos no início do ciclo, aumentando o valor conforme o acumulo de biomassa até atingir o pico vegetativo, mantendo o valor próximo ao máximo no período de maturação e com a queda na senescência e corte. Porém realizando analise com informações provenientes de campo, é constatado que a grande variedade de estágios de corte da planta, tratos culturais e variedades de cana de açúcar levam a um comportamento temporal de NDVI muito distintos e

próximos ao de pastagem (Figuras 1a e 1b). Tornando o mapeamento destas áreas de extrema dificuldade.



(a)



(b)

Figura 1: (a) Perfil temporal de NDVI de três áreas de cana de açúcar; (b) Perfis médios de NDVI de áreas de cana de açúcar e pasto.

Foram utilizadas 23 imagens, no período de 12 de julho de 2011 até 26 de Junho de 2012, de NDVI do produto MOD13Q1 do sensor MODIS para o estado do Mato Grosso, que é disponibilizado na Base Estadual Brasileira pela Embrapa (http://www.modis.cnptia.embrapa.br) (Esquerdo et al., 2011). A partir desta serie de imagens foram extraídas informações em relação a serie temporal de NDVI, Figura 2, estes dados foram trabalhados em conjunto com as imagens de NDVI, Tabela 1.

Para a seleção de atributos foram utilizados os seguintes métodos: CFS *(Correlation-based feature selection*), GainRatio (*Information Gaind Ratio*), InfoGain (*Information Gain*), PCA (*Principal Components Analysis*) com critério de corte sugerido por Jollife (1972), Wrapper e Qui-quadrado (x^2).

A quantidade de áreas que formaram o banco de dados para o estudo foram de 1348 instancias, onde 722 representaram a classe “cana de açúcar” e 626 a classe “pastagem”. O conjunto de dados foi divido em duas partes uma para treinamento com 1079 instancias (573 “cana de açúcar” e 506 “pastagem”) e outra para testes com 269 instancias (148 “cana de açúcar” e 121 “pastagem”).

A classificação ocorreu com a aplicação dos seguintes algoritmos: J48 (Arvore de decisão), MLP – *Multilayer Perceptron* (Redes Neurais) e SMO – *Sequential Miniaml Optimization* (SVM). Foi utilizado o software WEKA 3.6 (Waikato Environment for

Knowledge Analysis) na execução das técnicas seleção e reamostragem dos dados, assim como na obtenção dos modelos para os classificadores e seu teste.

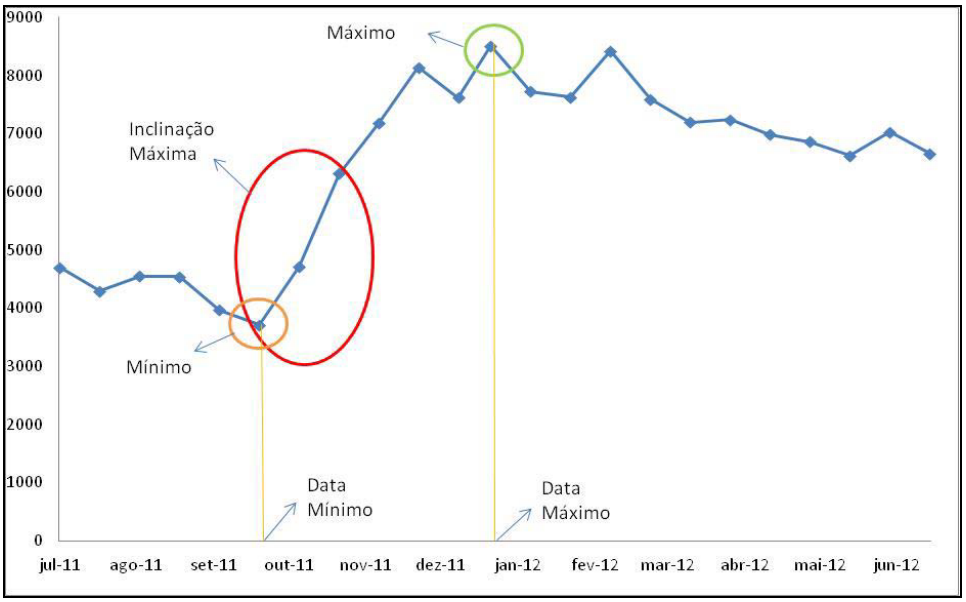


Figura 2: Indicação dos parâmetros extraídos de série temporal de NDVI.

