VOL.39 Nº4 2022
https://doi.org/102265\MECTA.9497.en

Evaluation of agronomic variables and Sentinel-2 satellite images to estimate sugar cane productivity using the Random Forest algorithm

Rafaella Pironato Amaro[a], Ana Cláudia dos Santos Luciano[b]

Department of Biosystems Engineering - USP
[a]rafaella.amaro@usp.br
[b]analuciano@usp.br

RESUMO: O objetivo deste trabalho, foi avaliar a importância de variáveis agronômicas e de imagens do satélite Sentinel-2 para estimativa da produtividade de cana-de-açúcar, utilizando o algoritmo Random Forest. Foram obtidos dados agronômicos referentes à variedade, estágio de corte, tipo de solo e relevo, além de dados provenientes das imagens de satélite referentes ao NDVI médio, máximo e o desvio padrão do NDVI de cada talhão. Foram criados dois modelos empíricos considerando: i) Variáveis agronômicas, ii) Variáveis agronômicas e imagens Sentinel-2. O modelo estimativo de produtividade apresentou R^2 igual a 0,64 e 0,83, RMSE de 10,17 e 7,0 ton/ha, para os modelos i e ii, respectivamente. A avaliação da importância das variáveis indicou que a variável estágio de corte foi a mais importante, seguida das variáveis variedade e NDVI médio do talhão. A combinação de variáveis agronômicas e de imagens de satélite trouxe melhorias na estimativa da produtividade de cana-de-açúcar.

PALAVRAS-CHAVE: Índice de vegetação, Modelo empírico, NDVI; Aprendizado de máquinas

ABSTRACT: The objective of this project was to evaluate the importance of agronomic variables and Sentinel-2 satellite images to estimate sugarcane yield using the Random Forest algorithm. We used agronomic data referring to the variety, cutting stage, soil type and relief, in addition to data from satellite images referring to the average, maximum NDVI and the standard deviation of the NDVI of each field. Two empirical models were created considering: i) Agronomic variables, ii) Agronomic variables and Sentinel-2 images. The model to estimate sugarcane yield showed R^2 equal to 0.64 and 0.83, RMSE of 10.17 and 7.0 ton/ha for models i and ii, respectively. The evaluation of the importance of the variables indicated that the variable cutting stage was the most important, followed by the variable variety and average NDVI of the field. The combination of agronomic variables and satellite images brought improvements to estimate sugarcane productivity.

KEYWORDS: Vegetation index, Empirical model, NDVI, Machine learning

1. Introduction

Sugar cane, which occupies approximately 8.7 million hectares in Brazil [1], is a global crop with importance not only for sugar production but is also considered one of the great alternatives for the biofuel sector due to its to the great potential in the production of ethanol and its by-products [1, 2]. In Brazil, culture has great social and economic importance for the country, as it generates jobs and exports abroad. In Brazil, culture has great social and economic importance for the country, as it generates jobs and exports abroad. Despite climate fluctuations, sugar cane production reached around 753 million tons in 10.1 million hectares in the 2021 harvest [3]. The Center-South region represented the largest productive axis in the country, accounting for 92% of the total sugar cane produced [1]. However, since sugar cane is a semi-perennial plant, it suffers from climate influences, which fluctuate during the crop's growth cycle. Such oscillations occur especially in the aspect of precipitation and its regularity of distribution, contrary to what happens with annual crops that are influenced by the climate in limited periods [1]. Therefore, sugar cane can suffer from occasional water deficits in some locations, impacting on the productive potentials that may vary depending on the interaction between the time of year in which they occur and the phase of the phenological cycle of the crop [3,4]. In this sense, the climate is a fundamental factor for the agricultural planning of

sugar cane production. The combined effects of natural climate variability, growing population conditions, soil loss and climate change require methods that provide timely and accurate assessment of crop growth and production and contribute to increased production sustainability. of agricultural foods [5, 6]. In this context, the need for adequate strategic planning, forecasting harvests of a given crop and knowledge of its distribution in geographic space is extremely important for the planning of the Brazilian sugar-energy sector. Furthermore, the monitoring of sugar cane production assists in the creation of public policies and food security, which directly impacts on improving the accuracy and robustness of crop monitoring systems [8].

