Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Estimativa de produtividade da cana-de-açúcar a partir de imagens do satélite Sentinel-2A e o algoritmo de aprendizagem de máquina *Random Forest*

Rafaella Pironato Amaro

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Engenharia de Sistemas Agrícolas

Piracicaba
2023

Rafaella Pironato Amaro
Engenheira Agrícola

Estimativa de produtividade da cana-de-açúcar a partir de imagens do satélite Sentinel-2A e o algoritmo de aprendizagem de máquina *Random Forest*

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientadora:
Profa. Dra. **ANA CLAUDIA DOS SANTOS LUCIANO**

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Engenharia de Sistemas Agrícolas

Piracicaba
2023

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Amaro, Rafaella Pironato

Estimativa de produtividade da cana-de-açúcar a partir de imagens do satélite Sentinel-2A e o algoritmo de aprendizagem de máquina *Random Forest* / Rafaella Pironato Amaro - - versão revisada de acordo com a Resolução CoPGr 6018 de 2011. - - Piracicaba, 2023.

67 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Monitoramento da cana-de-açúcar 2. Índices de vegetação 3. Red-edge 4. Seleção de variáveis 5. Sensoriamento remoto I. Título

AGRADECIMENTOS

Agradeço a Deus por preencher meus dias com força e bom ânimo.

Aos meus pais João e Rosana (*in memorian*), e a minha irmã que sempre me apoiaram e me deram todo o suporte para aprender e me desenvolver. Ao meu marido Bruno pelo amor incondicionar e paciencia diária. Com vocês tudo o que tenho produzido na vida é melhor.

Aos meus amigos e familia, que mesmo de longe se fizeram perto.

Agradeço à minha orientadora Ana pelo conhecimento compartilhado, compreensão e amizade.

Agradeco aos colegas de trabalho do CTC por todo o apoio, ajuda e amizade.

Agradeco à FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) pelo apoio ao projeto. Projeto financiado 'Imagens de satélite e aprendizado de máquina para estimativa de produtividade de cana-de-açúcar em regiões do estado de São Paulo' (número de concessão 2021/11183-5).

Agradeco à Coordenação de Aperfeiçoamento Pessoal de Nível Superior- Brasil (CAPES), que financiou em partes esse estudo (número de concessão 001).

Ao Centro de Tecnologia Canavieira (CTC) pelo tempo disponibilizado a fim de me dedicar ao meu desenvolvimento e por viabilizar parcerias com usinas de cana-de-açúcar.

Agradeço à Escola Superior de Agricultura "Luiz de Queiroz" ESALQ/USP e ao programa de pós-graduação em Engenharia de Sistemas agrícolas pela infraestrutura da qual eu dispus.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

EPÍGRAFE

“Se você não está falhando de vez em quando, é um sinal de que você não está fazendo nada de inovador.”

Woody Allen

RESUMO

Estimativa de produtividade da cana-de-açúcar a partir de imagens do satélite Sentinel-2A e o algoritmo de aprendizagem de máquina *Random Forest*

A cana-de-açúcar é uma das culturas mais importantes para a economia brasileira, por isso, técnicas de aprendizado de máquina são utilizadas como importantes ferramentas de estimativa da produtividade. O objetivo deste trabalho foi criar modelos empíricos utilizando dados agronômicos, climáticos e de imagens de satélite, a partir do algoritmo *Random Forest,* para estimar a produtividade da cana-de-açúcar antes da colheita, no estado de São Paulo (SP). Para isso, foram utilizadas imagens Sentinel-2A; dados agronômicos; balanço hídrico da cultura e dados climáticos. Para selecionar as variáveis preditoras mais importantes foram criados modelos de estimativa de produtividade com três conjuntos de dados de uma usina: i) o primeiro conjunto de dados utilizou as variáveis agronômicas, climáticas, o balanço hídrico da cultura, índices de vegetação e bandas espectrais; ii) no segundo conjunto de dados, as variáveis fortemente correlacionadas foram removidas; e iii) o terceiro conjunto de dados foi criado com base na seleção de varáveis mais importantes pelo índice de Gini. Os modelos criados com o conjuntos de dados i, ii, iii apresentaram R^2 entre 0,77 e 0,8, RMSE entre 8,2 e 8,6 ton ha $^{-1}$, MAE entre 4,9 e 5,26 ton ha $^{-1}$ e d-Willmott entre 0,93 e 0,94, sendo o melhor modelo com o conjunto de dados iii. As variáveis mais relevantes para estimar a produtividade da cana-de-açúcar foram o estágio de corte, o déficit hídrico, os indices NDVIRE e CIRE, além das bandas *Red-edge*, NIR-8A e SWIR1. A seleção das variáveis importantes reduziu a dimensionalidade dos dados e melhorou o desempenho do modelo. Após a identificação das variáveis preditoras mais importantes, foram criados três modelos operacionais para aplicação em escala regional, com 70% de dados para treino e 30% para teste. Para isso, foram utilizados dados de 3 usinas localizadas no estado de SP. O Modelo I (geral) considerou os dados de todas as usinas para treino e teste; o Modelo II foi similar ao I para o treino, porém foi testado em cada uma das usinas de forma separada; para o Modelo III o treinamento e teste foi feito com base em dois ciclos de produção da cana de açúcar (cana-planta e cana-soca). O Modelo I apresentou R^2 igual a 0,72 enquanto os R^2 do Modelo II ficaram entre 0,60 e 0,78, o RMSE para o Modelo I foi igual a 11,7 ton ha $^{-1}$enquanto o Modelo II de 8,62 a 15,56 ton ha $^{-1}$, rRMSE foi igual a 16,5% para o Modelo I e 12,4 a 21,6%, para o Modelo II. O Modelo III apresentou R^2 maior que 0,61, e RMSE entre 9,6 e 13,5 ton ha $^{-1}$. Quando se comparou o rendimento médio com os erros RMSE, obtém-se um melhor desempenho para o modelo III com rRMSE inferior a 15,3%. A utilização do *Random Forest* para a criação de modelos globais para estimativa da cana-de-açúcar no estado de São Paulo mostrou-se promissora quando calibrado com três usinas e, separados em ciclos de produção da cana-de-açúcar (cana-planta e cana-soca).

Palavras-chave: Monitoramento da cana-de-açúcar, Índices de vegetação, Borda vermelha, Seleção de variáveis, Sensoriamento remoto

ABSTRACT

Sugarcane yield estimation from Sentinel-2A satellite imagery and Random Forest machine learning algorithms

Sugarcane is a very important crop for the Brazilian economy, so machine learning techniques are being used as an important tool to improve yield estimation. This study aimed to create an empirical model using agronomic, climatic, and satellite images, by Random Forest algorithm, to estimated sugarcane yield before the harvest, in São Paulo state (SP). We used radiometric bands and vegetation indices from Sentinel-2 images; agronomic data; crop water balance and climatic data. To select the most important variables it were builted yield estimation models based on three datasets from one mill: i) the first dataset used agronomic data, climatic data, crop water balance, and remote sensing data); ii) in the second dataset, the most strongly correlated variables were removed; and iii) the third dataset was created with the variables selected by feature selection using the Gini index. The models created with the datasets i, ii, and iii showed R^2 from 0.77 to 0.8, RMSE from 8.2 to 8.6 ton ha $^{-1}$, MAE from 4.9 to 5.26 ton ha $^{-1}$ and d-Willmott from 0.93 to 0.94, where the best result was using dataset 3 (iii). The most relevant variables to estimated sugarcane productivity were number of harvests, water deficit, NDRE and CIRE vegetation indices and Red-edge, NIR-8A and SWIR1 bands. The variable selection reduced the dimensionality of the data and improved the models' performance. After the selection of the most important predictor variables, it was created three operational models for application on the regional scale, using 70% of data to train and 30% to test. For this, we used data from three mills located in SP. The Model I (general) considered data from all mills for training and testing; Model II was similar to I for training, however, it was tested in each mill independently; for Model III the training and testing were made based on two groups of the sugarcane production cycles (plant cane and sugarcane ratoons). The results for Model I showed R^2 equal to 0.72 while the R^2 of Model II were between 0.60 and 0.78, RMSE for Model I was equal to 11.7 ton ha $^{-1}$ while Model II from 8.62 to 15.56 ton ha $^{-1}$, rRMSE was equal to 16.5% for Model I and 12.4 to 21.6%, for Model II. Model III showed R^2 greater than 0.61, and RMSE between 9.6 and 13.5 ton ha $^{-1}$. When average yield was compared with RMSE errors, better performance is obtained for Model III with rRMSE less than 15.3%.The use of Random Forest to create general models for sugarcane yield estimation in the state of Sao Paulo showed promise when calibrated with three mills and, separated by sugarcane production cycles.

Keywords: Sugarcane monitoring, Vegetation indices, Red-edge, Variable selection, Remote sensing, Remote sensing

1. INTRODUÇÃO

No Brasil, a cana-de-açúcar ocupa aproximadamente 11,2 milhões de hectares (IBGE, 2023), sendo que a área plantada e produção da cultura duplicaram nos últimos 20 anos (IBGE, 2022). Na safra 2022/2023, a produção nacional de cana-de-açúcar apresentou um aumento de 3,4% em relação a 2021/2022 (CONAB, 2023), onde a região Centro-Sul foi responsável por 90% da cana-de-açúcar produzida, representando o maior eixo produtivo do país (CONAB, 2023).

A cana-de-açúcar pode ser usada para a produção de açúcar, parte indispensável na alimentação humana, e na produção de álcool tanto para a produção das bebidas alcoólicas, como a cachaça, quanto para a produção de combustíveis para veículos, também conhecido como etanol. No Brasil, a produção total de etanol (anidro e etílico) alcançou 29,9 bilhões de litros em 2021, destinado principalmente ao setor de transportes (EPE, 2022). O etanol é um biocombustível de grande importância, pois apresenta vantagens ambientais, relacionadas a menor emissão de gases de efeito estufa , maior eficiência no uso da terra e melhor método de descarte (Borrion et al., 2012). Desde o lançamento dos veículos *flex*, em março de 2003, até outubro de 2021, o uso do etanol evitou a emissão de 570 milhões de toneladas de CO_2 na atmosfera (UNICA, 2022).

Além da produção de etanol e açúcar, os subprodutos da cana-de-açúcar, como a palha e o bagaço, têm grande importância econômica, considerados uma das grandes alternativas para o setor de biocombustíveis (CONAB, 2021; Surendran et al., 2016). Os subprodutos da cana-de-açúcar podem ser utilizados na alimentação animal, fertilização do solo, produção de concreto ecologicamente correto, bioplásticos, biogás e biometano (UNICA, 2022; Yogitha et al., 2020). Além disso, o bagaço excedente obtido nas usinas de cana-de-açúcar e a palha deixada no campo durante a colheita, também podem ser coletados e utilizados para produção de bioeletrecidade, que auxilia no aumento da segurança energética do país, principalmente em épocas de pouca chuva. Ainda, o bagaço e a palha podem ser utilizados como matéria-prima para biocombustíveis de segunda geração, como o etanol 2G. Dessa forma, a utilização dos subprodutos da cana-açúcar possibilita o incremento na produção do biocombustível sem a necessidade de aumento da área cultivada (UNICA, 2022).

Nos últimos anos, a constante procura por combustíveis renováveis para substituição do petróleo e redução dos gases do efeito estufa (GEE), em conjunto com a crescente necessidade da diminuição do impacto de mudanças climáticas, influenciou o surgimento de políticas públicas, como as políticas que proíbem a queima da cana-de-açúcar e, a Política Nacional de Biocombustíveis (RENOVABIO- lei nº 13.576/2017). O RENOVABIO tem por objetivo ampliar a participação dos biocombustíveis na matriz energética brasileira, alinhado com o compromisso do país de descarbonização e, o cumprimento dos acordos mundiais contra as mudanças climáticas causadas pelo ser humano, acordados nas chamadas Contribuições Nacionalmente Determinadas (NDC, sigla em inglês para *Nationally Determined Contributions*) como as geradas pelo Acordo de Paris (MRE, 2015).

Em função da importância econômica, social e ambiental da cana-de-açúcar é de extrema importância a utilização de métodos que forneçam uma avaliação oportuna e precisa do desenvolvimento e da produção da cultura e, que contribuam para o aumento da sustentabilidade na produção de alimento e biocombustíveis (FAO, 2017; IPCC, 2018). Neste contexto, o monitoramento da produção de cana-de-açúcar auxilia no planejamento do setor sucroenergético e na criação de políticas de segurança alimentar, sendo crescente a necessidade de melhoria da precisão e robustez dos sistemas de monitoramento de culturas agrícolas no Brasil e no mundo (Holzman et al., 2014).

O monitoramento da produção agrícola nacional, de acordo com os métodos tradicionais, é conduzido por meio de pesquisas agrícolas, que são feitas mediante a realização de questionários e entrevistas, ou por especialistas,

com base em avaliações das condições visuais das culturas, produção histórica da área, manejo, ocorrência de pragas e doenças, além das condições ambientais (IBGE, 2018). Ressalta-se que os métodos tradicionais são subjetivos, demorados e, muitas vezes pouco representativos devido ao pequeno tamanho da amostra, já que de modo geral a avaliação da cultura é feita percorrendo somente o entorno da área cultivada, o que não leva em consideração toda a variabilidade espacial das áreas produtivas (Basso et al., 2013).

Para combater a subjetividade dos métodos tradicionais de predição da produtividade agrícola, a utilização de dados de sensoriamento remoto, com base em imagens de satélite, é uma alternativa promissora. Os dados provenientes do sensoriamento remoto, como as séries temporais de imagens de satélite, quando aplicados de forma adequada podem resultar em melhorias na detecção, monitoramento e previsão de áreas agrícolas, auxiliando os produtores agrícolas na tomada de decisão como o manejo das áreas, reduzindo custos e melhorando a produtividade da lavoura (Abdel-Rahman and Ahmed, 2008).

As imagens de satélite têm sido amplamente utilizadas no monitoramento de culturas agrícolas para avaliação geral do estado da cultura (Barbanti et al., 2018; Lukas et al., 2016) e estimativa de produtividade de culturas como trigo e milho (Lai et al., 2018; Peroni Venancio et al., 2020; Schwalbert et al., 2018) e da cana-de-açúcar (Luciano et al., 2021; Mulianga et al., 2013; Singla et al., 2020). De modo geral, estudos direcionados ao monitoramento agrícola utilizam imagens provenientes dos satélites multiespectrais LANDSAT e MODIS (Lai et al., 2018; Liao et al., 2019; Mirasi et al., 2021), além da combinação de ambos os satélites (Dubey et al., 2018; Liao et al., 2019). Dentre os dados mais utilizados no monitoramento de produtividade agrícola, estão as imagens multiespectrais com bandas espectrais distintas, como por exemplo da região do infravermelho e vermelho e, principalmente, os índices de vegetação (Lai et al., 2018; Peroni Venancio et al., 2020).

Os índices de vegetação mais usuais são o Índice de Vegetação por Diferença Normalizada (NDVI) (Rouse et al., 1973) e o Índice de Vegetação Aprimorado (EVI) (Huete et al., 1997). No entanto, ressalta-se que os sensores multiespectrais aplicados ao monitoramento agrícola, especialmente no que diz respeito a estimativa de produtividade, em sua maioria detectam a radiação em bandas que compreendem o intervalo do visível e infravermelho, mas não levam em consideração a banda do *Red-edge*, a qual está localizada entre a banda do vermelho e do infravermelho próximo.

O posicionamento das bandas do *Red-edge* são relevantes para detectar condições de vegetação que estão positivamente correlacionadas com o rendimento final da cultura, como o índice de área foliar (LAI) e biomassa (Dong et al., 2019; Kross et al., 2015), demonstrando resultados mais relevantes em comparação a índices espectrais derivados apenas das bandas posicionadas na região do visível e do infravermelho próximo (Dong et al., 2015; Nguy-Robertson et al., 2014; Viña et al., 2011). Satélites como o Sentinel-2, possuem sensores que são capazes de detectar informações nas regiões do espectro posicionadas na região do *Red-edge*.

