ESTIMATIVA DA PRODUTIVIDADE DE CANA-DE-AÇÚCAR UTILIZANDO IMAGENS LANDSAT E RANDOM FOREST

Ana Cláudia dos Santos Luciano [1,2], Daniel Garbellini Duft [1], Michelle Cristina Araújo Picoli [3], Jansle Vieira Rocha [2] e Guerric Le Maire[4,1]

[1]Laboratório Nacional de Ciência e Tecnologia do Bioetanol (CTBE), Centro Nacional de Pesquisa em Energia e Materiais (CNPEM), 13083-970, Campinas, São Paulo, Brasil, [ana.luciano; daniel.duft]@ctbe.cnpem.br; [2] Faculdade de Engenharia Agrícola (FEAGRI), Universidade Estadual de Campinas (UNICAMP), 13083-875 Campinas, Brasil, jansle@g.unicamp.br ; [3] Instituto Nacional de Pesquisas Espaciais (INPE), 12227-001, São José dos Campos, Brasil, michelle.picoli@inpe.br; [4] CIRAD, UMR Eco&Sols, Campinas, Brasil, guerric.le_maire@cirad.fr

RESUMO

O sensoriamento remoto tem contribuído para o monitoramento de área e produção da cana-de-açúcar. Neste trabalho, a produtividade de cana-de-açúcar foi estimada a partir de imagens dos satélites Landsat. Foram criados modelos calibrados e aplicados no mesmo ano e, um modelo global com calibração de 5 anos e aplicação em um ano de interesse. Os modelos foram criados com a série temporal de imagens Landsat e dados de campo, a partir do algoritmo *Random Forest*. Os dados de campo correspondem a produtividade, e preditores tipo de solos, data de colheita e variedade dos talhões. Os modelos anuais apresentaram melhores ajustes do que o modelo global (R^2=0,80 e *RMSE*=6,3ton/ha *versus* R^2=0,77 e *RMSE*=6,5 ton/ha). As principais variávies espectrais do modelo global foram índices de vegetação e bandas espectrais do infravermelho médio e infravermelho próximo. Os resultados apontam uma metodologia potencial de estimativa de produtividade da cana-de-açúcar com imagens de satélite.

Palavras-chave **—** estimativa de safra, sensoriamento remoto, aprendizado de máquina, índices de vegetação.

ABSTRACT

Remote sensing is essential for monitoring sugarcane area and production. In this study, the forecasting of sugarcane yield was done based on images from Landsat satellites. Models were calibrated and applied on the same year and a global model was calibrated on the 5 years and applied on one independent data from each year. The models were created with temporal series of Landsat images and field data with Random Forest algorithm. Field data are yield, soil type, harvest date and variety of sugarcane areas. The annual models showed better agreement than the global model (R^2=0,80 e RMSE=6,3ton/ha versus R^2=0,77 e RMSE=6,5 ton/ha). The most important spectral variables of the global model were vegetation index and bands from short wave infrared and near infrared spectral regions. The results show a potential methodology for forecasting sugarcane yield with Landsat images.

Key words **—** *yield forecasting, remote sensing, machine learning, vegetation index.*

1. INTRODUÇÃO

O Brasil é o maior produtor mundial de cana-de-açúcar e o segundo maior produtor de etanol [1]. Devido à crescente demanda do setor sucroenergético na produção de açúcar, bioetanol e bioeletricidade, análises da dinâmica espaço-temporal da cana-de-açúcar, tais como área plantada, área colhida, regiões de expansão e produção auxiliam nas tomadas de decisões gerenciais do setor.

Nas últimas décadas, o uso de dados de sensoriamento remoto tem contribuído para a extração de informações precisas da cultura de cana-de-açúcar, bem como têm permitido a gestão dessa cultura em tempo útil e adequado, com redução de custos e recursos [2,3]. As séries temporais de imagens de satélite representam uma oportunidade de avaliação da dinâmica da cobertura da terra e do monitoramento de culturas agrícolas em diferentes resoluções espaciais e temporais [3,4].