Among the means of agricultural monitoring, there is the estimation of area and productivity. The forecast of agricultural productivity according to traditional methods is conducted through agricultural surveys or by specialists, based on assessments of crop conditions, historical production in the area and environmental conditions [9],[10]. Such methods are subjective, time-consuming and often unrepresentative, due to the small sample sizes, which do not consider all the spatial variability of the production plots [11]. Another disadvantage of these traditional methods is that they are time-consuming and costly, given the large number of people involved [12].

To combat the subjectivity of traditional methods of predicting agricultural productivity, and enable the analysis of spatial and temporal variability, estimation based on empirical predictive models with satellite images is a promising alternative, which assists sugar cane producers in the assertive decision-making, helping in the management of the areas. Satellite images have been widely used in the monitoring of agricultural crops for the general assessment of the state of the sugar cane crop [13], such as, for example, in estimating its productivity [14]. The productivity estimate can be made based on agronomic and climatic data and the combination with satellite images, using conventional statistical techniques or machine learning [15-17].

Some advantages of using machine learning algorithms, such as Random Forest (RF), are related to the ability to use a large amount and variety of information, such as numerical and categorical data, arising from the combination of remote sensing data and agronomic data [18]. The integrated use of satellite images with machine learning algorithms, such as Random Forest, has shown promising results for predicting the production of crops such as wheat [19], soybean [20] and sugarcane [21] helping in accurately estimating productivity over the years and under different environmental conditions.

Considering the variability of environmental conditions, crop yield prediction is not trivial, so predictive models using data mining techniques and satellite images can accelerate the development and, improve the accuracy and robustness of these yield prediction systems in a way regional and temporal. Also, the use of productivity estimation models with satellite images assists sugar cane producers spatially and temporally in decision-making, helping in the management of areas, reducing costs and improving crop productivity.

The objective of this work is to evaluate the potential of agronomic variables and Sentinel-2 satellite images to estimate sugar cane productivity in the state of São Paulo, using the Random Forest machine learning algorithm.

2. Material And Methods

2.1 Data

The study area is located in the Catanduva region, in the center of the state of São Paulo, Brazil **figure 1**. The predominant type of soil is clayey, with the soil classes with greater predominance being Red Yellow Latosol and Red Latosol. In the study area, production environments C and D are predominant and the average productivity of the last three harvests was 72 ton/ha. In total there are 3447 plots with an average area of 9,15 hectares. The climate classification of the region is type AW, characterized by being tropical, with much more rainfall in summer than in winter. The average temperature is 23.3 °C and the average annual rainfall is 1444 mm [22].

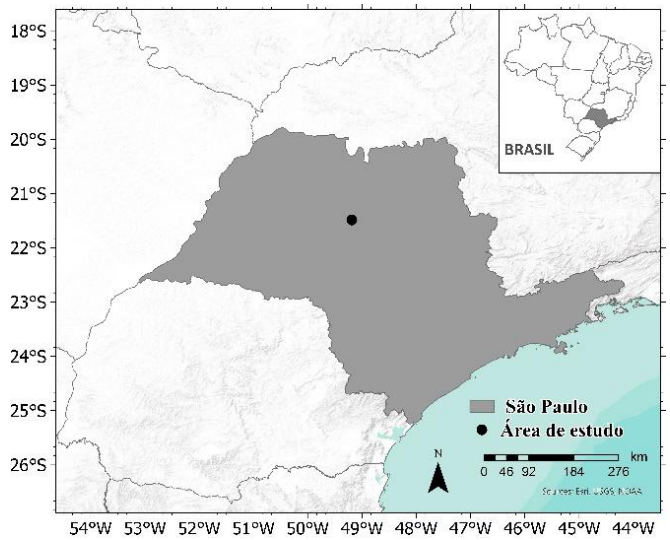
Agronomic data was obtained from a partner company for the plots in the study area. For each plot, data regarding the variety of sugar cane, cutting stage, TCH (ton of sugarcane per hectare), soil types and relief were obtained. In total there are 53 sugarcane varieties in the study area, the on-site cutting stage is between the 1st to

the 8th cut. The relief varies from smooth wavy to wavy. Data refer to the 2018-2019 sugar cane harvest, which comprises April 1, 2018, to March 31, 2019.

All data were organized to remove inconsistencies and failures (*outliers*) and, therefore, a standardization of information was performed. This consistency analysis was performed using the statistical software R [23]. After processing the data, a total of 2691 plots were obtained.