Shendryk et al. (2021) estimaram a produtividade da cana-de-açúcar a nível das áreas produtivas na Austrália utilizando imagens do satélite Sentinel-1 e Sentinel-2, além da altitude, o tipo de solo e informações climáticas. Os resultados indicaram que os índices espectrais derivados das bandas *Red-edge* e infravermelho próximo foram mais relevantes para a estimativa da produtividade do que os comumente usados NDVI e GNDVI (*Green* NDVI). De forma similar, Dimov *et al.* (2022) utilizaram dados do satélite Sentinel-2 entre os anos de 2018 e 2019 para realizar a estimativa da produtividade da cana-de-açúcar, em uma área de estudo na Etiópia. Ao utilizar o algoritmo *Random Forest*, os autores tiveram melhores resultados (R^2=0,65) para os preditores que utilizaram o *Red-edge* (NDRE e CIRE), em relação ao índice de vegetação NDVI (R^2=0,60).

Dentre as metodologias utilizadas para estimativa da produtividade com imagens de satélite, destaca-se o uso de modelos empíricos, os quais buscam relações estatísticas entre as características da cultura e outras variáveis que são determinadas por meio de dados observacionais (Thornley and Johnson, 1990). Os modelos estimativos de produtividade das culturas podem ser criados com base apenas em imagens de satélite (Cechim-Júnior et al., 2020), ou ainda dados climáticos (Verma et al., 2021), além da integração de dados climáticos, agronômicos e satélite (Luciano et al., 2021) e, podem usar técnicas estatísticas convencionais ou de aprendizado de máquina (Hammer et al., 2020).

Modelos empíricos de regressão são os mais populares para estimar a produtividade da cana-de-açúcar (Rahman and J. Robson, 2016), porém esses modelos detectam apenas correlações razoavelmente fortes entre as imagens de satélite e a produção de cana-de-açúcar, sendo necessário modelos mais confiáveis para uma estimativa mais assertiva. Neste sentido, pode-se utilizar algoritmos de aprendizado de máquina, como por exemplo, redes neurais artificiais, máquina de vetor de suporte (SVM - *Support Vector Machine*) e floresta aleatória (RF - *Random Forest*). Os algoritmos de aprendizagem possuem vantagens em relação aos modelos empíricos tradicionais, pois utilizam grande quantidade e variedade de informações, como dados numéricos e categóricos, advindos da combinação de dados de sensoriamento remoto, dados agronômicos e climáticos (Everingham et al., 2016). O uso integrado de imagens de satélite com algoritmos de aprendizado de máquina tem mostrados resultados promissores para estimar a produtividade da cultura do trigo (Kamir et al., 2020), soja (Schwalbert et al., 2020) e cana-de-açúcar (Luciano et al., 2021; Shendryk et al., 2021).

No contexto da cana-de-açúcar, a estimativa de produtividade por meio de imagens de satélite e algoritmos de aprendizagem de máquina, tem colaborado para uma estimativa de forma precisa e, localmente, ao longo dos anos e em diferentes condições ambientais. No entanto, em virtude da variabilidade espacial nos ambientes de produção agrícola, a estimativa da produtividade de cana-de-açúcar do rendimento da cultura não é trivial. Apesar disso, a utilização de imagens de satélite, dados climáticos e agronômicos em conjunto com algoritmos de aprendizado de máquina, pode contribuir com o desenvolvimento de modelos de estimativa de culturas de forma regional e temporal, buscando melhorias da precisão e robustez desses sistemas. Ainda assim, nota-se a necessidade de utilização de dados com maior detalhamento espacial e espectral, bem como o entendimento da integração das diversas variáveis influentes no desenvolvimento da cultura, ou seja, clima e manejo e, a integração destes dados para obtenção de modelos capazes de estimar a produtividade de forma regional, não apenas localmente. Neste contexto, a hipótese deste trabalho é que o uso de algoritmos de aprendizado de máquina aplicado à série temporal de dados do satélite Sentinel-2A, variáveis climáticas e agronômicas, permitem o monitoramento e estimativa da produtividade da cana-de-açúcar localmente e, de forma regional, considerando a utilização de séries temporais capazes de representar a variabilidade espaço-temporal do desenvolvimento da cultura.

1.1. Objetivo

O objetivo deste trabalho foi criar modelos de estimativa da produtividade de cana-de-açúcar antes da colheita, no estado de São Paulo, a partir de imagens do satélite Sentinel-2A, dados agronômicos e dados climáticos, utilizando o algoritmo de aprendizado de máquina *Random Forest*.

Como objetivos específicos espera-se avaliar a importância das variáveis agronômicas, climáticas e, principalmente, das imagens Sentinel-2A na estimativa de produtividade de cana-de-açúcar antes da colheita e avaliar o potencial de criação de um modelo regional (geral) de estimativa da produtividade da cana-de-açúcar antes da colheita, para as regiões do estado de São Paulo.

1.2. Estrutura da dissertação

A organização geral da dissertação consiste em um capítulo de revisão bibliográfica (Capítulo 2), dois capítulos subsequentes em formato de artigos científicos (Capítulo 3 e 4) e, por fim, um capítulo de considerações finais (Capítulo 5). Para o desenvolvimento de modelos estimativos da produtividade de cana-de-açúcar antes da colheita foram utilizadas três áreas de estudo localizadas em regiões edafoclimáticas distintas no estado de São Paulo, em quatro safras, a fim de verificar a capacidade de aplicação do modelo localmente, ao longo das safras e, de forma regional por meio de um modelo único. Para isso, foram utilizadas séries temporais de imagens do satélite Sentinel-2A (bandas e índices de vegetação), dados climáticos (radiação, precipitação, temperatura mensal e *déficit* hídrico) e dados agronômicos (variedade de cana-de-açúcar, estágio de corte, tipo de solo e relevo). O capítulo 3 teve como objetivo selecionar as principais variáveis provenientes do conjunto de dados das imagens de satélite, climáticos e agronômicos, em uma única usina, que melhor explicam a variabilidade de produtividade da cultura. A partir da seleção das variáveis foram desenvolvidos modelos empíricos para a estimativa de produtividade de cana-de-açúcar antes da colheita, de forma espacial e regional (modelo geral com as três usinas), utilizando o algoritmo *Random Forest* (Capítulo 4). O fluxo metodológico encontra-se a seguir (Figura 1).

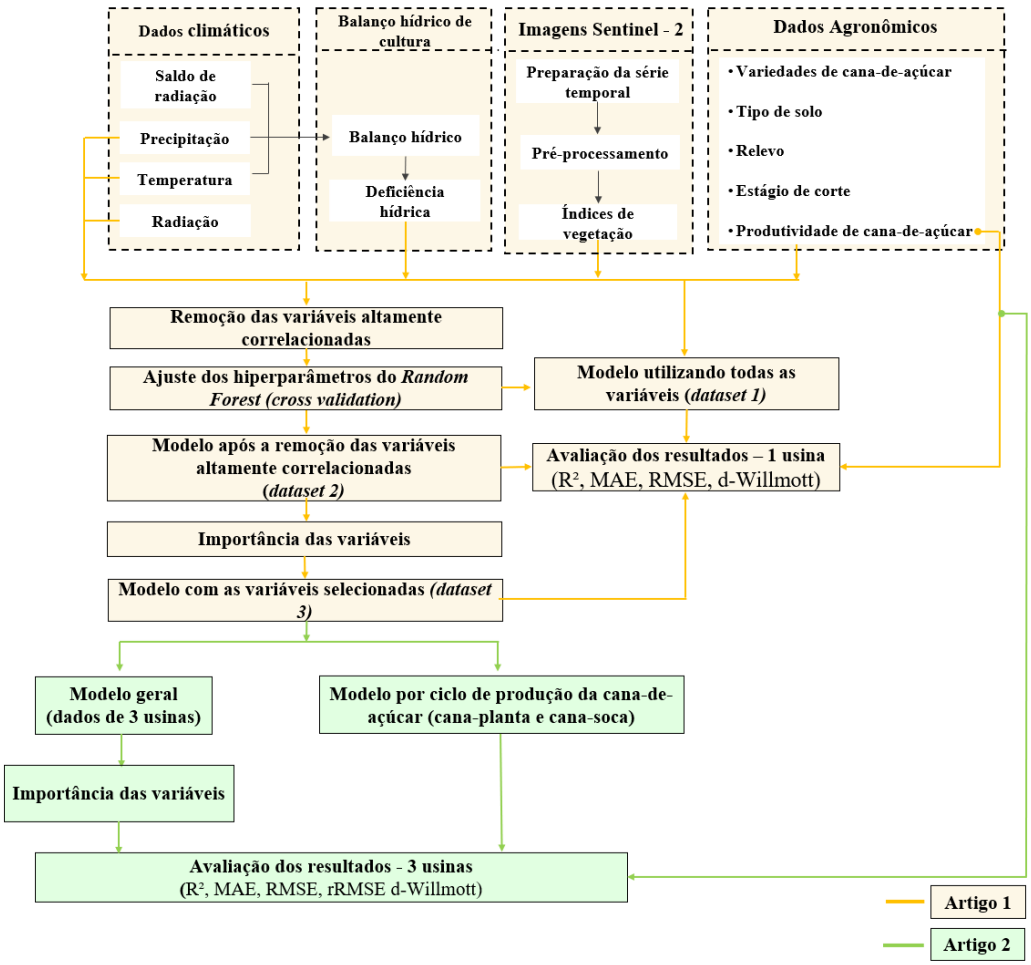

Figura 1. Fluxo metodológico da dissertação, contemplando o Artigo 1 (capítulo 3) e Artigo 2 (capítulo 4).

Referências

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

Barbanti, L., Adroher, J., Damian, J.M., Di Virgilio, N., Falsone, G., Zucchelli, M., Martelli, R., 2018. Assessing wheat spatial variation based on proximal and remote spectral vegetation indices and soil properties. Ital. J. Agron. 13, 21–30. https://doi.org/10.4081/ija.2017.1086

Basso, B., Cammarano, D., Carfagna, E., 2013. Review of Crop Yield Forecasting Methods and Early Warning Systems, in: The First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics. FAO, Rome, pp. 15–31.

Borrion, A.L., McManus, M.C., Hammond, G.P., 2012. Environmental life cycle assessment of lignocellulosic conversion to ethanol: A review. Renew. Sustain. Energy Rev. 16, 4638–4650. https://doi.org/10.1016/j.rser.2012.04.016

Cechim-Júnior, C., Johann, J.A., Antunes, J.F.G., Deppe, F.D., 2020. Sugarcane mapping in Paraná State Brazil using MODIS EVI images. Int. J. Adv. Remote Sens. GIS 9, 3205–3221. https://doi.org/10.23953/cloud.ijarsg.451

CONAB, 2023. Cana-de-açúcar. Análise Mensal. Cia. Nac. Abastecimento. Ministério da Agric. Pecuária e Abast. 5.

CONAB, 2021. Acompanhamento da Safra Brasileira de Cana-de-Açúcar – Quarto Levantamento da safra 2020/21. Cia. Nac. Abastecimento. . Ministério da Agric. Pecuária e Abast. 57.

Dimov, D., Uhl, J.H., Löw, F., Seboka, G.N., 2022. Sugarcane yield estimation through remote sensing time series and phenology metrics. Smart Agric. Technol. 2, 100046. https://doi.org/10.1016/j.atech.2022.100046

Dong, T., Liu, J., Shang, J., Qian, B., Ma, B., Kovacs, J.M., Walters, D., Jiao, X., Geng, X., Shi, Y., 2019. Assessment of red-edge vegetation indices for crop leaf area index estimation. Remote Sens. Environ. 222, 133–143. https://doi.org/10.1016/j.rse.2018.12.032

Dong, T., Meng, J., Shang, J., Liu, J., Wu, B., 2015. Evaluation of Chlorophyll-Related Vegetation Indices Using Simulated Sentinel-2 Data for Estimation of Crop Fraction of Absorbed Photosynthetically Active Radiation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 4049–4059. https://doi.org/10.1109/JSTARS.2015.2400134

Dubey, S.K., Gavli, A.S., Yadav, S.K., Sehgal, S., Ray, S.S., 2018. Remote Sensing-Based Yield Forecasting for Sugarcane (Saccharum officinarum L.) Crop in India. J. Indian Soc. Remote Sens. 46, 1823–1833. https://doi.org/10.1007/s12524-018-0839-2

EPE, 2022. Balanço Energético Nacional (BEN) 2022: Ano base 2021 - Relatório Final 264.

Everingham, Y., Sexton, J., Skocaj, D., Inman-Bamber, G., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agron. Sustain. Dev. 36, 27. https://doi.org/10.1007/s13593-016-0364-z

FAO, 2017. The Future of Food and Agriculture-Trends and Challenges.

Hammer, R.G., Sentelhas, P.C., Mariano, J.C.Q., 2020. Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. Sugar Tech 22, 216–225. https://doi.org/10.1007/s12355-019-00776-z

Holzman, M.E., Rivas, R., Piccolo, M.C., 2014. Estimating soil moisture and the relationship with crop yield using surface temperature and vegetation index. Int. J. Appl. Earth Obs. Geoinf. 28, 181–192. https://doi.org/10.1016/j.jag.2013.12.006

Huete, A.R., Liu, H.Q., van Leeuwen, W.J.D., 1997. Use of vegetation indices in forested regions: Issues of linearity and saturation. Int. Geosci. Remote Sens. Symp. 4, 1966–1968. https://doi.org/10.1109/igarss.1997.609169

IBGE, 2023. Produção Agrícola Municipal – PAM. IBGE- Ist. Bras. Geogr. e Estatística.

IBGE, 2022. Levantamento Sistemático da Produção Agrícola - LSPA. IBGE- Ist. Bras. Geogr. e Estatística.

IBGE, 2018. Levantamento Sistemático da Produção Agrícola - LSPA. IBGE- Ist. Bras. Geogr. e Estatística.

IPCC, 2018. An IPCC Special Report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, in: Masson-Delmotte, V., Zhai, P., Pörtner, H.O., Roberts, D., Skea, J., Shukla, P., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J., Chen, Y., Zhou, X., Gomis, M., Lonnoy, E., Maycock, T., Tignor, M., T. Waterfield, T. (Eds.), World Meteorological Organization. Geneva, Switzerland, p. 32.

