IMPORTÂNCIA DE ÍNDICES DE VEGETAÇÃO PARA MODELOS DE ESTIMATIVA DE PRODUTIVIDADE EM CANA-DE-AÇÚCAR

Eduardo Antonio Speranza[1], João Francisco Gonçalves Antunes[2], Luiz Antonio Falaguasta Barbosa[3], Geraldo Magela de Almeida Cançado[4], Julio Cezar Vansconcelos[5]

[1] Embrapa Agricultura Digital, Campinas, SP, eduardo.speranza@embrapa.br; [2] Embrapa Agricultura Digital, Campinas, SP joao.antunes@embrapa.br; [3]Embrapa Agricultura Digital, Campinas, SP, luiz.barbosa@embrapa.br; [4]Embrapa Agricultura Digital, Campinas, SP, geraldo.cancado@embrapa.br; [5]Embrapa Agricultura Digital, Campinas, SP julio.vanconcelos@colaborador.embrapa.br

RESUMO

Estimar a produtividade da cana-de-açúcar de forma confiável e com antecedência em relação à colheita é importante para a tomada de decisão do produtor. Neste contexto, este trabalho descreve o experimento realizado com o algoritmo de aprendizado de máquina Random Forest para identificação da importância de utilização de diferentes índices de vegetação obtidos de imagens suborbitais, nas diferentes fases do ciclo de desenvolvimento da cana-de-açúcar, como variáveis preditoras para a estimativa de produtividade. Foram utilizados índices de vegetação conhecidos por estimar o índice de área foliar, cobertura vegetal, volume de biomassa e clorofila presente nas plantas, e dados de produtividade de campo obtidos a partir de biometria em parcelas experimentais. Os resultados mostraram que os índices BI e BGI na fase de crescimento,e os índices NDVI e VARI na fase de maturação, possibilitaram a geração de modelos de estimativa de produtividade com menor erro dentre os índices estudados.

Palavras-chave **—** imagens suborbitais, aprendizado de máquina, florestas de decisão aleatória.

ABSTRACT

Estimating sugarcane productivity reliably and in advance of the harvest is important for the farmer's decision-making. In this context, this paper describes the experiment carried out with the Random Forest machine learning algorithm to identify the importance of using different vegetation indices obtained from suborbital images, at different stages of the sugarcane development cycle, as variables predictors for estimating yield. Vegetation indices known to estimate leaf area index, vegetation cover, volume of biomass and chlorophyll present in plants, and yield data obtained from biometry in experimental plots were used. The results showed that the BI and BGI indices, in the growth phase, and the NDVI and VARI indices, in the maturation phase, can enable the generation of yield estimation models with greater accuracy among the studied indices.

Key words **—** *suborbital imaging, machine learning, decision trees.*

1. INTRODUÇÃO

O Brasil é o maior produtor mundial de cana-de-açúcar, com uma produção de 654,5 milhões de toneladas produzidas na safra 2021/2022 [1], sendo o estado de São Paulo o responsável pela maior parte dessa produção (cerca de 54%). A cana-de-açúcar tornou-se importante para o agronegócio brasileiro devido aos recordes de exportação de açúcar e também por conta da produção do etanol para veículos a combustão [2]. Entretanto, a produtividade média da cana-de-açúcar tem se mantido estável nos últimos anos em cerca de 75 ton/ha, fazendo com que os produtores busquem soluções baseadas em tecnologias digitais para elevar a produtividade e conseguir atender às demandas previstas para os próximos anos. Nesse contexto, estimar a produtividade em áreas de produção com certa antecedência, pode garantir ao produtor tempo suficiente para tomada de decisão durante o ciclo ou até mesmo para o ciclo seguinte.