Satellite images were obtained using the MSI (Multispectral Imager) orbital multispectral sensor on board the Sentinel-2A satellite. These images have a temporal resolution of 5 days and 13 spectral bands ranging from 443 to 2190 nm, with a spatial resolution of 10 m for the red (B4) and near infrared (B8) bands. In addition, the images have atmospheric correction. The surface reflectance images of bands 4 and 8 were selected, which were processed for 7/9th/2018. Then, the spectral index Normalized Difference Vegetation Index - NDVI [24] was calculated in each plot. Finally, for each plot, the values of average NDVI, maximum NDVI and standard deviation of NDVI were calculated.

Fig. 1 - Location of the study area.



2.2 Sugar cane productivity modeling

Agronomic data and data from satellite images were integrated through the creation of empirical models, using the Random Forest (RF) regression algorithm, implemented in the statistical software R [23]. The RF technique is based on *bagging* decision trees, with an important extension - in addition to showing the records, the algorithm also shows the trees. In traditional decision trees, to determine how to create a subpartition of a partition, the algorithm chooses a variable and a division point through the minimization of a criterion to be chosen. However, in the case of RF, at each stage of the algorithm, the choice of a variable is limited to a random subset of variables. Thus, when compared to the basic tree algorithm, the RF algorithm adds two more steps: *bagging* and bootstrap sampling of the variables in each division [25]. RF has been used for productivity forecasting, due to its ability to handle high data dimensionality, outlier detection, robustness against overfitting and the possibility of studying the importance of the input variable in a calibrated model [26]. For the construction of the models, 80% of the data were used for calibration and 20% for validation. Two empirical models were created: i) Model 1 used only agronomic data, such as cutting stage, sugarcane variety, soil type and relief; ii) Model 2 considered the agronomic data of Model 1 (cutting stage, sugarcane variety, soil type and relief) and satellite image data was added, such as the average values of the NDVI vegetation index, NDVI standard deviation and maximum NDVI vegetation index values. The evaluation metrics used were the coefficient of determination (R^2) and Root Mean Square Error (RMSE). To determine the evaluation metrics, the TCH estimated by the model, using the 20% of the data from the validation set, was compared to the TCH measured by the plant. Finally, the importance of the variables was evaluated using the Random Forest algorithm [27].

4. Results And Discussion

Model 1, with agronomic variables, presented R^2 equal to 0.64 and RMSE equal to 10.2 ton/ha. The data dispersion can be seen in **figure 2a.** Similar results, R^2 of 0.73, were found for the study

using agronomic variables such as soil type, furrow width, plot yield in relation to the last year, sugarcane variety, irrigation, epidemic control, fertilization, and rainfall volume [28]. Other studies on sugarcane productivity obtained variation in the average absolute error obtained (MAE) between 4.6 and 7.5 ton/ha, that is, RMSE values between 2.1 and 2.7 ton/ha, close to that of the present study. The authors considered production and management variables, in addition to the climate that occurred during the analysis period, in order to evaluate models such as Artificial Neural Networks, Support Vector Machines, Driven Regression Trees and Random Forest [29]. Both authors associated productivity with climatological variables and achieved better results demonstrated by the evaluation metrics.

Among the variables evaluated, by model 1, the cutting stage was the most important variable, followed by the variety of sugar cane, type of soil and, finally, relief **table 1**. A similar result was found by other authors, who, carrying out a study with a decision tree for a single sugar mill unit in the west of the state of São Paulo, found that the number of cuts and clay content in the upper soil layer (up to 25 cm) are the main factors that affect the productivity of sugar cane [30]. Furthermore, the evaluation of more than one data mining technique, using agronomic variables and climate, to identify and order the main variables that condition sugar cane productivity, showed that the number of cuts was the most important factor by all data mining techniques [16].

As for the variety of sugar cane, it is known that it is a direct indicator of productivity, since the differences between the varieties contribute significantly to the variability of productivity [31]. This is because each variety of sugar cane has a specific characteristic. In addition, it is emphasized that there is a significant interaction between varieties and successive cuts in sugar cane productivity [32], which directly reflects on productivity.

The relief was the variable with less relevance in relation to the other variables. This fact can be explained due to the low variability of the terrain features, which remained between smooth wavy and wavy and ended up not influencing, in this study, in a significant way in the productivity variability.