Kamir, E., Waldner, F., Hochman, Z., 2020. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J. Photogramm. Remote Sens. 160, 124–135. https://doi.org/10.1016/j.isprsjprs.2019.11.008

Kross, A., McNairn, H., Lapen, D., Sunohara, M., Champagne, C., 2015. Assessment of RapidEye vegetation indices for estimation of leaf area index and biomass in corn and soybean crops. Int. J. Appl. Earth Obs. Geoinf. 34, 235–248. https://doi.org/10.1016/j.jag.2014.08.002

Lai, Y.R., Pringle, M.J., Kopittke, P.M., Menzies, N.W., Orton, T.G., Dang, Y.P., 2018. An empirical model for prediction of wheat yield, using time-integrated Landsat NDVI. Int. J. Appl. Earth Obs. Geoinf. 72, 99–108. https://doi.org/10.1016/j.jag.2018.07.013

Liao, C., Wang, J., Dong, T., Shang, J., Liu, J., Song, Y., 2019. Using spatio-temporal fusion of Landsat-8 and MODIS data to derive phenology, biomass and yield estimates for corn and soybean. Sci. Total Environ. 650, 1707–1721. https://doi.org/10.1016/j.scitotenv.2018.09.308

Luciano, A.C. dos S., Picoli, M.C.A., Duft, D.G., Rocha, J.V., Leal, M.R.L.V., le Maire, G., 2021. Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. Comput. Electron. Agric. 184. https://doi.org/10.1016/j.compag.2021.106063

Lukas, V., Novák, J., Neudert, L., Svobodova, I., Rodriguez-Moreno, F., Edrees, M., Kren, J., 2016. The combination of UAV survey and Landsat imagery for monitoring of crop vigor in precision agriculture. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 41, 953–957. https://doi.org/10.5194/isprsarchives-XLI-B8-953-2016

Mirasi, A., Mahmoudi, A., Navid, H., Valizadeh Kamran, K., Asoodar, M.A., 2021. Evaluation of sum-NDVI values to estimate wheat grain yields using multi-temporal Landsat OLI data. Geocarto Int. 36, 1309–1324. https://doi.org/10.1080/10106049.2019.1641561

Mulianga, B., Bégué, A., Simoes, M., Todoroff, P., 2013. Forecasting regional sugarcane yield based on time integral and spatial aggregation of MODIS NDVI. Remote Sens. 5, 2184–2199. https://doi.org/10.3390/rs5052184

Nguy-Robertson, A.L., Peng, Y., Gitelson, A.A., Arkebauer, T.J., Pimstein, A., Herrmann, I., Karnieli, A., Rundquist, D.C., Bonfil, D.J., 2014. Estimating green LAI in four crops: Potential of determining optimal spectral bands for a universal algorithm. Agric. For. Meteorol. 192–193, 140–148. https://doi.org/10.1016/j.agrformet.2014.03.004

Peroni Venancio, L., Chartuni Mantovani, E., do Amaral, C.H., Usher Neale, C.M., Zution Gonçalves, I., Filgueiras, R., Coelho Eugenio, F., 2020. Potential of using spectral vegetation indices for corn green biomass estimation based on their relationship with the photosynthetic vegetation sub-pixel fraction. Agric. Water Manag. 236. https://doi.org/10.1016/j.agwat.2020.106155

Rahman, M.M., J. Robson, A., 2016. A Novel Approach for Sugarcane Yield Prediction Using Landsat Time Series Imagery: A Case Study on Bundaberg Region. Adv. Remote Sens. 05, 93–102. https://doi.org/10.4236/ars.2016.52008

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERTS, in: Third ERTS-1 Symposium. NASA, Washington, DC, pp. 309–317.

Schwalbert, R.A., Amado, T., Corassa, G., Pott, L.P., Prasad, P.V.V., Ciampitti, I.A., 2020. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agric. For. Meteorol. 284, 107886. https://doi.org/10.1016/j.agrformet.2019.107886

Schwalbert, R.A., Amado, T.J.C., Nieto, L., Varela, S., Corassa, G.M., Horbe, T.A.N., Rice, C.W., Peralta, N.R., Ciampitti, I.A., 2018. Forecasting maize yield at field scale based on high-resolution satellite imagery. Biosyst. Eng. 171, 179–192. https://doi.org/10.1016/j.biosystemseng.2018.04.020

Shendryk, Y., Davy, R., Thorburn, P., 2021. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. F. Crop. Res. 260, 107984. https://doi.org/10.1016/j.fcr.2020.107984

Singla, S.K., Garg, R.D., Dubey, O.P., 2020. Ensemble machine learning methods to estimate the sugarcane yield based on remote sensing information. Rev. d'Intelligence Artif. 34, 731–743. https://doi.org/10.18280/RIA.340607

Surendran, U., Ramesh, V., Jayakumar, M., Marimuthu, S., Sridevi, G., 2016. Improved sugarcane productivity with tillage and trash management practices in semi arid tropical agro ecosystem in India. Soil Tillage Res. 158, 10–21. https://doi.org/10.1016/j.still.2015.10.009

Thornley, J.H.M., Johnson, I.R., 1990. Plant and crop modelling—A mathematical approach to plant and crop physiology, Clarendon Press. The Blackburn Press, Oxford.

UNICA, 2022. Outros produtos: Cana-de-açúcar, matéria-prima revolucionária.

Verma, A.K., Garg, P.K., Hari Prasad, K.S., Dadhwal, V.K., Dubey, S.K., Kumar, A., 2021. Sugarcane Yield Forecasting Model Based on Weather Parameters. Sugar Tech 23, 158–166. https://doi.org/10.1007/s12355-020-00900-4

Viña, A., Gitelson, A.A., Nguy-Robertson, A.L., Peng, Y., 2011. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. Remote Sens. Environ. 115, 3468–3478. https://doi.org/10.1016/j.rse.2011.08.010

Yogitha, B., Karthikeyan, M., Reddy, M., G., M., 2020. Progress of sugarcane bagasse ash applications in production of Eco-Friendly concrete – Review. Mater. Today Proc. v. 33, 695–699.

2. REVISÃO DE LITERATURA

2.1. Cultura da cana-de-açúcar

A cana-de-açúcar é uma gramínea tropical, cultivada nas regiões tropicais e subtropicais, que se desenvolve em forma de touceira com perfilho em sua base de 2 a 4 m de altura e diâmetro com cerca de 0,05 m (James, 2004). Os estádios de desenvolvimento da cana-de-açúcar são classificados em: brotação, perfilhamento, crescimento e maturação (Teare and Peet, 1983) (Figura 2).

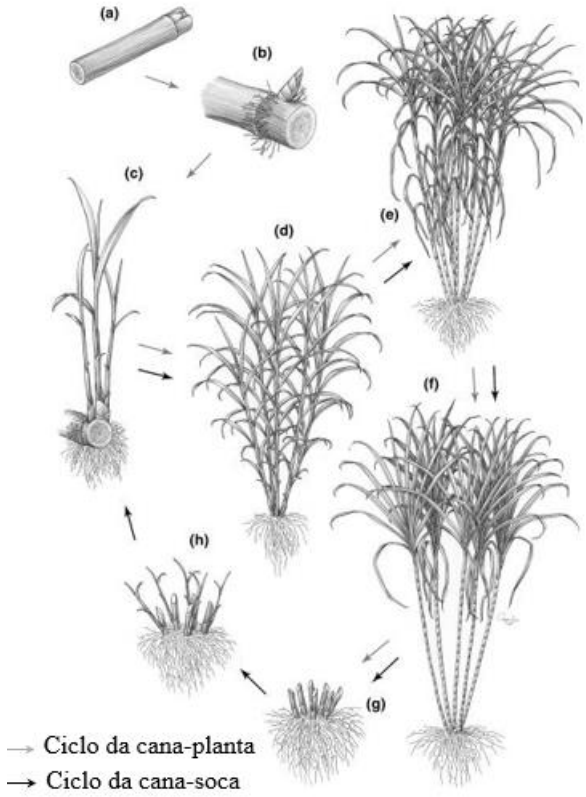

Figura 2. Representação esquemática das fases fenológicas da cana-de-açúcar. (a) Pedaços de caule utilizados no plantio; (b) Início da brotação e enraizamento das gemas; (c) Iniciação do perfilhamento; (d) Perfilhamento intenso; e Início da maturação; (f) Talos em concentração ótima de sacarose; (g) Colheita; (h) Rebrota. Adaptado de Cheavegatti-Gianotto et al. (2011).

O processo de brotação ocorre entre 20 e 30 dias após o plantio da cultura, porém pode sofrer a influência de fatores ambientais, genéticos e de tecnologia de plantio. O perfilhamento se inicia após 40 dias do plantio e é dependente da variedade, luminosidade, temperatura e umidade do solo (Casagrande and Casconcelos, 2008; Marafon, 2012; Teare and Peet, 1983). Já os estádios relacionados com o crescimento e maturação dos colmos, se iniciam em 120, 270 ou 360 dias após o plantio da cultura respectivamente (Marafon, 2012).

O desenvolvimento da cultura é determinado por dois ciclos de produção. O primeiro deles é chamado de cana-planta, já o segundo é chamado de cana-soca (ciclos de soqueira). O período de cana-planta ocorre quando a cultura ainda não teve o primeiro corte. Esse ciclo tem período de 12, 15 ou 18 meses, variando de acordo com a variedade da cultura e o momento na safra em que a cultura foi plantada. Segundo Barbosa (2012), os principais períodos de plantio e a denominação que a cultura recebe de acordo com o ciclo de desenvolvimento são chamados

de cana-de-ano, cana-de-ano-e-meio e cana-de-inverno. A cana-de-ano é plantada de setembro a início de dezembro, enquanto a cana-de-ano-e-meio é plantada de janeiro a abril, com produtividade esperada maior que a cana-de-ano, já que a cultura vegeta por um maior período, não tendo produção durante uma safra. Já a cana-de-inverno é plantada de maio a agosto, geralmente período seco na região Centro-Sul, período que a cultura mais necessita de irrigação devido à baixa disponibilidade de água no solo. Após o primeiro corte da cana-de-açúcar, encerra-se o ciclo da cana-planta e se inicia o ciclo da cana-soca. O período desse segundo ciclo é de 12 meses para todas as variedades da cultura (Casagrande and Vasconcelos, 2008). O número de safras e ciclos da cana-soca é de em média 6 anos e 5 ciclos. Ao longo dos ciclos ocorre a queda gradativa da produtividade, a qual depende principalmente do ambiente de produção, solo, manejo e tratos culturais, além da variedade plantada e condições climáticas.

As condições climáticas são as principais responsáveis pela variabilidade da produção de cana-de-açúcar, sendo um fator fundamental para o planejamento agrícola. Analisando a variabilidade espacial e temporal da eficiência produtiva da cana-de-açúcar em São Paulo, Marin *et al.* (2008) concluíram que os fatores climáticos explicaram 43% desta variabilidade enquanto os fatores do solo apenas 15%. Para as diferentes regiões climáticas brasileiras, sobretudo, na região Centro-Sul, que possui grandes áreas cultivadas com cana-de-açúcar, algumas condições climáticas limitantes devem ser analisadas, pois atuam diretamente na produtividade da cana-de-açúcar. Tais condições climáticas estão relacionadas à deficiência hídrica, temperatura média, precipitação, evapotranspiração e radiação (Henry and Kole, 2010; Machado et al., 2009; Marin et al., 2008).

A cana-de-açúcar, por ser uma planta C4, é considerada altamente eficiente na conversão de energia radiante em energia química, sendo que quanto maior for à saturação luminosa mais fotossíntese a planta realiza e, consequentemente, maior seu crescimento e acúmulo de açúcares (Rodrigues, 1995). Ainda, a temperatura do ar é extremamente importante, pois exerce grande influência no crescimento do caule, com uma faixa de temperatura ótima entre 25 °C e 35 °C e mínima em torno de 20 °C.

Outro fator de extrema relevância para que uma determinada cultura alcance sua produtividade potencial, é a necessidade hídrica. Para a cultura da cana-de-açúcar a necessidade hídrica pode variar entre 1500 e 2500 mm (Doorenbos and Kassam, 1979), distribuídas uniformemente ao longo do ciclo da cultura. As mudanças de necessidade hídrica podem afetar a produtividade da cultura, portanto, períodos em que a necessidade hídrica não é atendida são determinados como *déficit* hídrico. Os períodos de *déficit* hídrico podem ocorrer durante todo o ciclo da cultura, mas seu efeito sobre a produtividade de cana-de-açúcar varia muito em função da interação entre a época do ano em que ocorrem e, a fase do ciclo fenológico da cultura (Inman-Bamber et al., 2008; Machado et al., 2009). A determinação do *déficit* hídrico pode ser feita por meio do balanço hídrico.

Nos últimos 20 anos, a produção e área plantada de cana-de-açúcar no Brasil duplicaram (Figura 2). No ano de 2022 o país produziu cerca de 810 milhões de toneladas de cana-de-açúcar em uma área de 11,2 milhões de hectares (IBGE, 2022). Mais de 50% da produção de cana-de-açúcar do país está concentrada no estado de São Paulo, que produziu em 2022 cerca de 434 milhões de toneladas em uma área de 5,8 milhões de hectares (IBGE, 2022).

Na safra 2022/2023, a produção nacional de cana-de-açúcar apresentou um aumento de 3,4% em relação a 2020/2021 (CONAB, 2023), mesmo com as oscilações climáticas. A região Centro-Sul representou o maior eixo produtivo do país de acordo com a CONAB (2023), sendo responsável por 90% do total de cana-de-açúcar produzida, e devido ao incremento de produtividade obtido na safra 2022/2023 apresentou um aumento de 4% com relação à safra 2021/2022. Já as regiões Norte e Nordeste foram responsáveis pelo restante produzido, ou seja, apenas 10%. Os Estados de São Paulo, Goiás, Minas Gerais, Mato Grosso do Sul, Paraná e Mato Grosso, são os maiores produtores

nacionais de cana-de-açúcar, com destaque para o estado de São Paulo, com mais de 50% da produção nacional (Figura 3).

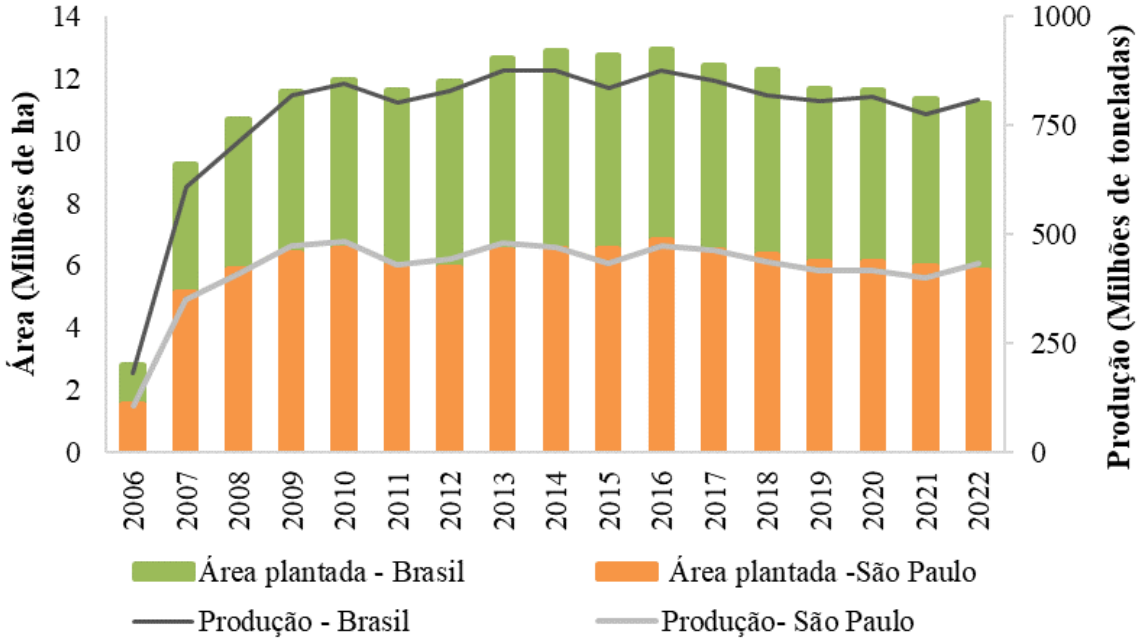

Figura 3. Histórico da área e produção de cana-de-açúcar no Brasil e no estado de São Paulo, com base em dados do IBGE-Levantamento Sistemático da produção agrícola (2023) (https://sidra.ibge.gov.br).

2.2. Técnicas de modelagem da produtividade de cana-de-açúcar

A produtividade agrícola é dependente das características intrínsecas da variedade da cultura em conjunto com as condições climáticas, manejo e práticas culturais (Gilbert et al., 2006). Uma das maneiras de se estimar a produtividade de culturas agrícolas é por meio de modelos matemáticos, os quais relacionam as variáveis responsáveis pelo crescimento e desenvolvimento da cultura com as condições ambientais diversas. Estes modelos podem ser classificados em determinísticos e empíricos.