Diversos estudos relacionaram variáveis espectrais obtidas de multissensores (p.e., sensores dos satélites Landsat, SPOT, Terra e Aqua), como índices de vegetação e bandas espectrais, com a produtividade da cana-de-açúcar [5–7]. Em geral, as variáveis espectrais mais utilizadas na estimativa de produtividade da cana-de-açúcar são as refletâncias nas regiões do visível e infravermelho próximo, sendo que o índice mais comum é o *Normalized Difference Vegetation Index* (NDVI).

A combinação de variáveis espectrais com variáveis agronômicas e climáticas também auxiliam na modelagem da produção agrícola [5,8]. A modelagem da produção agrícola tem sido feita a partir de técnicas estatísticas convencionais, como regressão e correlação [9]. No entanto, atualmente, a utilização de diferentes variáveis para desenvolver modelos que relacionem a produtividade aos fatores que influenciam o crescimento da cultura, envolve técnicas avançadas de modelagem baseadas em mineração de dados [10]. Estudos mostram que a utilização de técnicas de modelagem avançadas (*p.e., artificial neural networks*, *Random Forest*, *Suport Vector Machines*) para estimativa da produtividade

agrícola é mais representativa do que as técnicas convencionais [10,11].

Diante disto, verifica-se que a estimativa da produtividade agrícola, como no caso da cana-de-açúcar, tem apresentado ganhos significativos ao se utilizarem técnicas de modelagem mais avançadas. Apesar dessa importância, a análise integrada de dados de sensoriamento remoto em conjunto com técnicas de modelagem avançada para estimativa da produtividade da cana-de-açúcar, ainda permanece por ser melhor explorada. Neste contexto, este trabalho tem como objetivo estimar a produtividade da cana-de-açúcar a partir de séries temporais de imagens Landsat utilizando modelo *Random Forest*.

2. MATERIAIS E MÉTODOS

Para estimativa da produtividade de cana-de-açúcar foram utilizados dados de campo e imagens dos sensores OLI, ETM+ e TM dos satélites Landsat. Foi selecionada uma área teste com ~363 mil ha localizada no oeste do estado de São Paulo entre os municípios de Rancharia, Quatá, Paraguaçu Paulista e João Ramalho (Figura 1). A produtividade média de cana-de-açúcar na região foi de 75,5 ton/ha no ano de 2016.

A área de estudo apresenta dados históricos (2012 a 2016) de medições da produtividade por talhões (TCH, ton/ha), os quais representam 13% da área total. Além dos dados de produtividade, informações adicionais referentes ao tipo e textura de solos, variedade, estágio de corte, idade de corte e ambiente de produção também foram obtidas dos proprietários das áreas de estudo. Na Tabela 1 estão apresentados os dados descritivos dos talhões utilizados na modelagem de produtividade em cada ano de análise.

Tabela 1. Estatística dos talhões utilizados em cada ano de análise.

Ano	Amostras (talhões)	Média (TCH)	Desvio
2012	1791	77.2	18.2
2013	2485	68.2	20.4
2014	2717	54.9	14.9
2015	2520	70.5	18.9
2016	2238	64.9	19.4

Quanto as imagens de satélite, foram selecionadas imagens dos sensores TM, ETM+ e OLI dos satélites Landsat 5,7 e 8, durante o período de janeiro de 2011 a maio de 2016. As imagens foram obtidas do banco de dados da USGS (*United States Geological Survey* -https://www.usgs.gov/), a partir da plataforma do *Google Earth Engine* (GEE), com correção atmosférica (*Collection 1*). No GEE foi feita uma composição mensal das imagens Landsat disponíveis, para preenchimentos de falhas por nuvens e de imageamento dos sensores (*p.e.*, ETM+/Landsat-7). Os valores médios de refletância de 4 bandas e de 6 índices de vegetação foram extraídos para cada talhão (Tabela 2). Para cada ano de análise (*i*) foram utilizados os dados das imagens de janeiro do ano *i-1* a maio do ano *i*. Ao total, em cada ano, foram utilizadas 170 variáveis provenientes das imagens de satélite e 14 variáveis de campo.

