SBI Agro 2023

Peanut yield estimation models using Machine Learning techniques and Google Earth Engine.

Jarlyson Brunno Costa Souza[1], Franklin Daniel Inácio[2], Samira Luns Hatum de Almeida[1], Armando Lopes de Brito Filho[1], Adão Felipe dos Santos[2], Rouverson Pereira da Silva[1]

[1]Department of Engineering and Mathematical Sciences, School of Veterinarian and Agricultural Sciences, São Paulo State University (UNESP), São Paulo, Brazil.

[2]Department of Agriculture, School of Agricultural Sciences of Lavras, Federal University of Lavras (UFLA), Lavras 37200-900, Minas Gerais, Brazil.

Abstract. *The estimation of the yield for the peanut crop can be performed by visual methods, traditional destructive methods, which are time consuming, laborious and with imprecise predictions. Thus, the objective was to use SR techniques and Artificial Neural Networks (ANN) in the development of an innovative method for yield prediction in the peanut crop. The experiments were conducted in the state of São Paulo in the 2021/2022 crop season, in 6 commercial farms in sandy and clayey areas. The RBF and MLP networks were able to estimate peanut yield with an accuracy below 1000 kg/ha. GNDVI was a better vegetation index with estimation accuracy of 238.7 and 296.54 for the RBF and MLP networks, respectively.*

Introduction

The estimation of yields assists in planning, decision making and management of crop resources. In crops like peanuts, estimating yields considering the characteristics of cultivars and soil types allows the generation of more accurate estimation models, since it allows the analysis of the interaction between genotype and environment. As with other crops, the official estimates of peanut yields are based on visits by technicians to the production fields, which makes the estimate a subjective method and subject to high errors. Estimating productivity is fundamental for defining the management of agricultural harvest operations, as well as for planning transport and storage logistics, and can provide greater profits for producers.

Climate-based models are also one of the methods that exist for trying to estimate crop yields. These methods are generally a statistical approach in which linear regressions are used to relate yields to various climatic parameters acquired at various locations and dates of the year. Over the years several attempts have been made to generate yield estimation models using climatic variables and soil classification, but variations in climatic conditions and the unavailability of climatic data make this task complicated (Joshi et al. 2020). The creation of a simple peanut yield estimation model has opened space for tools such as Remote Sensing (RS) and Machine Learning to create methods that can estimate yields quickly and accurately.

Among the main tools of SR, one can mention the Vegetation Indices (IV's), which are simple and effective algorithms for quantitative and qualitative assessment of the condition of vegetation cover, among other approaches and applications in agriculture.

IR's transform spectral bands to a single variable and thus minimize the effects of soil, topography, and viewing angle on the spectral response of the desired feature. No yield estimation model has yet been created for the peanut crop, and the use of IR's along with the use of Artificial Neural Networks can be an alternative to assist the producer in the evaluation of several agronomic parameters of the peanut crop, as is the case of peanut maturity estimation (Santos., 2021, Souza et al., 2022).

Considering the monitoring of vegetation time series through vegetation indices (IV) and associate these data with productivity generating information for decision making and strategic planning, it is essential to use modern techniques of data analysis, such as Artificial Neural Networks (ANN's). This type of data analysis can process a large volume of data and transform it into reliable information, especially data that does not have a linear distribution model, i.e. that is extremely affected by spatial-temporal variability. The utilization of cloud computing (GEE) for the acquisition of satellite images from free platforms, associated with ANN, presents a great potential for the development of models for peanut yield prediction. Thus, this study aimed to develop and test peanut yield prediction models using Radial Base Function (RBF) and Multi-Layer Preceptor (MLP) ANNs with Sentinel images.

Material and Methods

The work was conducted in six production areas in the interior of the state of São Paulo, Brazil, located in the municipalities of Jaboticabal, Taquaritinga (Granja), Ibitinga, Monte Alto (Frutal and Santa Gertrudes), Taiúva (Santa Adélia), Guatapará (Capão) (Figure 1). In all areas, the long cycle (125-130 days) cultivar IAC OL3 (runner type) was used. The adopted sowing spacing was 0.90 m between rows for all areas.