By adding other variables to the empirical productivity model, such as mean values, standard deviation and maximum values of the NDVI vegetation index, the model presented better adjustments (R^2 equal to 0.83), see **figure 2b**. The R^2 obtained increased to 0.83 and the RMSE decreased to 7.0 ton/ha compared to model 1. This fact demonstrates that the variables from satellite images brought a gain in information in the creation of an empirical model for estimating sugar cane productivity. Similar studies found values of R^2 equal to 0.94 for sugarcane using images from the Landsat-8 satellite, and vegetation indices that combine the red and infrared bands such as the NDVI, Enhanced Vegetation Index (EVI) and the Soil-Adjusted Vegetation Index (SAVI), in addition to the Normalized Green Vegetation Index (GNDVI) [13]. Despite the differences in the results of the models, it is noteworthy that these studies considered more than one vegetation index over time series, rather than a single date, which was found by the authors to estimate more accurately compared to a single index of vegetation and a specific date of collection of the satellite image.

As for the importance of variables, for Model 2, the variables that were most important in estimating productivity were like Model 1, that is, the cutting stage and the variety of sugar cane. However, the variables related to the NDVI were more important than the type of soil and relief **table 1**. Vegetation indices such as the NDVI are primarily related to the abundance of green vegetation cover and biomass and are sensitive to variations in plant canopy stress responses, cultivars and management practices [33, 34]. In addition, data on vegetation indices obtained during the period of maximum growth are directly related to the productivity of the sugar cane crop [35]. Other studies have shown the importance of indices such as the NDVI for predicting sugar cane productivity. *Mulianga et all.* [36] obtained an RMSE of less than 5 ton/ha to estimate productivity using the NDVI

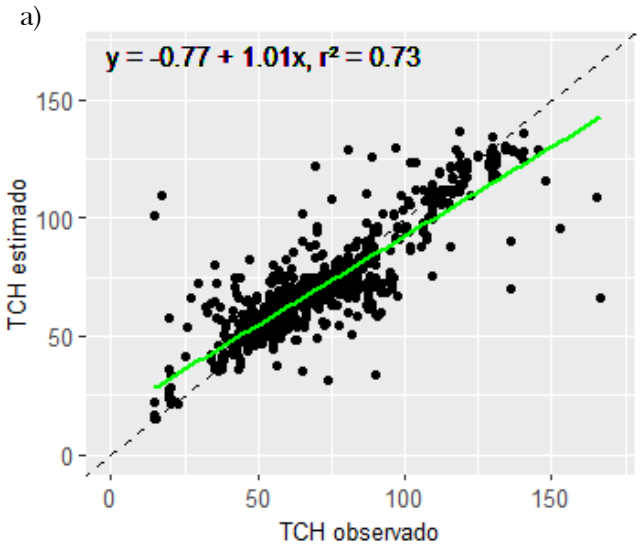
from the MODIS (Moderate Resolution Imaging Spectroradiometer) satellite and agronomic data, using a linear regression model *Fernandes et all*. [37] also showed the importance of the NDVI when using it to predict the productivity of sugar cane in the State of São Paulo, for this, images from the MODIS sensor and a model of a set of artificial neural networks were used. The R^2 obtained in this study was 0.61.

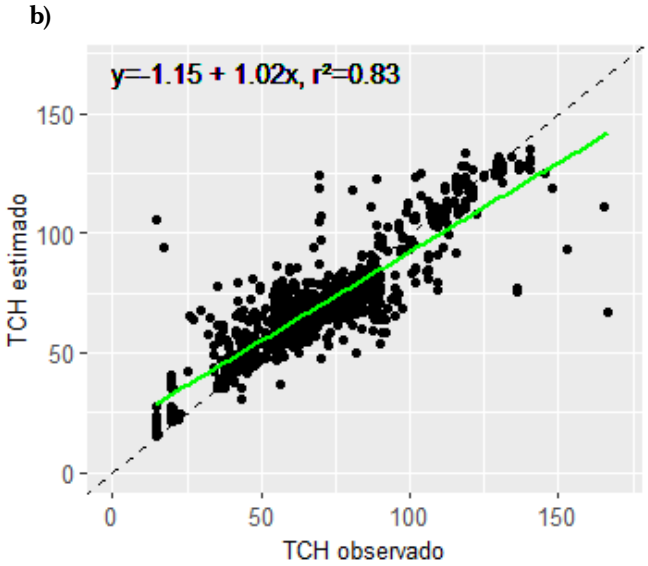
Tab. 1 - Importance of variables.

Order of importance	Model 1	Model 2
1	Cutting stage	Cutting stage
2	Variety of sugar cane	Variety of sugar cane
3	Soil type	Average NDVI
4	Relief	NDVI standard deviation
5	-	Max NDVI
6	-	Soil type
7	-	Relief

Source: Author.