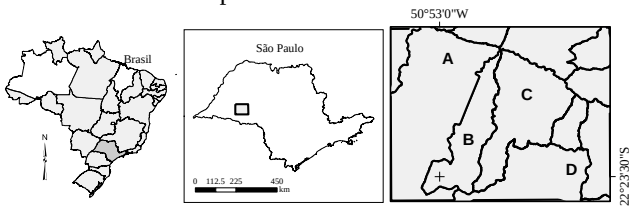

Figura 1. Localização da região de estudo no oeste do estado de SP com sobreposição dos limites municipais. A: Rancharia, B: João Ramalho, C: Quatá e D: Paraguaçu Paulista.

Tabela 2. Bandas e índices de vegetação utilizados para modelagem da produtividade

Banda/Índice	Sigla/Referência
Red	R
Near Infrared	NIR
Short wave infrared 1	$SWIR_1$
Short wave infrared 2	$SWIR_2$
Enhanced vegetation index	EVI [12]
Soil adjusted vegetation index	SAVI [13]
Normalized difference vegetation index	NDVI [14]
Normalized difference moisture index	NDMI [15]
Normalized difference water index 1	$NDWI_1$ [16]
Normalized difference water index 2	$NDWI_2$ [17]

A modelagem da produtividade foi feita com o pacote Ranger [18], implementado no software R, para calibração e aplicação do algoritmo *Random Forest*. Foram utilizados os parâmetros de árvore de decisão igual 500 (ntrees) e mtry igual ao *default*. Os modelos foram calibrados e aplicados ano a ano (modelos anuais) e, posteriormente foi feita uma calibração global com todos os anos de análise e aplicação ano a ano (modelo global). Para calibração dos modelos anuais (2012 a 2016) foram selecionadas amostras aleatórias de acordo com as variedades presentes nos talhões: 80% das amostras de cada variedade. Para o modelo global foram utilizadas amostras de treinamento de todos os anos de análise, similar a seleção feita para os modelos anuais. No entanto, a quantidade total de amostras do modelo global correspondeu a ~5 vezes o número de amostras de um único ano. Para a validação dos modelos anuais e global, foram utilizadas 20% das amostras não inseridas no processo de treinamento do algoritmo. O grau de importância das variáveis, obtido a partir do algoritmo *Random Forest*, foi calculado para o modelo global.

A capacidade de predição dos modelos anuais e do modelo global foi avaliada a partir do coeficiente de determinação R^2 entre a produtividade predita e real, o erro quadrático médio (*RMSE*) e o *Index of Agreement Modified* (d*mod*) que indica o grau de correspondência entre os valores preditos e reais.

3. RESULTADOS

Os modelos anuais apresentaram R^2 entre 0,77 e 0,83 (Tabela 3). O ano de 2012 foi o que apresentou menor ajuste em comparação aos outros anos (R^2 = 0,77). Em média, para os anos analisados, os modelos anuais apresentaram ajuste com R^2 igual a 0,80. A Figura 2 mostra a relação entre a produtividade obtida a partir do modelo anual e a produtividade real medida em campo. O grau de correspondência da predição e dos valores reais (*dmod*) variou entre 0,74 e 0,79, com média de 0,77 (Tabela 3). O erro da produtividade predita em comparação a produtividade real foi 6,3 ton/ha, em média, com erro mínimo para o ano de 2014 (*RMSE*= 5 ton/ha).

Tabela 3. Coeficientes de ajuste e erros dos modelos anuais e global aplicados nos anos de 2012 a 2016.