Os modelos determinísticos têm sido utilizados desde o final da década de 1960 para simular processos fisiológicos de culturas, além de prever o crescimento e desenvolvimento das plantas (Todorovic et al., 2009). Os modelos determinísticos são capazes de descrever todo o ciclo da cultura e explicam a relação e a influência de cada variável responsável pelo crescimento da cultura e a produtividade (Murthy, 2004). Dentre os modelos determinísticos, também chamados de modelos de crescimento de cultura, encontra-se o modelo de crescimento APSIM-*Sugarcane* desenvolvido na Austrália, o qual foi projetado como um simulador de sistemas agrícolas, que combina estimativas de rendimento das culturas de modo acurado e inerente ao sistema de gestão (Keating et al., 2003). Ainda, há o modelo de crescimento DSSAT/*Canegro,* desenvolvido na África do Sul com o intuito de melhorar o planejamento da produção de açúcar do país (Inman-Bamber, 2000).

No entanto, os modelos determinísticos são mais apropriados para estimar o rendimento das culturas em escala local (Hochman et al., 2009), isso porque requerem conhecimento intensivo sobre as práticas de manejo e o meio ambiente, levando a uma parametrização complexa baseada em pontos (Bazgeer et al., 2007), o que pode gerar incertezas associadas à distribuição espacial desses parâmetros (Doraiswamy et al., 2004). Para aplicações em grande

escala, a quantidade limitada de informações, acaba levando a uma estimativa imprecisa da produtividade pela maioria dos modelos de crescimento de cultura (Palosuo et al., 2011).

Os modelos empíricos são mais simples que os modelos determinísticos, pois utilizam menos parâmetros de entrada e a estrutura é mais flexível (Gonzalez-Sanchez et al., 2014). Os modelos empíricos buscam relações estatísticas entre as características da cultura e outras variáveis que são determinadas por meio de dados observacionais (Thornley and Johnson, 1990). Os dados observacionais podem ser oriundos de dados climáticos, incluindo variáveis como temperatura do ar, radiação solar e insolação (Barbieri et al., 2010), além de dados provenientes de imagens de satélite e dados agronômicos (Yue et al., 2018).

Os dados climáticos podem ser associados a dados agronômicos, como o modelo desenvolvido por Doorenbos e Kassam (1979), que obtiveram a relação entre a produtividade das culturas e o uso da água. Esse modelo propõe uma equação simples, na qual a redução relativa da produtividade está relacionada com a redução relativa da evapotranspiração. Os modelos empíricos também podem incluir informações sobre o desenvolvimento da planta, as quais podem ser medidas indiretamente por sensoriamento remoto. Ainda, os dados de sensoriamento remoto podem ser associados a dados climáticos (Mello et al., 2014) e também, em conjunto com dados agronômicos (Luciano et al., 2021) para o monitoramento da produtividade de cana-de-açúcar.

A análise de regressão, que simplesmente correlaciona as variáveis de entrada com as variáveis de saída, é o método mais popular para prever a produção de culturas agrícolas usando sensoriamento remoto e modelagem empírica (Karthikeyan et al., 2020). Porém, por meio da modelagem empírica é possível obter apenas correlações razoavelmente fortes entre as imagens de satélite e a produção de cana-de-açúcar, em um único local e safra (Duveiller et al., 2013; Mulianga et al., 2013; Rahman and J. Robson, 2016). Assim, faz-se necessário realizar a estimativa regional e em diferentes safras da produtividade de cana-de-açúcar com modelos empíricos mais robustos, como os métodos de aprendizado de máquinas.

As técnicas de aprendizado de máquina têm sido amplamente utilizadas para obter previsões de produtividade precisas para diferentes safras nos últimos anos. Algumas delas incluem redes neurais artificiais, máquinas de vetores de suporte e *Random Forest* (Hammer et al., 2020; Tuia et al., 2011; Wang et al., 2016).

Singla et al. (2020) utilizaram perfil temporal de dados multiespectrais Landsat-8 e compararam diferentes técnicas de aprendizado de máquina, para a estimativa da produtividade da cana-de-açúcar, como *Random Forest*, Máquinas de Vetor de Suporte e k-Vizinho mais próximo (KNN). O método *Random Forest* foi o que exibiu um desempenho significativo (RMSE= 1,0 ton ha $^{-1}$ e R^2= 0,94) em comparação com os outros métodos. Já Bocca e Rodrigues (2016) consideraram em seu estudo de estimativa de produtividade da cana-de-açúcar, as variáveis de produção e manejo no nível de talhões em conjunto com os dados de clima, a fim de avaliar alguns modelos como Redes Neurais Artificiais, Máquinas de Vetor de Suporte, Árvores de Regressão Impulsionada e *Random Forest*. A variação do erro médio absoluto obtido (MAE) ficou entre 4,57 e 7,53 ton ha $^{-1}$, sendo que o menor erro encontrado foi para o modelo *Random Forest*.

Dentre outros trabalhos existentes com modelagem de produtividade agrícola, os preditores mais usados são informações climáticas, tipo de solo e imagens provenientes de sensoriamento remoto, enquanto os algoritmos de aprendizado de máquina mais aplicados foram *Random Forest*, Redes Neurais Artificiais e Árvores de Aumento de Gradiente (Chlingaryan et al., 2018; van Klompenburg et al., 2020).

2.3. Sensoriamento remoto aplicado ao monitoramento de produtividade agrícola

O monitoramento de áreas com grandes extensões demanda considerável quantidade de esforços, principalmente quando se pretende captar a variabilidade do terreno, sendo o sensoriamento remoto (SR) uma das principais ferramentas utilizadas nesse processo. A fundamentação das técnicas de SR é baseada na interação da energia eletromagnética com a matéria ou alvo (Arco et al., 2003). Esta interação é dependente da estrutura atômica e molecular de cada alvo, sendo que as radiações incidentes podem ser refletidas, absorvidas e transmitidas. A quantidade refletida, absorvida e transmitida também é dependente do comprimento de onda do espectro eletromagnético (Figueiredo, 2005).

Os diferentes tipos de interação da energia com a matéria geram curvas características, as quais são chamadas de assinatura espectral. Sua obtenção se dá por meio dos sensores remotos embarcados em satélites, que são dispositivos capazes de detectar a energia eletromagnética proveniente de um objeto, transformá-las em um sinal elétrico e registrá-las, de tal forma que este possa ser armazenado ou transmitido em tempo real para posteriormente ser convertido em informações que descrevem as feições dos objetos que compõem a superfície terrestre (Moraes, 2002).

Nas últimas décadas, estudos têm utilizado informações provenientes de SR, como imagens de satélite e índices de vegetação para o monitoramento de culturas agrícolas. Os índices de vegetação são transformações matemáticas da reflectância com o propósito de se explorar as propriedades espectrais da vegetação. A quantificação de um índice de vegetação é influenciada principalmente pelo tipo de cultura, saturação da vegetação, solo e efeitos atmosféricos (Fang et al., 2019).

Os índices de vegetação podem ser usados como observações instantâneas ou em combinação ao longo do tempo, chamadas de séries temporais. Para monitoramento de culturas, estudos indicam que agregar o índice de vegetação ao longo de um período de tempo, ao invés de observações instantâneas, reduz o ruído devido a outros fatores, como solos e nuvens, principalmente quando se estuda a produtividade das culturas (Karthikeyan et al., 2020). Isto ocorre, pois a relação entre o rendimento da cultura e a refletância espectral varia com o crescimento da cultura (Rudorff and Batista, 1990).

Os índices de vegetação são obtidos por meio de fórmulas matemáticas, em geral, com base nas bandas das regiões do vermelho (R) e do infravermelho próximo (NIR) do espectro eletromagnético (Wiegand et al., 1991). Os índices de vegetação, que utilizam uma combinação das bandas R e NIR, estão principalmente relacionados à abundância de cobertura vegetal verde e biomassa (Silleos et al., 2006) e são muito utilizados para o monitoramento da vegetação. Por meio dos índices de vegetação é possível detectar atividades sazonais e fenológicas de culturas, duração do período de crescimento, mudanças fisiológicas, períodos de senescência (Ponzoni et al., 2012), avaliação geral do estado da cultura (Barbanti et al., 2018; Lukas et al., 2016) e, estimativa de produtividade de culturas, como a cana-de-açúcar (Mulianga et al., 2013; Peroni Venancio et al., 2020; Schwalbert et al., 2018). Dentre os índices que combinam as bandas R e NIR estão o Índice de Vegetação por Diferença Normalizada (NDVI) (Rouse et al., 1973), Índice de Vegetação Aprimorado (EVI) (Huete et al., 1997) e Índice de Vegetação Ajustado ao Solo (SAVI) (Huete, 1988).

Singla et al., (2020) estimaram a produtividade da cultura da cana-de-açúcar, por meio de séries temporais do satélite Landsat-8, utilizando índices de vegetação que combinam as bandas NIR e R, dentre eles SAVI, EVI e NDVI. Os autores utilizaram o modelo de aprendizado de máquina *Random Forest* e obtiveram RMSE igual a 1,51 ton ha^{-1} e R^2 igual a 0,94, sendo que os melhores desempenhos foram obtidos com os índices NDVI e Índice de Vegetação

Normalizada do Verde (GNDVI). Ainda, os autores concluíram que a utilização de mais de um índice de vegetação ao longo do tempo, possibilitou melhor estimativa da produtividade de cana-de-açúcar quando comparado ao uso de apenas um índice calculado em data específica.

Outros índices de vegetação, que combinam as bandas do NIR com a banda do infravermelho de ondas curtas (SWIR) também têm se mostrado importantes para o monitoramento de culturas agrícolas. A refletância NIR é útil para representar o dossel estrutural como índice de área foliar e biomassa (Dorigo et al., 2007). Já o aumento da refletância das regiões SWIR ocorre principalmente devido à diminuição do teor de água na planta e no solo, que pode estar relacionado ao desenvolvimento e senescência da cultura, isso porque a banda SWIR é fortemente influenciada pela água e estruturas de dossel (Gao, 1996). Dentre os índices que consideram em seus cálculos o uso das bandas do NIR e SWIR, os mais utilizados são o Índice de Umidade de Diferença Normalizado (NDMI) (Wilson and Sader, 2002), Índice de Diferença Normalizada da Água 1 ($NDWI_1$) (McFeeters, 1996) e Índice de Diferença Normalizada da Água 2 ($NDWI_2$) (Rogers and Kearney, 2004).

Luciano *et al.* (2021) utilizaram índices de vegetação que correlacionam a banda do NIR e SWIR, do Landsat-8, para estimativa da produtividade da cana-de-açúcar e constataram que o NDMI é um índice de extrema importância para o monitoramento da produtividade da cultura. Dong *et al.* (2020) avaliaram o potencial de assimilação do índice de área foliar dos dados Sentinel-2 e Landsat-8, em um modelo simples de crescimento de safra, para estimar a biomassa de culturas como soja, milho, canola, feijão e aveia e constataram que o $NDWI_1$ teve uma forte relação com o índice de área foliar para todas as culturas.

Nos últimos anos, alguns índices de vegetação foram desenvolvidos usando reflectância de borda vermelha (*Red-edge*), como o NDVI calculado com o *Red-edge* ($NDVI_{RE}$) (Gitelson and Merzlyak, 1994), o Índice de Clorofila *Red-edge* (CI_{RE}) (Gitelson et al., 2003) e o *Modified simple ratio Red-edge* (MSR_{RE}) (Wu et al., 2008). Esses índices vêm sendo aplicados na estimativa do Índice de Área Foliar e, portanto, impactam na produtividade, demonstrando melhor resultado em comparação com os índices de vegetação calculados com a banda do vermelho (Dong et al., 2015; Nguy-Robertson et al., 2014; Shang et al., 2014; Viña et al., 2011).

Yu et al. (2020) avaliaram a estimativa de Índice de Área Foliar para a cultura do milho por meio de dados dos satélites Landsat-8 e Sentinel-2 e dados medidos em campo. Foram utilizados os índices de vegetação NDVI, EVI, CI_{Green} (Índice de Clorofila Verde) e o índice CI_{RE}. O IAF estimado teve boa consistência com as medições em campo, sendo que o LAI baseado no *Red-edge* (CI_{RE}), resultou em menores erros (RMSE=0,64) do que LAI derivado de CI_{Green} (RMSE=0,72). Os autores reforçam que a combinação de várias bandas espectrais pode melhorar o desempenho da estimativa do LAI, diminuindo efeitos que afetam a reflectância espectral, como absorção de clorofila pela folha e impactos da reflectância do solo.

De forma similar, Shendryk et al. (2021) também indicaram em seus estudos que os índices espectrais derivados das bandas *Red-edge* e NIR, do satélite Sentinel-2, foram mais relevantes para a estimativa da produtividade da cana-de-açúcar do que os frequentemente usados, como o NDVI. Como conclusões, os autores mostraram a importância de utilização de bandas espectrais e índices de vegetação calculados com o *Red-edge* e, ressaltaram que estes índices e bandas apresentam potencial para serem utilizados em estudos futuros sobre a estimativa da produtividade de culturas agrícolas, como a cana-de-açúcar.

Outras fontes de dados de sensoriamento remoto que podem ser utilizadas para o monitoramento de produtividade agrícola são as que fornecem variáveis meteorológicas. Alguns modelos matemáticos, assumem em seus cálculos dados de satélites, associados a dados de radares meteorológicos e estações meteorológicas, aplicando técnicas de interpolação espacial e método de assimilação de dados, que geram produtos climáticos em grade.

Existem vários conjuntos de dados climáticos em grade disponíveis em todo o mundo, como a reanálise do Centro Nacional de Previsão Ambiental/Centro Nacional de Pesquisa Atmosférica (NCEP/NCAR), a Medição de Precipitação Global (GPM), além da Análise Retrospectiva da Era Moderna para Pesquisa e Aplicações (MERRA) pela Administração Nacional de Aeronáutica e Espaço e Previsão de Recursos Energéticos Mundiais da NASA (NASAPOWER), e o Centro Europeu de Previsões meteorológicas de médio alcance (ECMWF). Esses dados espacializados são úteis principalmente em locais onde há escassez de equipamentos que fazem as medições dessas variáveis climáticas. Tal fato ocorre, pois os dados climáticos especializados tendem a ser temporalmente e espacialmente semelhantes aos dados medidos por estações meteorológicas, como as estações do INMET (Instituto Nacional de Meteorologia) (Aparecido et al., 2020).

Os dados climáticos provenientes do ECMWF, GPM e NASAPOWER podem ser combinados com imagens de satélite, com base no cálculo de índices de vegetação e bandas espectrais, para estimar a produtividade de culturas agrícolas. Salvador et al. (2020) estimaram a produtividade da batata sobre o México em nível municipal, utilizando dados meteorológicos fornecidos pelo conjunto de dados ERA5 (ECMWF *Re-Analysis*), imagens de satélite da plataforma MODIS/TERRA e informações de campo, integrados por meio de algoritmos de aprendizado de máquinas, como o algoritmo *Random Forest* e obtiveram resultados de R^2= 0,757 e RMSE = 18,9 ton ha $^{-1}$

Para a cultura da cana-de-açúcar, dados climáticos oriundos do NASAPOWER, foram utilizados para a estimativa da produtividade potencial da cultura, por meio de um modelo genérico da FAO (Monteiro et al., 2018). Os autores obtiveram erros baixos (RMSE<30 ton ha $^{-1}$) na maior parte do território brasileiro. Em abordagem similar, Zhao e Justina (2020) combinaram dados de precipitação do GPM e ECMWF e dados espectrais derivados do sensor Landsat-8/OLI para estimativa da produtividade da cana-de-açúcar. Os autores constataram uma melhora no desempenho dos modelos avaliados, quando adicionadas as variáveis climáticas (R^2=0,48), isso porque as variáveis climáticas são grandes responsáveis pelas alterações da produtividade.