Tabela 1: Dados extraídos para seleção de atributos.

NDVI_minimo – NDVI mínimo da série	NDVI_maximo – NDVI máximo da série
NDVI_diferenca – Diferença entre NDVI mínimo e máximo	NDVI_media – Média de NDVI da série
Data_minimo – Data do valor de NDVI mínimo □	Data_maximo – Data do valor de NDVI máximo
Inclinacao– Inclinação máxima da série	Dado_ 1 – NDVI dia 12/07/11
Dado_2 – NDVI dia 28/07/11	Dado_3 – NDVI dia 13/08/11
Dado_4 – NDVI dia 29/08/11	Dado_5 – NDVI dia 14/09/11
Dado_6 – NDVI dia 30/09/11	Dado_7 – NDVI dia 16/10/11
Dado_8 – NDVI dia 01/11/11	Dado_9 – NDVI dia 17/11/11
Dado_10 – NDVI dia 03/12/11	Dado_11 – NDVI dia 19/12/11
Dado_12 – NDVI dia 01/01/12	Dado_13 – NDVI dia 17/01/12
Dado_14 – NDVI dia 02/02/12	Dado_15 – NDVI dia 18/02/12
Dado_16 – NDVI dia 06/03/12	Dado_17 – NDVI dia 22/03/12
Dado_18 – NDVI dia 07/04/12	Dado_19 – NDVI dia 23/04/12
Dado_20 – NDVI dia 09/05/12	Dado_21 – NDVI dia 25/05/12
Dado_22 – NDVI dia 10/06/12	Dado_23 – NDVI dia 26/06/12

3. Resultados e Discussão

A etapa de seleção de atributos foi realizada para a eliminação daqueles que não apresentam grande poder preditivo para as classes. No Tabela 2 são apresentados os atributos selecionados através dos métodos mostrados anteriormente. De modo geral, foram mantidas as configurações padrões presentes no software WEKA para a execução desta etapa, apenas para o Wrapper com os algoritmos J48 e SMO foram feitas alterações, utilizando para o algoritmo J48 o numero mínimo de instancias por folha de 30, e para o SMO valores de custo de 4,0 e de gamma de 0,125.

Ao analisar os resultados da seleção de atributos, Dado_12, Dado_17 e Dado_18 foram os únicos a não aparecer em nenhum dos métodos, podendo ser considerados os atributos com menor poder preditivo. Em contrapartida o atributo Dado_1 foi o que apresentou maior incidência na seleção, não estando presente apenas no método Wrapper para SMO, sendo assim considerado aquele com maior poder preditivo. Dos atributos gerados a partir da série temporal, "NDVI_diferenca" se apresentou em maior quantidade, porém estes atributos não apareceram em grande □ndice□cia durante a seleção.

Ao analisar somente os parâmetros escolhidos relacionados aos valores NDVI em datas especificas é possível constatar que há três épocas onde a diferenciação dos dois alvos é mais representativa, ou seja, datas onde os parâmetros foram considerados importantes na seleção de atributos. Estas etapas correspondem a época de final do ciclo passado e inicio do ciclo estudado da cana-de-açúcar (12 de Julho a 29 de Agosto de 2011), inicio do pico vegetativo (17 de Novembro a 19 de Dezembro de 2011) e o final do ciclo, na fase final de senescência e colheita da cana-de-açúcar (05 de Maio a 26 de Junho de 2012).