Fig. 2 – Scatter plot for model 1 considering only agronomic variables (a) and considering agronomic and satellite variables -model 2 (b).





5. Conclusions

Empirical productivity models were created using agronomic information and images from the Sentinel-2 satellite, based on the Random Forest algorithm. The results demonstrated that the combination of agronomic variables and satellite images brought improvements in the estimation of sugar cane productivity. The most important variables for the empirical models were the cutting stage, sugar cane variety and average NDVI of the plot. Future studies should be carried out to integrate climate information and other vegetation indices. Finally, the use of temporal series of images to estimate and monitor productivity can bring gains to the monitoring of sugar cane.

References

[1] CONAB. Acompanhamento da Safra Brasileira de Cana-de-Açúcar – Quarto Levantamento da safra 2020/21. Companhia Nacional de Abastecimento, v.7, 2021, p. 57

[2] Surendran, U.; Ramesh, V.; Jayakumar, M.; Marimuthu, S.; Sridevi, G. Improved sugarcane productivity with tillage and trash management practices in semi-arid tropical agro ecosystem in India Soil Tillage Res., 158, 2016.pg 10-21.

[3] INMAN-BAMBER, N.G.; BONNETT, G.D.; SPILLMAN, M.F.; HEWITT, M.L.; JACKSON, J. Increasing sucrose accumulation in sugarcane by manipulating leaf extension and photosynthesis with irrigation. Australian Journal of Agricultural Research, v.59, p.13-26, 2008.

[4] MACHADO, R.S.; RIBEIRO, R.V.; MARCHIORI, P.E.R.; MACHADO, D.F.S.P.; MACHADO, E.C.; LANDELL, M.G.A. Respostas biométricas e fisiológicas ao déficit hídrico em cana-de-açúcar em diferentes fases fenológicas. Pesquisa Agropecuária Pesquisa Agropecuária Brasileira, Brasília. v.44, n.12, p.1575-1582, 2009.1582, 2009.

[5] IBGE. Produção Agrícola Municipal – PAM. 2019. Disponível em: <https://sidra.ibge.gov.br>. Acesso em: 10 ago. 2021.

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

[7] IPCC Summary for policymakers. 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.), Global Warming of 1.5 °C. 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, and efforts to eradicate poverty. World Meteorological Organization, Geneva, Switzerland, 2018, 32pg.

[8] Holzman, M.E.; Rivas, R.; Piccolo, M.C. Estimating soil moisture and the relationship with crop yield using surface temperature and vegetation index. Int. J. Appl. Earth Obs. Geoinf., 28, 2014, pg. 181-192. 10.1016/j.jag.2013.12.006

[9] Bocca, F.F.; Rodrigues, L.H.A. Arraes, N.A.M. When do I want to know and why? Different demands on sugarcane yield predictions. Agric. Syst., 135, 2015, pg. 48-56.

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

[11] Basso, B., Cammarano, D., Carfagna, E. Review of Crop Yield Forecasting Methods and Early Warning Systems, in: Report Presented to First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics FAO Headquarters, Rome, Italy. 2, 2013, pg. 1-56. 10.1017/CBO9781107415324.004

[12] Picoli, M.C. A Estimativa da produtividade da cana-de-aúcar utilizando agregados de redes neurais artificiais: estudo de caso Usina Catanduva. 2007. 90p. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2007.

[13] Singla, S.K., Garg, R.D., Dubey, O.P. Ensemble machine learning methods to estimate the sugarcane yield based on remote sensing information. Revue d'Intelligence Artificielle, v. 34, n.6, 2020, pg. 731-743. https://doi.org/10.18280/ria.340607

[14] Cechin-Júnior, C. Johann, J. A. Antunes, J. F. G.; Deppe, F. Sugarcane mapping in Paraná State Brazil using MODIS EVI images.International Journal of Advanced Remote Sensing and GIS, v.9, n.1, pg. 3205-3221, 2020.

[15] Verma, A.M.; GaRG, P. K.; Prasad, K. S. H.; Dadhwal, V. K.; Dubey, S. K.; Kumar, A. Sugarcane Yield Forecasting Model Based on Weather Parameters. Sugar Tech. v.23, n.1, 2021, pg.158–166.

[16] Hammer, R. G.; Sentelhas, P. C.; Mariano, J. C. Q. Sugarcane yield prediction through data mining and crop simulation models. Sugar Tech., 22, 2020, pg. 216-225. 10.1007/s12355-019-00776-z.