2.4. Algoritmo de aprendizado de máquina *Random Forest*

Modelos estatísticos são frequentemente usados para estimar o rendimento das culturas em um ambiente único, a fim de fornecer informações úteis aos formuladores de políticas sobre opções de manejo e produção (Di Paola et al., 2016). Dentre os algoritmos de aprendizagem de máquina mais utilizados na estimativa de produtividade de culturas encontram-se: *Neural Networks, Linear Regression, Random Forest* (RF) e *Support Vector Machines* (SVM) (van Klompenburg et al., 2020). Muitos dos algoritmos de aprendizagem de máquina tem apresentado resultados promissores para a estimativa da produtividade de cana-de-açúcar (Sunil Kumar et al., 2015; Medar et al., 2019; Shendryk et al., 2021). Singla et al. (2020) compararam diferentes algoritmos para estimar a produtividade de cana-de-açúcar com base em dados de sensoriamento remoto. Os resultados mostraram que o RF teve melhor performance do que os outros métodos como árvores de decisão e SVM Charoen-Ung (2018) também comparou algoritmos de aprendizado de máquinas para a estimativa da produtividade de cana-de-açúcar, utilizando dados agronômicos, e precipitação fornecidos por uma usina na Tailândia e os resultados indicaram que o RF apresentou melhor acurácia do que *Gradient Boosting*, com R^2 igual a 71,8%.

O RF tem sido usado para estimativa da produtividade em grandes áreas, devido à sua capacidade de lidar com alta dimensionalidade de dados, detecção de *outliers*, robustez contra *overfitting* e a possibilidade de estudar a importância da variável de entrada em um modelo calibrado (Gislason et al., 2006). O RF é uma técnica estatística e de análise de dados não linear e não paramétrica baseada na abordagem de aprendizado de máquina por *ensemble*, que

realiza a regressão crescendo uma infinidade de árvores de decisão no momento do treinamento do modelo e, em seguida, produz uma saída calculando a média das previsões de todas as árvores individuais (Breiman, 2001; Yue et al., 2018).

A potência do algoritmo RF é evidente quando se constrói modelos preditivos com grande quantidade de dados, uma vez que o algoritmo tem a habilidade de determinar automaticamente quais variáveis preditoras são importantes. No entanto, faz-se necessário ajustar o funcionamento dos hiperparâmetros do algoritmo. O desempenho de muitos algoritmos de aprendizado de máquinas depende das configurações de hiperparâmetros, sendo seu ajuste fundamental para seu melhor desempenho. Os hiperparâmetros não são otimizados como parte do processo de treinamento e, diferentemente dos modelos estatísticos tradicionais, os hiperparâmetros para RF não exigem escolhas (por exemplo, a escolha dos preditores a serem usados em um modelo de regressão) e são mais críticos, especialmente para evitar sobre ajuste (Bruce et al., 2020). O hiperparâmetro mais importante para RF é o *Nodesize* ou o tamanho mínimo para os nós terminais (folhas na árvore). O desempenho de muitos métodos de aprendizado de máquina depende criticamente das configurações de hiperparâmetros.

A técnica RF vem sendo amplamente utilizada para mapeamento e monitoramento do uso da terra (Belgiu and Drăgu, 2016) e para a estimativa da produtividade em grandes áreas agrícolas, como milho e soja (Sakamoto, 2020), algodão (Filippi et al., 2020) e cana-de-açúcar (Shendryk et al., 2020; Singla et al., 2020). Isto devido à capacidade do RF em modelar interações complexas entre variáveis de entrada, advindas de dados observacionais da terra, robustez contra *overfitting e outliers* (Verrelst et al., 2015). Ainda, o *RF* é adequado para utilização com dados de dimensionalidade elevada, com grande proporção de preditores colineares, já que as árvores de decisão são por natureza imunes à multicolinearidade (Piramuthu, 2008).

De modo geral, a utilização de séries temporais de imagens de satélite em conjunto com informações de manejo e clima são essenciais no monitoramento da produtividade de cana-de-açúcar e até mesmo de outras culturas agrícolas de forma local e em períodos pré-determinados. No entanto, ainda é um desafio a utilização de séries temporais de imagens de satélite e dados climáticos para o monitoramento da produtividade de cana-de-açúcar em escala regional, ou seja, grandes áreas onde há a necessidade de uma quantidade expressiva de dados para calibração de modelos empíricos e até mesmo determinísticos.

Referências

Aparecido, L.E. de O., Rolim, G. de S., Moraes, J.R. da S.C. de, 2020. Validation of ECMWF climatic data, 1979–2017, and implications for modelling water balance for tropical climates. Int. J. Climatol. 40, 6646–6665. https://doi.org/https://doi.org/10.1002/joc.6604

Arco, E.D., Alvarenga, B.S., Moura, P., Teixeira, C.G., 2003. Estudos de refletância de amostras de 5 tipos de solos brasileiros, em condições de laboratório, in: INPE (Ed.), Simpósio Brasileiro de Sensoriamento Remoto. Belo Horizonte, pp. 2327–2334.

Barbanti, L., Adroher, J., Damian, J.M., Di Virgilio, N., Falsone, G., Zucchelli, M., Martelli, R., 2018. Assessing wheat spatial variation based on proximal and remote spectral vegetation indices and soil properties. Ital. J. Agron. 13, 21–30. https://doi.org/10.4081/ija.2017.1086

Barbieri, V., Da SIlva, F.C., Dias-Ambrona, C.G.H., 2010. Modelagem de cana-de-açúcar para previsão de produtividade de canaviais no Brasil e na Austrália. 39Jaiio - Cai 2010 745–762.

Barbosa, V.F.A.M., 2012. Plantio, in: Santos, F., Borém, A., Caldas, C. (Eds.), Cana-de-Açúcar: Bionergia, Açúcar e Etanol: Tecnologias e Perspectivas. Viçosa, MG, pp. 637, 2012.

Bazgeer, S., Kamali, G., Mortazavi, A., 2007. Wheat yield prediction through agrometeorological indices for Hamedan, Iran. Biaban 12, 33–38.

Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

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

Breiman, L., 2001. Random Forests. Mach. Learn. 5–32.

Bruce, P., Bruce, A., Gedeck, P., 2020. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. Technometrics 63, 363. https://doi.org/10.1080/00401706.2021.1904738

Casagrande, A.A., Vasconcelos, C.M., 2008. Fisiologia da Parte Aérea, in: Miranda, L.L.D., Vasconcelos, A.C.M., Landell, M.G. DE (Eds.), Cana-de-Açúcar. Campinas: Instituto Agronômico, p. 882.

Charoen-ung, P., Mittrapiyanuruk, P., 2018. Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques.

Cheavegatti-Gianotto, A., de Abreu, H.M.C., Arruda, P., Bespalhok Filho, J.C., Burnquist, W.L., Creste, S., di Ciero, L., Ferro, J.A., de Oliveira Figueira, A.V., de Sousa Filgueiras, T., Grossi-de-Sá, M. de F., Guzzo, E.C., Hoffmann, H.P., de Andrade Landell, M.G., Macedo, N., Matsuoka, S., de Castro Reinach, F., Romano, E., da Silva, W.J., de Castro Silva Filho, M., César Ulian, E., 2011. Sugarcane (Saccharum X officinarum): A Reference Study for the Regulation of Genetically Modified Cultivars in Brazil. Trop. Plant Biol. 4, 62–89. https://doi.org/10.1007/s12042-011-9068-3

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

CONAB, 2023. Cana-de-açúcar. Análise Mensal. Cia. Nac. Abastecimento. Ministério da Agric. Pecuária e Abast. 5.

Di Paola, A., Valentini, R., Santini, M., 2016. An overview of available crop growth and yield models for studies and assessments in agriculture. J. Sci. Food Agric. 96, 709–714. https://doi.org/10.1002/jsfa.7359

Dong, T., Liu, Jiangui, Qian, B., He, L., Liu, Jane, Wang, R., Jing, Q., Champagne, C., McNairn, H., Powers, J., Shi, Y., Chen, J.M., Shang, J., 2020. Estimating crop biomass using leaf area index derived from Landsat 8 and Sentinel-2 data. ISPRS J. Photogramm. Remote Sens. 168, 236–250. https://doi.org/10.1016/J.ISPRSJPRS.2020.08.003

Dong, T., Meng, J., Shang, J., Liu, J., Wu, B., 2015. Evaluation of Chlorophyll-Related Vegetation Indices Using Simulated Sentinel-2 Data for Estimation of Crop Fraction of Absorbed Photosynthetically Active Radiation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 4049–4059. https://doi.org/10.1109/JSTARS.2015.2400134

Doorenbos, J., Kassam, A.H., 1979. Yield Response to Water, Irrigation and Agricultural Development. Food And Agriculture Organization Of The United Nations, Rome. https://doi.org/10.1016/b978-0-08-025675-7.50021-2

Doraiswamy, P.C., Hatfield, J.L., Jackson, T.J., Akhmedov, B., Prueger, J., Stern, A., 2004. Crop condition and yield simulations using Landsat and MODIS. Remote Sens. Environ. 92, 548–559. https://doi.org/10.1016/j.rse.2004.05.017

Dorigo, W.A., Zurita-Milla, R., de Wit, A.J.W., Brazile, J., Singh, R., Schaepman, M.E., 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. Int. J. Appl. Earth Obs. Geoinf. 9, 165–193. https://doi.org/10.1016/j.jag.2006.05.003

Duveiller, G., López-Lozano, R., Baruth, B., 2013. Enhanced processing of 1-km spatial resolution fAPAR time series for sugarcane yield forecasting and monitoring. Remote Sens. 5, 1091–1116. https://doi.org/10.3390/rs5031091

Fang, H., Baret, F., Plummer, S., Schaepman-Strub, G., 2019. An Overview of Global Leaf Area Index (LAI): Methods, Products, Validation, and Applications. Rev. Geophys. 57, 739–799. https://doi.org/10.1029/2018RG000608

Figueiredo, D., 2005. Conceitos básicos de sensoriamento remoto. Cia. Nac. Abast.

Filippi, P., Whelan, B.M., Vervoort, R.W., Bishop, T.F.A., 2020. Mid-season empirical cotton yield forecasts at fine resolutions using large yield mapping datasets and diverse spatial covariates. Agric. Syst. 184, 102894. https://doi.org/10.1016/j.agsy.2020.102894

Gao, B.-C., 1996. NDWI A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water From Space. Remote Sens. Environ. 58, 257–266.

Gilbert, R.A., Shine, J.M., Miller, J.D., Rice, R.W., Rainbolt, C.R., 2006. The effect of genotype, environment and time of harvest on sugarcane yields in Florida, USA. F. Crop. Res. 95, 156–170. https://doi.org/10.1016/j.fcr.2005.02.006

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognit. Lett. 27, 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Gitelson, A., Merzlyak, M.N., 1994. Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus hippocastanum L. and Acer platanoides L. Leaves. Spectral Features and Relation to Chlorophyll Estimation. J. Plant Physiol. 143, 286–292. https://doi.org/https://doi.org/10.1016/S0176-1617(11)81633-0

Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. J. Plant Physiol. 160, 271–282. https://doi.org/10.1078/0176-1617-00887

Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction. Spanish J. Agric. Res. 12, 313–328. https://doi.org/10.5424/sjar/2014122-4439

Hammer, R.G., Sentelhas, P.C., Mariano, J.C.Q., 2020. Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. Sugar Tech 22, 216–225. https://doi.org/10.1007/s12355-019-00776-z

Henry, R., Kole, C., 2010. Genetics, Genomics and Breeding of Sugarcane. Science Publishers, Boca Raton.

Hochman, Z., van Rees, H., Carberry, P.S., Hunt, J.R., McCown, R.L., Gartmann, A., Holzworth, D., van Rees, S., Dalgliesh, N.P., Long, W., Peake, A.S., Poulton, P.L., McClelland, T., 2009. Re-inventing model-based decision support with Australian dryland farmers. 4. Yield Prophet<sup/> helps farmers monitor and manage crops in a variable climate. Crop Pasture Sci. 60, 1057–1070.

Huete, A.R., 1988. A Soil-Adjusted Vegetation Index (SAVI). Remote Sens. Appl. Soc. Environ. 25, 295–309.

Huete, A.R., Liu, H.Q., van Leeuwen, W.J.D., 1997. Use of vegetation indices in forested regions: Issues of linearity and saturation. Int. Geosci. Remote Sens. Symp. 4, 1966–1968. https://doi.org/10.1109/igarss.1997.609169

IBGE, 2022. Levantamento Sistemático da Produção Agrícola - LSPA. IBGE- Ist. Bras. Geogr. e Estatística.

Inman-Bamber, N.G., 2000. History of the Canegro Model, in: IV International Workshop. Mount Edgecombe, South Africa, pp. 5–8.

Inman-Bamber, N.G., Bonnett, G.D., Spillman, M.F., Hewitt, M.L., Jackson, J., 2008. Increasing sucrose accumulation in sugarcane by manipulating leaf extension and photosynthesis with irrigation. Aust. J. Agric. Res. 59, 13–26.

James, G., 2004. Sugarcane, 2nd ed. Blackwell Science, Oxford.

Karthikeyan, L., Chawla, I., Mishra, A.K., 2020. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. J. Hydrol. 586, 124905. https://doi.org/10.1016/j.jhydrol.2020.124905

Keating, B.., Carberry, G.., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzwarth, D., Huth, N.I. Hargreaves, J.N.G., Meinke, H., Hochman, Z., Mclean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM: a model designed for farming systems simulation. Eur. J. Agron. 18, 267–288.

Kumar, S., Kumar, V., R.K, S., 2015. Sugarcane Yield Forecasting using Artificial Neural Network Models. Int. J. Artif. Intell. Appl. 6, 51–68. https://doi.org/10.5121/ijaia.2015.6504

Luciano, A.C. dos S., Picoli, M.C.A., Duft, D.G., Rocha, J.V., Leal, M.R.L.V., le Maire, G., 2021. Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. Comput. Electron. Agric. 184. https://doi.org/10.1016/j.compag.2021.106063

Lukas, V., Novák, J., Neudert, L., Svobodova, I., Rodriguez-Moreno, F., Edrees, M., Kren, J., 2016. The combination of UAV survey and Landsat imagery for monitoring of crop vigor in precision agriculture. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 41, 953–957. https://doi.org/10.5194/isprsarchives-XLI-B8-953-2016

Machado, R.S., Ribeiro, R.V., Marchiori, P.E.R., Machado, D.F.S.P., Machado, E.C., Landell, M.G. de A., 2009. Respostas biométricas e fisiológicas ao deficit hídrico em cana-de-açúcar em diferentes fases fenológicas. Pesqui. Agropecuária Bras. 44, 1575–1582. https://doi.org/10.1590/s0100-204x2009001200003

Marafon, A.C., 2012. Análise quantitativa de crescimento em Cana-de-açúcar: Uma introducao ao procedimento práctico. Embrapa Tabuleiros Costeiros 168, 31.

Marin, F.R., Lopes-Assad, M.L., Assad, E.D., Vian, C.E., Santos, M.C., 2008. Sugarcane crop efficiency in two growing seasons in São Paulo State, Brazil. Pesqui. Agropecuária Bras. 43, 1449–1455. https://doi.org/10.1590/s0100-204x2008001100002

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

Medar, R.A., Rajpurohit, V.S., Ambekar, A.M., 2019. Sugarcane Crop Yield Forecasting Model Using Supervised Machine Learning. Int. J. Intell. Syst. Appl. 11, 11–20. https://doi.org/10.5815/ijisa.2019.08.02

Mello, F.F.C., Cerri, C.E.P., Davies, C.A., Holbrook, N.M., Paustian, K., Maia, S.M.F., Galdos, M. V, Bernoux, M., Cerri, C.C., 2014. {P}ayback time for soil carbon and sugar-cane ethanol. {N}ature {C}limate {C}hange 4, 605–609. https://doi.org/10.1038/nclimate2239

Monteiro, L.A., Sentelhas, P.C., Pedra, G.U., 2018. Assessment of NASA/POWER satellite-based weather system for Brazilian conditions and its impact on sugarcane yield simulation. Int. J. Climatol. 38, 1571–1581. https://doi.org/https://doi.org/10.1002/joc.5282

Moraes, E.C. De, 2002. Fundamentos de sensoriamento remoto. Inst. Nac. Pesqui. Espac. Ministério da Ciência e Tecnol. Capítulo 1, 3–12.