Tabela 2: Atributos escolhidos após a etapa de seleção de atributos.

	CFS	GainRatio	InfoGain	PCA	Wrapper J48	Wrapper MLP	Wrapper SMO	x^2
NDVI_minimo					x			x
NDVI_maximo								x
NDVI_diferenca			x	x		x	x	x
NDVI_media								x
inclinacao	x						x	x
Data_maximo								x
Data_minimo				x				
Dado_1	x	x	x	x	x	x		x
Dado_2	x	x	x			x	x	x
Dado_3	x	x	x					x
Dado_4	x	x	x		x		x	x
Dado_5								x
Dado_6					x			x
Dado_7				x				x
Dado_8	x							x
Dado_9	x	x	x					x
Dado_10				x				x
Dado_11	x	x	x		x			x
Dado_12								
Dado_13					x			
Dado_14								x
Dado_15								x
Dado_16	x							x
Dado_17								
Dado_18								
Dado_19								x
Dado_20				x		x		
Dado_21				x		x		x
Dado_22		x	x	x				x
Dado_23	x	x	x					x

Para a modelagem dos classificadores e na etapa de seleção de atributos, após testes realizados, foi utilizado para o algoritmo J48 o numero mínimo de instancias por folha de 30,

e para o SMO valores de custo de 4,0 e de gamma de 0,125. Para os demais se utilizaram as configurações padrões disponíveis no WEKA 3.6. Na Tabela 3 são apresentados os resultados dos testes realizados para os classificadores J48, SMO e MLP.

Tabela 3: Comparação das técnicas de seleção de atributos.

	J48		SMO		MLP	
Métodos	**T. Acerto**	**I. Kappa**	**T. Acerto**	**I. Kappa**	**T. Acerto**	**I. Kappa**
Sem Seleção	66,54	0,32	69,52	0,38	67.29	0,34
PCA	63,57	0,28	64,68	0,29	62,45	0,26
x^2	66,91	0,33	70,26	0,40	**71,75**	**0,43**
Wrapper	66,54	0,33	71,00	0,42	70,26	0,39
Infogain	68,65	0,35	69,15	0,37	68,03	0,36
Gainratio	**69,15**	**0,36**	**71,75**	**0,42**	70,26	0,40
CFS	68,52	0,35	71,00	0,41	69,52	0,39

Na comparação entre os valores encontrados para os modelos sem seleção atributos e aqueles com seleção, é notada uma melhora significativa na Taxa de Acerto e no índice Kappa em alguns casos. A classificação com o algoritmo J48 com o método Gainratio apresentou um índice Kappa de 0,36. Os melhores resultados foram para SMO com Gainratio e MLP com x^2, que tiveram resultados de índice Kappa de 0,42 e 0,43 respectivamente. Sendo assim, a seleção de atributos contribui para a melhora dos classificadores.

Ao analisar a Tabela 3, logo foi notado que os valores de taxa de acerto e índice Kappa de todos os modelos gerados são baixos. Porém tendo em vista a dificuldade de classificação destas duas classes exclusivamente, os resultados encontrados apresentam uma boa contribuição na descoberta de conhecimento.

Outros parâmetros importantes para a análise são a Taxa de Verdadeiro Positivo, a Taxa de Falso Positivo e a Precisão para as classes. Na Tabela 4 são apresentados estes valores para o melhor método de seleção de atributos em cada classificador. Os valores da Taxa de Verdadeiro Positivo para a classe cana de açúcar foram maiores na comparação com a pastagem para os classificadores J48 e SMO, da mesma forma que a Taxa de Falso Positivo. O que indica que os modelos gerados tenderam a classificar de maneira mais agressiva a classe cana de açúcar.

Tabela 4: Métricas de avaliação dos modelos de classificação.