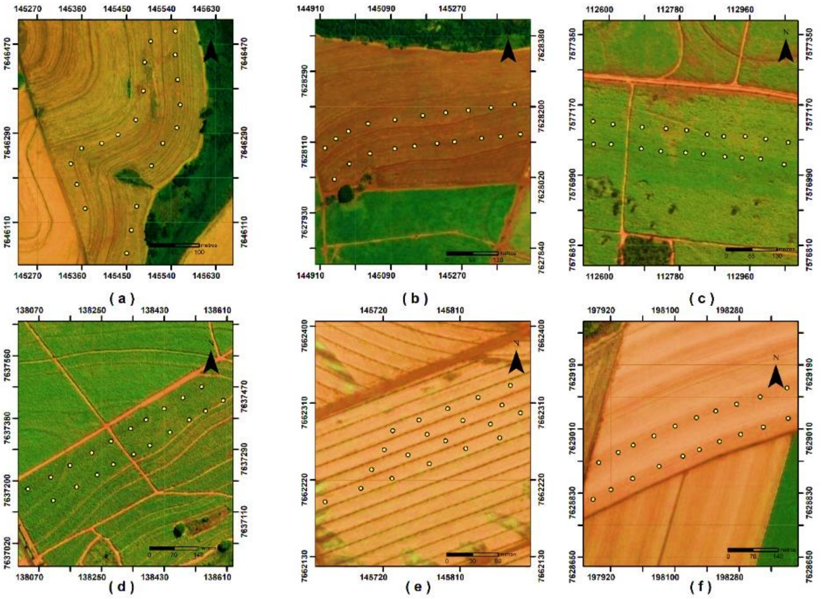


Figure 1. Distribution of yield sample points in A) Frutal, B) Granja, C) Ibitinga, D) Santa |Gertrudes, E) Santa Adélia, F) Capão.

Field Samples

In each area, 20 georeferenced sampling points were installed, spaced approximately 50 meters apart (Figure 2). Monitoring using satellite images began after 95 DAS and were performed weekly until the harvest (uprooting) of the experimental areas.

At each sampling point, a 2 m^2 frame was used to measure productivity; all plants within the frame were pulled, bagged and identified to then calculate productivity at each sampling point. After separation of the pods from the plants, they were weighed and later placed in a forced air oven at 65ºC for 72 hours to measure the dry weight of the pods. After this period, the humidity value was corrected to 8%, extrapolating the values to kg ha^{-1}.

Acquisition and processing of satellite images

For acquisition of spectral data of the peanut crop, satellite images from the Sentinel-2A platform were downloaded coinciding with the dates for monitoring. Sentinel-2 is a European multispectral satellite, which acquires images in 13 spectral bands, such as visible (band 2-4), Red Edge (RE, band 5-7), Near Infrared (NIR, band 8) and Shortwave Infrared (SWIR, band 11-12). For this work, cloud-free images from 95 to 125 Days After Sowing (DAS) were used. All the necessary corrections were performed, such as top-of-atmosphere reflectance (TOA) values scaled to 10,000 by means of radiometric calibration (Sentinel-2User, 2022). After image pre-processing, vegetation indices (Table 1) were calculated. For each area, a file in "shape" format was imported into GEE so that the average values of vegetation indices were extracted for each of the dates.