Mulianga, B., Bégué, A., Simoes, M., Todoroff, P., 2013. Forecasting regional sugarcane yield based on time integral and spatial aggregation of MODIS NDVI. Remote Sens. 5, 2184–2199. https://doi.org/10.3390/rs5052184

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

Nguy-Robertson, A.L., Peng, Y., Gitelson, A.A., Arkebauer, T.J., Pimstein, A., Herrmann, I., Karnieli, A., Rundquist, D.C., Bonfil, D.J., 2014. Estimating green LAI in four crops: Potential of determining optimal spectral bands for a universal algorithm. Agric. For. Meteorol. 192–193, 140–148. https://doi.org/10.1016/j.agrformet.2014.03.004

Palosuo, T., Kersebaum, K.C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J.E., Patil, R.H., Ruget, F., Rumbaur, C., Takáč, J., Trnka, M., Bindi, M., Çaldağ, B., Ewert, F., Ferrise, R., Mirschel, W., Şaylan, L., Šiška, B., Rötter, R., 2011. Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. Eur. J. Agron. 35, 103–114. https://doi.org/10.1016/j.eja.2011.05.001

Peroni Venancio, L., Chartuni Mantovani, E., do Amaral, C.H., Usher Neale, C.M., Zution Gonçalves, I., Filgueiras, R., Coelho Eugenio, F., 2020. Potential of using spectral vegetation indices for corn green biomass estimation based on their relationship with the photosynthetic vegetation sub-pixel fraction. Agric. Water Manag. 236. https://doi.org/10.1016/j.agwat.2020.106155

Piramuthu, S., 2008. Input data for decision trees. Expert Syst. Appl. 34, 1220–1226. https://doi.org/10.1016/j.eswa.2006.12.030

Ponzoni, F.J., Shimabukuro, Y.E., Kuplich, T.M., 2012. Sensoriamento remoto da vegetação, 2nd ed. Oficina de Textos, São José dos Campos.

Rahman, M.M., J. Robson, A., 2016. A Novel Approach for Sugarcane Yield Prediction Using Landsat Time Series Imagery: A Case Study on Bundaberg Region. Adv. Remote Sens. 05, 93–102. https://doi.org/10.4236/ars.2016.52008

Rodrigues, J.D., 1995. Fisiologia da cana-de-açúcar. Botucatu: Unesp.

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

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERTS, in: Third ERTS-1 Symposium. NASA, Washington, DC, pp. 309–317.

Rudorff, B.F.T., Batista, G.T., 1990. Spectral response of wheat and its relationship to agronomic variables in the tropical region. Remote Sens. Environ. 31, 53–63. https://doi.org/10.1016/0034-4257(90)90076-X

Sakamoto, T., 2020. Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. ISPRS J. Photogramm. Remote Sens. 160, 208–228. https://doi.org/10.1016/j.isprsjprs.2019.12.012

Schwalbert, R.A., Amado, T.J.C., Nieto, L., Varela, S., Corassa, G.M., Horbe, T.A.N., Rice, C.W., Peralta, N.R., Ciampitti, I.A., 2018. Forecasting maize yield at field scale based on high-resolution satellite imagery. Biosyst. Eng. 171, 179–192. https://doi.org/10.1016/j.biosystemseng.2018.04.020

Shang, J., Liu, J., Huffman, T., Qian, B., Pattey, E., Wang, J., Zhao, T., Geng, X., Kroetsch, D., Dong, T., Lantz, N., 2014. Estimating plant area index for monitoring crop growth dynamics using Landsat-8 and RapidEye images. J. Appl. Remote Sens. 8, 085196. https://doi.org/10.1117/1.jrs.8.085196

Shendryk, Y., Davy, R., Thorburn, P., 2021. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. F. Crop. Res. 260, 107984. https://doi.org/10.1016/j.fcr.2020.107984

Shendryk, Y., Pan, L., Craigie, M., Stasolla, M., Ticehurst, C., Thorburn, P., 2020. A Satellite-Based Methodology for Harvest Date Detection and Yield Prediction in Sugarcane. Int. Geosci. Remote Sens. Symp. 5167–5170. https://doi.org/10.1109/IGARSS39084.2020.9323418

Silleos, N.G., Alexandridis, T.K., Gitas, I.Z., Perakis, K., 2006. Vegetation Indices: Advances Made in Biomass Estimation and Vegetation Monitoring in the Last 30 Years. Geocarto Int. 21, 21–28. https://doi.org/10.1080/10106040608542399

Singla, S.K., Garg, R.D., Dubey, O.P., 2020. Ensemble machine learning methods to estimate the sugarcane yield based on remote sensing information. Rev. d'Intelligence Artif. 34, 731–743. https://doi.org/10.18280/RIA.340607

Teare, I.D., Peet, M.M., 1983. Crop-water relations. Wiley, New York.

Thornley, J.H.M., Johnson, I.R., 1990. Plant and crop modelling—A mathematical approach to plant and crop physiology, Clarendon Press. The Blackburn Press, Oxford.

Todorovic, M., Albrizio, R., Zivotic, L., Saab, M.-T.A., Stöckle, C., Steduto, P., 2009. Assessment of AquaCrop, CropSyst, and WOFOST Models in the Simulation of Sunflower Growth under Different Water Regimes. Agron. J. 101, 509–521. https://doi.org/https://doi.org/10.2134/agronj2008.0166s

Tuia, D., Verrelst, J., Alonso, L., Perez-Cruz, F., Camps-Valls, G., 2011. Multioutput support vector regression for remote sensing biophysical parameter estimation. IEEE Geosci. Remote Sens. Lett. 8, 804–808. https://doi.org/10.1109/LGRS.2011.2109934

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J.P., Veroustraete, F., Clevers, J.G.P.W., Moreno, J., 2015. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties - A review. ISPRS J. Photogramm. Remote Sens. 108, 273–290. https://doi.org/10.1016/j.isprsjprs.2015.05.005

Viña, A., Gitelson, A.A., Nguy-Robertson, A.L., Peng, Y., 2011. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. Remote Sens. Environ. 115, 3468–3478. https://doi.org/10.1016/j.rse.2011.08.010

Wang, L., Zhou, X., Zhu, X., Dong, Z., Guo, W., 2016. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. Crop J. 4, 212–219. https://doi.org/10.1016/j.cj.2016.01.008

Wiegand, C.L., Richardson, A.J., Escobar, D.E., Gerbermann, A.H., 1991. Vegetation indices in crop assessments. Remote Sens. Environ. 35, 105–119. https://doi.org/10.1016/0034-4257(91)90004-P

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

Wu, C., Niu, Z., Tang, Q., Huang, W., 2008. Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. Agric. For. Meteorol. 148, 1230–1241. https://doi.org/10.1016/j.agrformet.2008.03.005

Yue, J., Feng, H., Yang, G., Li, Z., 2018. A comparison of regression techniques for estimation of above-ground winter wheat biomass using near-surface spectroscopy. Remote Sens. 10. https://doi.org/10.3390/rs10010066

3. PERFORMANCE EVALUATION OF SENTINEL-2 IMAGERY, AGRONOMIC AND CLIMATIC DATA FOR SUGARCANE YIELD ESTIMATION

Abstract

Given the importance of the sugarcane sector, machine learning techniques are being used as an important tool to improve yield estimation. This study aims to select the most relevant predictors from Sentinel-2 imagery, agronomic and climatic data, using Random Forest algorithm, to estimate sugarcane yield before the harvest in a mill in the west of Sao Paulo state. We used radiometric bands (Red-edge 1 to Red-edge 3, Red, NIR, SWIR1, and SWIR2) and vegetation indices from Sentinel-2 imagery (NDVIRE1 to NDVIRE3, EVI, CIRE1 to CIRE3, NDVI, NDWI1, NDWI2, SIWSI, NDMI, SAVI); agronomic data (soil type, number of harvests, variety, slope); climatic and agroclimatic data (temperature, precipitation, radiation, and crop water balance). We built models based on three datasets to create yield estimation models for the mill: i) the first dataset was included all variables (agronomic, climatic and agroclimatic, and remote sensing-based variables); ii) in the second dataset, the most strongly correlated variables were removed; and iii) the third dataset included the variables identified by feature selection within the 2nd dataset, based on the Gini index. The models showed R^2 values ranging from 0.58 to 0.70 with dataset 3, and d-Willmott index ranged from 0.83 to 0.89. The most relevant variables to estimate sugarcane yield were the number of harvests, water deficit, sugarcane varieties, temperature, precipitation data, NDRE2, NDRE3, CIRE2 and CIRE3 vegetation indices, and Red-edge, near-infrared narrow and SWIR1. The climatic data, agronomic and remote sensing data improved the model's performance.

Keywords: vegetation index, variable selection, sugarcane monitoring, machine learning, Random Forest

3.1. Introduction

Sugarcane is a perennial crop that plays a major role in the Brazilian economy. It is a main source of sugar and clean energy. In the last sugarcane crop season, the country produced 780 million tons of sugarcane (IBGE, 2022), accounting for more than 49.8 % of sugar and 50.2% of bioethanol production (CONAB, 2022). Sugarcane is also a key crop in sustainability and offers a vast potential for environmental benefits, especially through bioenergy. However, the knowledge of crop management practices, such as the cultivation practices, and the influence of many biophysical variables in its development, as climatic conditions are crucial to consolidate the potential of sugarcane yield (Bordonal et al., 2018; Henry and Kole, 2010).

In the last decades, many studies showed remote sensing data's relevance for estimate sugarcane yield at local and regional scale levels (Mulianga et al., 2013; Rao et al., 2002; Rudorff and Batista, 1990; Singla et al., 2020). The most straightforward approach to estimate crop yields using remote sensing is using empirical relationships between ground-based yield observations and vegetation indices computed at a single date or integrated during crop growth. Several studies have reported the potential for linear correlation between the vegetation indices obtained from images and the sugarcane yield estimation (Bégué et al., 2010; Mulianga et al., 2013; Rahman and J. Robson, 2016).

In most cases, yield estimation is based on remote sensing data from moderate -resolution satellite images (such as MODIS or Landsat). Typically, these studies use NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index) vegetation indices and achieve results with R^2 higher than 0.60, and RMSE less than 5.3

ton ha^{-1} (Fernandes et al., 2017; Lofton et al., 2012; Pandey et al., 2019; Rahman et al., 2017). Furthermore, some studies have considered the assimilation of remote sensing data into crop growth models. Despite accepting phenological variables as input data, these models cannot accurately estimate large-scale yields due to they need local data (Dorigo et al., 2007; Ma et al., 2022; Xie et al., 2017). A way to overcome this limitation is to consider the combination of remote sensing with climatic, agronomic, and management data (Verma et al., 2021).

Climatic data is essential for crop models since climate is the driver of the agricultural production variability (Everingham et al., 2002; Muchow et al., 1996). To improve yield estimating from remote-sensing data, Zhao and Justina (2020) combined precipitation data from the GPM (Global Precipitation Measurement), ECMWF (European Centre for Medium-Range Weather Forecasts), and spectral data from Landsat-8/OLI sensor to estimate sugarcane yield. Other studies also showed that using climatic data in crop models improved the sugarcane yield prediction compared to using only agronomic data in crop models (Bocca and Rodrigues, 2016; Charoen-ung and Mittrapiyanuruk, 2018; Luciano et al., 2021; Oliveira et al., 2017). In addition to climatic and remote sensing data, many variables can affect sugarcane yield, in a complex way, including field management, soil type, weeds, pests, diseases, and varieties (Alvarez et al., 1982).

Machine learning techniques are an opportunity to use large data sources in empirical crop yield estimation systems. They have been implemented successfully over the past several decades in many studies. The most used algorithms to crop yield estimation are artificial neural networks (ANN), random forest (RF), support vector machine (SVM), and gradient boosting tree (Van Klompenburg et al., 2020).

One of the advantages of using machine learning algorithms to estimate crop yield is the ability to work with many variables with no assumptions about their independence or linearity. In general, many predictors increase the complexity of the model, multicollinearity, potential instability, and overfitting (Han et al., 2016). Furthermore, irrelevant predictors degrade the model's performance, both in learning speed (due to high dimensionality) and predictive accuracy (due to irrelevant information), as well as the cost of collection and data storage. Therefore, selecting an optimal number of predictors is crucial to ensure the model's performance and accuracy (Maya Gopal and Bhargavi, 2019).

Many methodologies and algorithms are used to select a set of relevant predictors. These methodologies are called "Feature Selection". Among them are the exhaustive search algorithm, relief algorithm, and variable importance random forest algorithm. The exhaustive search algorithm performs a heuristic search exhaustively over all subsets, which makes it intractable due to the time and processing power required to execute the method (Kira and Rendell, 1992). Other algorithms avoid the exhaustive heuristic search, such as the Relief algorithm, which uses a statistical method. Oliveira et al., (2017) performed the feature selection using the Relief algorithm to predict the Total Recoverable Sugar in a mill in the state of São Paulo. The authors had good results with a 3.6 kg ton^{-1} Root Mean Square error (RMSE) and 2.02 kg ton^{-1} Mean Absolute error (MAE), using the learning technique of Random Forest after selecting the most important variables. One of the main disadvantages of the Relief algorithm is that it is likely to select redundant variables (Kira and Rendell, 1992).

The RF-based variable selection method computes an importance index of the variables with respect to predicted variables, such as the Gini index. The Gini index is a measure that reflects how each variable contributes to the homogeneity of nodes and leaves in the resulting Random Forest. It relates the score of each variable in relation to the subset of the most important variables (Breiman, 2001). Maya Gopal and Bhargavi (2019) compared the most usual features selection techniques and concluded that the RF variable importance algorithm leads to the best metrics (RMSE, MAE, and R^2). Although machine learning algorithms for crop yield estimation can use large datasets using

remote sensing and climatic variables (Abdel-Rahman and Ahmed, 2008; Weiss et al., 2020), it is crucial to pay attention to the feature selection process to improve the model's performance.

The objective of this work is to select the most relevant predictors from Sentinel-2 imagery, agronomic and climatic data, using the RF to estimate sugarcane yield before the harvest in a mill located in the west part of Sao Paulo state. We paid particular attention to algorithm parameterization and the size of representative samples, to compare the importance of variables that are needed to estimate yield accurately before the harvest.

3.2. Conclusions

We created empirical models for sugarcane yield estimation before the harvest using agronomic, climatic, crop water balance data and Sentinel-2 imagery. Removing the highly correlated variables and choosing the main variables that explain sugarcane yield variability improved the model's performance (R^2 =0.80) and reduced the data dimensionality, decreasing the time processing of the random forest model. The most important variables for sugarcane yield estimation were the number of harvests (agronomic data), crop water deficit, vegetation indices NDVIRE and CIRE, and Red-edge, NIR-8A, and $SWIR_1$ bands. Sentinel-2 imagery showed potential to estimate sugarcane yield, especially regarding vegetation indices that use Red-edge. The combination of satellite images with climatic and agronomic data improved the sugarcane yield estimation. Future study will include the model calibration using the most important variables from agronomic, climatic and Sentinel-2 imagery for sugarcane yield estimation in regional and temporal scale. Moreover, the creation of models for each sugarcane production cycles represents an important step for sugarcane yield estimation in local and regional areas.

References

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

Abebe, G., Tadesse, T., Gessesse, B., 2022. Combined Use of Landsat 8 and Sentinel 2A Imagery for Improved Sugarcane Yield Estimation in Wonji-Shoa, Ethiopia. J. Indian Soc. Remote Sens. 50, 143–157. https://doi.org/10.1007/s12524-021-01466-8

Alvarez, J., Crane, D.R., Spreen, T.H., Kidder, G., 1982. A yield prediction model for Florida sugarcane. Agric. Syst. 9, 161–179. https://doi.org/10.1016/0308-521X(82)90018-X

Arnt, W.R., 2016. Desempenho de variedades de cana-de-açúcar no pontal do paranapanema. Universidade Federal da Grande Dourados.