	Taxa VP			Taxa FP			Precisão		
Classes	**J48**	**SMO**	**MLP**	**J48**	**SMO**	**MLP**	**J48**	**SMO**	**MLP**
Cana	0,85	0,76	0,72	0,50	0,36	0,29	0,67	0,72	0,75
Pasto	0,50	0,64	0,71	0,15	0,24	0,28	0,73	0,68	0,68

Isto pode ter ocorrido devido ao desbalanceamento de classes, uma vez que existe uma maior quantidade instancias da classe cana de açúcar. Para eliminar tal problema, foi realizado no software WEKA o balanceamento de classes, com a reamostragem para 90% do conjunto de dados, sem a realocação de instancias selecionada. Foram testados os valores 0,5 e 1,0 para o parâmetro *biasToUniformClass*, sendo mantido os outros parâmetros padrões. Com esta etapa o conjunto de dados ficou com total de 971 registros, para *biasToUniformClass* de 0,5 foi obtido 486 instancias da classe cana de açúcar e 486 da classe pastagem, e para *biasToUniformClass* 1,0 foram selecionadas 465 instancias da classe cana de açúcar e 506 da classe pastagem. Então foram aplicados novamente os métodos de seleção de atributos, onde se encontrou os mesmos resultados apresentados anteriormente.

Então como base os melhores resultados de Taxa de Acerto e Índice Kappa apresentados anteriormente, foi selecionado o melhor método para cada classificador e aplicado na avaliação dos conjuntos de dados balanceado, Tabela 5.

Tabela 5: Comparação de classes desbalanceadas e balanceadas.

	J48		SMO		MLP	
Classes	**T. Acerto**	**I. Kappa**	**T. Acerto**	**I. Kappa**	**T. Acerto**	**I. Kappa**
Desbalanceadas	69,15	0,36	**71,75**	**0,42**	71,75	0,43
biasToUniformClass 0.5	**69.89**	**0,38**	70,26	0,40	**72,49**	**0,45**
biasToUniformClass 1.0	60,97	0,25	68,77	0,38	68,40	0,37

Ocorreu um aumento pouco significativo para taxa de acerto e índice kappa para os algoritmos J48 e MLP com o balanceamento com *biasToUniformClass* 0,5, já para o balanceada com *biasToUniformClass* 1,0 ocorreu a piora do modelo. Sendo o melhor resultado obtido de 72,49% de Taxa de Acerto e índice Kappa de 0,45 para o MLP de *biasToUniformClass* de 0,5.

Analisando as taxas de Verdadeiro Positivo, Falso Positivo e Precisão para os dados balanceados de melhor resultado, *biasToUniformClass* de 0,5, Tabela 6. Foi notado um maior equilíbrio entre os acertos e os erros entre as classes, para os classificadores SMO e MLP. Ocorrendo para J48 e SMO, uma diminuição da Taxa de Verdadeiro Positivo em relação ao conjunto desbalanceado para a cana de açúcar, porém a precisão é mantida no mesmo patamar.

Tabela 6: Métricas de avaliação dos modelos de classificação balanceados.

	Taxa VP			Taxa FP			Precisão		
Classes	**J48**	**SMO**	**MLP**	**J48**	**SMO**	**MLP**	**J48**	**SMO**	**MLP**
Cana	0,83	0,70	**0,73**	0,46	0,29	**0,28**	0,69	0,75	**0,76**
Pasto	0,54	0,71	**0,72**	0,17	0,30	**0,27**	0,72	0,66	**0,69**

O algoritmo MLP com classes balanceadas alcançou valores superiores aos outros, apresentando valores bons e balanceados para os três parâmetros de análise de balanceamento. Podendo ser considerado o melhor classificador testado para a diferenciação de cana de açúcar e pastagem.

4. Conclusões

Os resultados encontrados indicam potencial na aplicação de técnicas de mineração de dados para a classificação de culturas agrícolas, utilizando séries temporais de índice de vegetação provenientes de imagens de satélites.