Por se tratar de uma cultura agrícola onde a produtividade é medida essencialmente pelo nível de biomassa disponível, os índices de vegetação calculados a partir de imagens orbitais e suborbitais são bons indicativos para estimativa de produtividade. No que se refere à imagens orbitais, foi desenvolvido por [3] um modelo empírico para estimativa de produtividade da cana-de-açúcar a nível regional, utilizando séries temporais de índices de vegetação gerados a partir de imagens Landsat, dados coletados em campo e modelos baseados no algoritmo de aprendizado de máquina Random Forest, atingindo estimativas com erro quadrático médio (RMSE) aproximado de 10 ton/ha. O trabalho de [4] utilizou imagens de radar com três bandas, geradas a partir de um sensor embarcado em drone, para estimar a biomassa de cana-de-açúcar usando correlação direta com um índice de amadurecimento (*Ripening Index*), e obtiveram estimativas com erro de aproximadamente 2 kg/m^2.

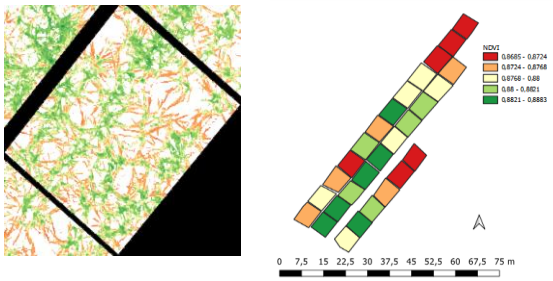
O objetivo principal deste trabalho é apresentar um estudo da viabilidade de utilização de diferentes índices de vegetação disponíveis na literatura para estimar a produtividade da cana-de-açúcar, considerando as fases fenológicas mais importantes dessa cultura, utilizando

imagens suborbitais coletadas com drone em áreas experimentais controladas em três localidades distintas, e dados coletados em campo a partir de biometria.

2. MATERIAL E MÉTODOS

As três áreas experimentais utilizadas neste trabalho são formadas por 28 parcelas de 70 m^2, totalizando cerca de 0,2 ha por experimento. As 28 parcelas possuem variações ao acaso de quatro variedades distintas de cana-de-açúcar. Os experimentos são idênticos nas três áreas, sendo duas delas localizadas no município de Piracicaba e uma no município de Charqueada, no estado de São Paulo. Em todas as áreas foi realizado o monitoramento a partir de imagens coletadas por drones, com resolução espacial entre 2 e 5 cm/pixel e, também, a biometria da produtividade em campo nas safras 2020/2021 e 2021/2022.

Devido a disponibilidade de imagens em todas as fases de desenvolvimento da cultura, inicialmente foi verificada a importância dos índices de vegetação nas principais fases de crescimento da cana-de-açúcar: germinação (0-45 dias após o plantio/corte); perfilhamento (45-120 dias após o plantio/corte); crescimento (120-250 dias após o plantio/corte); e maturação (250-360 dias após o plantio/corte) . Os índices de vegetação utilizados em cada fase, em cada área, também dependeram da disponibilidade de imagens, uma vez que em alguns casos apenas imagens no espectro visível (RGB) estavam disponíveis. Os índices de vegetação utilizados neste experimento, todos eles citados na literatura por serem bons estimadores de cobertura vegetal, área foliar, clorofila e biomassa, foram: Bright Index (BI), Green Leaf Index (GLI), Normalized Green Red Difference Index (NGRDI), Visible Atmospherically Resistant Index (VARI), Blue Green Pigment Index (BGI), Normalized Difference Vegetation Index (NDVI), Ratio Vegetation Index (RVI), Normalized Difference RedEdge Index (NDRE), Green Normalized Difference Vegetation Index (GNDVI), Triangular Vegetation Index (TVI), Enhanced Vegetation Index (EVI), Chlorophyll index – green (CIG), Chlorophyll index – red edge (CIRE) e Difference Vegetation Index (DVI). Esses índices foram gerados a partir da utilização do pacote *FieldImageR* [5] do software *R*, considerando etapas de pré-processamento para que apenas pixels referentes à vegetação fossem utilizados. Os valores médios de cada índice de vegetação em cada parcela experimental foram calculados com o QGIS. A Figura 1 representa uma das áreas experimentais com índice de vegetação calculado em sua resolução espacial original (parcela) e com os valores médios após o tratamento.