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

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

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

Boiarskii, B., 2019. Comparison of NDVI and NDRE Indices to Detect Differences in Vegetation and Chlorophyll Content. J. Mech. Contin. Math. Sci. https://doi.org/10.26782/jmcms.spl.4/2019.11.00003

Bordonal, R. de O., Carvalho, J.L.N., Lal, R., de Figueiredo, E.B., de Oliveira, B.G., La Scala, N., 2018. Sustainability of sugarcane production in Brazil. A review. Agron. Sustain. Dev. 38. https://doi.org/10.1007/s13593-018-0490-x

Breiman, L., 2001. Random Forests. Mach. Learn. 5–32.

Canata, T.F., Wei, M.C.F., Maldaner, L.F., Molin, J.P., 2021. Sugarcane yield mapping using high-resolution imagery data and machine learning technique. Remote Sens. https://doi.org/10.3390/rs13020232

Charoen-ung, P., Mittrapiyanuruk, P., 2018. Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques.

CONAB, 2022. Acompanhamento da Safra Brasileira de Cana-de-Açúcar – Primeiro Levantamento da safra 2022/23. Cia. Nac. Abastecimento. Ministério da Agric. Pecuária e Abastecimento.

Copernicus Sentinel-2, 2021. MSI Level-2A BOA Reflectance Product. Collection 1. https://doi.org/https://doi.org/10.5270/S2_-znk9xsj

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., Van den Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597. https://doi.org/10.1002/qj.828

Dimov, D., Uhl, J.H., Löw, F., Seboka, G.N., 2022. Sugarcane yield estimation through remote sensing time series and phenology metrics. Smart Agric. Technol. 2, 100046. https://doi.org/10.1016/j.atech.2022.100046

Doorenbos, J., Kassam, A.H., 1979. Yield Response to Water, Irrigation and Agricultural Development. Food And Agriculture Organization of The United Nations, Rome. https://doi.org/10.1016/b978-0-08-025675-7.50021-2

Dorigo, W.A., Zurita-Milla, R., de Wit, A.J.W., Brazile, J., Singh, R., Schaepman, M.E., 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. Int. J. Appl. Earth Obs. Geoinf. 9, 165–193. https://doi.org/10.1016/j.jag.2006.05.003

Embrapa, 1979. Clases de Declividade. Embrapa. Serviço Nac. Levant. e Conserv. Solos 83.

Everingham, Y.L., Muchow, R.C., Stone, R.C., Inman-Bamber, N.G., Singels, A., Bezuidenhout, C.N., 2002. Enhanced risk management and decision-making capability across the sugarcane industry value chain based on seasonal climate forecasts. Agric. Syst. 74, 459–477. https://doi.org/10.1016/S0308-521X(02)00050-1

Fensholt, R., Sandholt, I., 2003. Derivation of a shortwave infrared water stress index from MODIS near- and shortwave infrared data in a semiarid environment. Remote Sens. Environ. 87, 111–121. https://doi.org/10.1016/j.rse.2003.07.002

Fernandes, J.L., Ebecken, N.F.F., Esquerdo, J.C.D.M., 2017. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. Int. J. Remote Sens. 38, 4631–4644. https://doi.org/10.1080/01431161.2017.1325531

Gitelson, A., Merzlyak, M.N., 1994. Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus hippocastanum L. and Acer platanoides L. Leaves. Spectral Features and Relation to Chlorophyll Estimation. J. Plant Physiol. 143, 286–292. https://doi.org/https://doi.org/10.1016/S0176-1617(11)81633-0

Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. J. Plant Physiol. 160, 271–282. https://doi.org/10.1078/0176-1617-00887

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Han, H., Guo, X., Yu, H., 2016. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS 0, 219–224. https://doi.org/10.1109/ICSESS.2016.7883053

Hapfelmeier, A., Ulm, K., 2013. A new variable selection approach using Random Forests. Comput. Stat. Data Anal. 60, 50–69. https://doi.org/10.1016/j.csda.2012.09.020

Henry, R., Kole, C., 2010. Genetics, Genomics and Breeding of Sugarcane. Science Publishers, Boca Raton.

Huete, A.R., 1988. A Soil-Adjusted Vegetation Index (SAVI). Remote Sens. Appl. Soc. Environ. 25, 295–309.

Huete, A.R., Liu, H.Q., van Leeuwen, W.J.D., 1997. Use of vegetation indices in forested regions: Issues of linearity and saturation. Int. Geosci. Remote Sens. Symp. 4, 1966–1968. https://doi.org/10.1109/igarss.1997.609169

IBGE, 2022. Levantamento Sistemático da Produção Agrícola - LSPA. IBGE- Ist. Bras. Geogr. e Estatística.

Kira, K., Rendell, L.A., 1992. A Practical Approach to Feature Selection, Machine Learning Proceedings 1992. Morgan Kaufmann Publishers, Inc. https://doi.org/10.1016/b978-1-55860-247-2.50037-1

Kira, O., Nguy-Robertson, A.L., Arkebauer, T.J., Linker, R., Gitelson, A.A., 2016. Informative spectral bands for remote green LAI estimation in C3 and C4 crops. Agric. For. Meteorol. 218–219, 243–249. https://doi.org/10.1016/j.agrformet.2015.12.064

Köppen, W., Geiger, R., 1928. Klimate der Erde. Gotha Verlag Justus Perthes. Wall-map 150 x 200cm.

Kuhn, M., 2021. caret: Classification and Regression Training.

Lofton, J., Tubana, B.S., Kanke, Y., Teboh, J., Viator, H., Dalen, M., 2012. Estimating sugarcane yield potential using an in-season determination of normalized difference vegetative index. Sensors (Switzerland) 12, 7529–7547. https://doi.org/10.3390/s120607529

Luciano, A.C. dos S., Picoli, M.C.A., Duft, D.G., Rocha, J.V., Leal, M.R.L.V., le Maire, G., 2021. Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. Comput. Electron. Agric. 184. https://doi.org/10.1016/j.compag.2021.106063

Ma, C., Liu, M., Ding, F., Li, C., Cui, Y., Chen, W., Wang, Y., 2022. Wheat growth monitoring and yield estimation based on remote sensing data assimilation into the SAFY crop growth model. Sci. Rep. 12, 5473. https://doi.org/10.1038/s41598-022-09535-9

Machado, R.S., Ribeiro, R.V., Marchiori, P.E.R., Machado, D.F.S.P., Machado, E.C., Landell, M.G. de A., 2009. Respostas biométricas e fisiológicas ao deficit hídrico em cana-de-açúcar em diferentes fases fenológicas. Pesquisa Agropecuária Bras. 44, 1575–1582. https://doi.org/10.1590/s0100-204x2009001200003

Marin, F.R., Lopes-Assad, M.L., Assad, E.D., Vian, C.E., Santos, M.C., 2008. Sugarcane crop efficiency in two growing seasons in São Paulo State, Brazil. Pesqui. Agropecuária Bras. 43, 1449–1455. https://doi.org/10.1590/s0100-204x2008001100002

MapBiomas. (2023). Coleção 8 da Série Anual de Mapas de Cobertura e Uso da Terra do Brasil. <https://mapbiomas.org/>

Maya Gopal, P.S., Bhargavi, R., 2019. A novel approach for efficient crop yield prediction. Comput. Electron. Agric. 165, 104968. https://doi.org/10.1016/j.compag.2019.104968

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

Muchow, R.C., Robertson, M.J., Keating, B.A., 1996. Limits to the Australian sugar industry: climatic and biological factors, in: Intensive Sugarcane Production: Meeting the Challenges Beyond 2000. Proceedings of the Sugar 2000 Symposium. CSIRO Tropical Agriculture, Brisbane, QLD, Australia, pp. 37–54.

Mulianga, B., Bégué, A., Simoes, M., Todoroff, P., 2013. Forecasting regional sugarcane yield based on time integral and spatial aggregation of MODIS NDVI. Remote Sens. 5, 2184–2199. https://doi.org/10.3390/rs5052184

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D.G., Piles, M., Rodríguez-Fernández, N.J., Zsoter, E., Buontempo, C., Thépaut, J.N., 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Oliveira, M.P.G. de, Bocca, F.F., Rodrigues, L.H.A., 2017. From spreadsheets to sugar content modeling: A data mining approach. Comput. Electron. Agric. 132, 14–20. https://doi.org/10.1016/j.compag.2016.11.012

Pandey, S., Patel, N.R., Danodia, A., Singh, R., 2019. Discrimination of sugarcane crop and cane yield estimation using Landsat and IRS resourcesat satellite data. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 42, 229–233. https://doi.org/10.5194/isprs-archives-XLII-3-W6-229-2019

Peloia, P.R., Bocca, F.F., Rodrigues, L.H.A., 2019. Identification of patterns for increasing production with decision trees in sugarcane mill data. Sci. Agric. 76, 281–289. https://doi.org/10.1590/1678-992X-2017-0239

Pereira, H.R., Meschiatti, M.C., Pires, R.C. de M., Blain, G.C., 2018. On the performance of three indices of agreement: An easy-to-use r-code for calculating the willmott indices. Bragantia 77, 394–403. https://doi.org/10.1590/1678-4499.2017054

Person, A.; Grazziani, F., 2007. User guide to ECMWF forecast products. Meteorol. Bull. v. 4,.

Prado, H. do, 2005. Ambientes de produção de cana-de-açúcar na região centro-sul do Brasil. Encarte do Informações agronômicas 110, 121–17.

Priestley, C.H.B., Taylor, R.J., 1972. On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. Mon. Weather Rev. 100, 81–92. https://doi.org/10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2

R Core Team, 2020. R: A Language and Environment for Statistical Computing.

Rahman, M.M., J. Robson, A., 2016. A Novel Approach for Sugarcane Yield Prediction Using Landsat Time Series Imagery: A Case Study on Bundaberg Region. Adv. Remote Sens. 05, 93–102. https://doi.org/10.4236/ars.2016.52008

Rahman, M.M., Muir, J., Robson, A.J., 2017. Multi-temporal Landsat algorithms for the yield prediction of sugarcane crops in Australia. 7th Asian-Australasian Conf. Precis. Agric. 1–6.

4. OPERATIONAL SUGARCANE YIELD ESTIMATION IN THE STATE OF SAO PAULO USING REMOTE SENSING DATA

Abstract

Sugarcane yield estimation before the harvest is an important tool to support the sugar - energy sector. This aimed to create an operational empirical model, using the Random Forest algorithm, to estimate the sugarcane yield before the harvest, in the state of Sao Paulo (general model). For this, we used Sentinel-2 imagery, agronomic data, climatic data (temperature, precipitation), and crop water deficit data from three mills. The Sentinel-2 imagery were vegetation indices CIRE2, CIRE3, NDVIRE2 and NDVIRE3, spectral bands Red-edge 1 to 3, and near-infrared narrow (NIR-8A). The agronomic data referred to the sugarcane number of harvests and sugarcane variety. It was created three models based on training and testing data: I) Model I used 70% of the dataset for training and 30% for testing, II) Model II used 70% of the dataset for training and 30 % for testing in each mill, III) Model III considered the sugarcane production cycles, using 70% of the dataset for Plant cane and 70% for ratoon stages and testing on 30% of each sugarcane production cycle. Model I showed R^2 equal to 0.72 while Model II R^2 was between 0.60 and 0.78, the RMSE for Model I was 11.7 ton ha^{-1}, while for Model II from 8.62 to 15.56 ton ha^{-1}. The rRMSE was 16.5% for Model I and from 12.4% to 21.6%, for Model II. Model III showed R^2 greater than 0.61, and RMSE between 9.6 and 13.5 ton ha^{-1}. When comparing the average yield with RMSE errors we have better performance for model III with rRMSE less than 15.3%. The conception model by the sugarcane production cycle (Model III) could be improved by considering the harvest date. The general model showed potential to be applied to different locals using data from three mills.

Keywords: vegetation index, random forest, sugarcane monitoring, sugarcane production cycle.

4.1. Introduction

The estimation of the sugarcane yield is essential to have a suitable management production as the application of inputs, maintenance schedules, and labor throughout the chain, and for the decision-making of managers and producers in the mill (Gunnula et al., 2012). The most common methodology for estimating crop yield in Brazil is made by official institutions considering historical data and the experience of managers and producers in local areas, it could bring some uncertainties and subjectivity (IBGE, 2011). However, there are many methodologies to estimate crop yield considering climatic data; agronomic information such as variety, and soil type; remote sensing data; and the spatial scale and time series.

In recent decades, many studies focus on the relevance of remote sensing data to estimate the sugarcane yield at local levels (Luciano et al., 2021; Mulianga et al., 2013; Rudorff and Batista, 1990; Singla et al., 2020). In general, the sugarcane yield estimation in these studies is based on remote sensing data obtained from moderate-resolution satellite images, such as from sensor MODIS and sensors from Landsat satellite. Despite that, the spatial scale, planting, and estimation date as well as the details of practice management can affect the estimate of yield (Mavromatis, 2016). Regarding sugarcane yield estimate, the methodologies to estimate sugarcane yield are based on empirical models or crop simulation models, it is in general tested in local areas. The crop simulation models have disadvantages regarding input data at a regional scale and it depends on a lot of specific parameters to calibrate (Pagani et al., 2017). Furthermore, these simulation models need point location measurements, which it makes difficult to provide an estimate yield on a regional scale. Otherwise, the empirical models use fewer parameters, which makes them simple,

and when combined with remote sensing data it gives a better estimation of yield in spatial resolution and in regional areas (Filippi et al., 2020).

There are few studies developed to estimate sugarcane yield in large areas using remote sensing data and empirical models since the majority of them are conducted with few data and in local places (Abebe et al., 2022; Canata et al., 2021). In addition, these studies are based on the construction of a linear relationship between yield observations and vegetation indices calculated on a single date or a time series for crop growth (Bégué et al., 2010; Mulianga et al., 2013; Rahman and J. Robson, 2016). These methodologies provide a sugarcane yield estimation based on few variables or only one variable, in a single study site, which may not represent well the variability of the sugarcane field in other areas. Several factors have an effect on sugarcane yield, for example climatic conditions, such as crop water deficit, average temperature, precipitation, evapotranspiration, and radiation, which often implies temporal analyses over the time (Henry and Kole, 2010; Machado et al., 2009; Marin et al., 2008). Despite that, the amount of spatial and temporal data from remote sensing provides a promising opportunity to integrate these factors and remote sensing data to estimate sugarcane yield over time and on regional scale using, for example, machine learning techniques.

Machine learning techniques (ML) can be used to create operational empirical models to estimate crop yield since it operates with large amounts of data and can help intelligent system decision-making (Iniyan et al., 2023). There are many ML such as neural network techniques, deep learning, support vector regression (SVR), gradient boosted trees (GBT), and Random Forest (RF) (van Klompenburg et al., 2020). For sugarcane yield estimation RF has been widely used (Everingham et al., 2009; Felipe Maldaner et al., 2021; Krupavathi et al., 2022; Sunir Kumar et al., 2015). RF algorithm has the ability to integrate and process a large number of inputs derived from different variables such as satellite data, climate, agronomic and management data, and it is possible to investigate non-linear and hierarchical relationships between the predictors and the response using a joint learning approach (Breiman, 2001; Everingham et al., 2009). Furthermore, the RF allows to realize yield estimations with better results than single model approaches, while avoiding model overfitting, which it is interesting to develop operational empirical models, in particular for sugarcane yield.

In this paper we present an operational empirical model to estimate sugarcane yield before the harvest on a regional scale based on remote sensing data and datasets from three mills, located in Sao Paulo state, using the Random Forest algorithm. The aim is to estimate sugarcane yield before the harvest in different local conditions to guide strategy actions and improve the traditional estimation by mills. Moreover, we evaluated empirical models to estimate sugarcane yield before the harvest for the sugarcane production cycles (plant cane and sugarcane ratoon stages) to better support analysis during the growth cycle compared to only one general model.