Table 1. Vegetation indices used

Indice	Equation	Reference
Normalized Difference Vegetation Index (NDVI)	(NIR – Red) / (NIR + Red)	Rouse et al. (1974)
Normalized Difference Water Index (NDWI)	(NIR – SWIR) / (NIR + SWIR)	(Gao, 1996)
Green Normalized Difference Vegetation Index (GNDVI)	(NIR –Green) / (NIR + Green)	Gitelson et al. (1996)
Simple Ratio (SR)	NIR / Red	Jordan (1969)
Non-Linear Index (NLI)	(NIR^2 – Red) / (NIR^2 + Red)	Goel and Qin (1994)

Data Analysis

To estimate the peanut yield, non-linear models were analyzed using Artificial Neural Networks (ANNs) to estimate and predict the yield. The Multilayer Perceptron (MLP) and Radial Basis Function (RBF) networks were tested. ANN's are tools to describe, substantiate and elucidate highly complex issues in the modeling field. Compared to other statistical modeling techniques, ANNs present better performance because they have universal adjustment functions, admit data loss, are nonparametric, and do not require previous information of the phenomenon to be modeled, being applied for several purposes in agriculture. Therefore, ANNs have input layers, neurons, hidden layers, and output layers.

The networks were trained with the 2020-2021 crop data using all the input variables (vegetation indices) to estimate the productivity (output layer) in each area, to generate prediction models and verify which is the best date and which are the best IV's to predict the productivity.

RBF nets have only one hidden layer, and each neuron contains a radial basis activation function. In each neuron the Gaussian or normal function was used as the radial basis function (Bishop 1995), and distance (offset) values from this function increase or decrease the ratio to the center point (Haykin and Principe 1998). As with the MLP neural network, the input values were normalized, and the values in the output layer provided the productivity at each point.

The RBF neural nets were trained using the k-means algorithm (Bishop, 1995). This algorithm attempts to select the optimal set of points that are placed at the centroids in the patterns in the training data.

Database for training and validation

For model training and testing, the database was divided into 80% for training and 20% for testing. During the training phase the ANNs were trained 1000 times, with the network selecting the 10 best models (5 MLP and 5 RBF). The procedures for training and testing the neural models were implemented in the Neural Networks package of the Statistica data analysis software (Statistica 7.0, Statsoft Inc, Tulsa).

The efficiency of the networks was analyzed by means of graphs as a function of accuracy, obtained from the error of the predictions by the MAE (Mean Absolute Error), demonstrating through these calculations the reliability of the data obtained from the variable predictions.

Results

The average productivity of the study areas is shown in Figure 2. The highest yields were observed in Frutal and Capão, which are areas with characteristics of more clayey soils. However, just the fact that the areas have more clayey soils cannot be considered a determining factor for obtaining high yields, considering that Santa Adélia has soil characteristics similar to the areas of high productivity, but had lower productivity than these areas. Another issue is that the less clayey areas, such as Granja and Santa Gertrudes, also showed similar productivity averages to Frutal and Capão.

The Ibitinga area stood out negatively in relation to the other areas, presenting low productivity, being, curiously, the area with the lowest percentage of clay (12%). The best conditions for high yields do not depend exclusively on the type of soil, but on the appropriate management throughout the cycle. It is known that the peanut plant needs lighter soils (sandy and sandy loam) for better development of the crop, but, however, this study shows that it is possible to obtain high yields in heavier soils (clayey). Another important factor to be highlighted is the high variability between areas. Analyzing areas with distinct characteristics and management is important to generate more robust estimation models.