[17] Luciano, A. C. S.; Picoli, M. C. A.; Duft, D. G.; Rocha, J. V.; Leal, M. R. L. V.; Maire, G. Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. Computers and Electronics in Agriculture. 184, 2021, 106063. https://doi.org/10.1016/j.compag.2021.106063.

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

[19] Kamir, E.; Waldner, F.; Hochman, Z. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. Journal of Photogrammetry and Remote Sensing, v.160, 2020, pg. 124-135.

[20] Schwalbert, R.A.; Amado, T.; Corassa, G.; Pott, L.P.; Prasad, P.V.V.; Ciampittia, I.A. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology. v.284, 2020, p. 07886.

[21] Shendryk, Y.; Davy, R.; Thorburn, P. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. Field Crops Research. v.260, 2021, p.107984. https://doi.org/10.1016/j.fcr.2020.107984

[22] Köppen W, Geiger R (1928) Klimate der Erde. Justus Perthes, Gotha. 1928. Wall-Map 150 cm x 200 cm.

[23] RSTUDIO. Studio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. 2020. URL http://www.rstudio.com/rstudio

[24] Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation system in the great plains with ERTS Earth Resources Technology Satellite-1 Symposium, 2. Washington, D.C., Proceeding,1, NASA. Goddart Space Flight Center, Washington, D.C., 1973, pg.309-317.

[25] BRUCE, P; BRUCE, A. Estatística Prática para Cientistas de Dados: 50 conceitos essenciais. 1. ed. Rio de Janeiro: Alta Books, 2019.

[26] Gislason, P.O.; Benediktsson, J.A.; Sveinsson, J.R. Random forests for land cover classification Pattern Recognit. Lett., 27, 2006, pg. 294-300. 10.1016/j.patrec.2005.08.011.

[27] Breiman, L. Random forests. Mach. Learn., 45, 2001, pg. 5-32. 10.1023/A:1010933404324.

[28] Charoen-Ung, P. Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques.15th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2018, pg.1-6.

[29] Bocca, F. F.; Rodrigues, L. H. A. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Computers and Electronics in Agriculture, v. 128, 2016, p. 67–76.

[30] Peloia, P.R., F.F. Bocca, and L.H.A. Rodrigues. 2019. Identification of patterns for increasing production with decision trees in sugarcane mill data. Scientia Agricola 76. 2019, pg.281–289.

[31] Ramburan, S.; Wettergreen, T.; Berru, S.D.; Shongwe, B. Genetic, environmental and management contributions to ratoon decline in sugarcane. Field Crops Research, Amsterdam, v. 146, 2013, pg.105-112.

[32] Arnt, W. R. Desempenho de variedades de cana-de-açúcar em duas épocas de colheita no pontal do Paranapanema. Orientador: 2016. 55p. Dissertação (Mestrado) – Agronomia. Universidade Federal da Grande Dourados. Dourados-MT.

[33] Rudorff, B.F.T.; Batista, G.T. Yield estimation of sugarcane based on agrometeorological spectral models. Remote Sens. Environ., 33, 1990, pg. 183-192, 10.1016/0034-4257(90)90029-L

[34] Silleos, N.G.; Alexandridis, T.K.; Gitas, I.Z.; Perakis, K. Vegetation indices: advances made in biomass estimation and vegetation monitoring in the last 30 years. Geocarto Int., 21, 2006, pg. 21-28. 10.1080/10106040608542399

[35] Pinheiro Lisboa, I.; Melo Damian, J. Roberto Cherubin, M. Silva Barros, P.P. Ricardo Fiorio, P. Cerri, C.C. Eduardo Pellegrino Cerri, C. Prediction of Sugarcane Yield Based on NDVI and Concentration of Leaf-Tissue Nutrients in Fields Managed with Straw Removal. Agronomy 8, 2018, pg.196. https://doi.org/10.3390/agronomy8090196

[36] Mulianga, B.; Begue, A.; Simoes, M. Todoroff, P. Forecasting regional sugarcane yield based on time integral and spatial aggregation of MODIS NDVI. Remote Sens., v.5, n.5, 2013, pg. 2184-2199. 10.3390/rs5052184.

[37] Fernandes, J.L.; Esquerdo, J.C.D.; Favilla, N.F. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. Taylor & Francis, v.38, n.16, 2017, pg. 4631-4644.