4.2. Conclusion

We created operational empirical models to estimate sugarcane yield before the harvest on a regional scale using agronomic, climatic data, water deficit, and Sentinel-2 imagery. The creation of a general model calibrated with data from three mills allowed the estimation of sugarcane yield with R^2 of 0.72 and a relative error of 16%. The general model improves the yield response when applied to all mills (Model I) than to a single mill (Model II). Model II was able to estimate new sugarcane yields in different regions, even though their characteristics are not necessarily similar to the calibration model, proving the good capability of the training model. The sugarcane number of harvests was the most important variable to estimate sugarcane yield before the harvest, followed by CIRE and NDVIRE vegetation

indices, crop water deficit and precipitation. The creation of a model by sugarcane cycle of production (Model III) could be improved by considering the harvest date.

References

Abebe, G., Tadesse, T., Gessesse, B., 2022. Combined Use of Landsat 8 and Sentinel 2A Imagery for Improved Sugarcane Yield Estimation in Wonji-Shoa, Ethiopia. J. Indian Soc. Remote Sens. 50, 143–157. https://doi.org/10.1007/s12524-021-01466-8

Arnt, W.R., 2016. Desempenho de variedades de cana-de-açúcar no pontal do paranapanema. Universidade Federal da Grande Dourados.

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

Bordonal, R. de O., Carvalho, J.L.N., Lal, R., de Figueiredo, E.B., de Oliveira, B.G., La Scala, N., 2018. Sustainability of sugarcane production in Brazil. A review. Agron. Sustain. Dev. 38. https://doi.org/10.1007/s13593-018-0490-x

Breiman, L., 2001. Random Forests. Mach. Learn. 5–32.

Brüggemann, E., Klug, J.R., Greenfield, P., Dicks, H., 2001. Empirical modelling and prediction of sugarcane yields from field records, in: Proceedings of South Africa Sugarcane Technologists Association. pp. 204–210.

Canata, T.F., Wei, M.C.F., Maldaner, L.F., Molin, J.P., 2021. Sugarcane yield mapping using high-resolution imagery data and machine learning technique. Remote Sens. https://doi.org/10.3390/rs13020232

Cardozo, N.P., Sentelhas, P.C., 2013. Climatic effects on sugarcane ripening under the infl uence of cultivars and crop age. Sci. Agric. 70, 449–456. https://doi.org/10.1590/S0103-90162013000600011

Copernicus Sentinel-2, 2021. MSI Level-2A BOA Reflectance Product. Collection 1 [WWW Document]. https://doi.org/https://doi.org/10.5270/S2_-znk9xsj

Dimov, D., Uhl, J.H., Löw, F., Seboka, G.N., 2022. Sugarcane yield estimation through remote sensing time series and phenology metrics. Smart Agric. Technol. 2, 100046. https://doi.org/10.1016/j.atech.2022.100046

Dlamini, N.E., Zhou, M., 2022. Soils and seasons effect on sugarcane ratoon yield. F. Crop. Res. 284, 108588. https://doi.org/10.1016/j.fcr.2022.108588

Dong, T., Liu, Jiangui, Qian, B., He, L., Liu, Jane, Wang, R., Jing, Q., Champagne, C., McNairn, H., Powers, J., Shi, Y., Chen, J.M., Shang, J., 2020. Estimating crop biomass using leaf area index derived from Landsat 8 and Sentinel-2 data. ISPRS J. Photogramm. Remote Sens. 168, 236–250. https://doi.org/10.1016/J.ISPRSJPRS.2020.08.003

Doorenbos, J., Kassam, A.H., 1979. Yield Response to Water, Irrigation and Agricultural Development. Food And Agriculture Organization of The United Nations, Rome. https://doi.org/10.1016/b978-0-08-025675-7.50021-2

Dorigo, W.A., Zurita-Milla, R., de Wit, A.J.W., Brazile, J., Singh, R., Schaepman, M.E., 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. Int. J. Appl. Earth Obs. Geoinf. 9, 165–193. https://doi.org/10.1016/j.jag.2006.05.003

Embrapa, 1979. Clases de Declividade. Embrapa. Serviço Nac. Levant. e Conserv. Solos 83.

Everingham, Y., Sexton, J., Skocaj, D., Inman-Bamber, G., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agron. Sustain. Dev. 36, 27. https://doi.org/10.1007/s13593-016-0364-z

Everingham, Y.L., Smyth, C.W., Inman-Bamber, N.G., 2009. Ensemble data mining approaches to forecast regional sugarcane crop production. Agric. For. Meteorol. 149, 689–696. https://doi.org/10.1016/j.agrformet.2008.10.018

Felipe Maldaner, L., de Paula Corrêdo, L., Fernanda Canata, T., Paulo Molin, J., 2021. Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. Comput. Electron. Agric. 181, 1–9. https://doi.org/10.1016/j.compag.2020.105945

Filippi, P., Whelan, B.M., Vervoort, R.W., Bishop, T.F.A., 2020. Mid-season empirical cotton yield forecasts at fine resolutions using large yield mapping datasets and diverse spatial covariates. Agric. Syst. 184, 102894. https://doi.org/10.1016/j.agsy.2020.102894

Gao, B.-C., 1996. NDWI A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space. Remote Sens. Environ. 58, 257–266.

Gitelson, A., Merzlyak, M.N., 1994. Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus hippocastanum L. and Acer platanoides L. Leaves. Spectral Features and Relation to Chlorophyll Estimation. J. Plant Physiol. 143, 286–292. https://doi.org/https://doi.org/10.1016/S0176-1617(11)81633-0

Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. J. Plant Physiol. 160, 271–282. https://doi.org/10.1078/0176-1617-00887

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Gunnula, W., Kosittrakun, M., Righetti, T.L., Weerathaworn, P., Prabpan, M., 2012. Evaluating sugarcane growth and maturity using ground-based measurements and remote sensing data. Thai J. Agric. Sci. 45, 17–28.

Henry, R., Kole, C., 2010. Genetics, Genomics and Breeding of Sugarcane. Science Publishers, Boca Raton.

IBGE, 2011. Levantamento Sistemático Da Produção Agrícola. IBGE- Ist. Bras. Geogr. e Estatística.

Iniyan, S., Akhil Varma, V., Teja Naidu, C., 2023. Crop yield prediction using machine learning techniques. Adv. Eng. Softw. 175, 103326. https://doi.org/10.1016/j.advengsoft.2022.103326

Inman-Bamber, N.G., Bonnett, G.D., Spillman, M.F., Hewitt, M.L., Jackson, J., 2008. Increasing sucrose accumulation in sugarcane by manipulating leaf extension and photosynthesis with irrigation. Aust. J. Agric. Res. 59, 13–26.

Köppen, W., Geiger, R., 1928. Klimate der Erde. Gotha Verlag Justus Perthes. Wall-map 150 x 200cm.

Krupavathi, K., Raghubabu, M., Mani, A., Parasad, P.R.K., Edukondalu, L., 2022. Field-Scale Estimation and Comparison of the Sugarcane Yield from Remote Sensing Data: A Machine Learning Approach. J. Indian Soc. Remote Sens. 50, 299–312. https://doi.org/10.1007/s12524-021-01448-w

Kumar, S., Kumar, V., Sharma, R.K., 2015. Sugarcane yield forecasting using artificial neural network models. Int. J. Artif. Intell. Appl. 6.

Luciano, A.C. dos S., Picoli, M.C.A., Duft, D.G., Rocha, J.V., Leal, M.R.L.V., le Maire, G., 2021. Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. Comput. Electron. Agric. 184. https://doi.org/10.1016/j.compag.2021.106063

Machado, R.S., Ribeiro, R.V., Marchiori, P.E.R., Machado, D.F.S.P., Machado, E.C., Landell, M.G. de A., 2009. Respostas biométricas e fisiológicas ao deficit hídrico em cana-de-açúcar em diferentes fases fenológicas. Pesqui. Agropecuária Bras. 44, 1575–1582. https://doi.org/10.1590/s0100-204x2009001200003

MapBiomas. (2023). Coleção 8 da Série Anual de Mapas de Cobertura e Uso da Terra do Brasil. <https://mapbiomas.org/>

Marin, F.R., Lopes-Assad, M.L., Assad, E.D., Vian, C.E., Santos, M.C., 2008. Sugarcane crop efficiency in two growing seasons in São Paulo State, Brazil. Pesqui. Agropecuária Bras. 43, 1449–1455. https://doi.org/10.1590/s0100-204x2008001100002

Marin, F.R., Rattalino Edreira, J.I., Andrade, J., Grassini, P., 2019. On-farm sugarcane yield and yield components as influenced by number of harvests. F. Crop. Res. 240, 134–142. https://doi.org/10.1016/j.fcr.2019.06.011

Mavromatis, T., 2016. Spatial resolution effects on crop yield forecasts: An application to rainfed wheat yield in north Greece with CERES-Wheat. Agric. Syst. 143, 38–48. https://doi.org/10.1016/j.agsy.2015.12.002

Mulianga, B., Bégué, A., Simoes, M., Todoroff, P., 2013. Forecasting regional sugarcane yield based on time integral and spatial aggregation of MODIS NDVI. Remote Sens. 5, 2184–2199. https://doi.org/10.3390/rs5052184

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D.G., Piles, M., Rodríguez-Fernández, N.J., Zsoter, E., Buontempo, C., Thépaut, J.N., 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Pagani, V., Stella, T., Guarneri, T., Finotto, G., van den Berg, M., Marin, F.R., Acutis, M., Confalonieri, R., 2017. Forecasting sugarcane yields using agro-climatic indicators and Canegro model: A case study in the main production region in Brazil. Agric. Syst. 154, 45–52. https://doi.org/10.1016/j.agsy.2017.03.002

Pandey, S., Patel, N.R., Danodia, A., Singh, R., 2019. Discrimination of sugarcane crop and cane yield estimation using Landsat and IRS resourcesat satellite data. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 42, 229–233. https://doi.org/10.5194/isprs-archives-XLII-3-W6-229-2019

Pereira, H.R., Meschiatti, M.C., Pires, R.C. de M., Blain, G.C., 2018. On the performance of three indices of agreement: An easy-to-use r-code for calculating the willmott indices. Bragantia 77, 394–403. https://doi.org/10.1590/1678-4499.2017054

Prado, H. do, 2005. Ambientes de produção de cana-de-açúcar na região centro-sul do Brasil. Encarte do Informações agronômicas 110, 121–17.

Priestley, C.H.B., Taylor, R.J., 1972. On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. Mon. Weather Rev. 100, 81–92. https://doi.org/10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2

R Core Team, 2020. R: A Language and Environment for Statistical Computing.

Rahman, M.M., J. Robson, A., 2016. A Novel Approach for Sugarcane Yield Prediction Using Landsat Time Series Imagery: A Case Study on Bundaberg Region. Adv. Remote Sens. 05, 93–102. https://doi.org/10.4236/ars.2016.52008

Rudorff, B.F.T., Batista, G.T., 1990. Spectral response of wheat and its relationship to agronomic variables in the tropical region. Remote Sens. Environ. 31, 53–63. https://doi.org/10.1016/0034-4257(90)90076-X

Sanches, G.M., de Paula, M.T.N., Magalhães, P.S.G., Duft, D.G., Vitti, A.C., Kolln, O.T., Borges, B.M.M.N., Franco, H.C.J., 2019. Precision production environments for sugarcane fields. Sci. Agric. 76, 10–17. https://doi.org/10.1590/1678-992x-2017-0128

Shendryk, Y., Davy, R., Thorburn, P., 2021. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. F. Crop. Res. 260, 107984. https://doi.org/10.1016/j.fcr.2020.107984

Singla, S.K., Garg, R.D., Dubey, O.P., 2020. Ensemble machine learning methods to estimate the sugarcane yield based on remote sensing information. Rev. d'Intelligence Artif. 34, 731–743. https://doi.org/10.18280/RIA.340607

Thornthwaite, C.W., Mather, J.R., 1955. The water balance, 1st ed, Stanford Libraries. Centerton, New Jersey. https://doi.org/10.1201/9780203751435-9

Tukey, J.W., 1977. Exploratory Data Analysis by John W. Tukey, Biometrics. Reading, Mass: Addison-Wesley Pub. Co.

Van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

Vincini, M., Amaducci, S., Frazzi, E., 2014. Empirical estimation of leaf chlorophyll density in winter wheat canopies using Sentinel-2 spectral resolution. IEEE Trans. Geosci. Remote Sens. 52, 3220–3235. https://doi.org/10.1109/TGRS.2013.2271813

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90, 8995. https://doi.org/10.1029/jc090ic05p08995

Wright, M.N., Ziegler, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw. 77. https://doi.org/10.18637/jss.v077.i01

Yu, L., Shang, J., Cheng, Z., Gao, Z., Wang, Z., Tian, L., Wang, D., Che, T., Jin, R., Liu, J., Dong, T., Qu, Y., 2020. Assessment of Cornfield LAI Retrieved from Multi-Source Satellite Data Using Continuous Field LAI Measurements Based on a Wireless Sensor Network. Remote Sens. 12, 3304. https://doi.org/10.3390/rs12203304

Zhao, D., Li, Y.R., 2015. Climate Change and Sugarcane Production: Potential Impact and Mitigation Strategies. Int. J. Agron. 2015. https://doi.org/10.1155/2015/547386

5. CONSIDERAÇÕES FINAIS

Foram criados modelos empíricos para realizar a estimativa da produtividade da cana-de-açúcar antes da colheita usando dados agronômicos, climáticos, de déficit hídrico e imagens Sentinel-2A. A remoção das variáveis altamente correlacionadas e a escolha das principais variáveis para estimativa da produtividade da cana-de-açúcar melhoraram o desempenho dos modelos e, reduziram a dimensionalidade dos dados, diminuindo o tempo de processamento do *Random Forest*, o que pode facilitar a aquisição dos dados e manutenção do modelo. As variáveis mais importantes para a estimativa da produtividade da cana-de-açúcar foram o estágio de corte da cana-de-açúcar (dados agronômicos), o déficit hídrico da cultura (dados climáticos), os índices de vegetação NDVIRE e CIRE, e as bandas *Red-edge*, NIR-8A e SWIR1. As imagens Sentinel-2A mostraram potencial para estimar a produtividade da cana-de-açúcar, especialmente em relação aos índices de vegetação calculados com a banda *Red-edge*, que tiveram maior importância que os índices de vegetação mais utilizados na literatura, como o NDVI, EVI e SAVI. Para estudos futuros é recomendado a utilização desses preditores, para a estimativa da produtividade da cana-de-açúcar antes da colheita.

Os modelos empíricos regionais para estimar a produtividade de cana-de-açúcar mostram que utilizando o modelo *Random Forest* foi possível estimar a produtividade da cana-de-açúcar com R^2 de 0,72 e RMSE de 11,7 ton ha^{-1}. A criação de um modelo geral calibrado com dados das três usinas possibilitou maior capacidade de aplicação do modelo em regiões distintas. O modelo geral apresentou melhor resultado para estimativa da produtividade quando aplicado nas três usinas (Modelo I) do que ao ser aplicado em apenas uma das usinas (Modelo II). Embora o Modelo II tenha apresentado boa capacidade para estimar a produtividade de cana-de-açúcar antes da colheita, em diferentes regiões, as características gerais da região e, não necessariamente da mesma usina, foi essencial no conjunto de dados de treinamento do modelo. A criação de um modelo por ciclo de produção da cana-de-açúcar (cana-planta e estágios de soqueira - Modelo III) apresentou R^2 de 0,63, RMSE 13,5 ton ha^{-1} e rRMSE 15,3% para a cana-planta e R^2 de 0,61, RMSE 9,6 ton ha^{-1} e rRMSE de 14,8% para a cana-soca. O Modelo por ciclo de produção da cana-de-açúcar (Modelo III) é uma boa opção para análises futuras, principalmente porque o número de cortes tem grande influência na produtividade da cana-de-açúcar. Os modelos propostos neste estudo representam ferramentas de auxílio ao planejamento do setor canavieiro, possibilitando estimativa da produtividade e da produção de cana-de-açúcar e, incentivando aplicações semelhantes para outras culturas.