(a) **(b)**

Figura 1. Índice de vegetação NDVI calculado a partir de imagem suborbital (detalhe de uma parcela) (a); e calculado a partir de valores médios de cada parcela (b).

Para avaliar a importância de cada índice de vegetação listado na Tabela 1, em cada fase de desenvolvimento da cana-de-açúcar, foi utilizado o algoritmo de *ensemble* Random Forest [6], disponível no pacote *randomForest* do software *R*. O modelo de *ensemble* foi configurado para utilizar 501 árvores de decisão, sendo o conjunto de dados formado apenas pelos atributos - índice de vegetação médio de cada parcela em cada área, em cada período - que não possuíam valores ausentes; e pelo valor de produtividade (toneladas de cana por hectare) em cada safra em cada área. Esse conjunto de dados teve suas amostras divididas em 70% para treinamento e 30% para teste. As medidas utilizadas para detectar a importância de cada índice de vegetação foram: *%IncMSE*, que indica a percentagem de erro do modelo, caso determinado índice de vegetação não seja utilizado; e *IncNodePurity*, que indica o aumento do erro do modelo quando determinado índice de vegetação é permutado aleatoriamente nas árvores de decisão. Para avaliar o erro de possíveis modelos de estimativa de produtividade, foram utilizadas as medidas *RMSE* (erro quadrático médio) e *MAE* (erro médio absoluto).

3. RESULTADOS

Inicialmente foi verificada a importância de cada índice de vegetação gerado, em cada fase de desenvolvimento da cana-de-açúcar, tomando como base o modelo e medidas previamente definidas, considerando cada uma das três áreas experimentais estudadas de maneira diferenciada. Devido à grande quantidade de pixels com ausência de vegetação, ou, seja, com solo exposto nas fases de germinação e perfilhamento, a importância dos índices de vegetação coletados nessas fases foi muito baixa. Assim, foi observado maior impacto nas medidas *%IncMSE* e *IncNodePurity* para valores coletados nas fases de crescimento e maturação. A partir dessas medidas, e considerando o maior impacto de cada índice de vegetação considerando essas duas fases de crescimento nas três áreas experimentais estudadas, os índices BI [7] e BGI [8] na fase de crescimento; e VARI [9] e NDVI [10] na fase de maturação, foram escolhidos como variáveis preditoras para possíveis modelos de estimativa de produtividade.

A partir dessa seleção, novos modelos foram gerados, considerando os índices de vegetação selecionados.

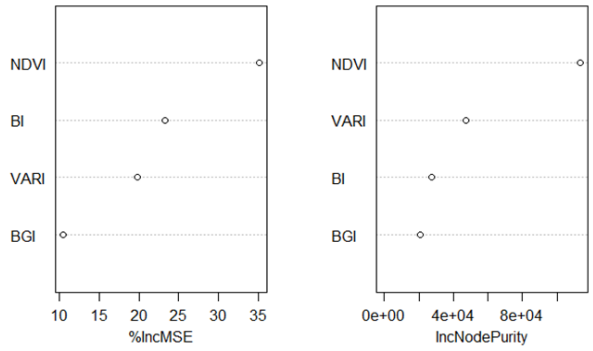


Figura 2. *Valores de %IncMSE* e *IncNodePurity* para os índices de vegetação NDVI, VARI, BI e BGI.