Ano	Anual			Global		
	R^2	*dmod*	*RMSE (ton/ha)*	R^2	*dmod*	*RMSE (ton/ha)*
2012	0,77	0,74	6,5	0,74	0,71	6,5
2013	0,83	0,77	6,5	0,80	0,73	6,5
2014	0,80	0,75	5,1	0,76	0,71	5,6
2015	0,80	0,78	6,7	0,77	0,75	6,9
2016	0,82	0,79	6,7	0,80	0,77	6,9
Média	0,80	0,77	6,3	0,77	0,73	6,5

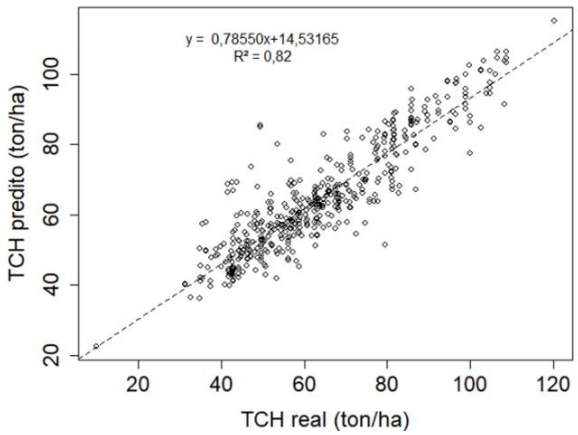

Figura 2. Dispersão dos dados preditos em comparação aos dados reais para o modelo do ano de 2016.

O modelo global apresentou R^2 médio igual a 0,77 e erro quadrático médio de 6,5 ton/ha (Tabela 3). O ano de 2012 foi o que apresentou o menor ajuste (R^2=0,74) ao se aplicar o modelo global, similiar ao que ocorreu na aplicação do modelo anual para este ano. O ajuste do modelo global (*dmod*) para cada ano variou entre 0,71 e 0,77, com média de 0,73 (Tabela 3). Em comparação aos modelos anuais, foi possível verificar menores valores do ajuste do modelo global. Ao avaliar ano a ano os erros estimativos (*RMSE*), foi verificado que os erros foram maiores para o modelo global.

As variáveis mais importantes para criação dos modelo global foram o estágio de corte, a idade e a duração do ciclo dos talhões (Figura 3). Quanto as variáveis provenientes das imagens de satélite foi possível verificar que os índices EVI e NDMI (ver Tabela 2), índices associados as bandas na região espectral do SWIR e NIR, foram os mais importantes na estimativa de produtividade. A banda $SWIR_2$ também apresentou importância na estimativa de produtivdade. Os meses mais representativos para o modelo foram do final do ano *i-1* (novembro e dezembro) e início do ano *i* (janeiro e fevereiro) (Figura 3).

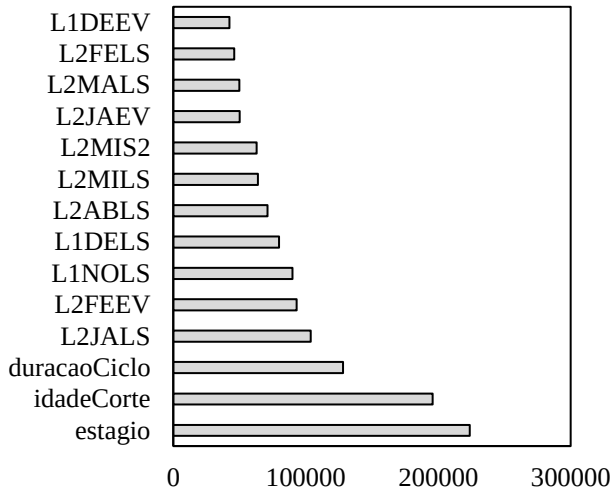

Figura 3. Variávies importantes para criação do modelo global. L1: ano *i-1*, L2: ano *i*, JA: janeiro, FE: fevereiro, MA: março, AB:abril, MI:maio, NO: novembro, DE: dezembro. LS: índice NDMI, EV: índice EVI, S2: banda $SWIR_2$.