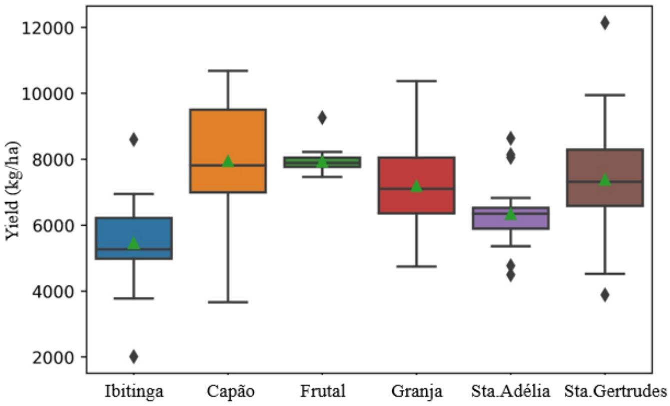
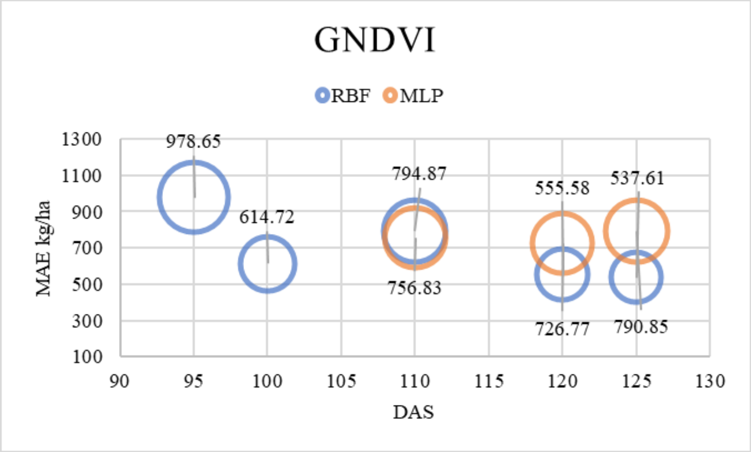


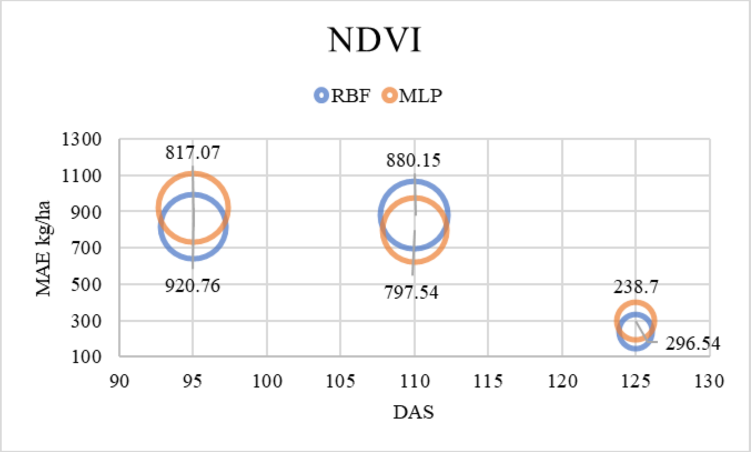
Figure 2. Descriptive statistics with BoxPlot, of the productivity of the study areas.

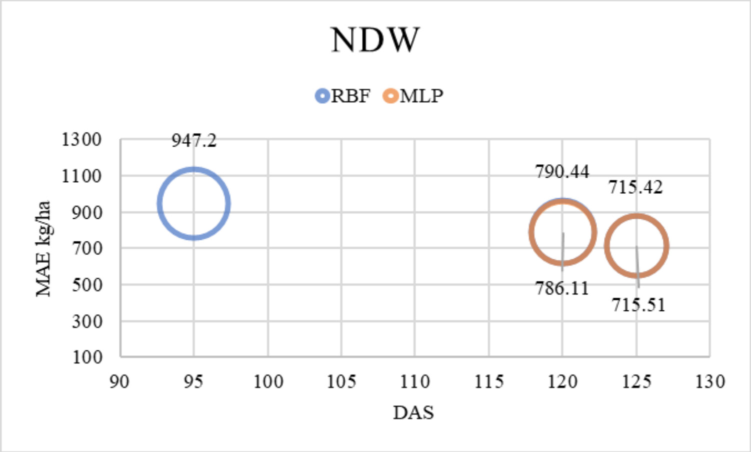
The estimation errors presented in figure 3, show the models created on each date (95, 100, 110,115,120 and 125 DAS), however, for some Vegetation Indices (IV's) some dates were discarded, because models with errors above 1000 kg/ha, were considered inadequate models. Thus, the only index that was able to estimate yields on the five dates evaluated was the GNDVI.