Entretanto, as amostras das três áreas em questão para as duas safras estudadas foram utilizadas como um conjunto de dados único, gerando 168 amostras com os atributos BI, BGI, VARI, NDVI como variáveis preditoras; e TCH (toneladas de cana por hectare) como variável resposta. Novamente, o conjunto de dados gerados foi dividido em 70% para treinamento e 30% para teste e a quantidade de 501 árvores de decisão foi mantida. A análise de importância das quatro variáveis preditoras selecionadas foi realizada considerando esse novo conjunto de dados (Figura 2)

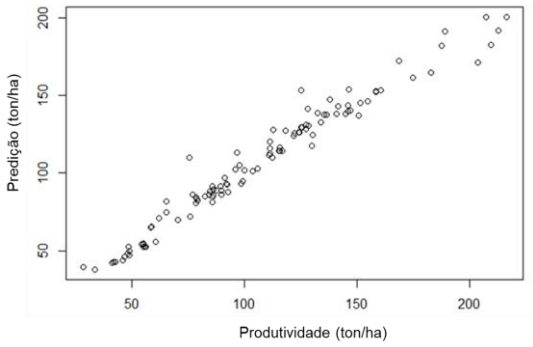
Considerando a importância das variáveis conforme os gráficos da Figura 2, foram geradas quatro configurações distintas de variáveis preditoras para geração de modelos para estimativa de produtividade. Os resultados de treinamento e teste, considerando as medidas de erro *RMSE* e *MAE* para essas 4 configurações são exibidos na Tabela 1.

Fase (C)	Fase (M)	RMSE (Tr.)	MAE (Tr.)	RMSE (Te.)	MAE (Te.)
BI e BGI	NDVI e VARI	8,76	5,82	17,9	13
BI	NDVI	9,56	6,37	21,4	16,1
-	NDVI e VARI	10,6	7,2	22,4	16,3
BI	NDVI e VARI	9,71	6,43	19,0	14,1

Tabela 1. Erro quadrático médio nas fases de treinamento e teste (RMSE Tr. e RMSE Te.) e erro médio absoluto nas fases de treinamento e teste (MAE Tr. e MAE Te.) para modelos de estimativa de produtividade considerando diferentes variáveis preditivas nas fases de crescimento (C) e maturação (M) da cana-de-açúcar. Valores em ton/ha.

Os valores de RMSE e MAE obtidos na Tabela 1 mostram melhor resultado para o modelo que utilizou as quatro variáveis selecionadas (duas para cada fase fenológica), tanto na fase de treinamento, quanto na fase de teste. A Figura 3 exibe os gráficos de correlação entre a produtividade real e a produtividade predita para esse modelo para as fases de treinamento e teste, respectivamente..

(a)



(b)

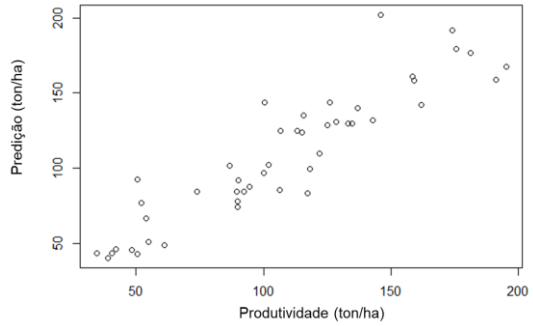


Figura 3. Gráficos de correlação entre valores reais de produtividade e valores preditos, em ton/ha, obtidos nas fases de treinamento e teste de modelo gerado.

4. DISCUSSÃO

Os gráficos da Figura 2 mostram que os índices de vegetação BI e BGI, na fase de crescimento e os índices NDVI e VARI, na fase de maturação, são variáveis preditoras capazes de proporcionar a geração de modelos de estimativa de produtividade com acurácia satisfatória. A não utilização de algum desses índices como variável preditora para a geração do modelo, conforme exibido na Tabela 1, proporciona o aumento do erro dos modelos gerados, indicando que cada um dos quatro índices, em cada período de coleta, tem relativa importância para a estimativa da produtividade.

Com relação ao erro da predição, é possível verificar uma diferença entre 7 e 9 ton/ha entre nos valores de RMSE e MAE nas fases de treinamento e teste, considerando o

melhor modelo. Porém, a partir dos gráficos da Figura 3, é possível observar que a maior parte desse erro está concentrada em valores acima de 150 ton/ha, onde a quantidade de amostras disponibilizadas para treinamento e teste é menor. Na prática, valores acima de 100 ton/ha são raros atualmente em áreas comerciais, e que, se forem excluídos da base de dados, podem diminuir o erro do modelo.