4. DISCUSSÃO

Os resultados apresentados neste estudo mostraram que a estimativa da produtividade de cana-de-açúcar, utilizando imagens de satélite e o algoritmo *Random Forest*, teve bons ajustes entre os valores preditos e reais, com R^2 médio de 0,80 e *RMSE* médio de 6,3 ton/ha para os modelos anuais. Outros estudos estimaram a produtividade da cana-de-açúcar e obtiverem ajustes médio entre 0,64 e 0,78 com erros acima de 7 ton/ha [6,8].

As variações de ajuste dos modelos entre os anos, principalmente em relação ao ano de 2012, pode ter ocorrido em função das condições climáticas no período de análise. Isto porque, no ano de 2012, entre os meses de fevereiro e março, houve um atípico deficit de precipitação na região. Ainda, é possível que a presença de nuvens, sombras e falhas das imagens possam ter ocasionado variações nos ajustes dos modelos, em virtude da falta de informações espectrais.

A utilização de um modelo global, com calibração de vários anos, mostrou ajustes próximos aos modelos anuais. No entanto, apesar dos modelos anuais apresentarem melhores ajustes do que o modelo global (R^2 médio de 0,80 *vesus* 0,77), foi possível verificar a capacidade de se utilizar um modelo global para aplicações futuras. No presente trabalho foram utilizadas quatro bandas espectrais e seis índices de vegetação para estimar a produtividade de cana-

de-açúcar. Apesar da maioria dos estudos utilizarem apenas o índice NDVI, a utilização de outros índices de vegetação e bandas espectrais são bastante importantes para a evolução das metodologias de análise do crescimento da cana-de-açúcar [19].

As principais variáveis espectrais para a estimativa de produtividade da cana-de-açúcar foram índices associados as regiões espectrais do SWIR e NIR, os quais tem potencial para detecção de vigor vegetativo [2]. Estas variáveis apresentaram maior importância quando associadas aos meses chuvosos, nos quais a cana-de-açúcar apresenta maior vigor vegetativo.

5. CONCLUSÕES

A produtividade de cana-de-açúcar foi estimada neste estudo a partir de imagens do satélite Landsat. O uso de séries temporais de imagens reduz a possibilidade de perdas de informações por nuvens e/ou sombras para a modelagem da produtividade. Os índices de vegetação e bandas das regiões espectrais do infravermelho próximo (NIR) e infravermelho médio (SWIR) agregam informação a estimativa da produtividade de cana-de-açúcar. Os modelos *Random Forest* quando calibrados anualmente tendem a ter melhores ajustes do que os modelos calibrados com vários anos (modelo global). Para aplicações futuras, o modelo global se apresenta como uma excelente alternativa. A abordagem utilizada neste estudo para estimativa da produtividade de cana-de-açúcar mostrou-se promissora em escala local.

6. REFERÊNCIAS

[1] OCDE/FAO Brazilian agriculture: prospects and challenges. *OECD-FAO Agric. Outlook 2015* **2015**, 61–108, doi:10.1787/agr_outlook-2015-en.

[2] Abdel-Rahman, E. M.; Ahmed, F. B. The application of remote sensing techniques to sugarcane (Saccharum spp. hybrid) production: a review of the literature. *Int. J. Remote Sens.* **2008**, *29*, 3753–3767, doi:10.1080/01431160701874603.

[3] Rahman, M. M.; Robson, A. J. A Novel Approach for Sugarcane Yield Prediction Using Landsat Time Series Imagery : A Case Study on Bundaberg Region. **2016**, 93–102, doi:10.4236/ars.2016.52008.