Regarding the results acquired by each algorithm (RBF and MLP), we highlight the RBF algorithm, with 18 models with errors below 1000 kg/ha, while the MLP network presented 12 models. Still highlighting the efficiency of neural networks, it can be seen that the RBF algorithm showed the lowest estimation errors, especially for the NDVI index at 125 DAS, which obtained an error of 238.7 kg/ha.

The best date to perform estimation was at 125 DAS, i.e. one day before the first stage of peanut harvest. All indices and the two algorithms were able to estimate at 125 DAS with NDVI presenting MAE of 238.7 and 296.54 for the RBF and MLP networks, respectively. It can be observed that the more distant is the date of estimation the lower will be the error of the models. At 115 DAS it was not possible to estimate peanut yield within the optimal estimate value (<1000 kg/ha) with any of the neural networks, while at 100 DAS, only GNDVI and SR were able to estimate yield with errors of 614.72 and 995.87 kg/ha with RBF and MLP networks, respectively.


GNDVI
RBF
MLP
MAE kg/ha
1300
1100
900
700
500
300
100
90
95
100
105
110
115
120
125
130
DAS
978.65
614.72
794.87
756.83
555.58
726.77
537.61
790.85


NDVI
RBF
MLP
MAE kg/ha
1300
1100
900
700
500
300
100
90
95
100
105
110
115
120
125
130
DAS
817.07
920.76
880.15
797.54
238.7
296.54


NDW
RBF
MLP
MAE kg/ha
1300
1100
900
700
500
300
100
90
95
100
105
110
115
120
125
130
DAS
947.2
790.44
786.11
715.42
715.51

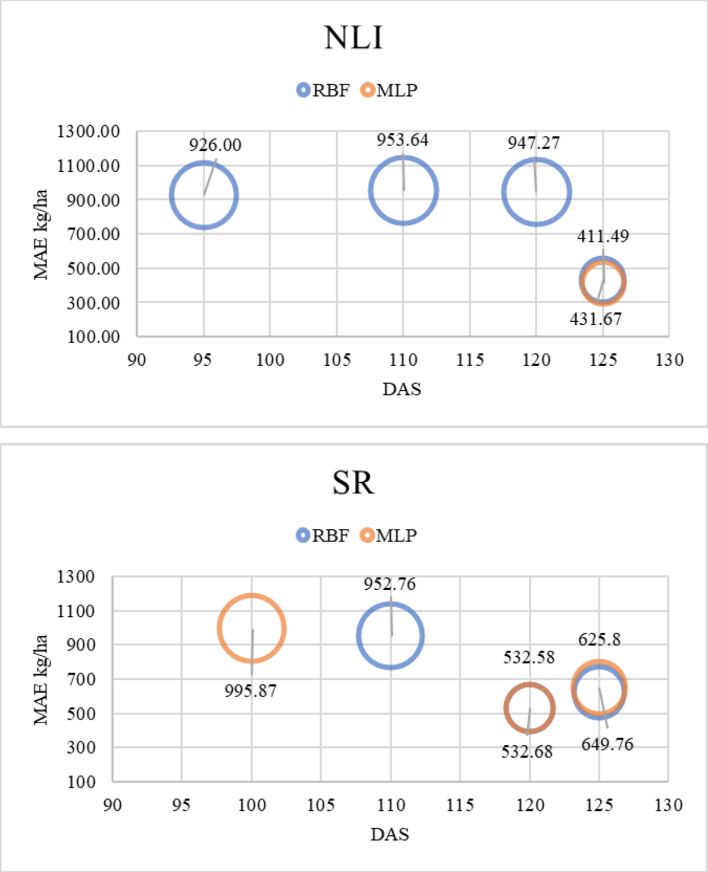


Figure 3. Plots of the mean absolute error (MAE) of peanut yield estimates at 95, 100,115, 120 and 125 days after sowing (DAS).