5. CONCLUSÕES

Este trabalho teve o objetivo de identificar a capacidade de diferentes índices de vegetação em diferentes fases de desenvolvimento da cultura atuarem como variáveis preditoras para a produtividade de cana-de-açúcar. Os índices utilizados foram gerados a partir de imagens obtidas com drones e biometria da produtividade em campo, em experimentos controlados com parcelas com dosagens de insumos e variedades escolhidas ao acaso. Os resultados preliminares mostram que os índices de vegetação relacionados à área foliar, cobertura vegetal e clorofila, como BI e BGI, na fase de crescimento; e índices relacionados à quantificação de biomassa, como NDVI e VARI, na fase de maturação, são potenciais candidatos a serem utilizados como variáveis preditoras para a estimativa de produtividade em cana-de-açúcar, considerando as medidas de importância de variável e erro dos modelos identificadas pelo algoritmo de *ensemble* Random Forest. Uma nova etapa de testes dos modelos produzidos, a partir de índices de vegetação e dados de produtividade obtidos de áreas comerciais, precisa ser realizada para que a indicação dos quatro índices de vegetação obtidos neste trabalho seja validada.

6. REFERÊNCIAS

[1] CONAB. *Boletim de Safra de Cana-de-Açúcar. 2º. Levantamento – Safra 2022/2023*. Disponível em <https://www.conab.gov.br/info-agro/safras/cana/boletim-da-safra-de-cana-de-acucar>. Acesso em 13 out. 2022.

[2] EMBRAPA. *Visão de Futuro do Agro Brasileiro*. Disponível em <https://www.embrapa.br/visao-de-futuro-do-agro-brasileiro>. Acesso em 13 out. 2022.

[3] Luciano, A. C. S., Picoli, M. C. A., Duft, D. G., Rocha, J. V., Leal, M. R. L. V., and Le Maire, G. . Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. *Computers and Electronics in Agriculture*, *184*, 106063, 2021.

[4] Oré, G, Alcântara, M.S., Góes, J.A., Teruel, B., Oliveira, L.P, Yepes, J., Castro, V., Bins, L.S., Castro, F., Luebeck, D., Moreira, L.F., Cintra, R., Gabrielli, L.H., Hernandez-Figueroa, H.E. Predicting Sugarcane Harvest Date and Productivity with a Drone-Borne Tri-Band SAR. *Remote Sensing*, v. 14, n. 7, p. 1734, 2022.

[5] Matias, F. I., Caraza-Harter, M.V., Endelman, J. B. FIELDimageR: An R package to analyze orthomosaic images from agricultural field trials. *The Plant Phenome Journal*, v. 3, n. 1, p. e20005, 2020.

[6] Ho, T. K. Random decision forests. *In: Proceedings of 3rd international conference on document analysis and recognition*. IEEE, 1995. p. 278-282.

[7] Richardson, A. J.; Wiegand, C. L. Distinguishing vegetation from soil background information. *Photogrammetric engineering and remote sensing*, v. 43, n. 12, p. 1541-1552, 1977.

[8] Zarco-Tejada, P. J., Berjón, A., López-Lozano, R., Miller, J.R., Martín, P., Cachorro, V., González, M. R., de Frutos, A. Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy. *Remote Sensing of Environment*, v. 99, n. 3, p. 271-287, 2005.

[9] Gitelson, A. A., Kaufman, Y. J., Stark, R., Rundquist, D. Novel algorithms for remote estimation of vegetation fraction. *Remote sensing of Environment*, v. 80, n. 1, p. 76-87, 2002.

[10] Rouse, J. W., Haas R.H., Schell, J.A., Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS, *In: S.C. Freden, E.P. Mercanti, and M. Becker (eds) Third Earth Resources Technology Satellite–1 Syposium.* Volume I: Technical Presentations, NASA SP-351, NASA, Washington, D.C., pp. 309-317, 1973.