[4] Luciano, A. C. dos S.; Picoli, M. C. A.; Rocha, J. V.; Franco, H. C. J.; Sanches, G. M.; Leal, M. R. L. V.; le Maire, G. Generalized space-time classifiers for monitoring sugarcane areas in Brazil. *Remote Sens. Environ.* **2018**, *215*, 438–451, doi:10.1016/j.rse.2018.06.017.

[5] Mutanga, S.; Schoor, C. Van; Olorunju, P. L.; Gonah, T.; Ramoelo, A. Determining the best optimum time for predicting sugarcane yield using hyper-temporal satellite imagery. *Adv. Remote Sens.* **2013**, *2*, 269–275, doi:10.4236/ars.2013.23029.

[6] Bégué, a.; Lebourgeois, V.; Bappel, E.; Todoroff, P.; Pellegrino, a.; Baillarin, F.; Siegmund, B. Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI. *Int. J. Remote Sens.* **2010**, *31*, 5391–5407, doi:10.1080/01431160903349057.

[7] Simões, M. D. S.; Rocha, J. V.; Lamparelli, R. A. C. Growth indices ans productivity in sugarcane. *Sci. Agric.* **2005**, *62*, 23–30, doi:10.1590/S0103-90162005000100005.

[8] Andrade, R. G.; Sediyama, G.; Soares, V. P.; Gleriani, J. M.; Menezes, S. J. M. da C. Estimativa da produtividade da cana-de-açúcar utilizando o Sebal e imagens Landsat. *Rev. Bras. Meteorol.* **2014**, *29*, 433–442, doi:10.1590/0102-778620130022.

[9] Murthy, V. R. K. Crop Growth Modeling and Its Applications in Agricultural Meteorology. *Satell. Remote Sens. GIS Appl. Agric. Meteorol.* **2004**, 235–261.

[10] Bocca, F. F.; Rodrigues, L. H. A. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Comput. Electron. Agric.* **2016**, *128*, 67–76, doi:10.1016/j.compag.2016.08.015.

[11] Everingham, Y.; Sexton, J.; Skocaj, D.; Inman-Bamber, G. Accurate prediction of sugarcane yield using a random forest algorithm. *Agron. Sustain. Dev.* **2016**, *36*, doi:10.1007/s13593-016-0364-z.

[12] Huete, A. R.; HuiQing Liu; van Leeuwen, W. J. D. The use of vegetation indices in forested regions: issues of linearity and saturation. *IGARSS'97. 1997 IEEE Int. Geosci. Remote Sens. Symp. Proceedings. Remote Sens. - A Sci. Vis. Sustain. Dev.* **1997**, *4*, 1966–1968, doi:10.1109/IGARSS.1997.609169.

[13] Huete, A. R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309, doi:10.1016/0034-4257(88)90106-X.

[14] Rouse, J. W.; Hass, R. H.; Schell, J. A.; Deering, D. W. Monitoring vegetation systems in the great plains with ERTS. *Third Earth Resour. Technol. Satell. Symp.* **1973**, *1*, 309–317, doi:citeulike-article-id:12009708.

[15] Wilson, E. H.; Sader, S. A. Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sens. Environ.* **2002**, *80*, 385–396, doi:10.1016/S0034-4257(01)00318-2.

[16] McFEETERS, S. K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432, doi:10.1080/01431169608948714.

[17] Rogers, A. S.; Kearney, M. S. Reducing signature variability in unmixing coastal marsh Thematic Mapper scenes using spectral indices. *Int. J. Remote Sens.* **2004**, *25*, 2317–2335, doi:10.1080/01431160310001618103.

[18] Wright, M. N.; Ziegler, A. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C ++ and R. **2017**, *77*, doi:10.18637/jss.v077.i01.

[19] Robson, A.; Abbott, C.; Lamb, D.; Bramley, R. O. B. Developing Sugar Cane Yield Prediction. In *34th Annual Conference Australian Society of Sugar Cane Technologists*; 2012; Vol. 34, pp. 1–11.