A seleção de atributos contribuiu para a melhora no desempenho dos modelos de classificação gerados. Tendo sido o método GainRatio o melhor para os algoritmos J48 e SMO e o método chi-quadrado para o algoritmo MLP.

A realização do balanceamento de classes contribuiu para uma melhora nos classificadores, tendo como principal contribuição a equiparação nos valores de Taxa de Verdadeiros Positivos para a classes de cana de açúcar e pasto.

O melhor resultado encontrado foi para o classificador MLP com seleção de atributos e balanceamento de classes, tendo um resultado de 72,49% de Taxa de Acerto e 0,45 de Índice Kappa.

Com isto, o próximo passo para a classificação de áreas de cana de açúcar é a aplicação deste conhecimento gerado em software SIG para a geração de máscaras de cultivo, podendo assim realizar uma análise espacial dos resultados.

5. Referências Bibliográficas

Esquerdo, J. C. D. M. ; Antunes, J. F. G. ; Andrade, J. C. . Desenvolvimento do Banco de Produtos MODIS na Base Estadual Brasileira. In: XV Simpósio Brasileiro de Sensoriamento Remoto, 2011, Curitiba-PR. **Anais**... São José dos Campos: INPE, 2011. p. 7596-7602.

FAO - Food and Drug Organization of the United Nations. **Multiple frame agricultural surveys: agricultural survey programs based on area frame or dual frame (area and list) sample design**.(FAO, Statistical Development Series, 10). Rome, v.2, 242p., 1998.

Fernandes, J. L.; Rocha, J. V.; Lamparelli, R. A. C. Sugarcane yield estimates using time series analysis of SPOT vegetation images. **Scientia Agricola**, v. 68, p. 139-146, 2011.

Foody, G. M., Mathur, A. The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM. **Remote Sensing of Environment**, 103, 179–189, 2006.

Han, J.; Kamber, M. Data **mining: concepts and techniques.** 2nd ed., 770 p., San Francisco: Morgan Kaufmann, 2006.

Hansen, M. C., Defries, R. S., Townshend, J. R. G., Sohlberg, R. **Global land cover classification at 1 km spatial resolution using a classification tree approach.**

Jolliffe, I. T. Discarding variables in a principal component analysis. I: Artificial data. Applied Statistics, v. 21, n. 2, p. 160-173, 1972.**International Journal of Remote Sensing**, v. 21, n. 6, p. 1331-1364, 2000.

Latorre, M.L.; Shimabukuro, Y.E.; Anderson, L.O. **Produtos para ecossistemas terrestres – MODLAND.** p.23-35. Org: RUDORFF, B.F.T.; SHIMABUKURO, Y.E.; CEBALLOS, J.C. O sensor *MODIS* e suas aplicações ambientais no Brasil. São José dos Campos/SP: Parêntese, 423p, 2007.

Moraes, R. A. **Monitoramento e estimativa da produção da cultura de cana-de-açúcar no estado de São Paulo por meio de dados espectrais e agrometeorológicos**. Campinas, SP. 115p. Tese (Doutorado em Engenharia Agrícola). UNICAMP – Universidade Estadual de Campinas. 2012.

Peña-Barragán, J. M., Ngugi, M. K., Plant, R. E.. Johan Six Object-based crop identification using multiple vegetation indices, textural features and crop phenology. Remote Sensing of Environment, Volume 115, Issue 6, 15 June 2011, Pages 1301–1316

Rudorff, B.F.T.; Aguiar, D.A.; Silva, W.F.; Sugawara, L.M.; Adami, M.; Moreira, M.A. Studies on the Rapid Expansion of Sugarcane for Ethanol Production in São Paulo State (Brazil) Using Landsat Data. **Remote Sensing,** v. *2*, p. 1057-1076. 2010.

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**, v. 123, p. 553-562, 2012.