Because of the distinct characteristics of the peanut culture, evaluations under different conditions are ideal to create more robust models. Therefore, the prediction performed in this study allows to generate good expectations regarding the models developed, considering that the areas presented heterogeneous characteristics of soil, management and climatic conditions. It is important to emphasize that these results are promising, but further studies are needed for the method to become robust, in order to validate the best models taking into account characteristics of other regions. The first results are encouraging, however, at the moment they are not sufficient for these models to be applied throughout the state of São Paulo or even in other countries.

Discussion

Some studies on productivity show that there is a strong relationship between the curves of time series of vegetation indices such as NDVI and crop growth, with satisfactory results that correlate the VI's with the productivity of a given crop (Zhang et al., 2022). The use of remote sensing techniques combined with machine learning algorithms has been widely used to estimate agronomic parameters, such as the productivity of crops like alfalfa, pastures, and soybean (Feng et al., 2020, Rocha et al., 2018 and Yoosefzadeh-Najafabadi et al., 2021). The peanut crop presents distinct characteristics from other major crops, such as the indeterminate growth habit and the development of fruits below ground. These characteristics make difficult the analysis of important parameters of the peanut crop, such as the optimal harvest time, which is directly related to pod maturity index

(PMI). Studies indicate that the use of SR and ANN tools are strongly capable of predicting important variables of the peanut culture, such as the MIP, helping the producer in the assetivity of the ideal point of harvest, which reduces the quantitative and qualitative losses (Souza et al., 2022). For yields this estimate becomes even more difficult due to the high variability of this variable in the production fields. Anyway, the results found in this study are promising, considering that the current method to estimate yields in peanut production fields in the state of São Paulo is the visual estimation method.

Conclusion

This paper presented the first studies of peanut yield estimation using Remote Sensing and Artificial Intelligence. The RBF and MLP algorithms were able to estimate peanut yields with errors below 1000 kg/ha. The GNDVI estimated yields on all dates analyzed in this study. The best date for peanut yield estimation is at 125 DAS.

References

Joshi, Vijaya R. et al. In-season weather data provide reliable yield estimates of maize and soybean in the US central Corn Belt. International Journal Of Biometeorology, [S.L.], v. 65, n. 4, p. 489-502, 21 nov. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00484-020-02039-z.

Souza, J. B. C., de Almeida, S. L. H., Freire de Oliveira, M., Santos, A. F. D., Filho, A. L. D. B., Meneses, M. D., & Silva, R. P. D. (2022). Integrating Satellite and UAV Data to Predict Peanut Maturity upon Artificial Neural Networks. Agronomy, 12(7), 1512.

Santos, A. F., Lacerda, L. N., Rossi, C., Moreno, L. D. A., Oliveira, M. F., Pilon, C., ... & Vellidis, G. (2022). Using UAV and Multispectral Images to Estimate Peanut Maturity Variability on Irrigated and Rainfed Fields Applying Linear Models and Artificial Neural Networks. Remote Sensing, 14(1), 93.

Sentinel2UserHandbook.Availableonline:https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook (Acessado em 06/04/2022).

Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.

Haykin, S., & Principe, J. (1998). Making sense of a complex world [chaotic events modeling]. IEEE Signal Processing Magazine, 15(3), 66-81.

Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., & Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. Frontiers in plant science, 11, 624273.

Zhang, J., Tian, H., Wang, P., Tansey, K., Zhang, S., & Li, H. (2022). Improving wheat yield estimates using data augmentation models and remotely sensed biophysical indices within deep neural networks in the Guanzhong Plain, PR China. Computers and Electronics in Agriculture, 192, 106616.

Feng, L., Zhang, Z., Ma, Y., Du, Q., Williams, P., Drewry, J., & Luck, B. (2020). Alfalfa yield prediction using UAV-based hyperspectral imagery and ensemble learning. Remote Sensing, 12(12), 2028.

Rocha, A. D., Groen, T. A., Skidmore, A. K., Darvishzadeh, R., & Willemen, L. (2018). Machine learning using hyperspectral data inaccurately predicts plant traits under spatial dependency. Remote sensing, 10(8), 1263.