Advancing Peanut Yield and Maturity Estimation through Multi-Target Regression and Remote Sensing

J. B. C Souza[1], J. L. S. Ferreira[3], S. L. H. Almeida[1], M. F. Oliveira[2], S. S. Washington[3] and R. P. Silva[1]
[1] São Paulo State University "Júlio de Mesquita Filho" – Jaboticabal – SP, Brazil.
[2] Extension Educator, Nebraska Extension in Dodge County – Nebraska, USA.
[3] Maranhão Federal University – MA, Brazil.

Abstract

Peanut harvesting poses significant challenges due to the difficulty in estimating yield and maturity directly in the field. This study aimed to estimate both variables simultaneously using remote sensing and multi-target regression. The experiment was conducted across 11 commercial areas, utilizing PlanetScope satellite imagery to calculate seven vegetation indices. The multi-target regression approach employed the Random Forest algorithm. Results showed high accuracy, with a mean absolute error of 0.07 for maturity and 685.32 kg ha^{-1} for yield, highlighting the model's robustness for smart peanut harvesting applications.

Introduction

Accurate estimates of yield and maturity are essential for optimizing harvest management in low-technology crops such as peanuts. However, the spatial and temporal variability of these variables poses significant challenges to conventional methods, which often lack scalability and representativeness. Furthermore, specific characteristics of peanuts, such as their indeterminate growth and geocarpic fruit development, further complicate the application of traditional approaches.

Yield determination can be performed using manual methods, which are labor-intensive and imprecise, or through sensors mounted on harvesters (Barbosa Júnior et al., 2024; Shahi et al., 2023). However, both approaches have limitations: manual methods require substantial labor, while crop-specific sensors for peanuts are not yet widely available. For maturity assessment, the traditionally employed method is the maturity profile board, which identifies the optimal harvest time based on the exocarp coloration (Williams & Drexler, 1981). Nevertheless, this method is subjective and lacks representativeness, particularly in large-scale production systems.

In recent years, digital technologies such as Remote Sensing (RS) and Artificial Intelligence (AI) have been explored to overcome these limitations. Souza et al. (2022) utilized satellite data and Remotely Piloted Aircraft Systems (RPAS) in combination with Artificial Neural Networks (ANN) to estimate peanut maturity. Oliveira et al. (2024) further advanced this approach by simultaneously estimating maturity and biomass using machine learning (ML) algorithms. However, the literature on yield estimation remains limited, with Shahi et al. (2023) presenting the only study that employed RPAs to estimate peanut yield.

Despite recent advances, there is a lack of studies that simultaneously integrate the prediction of yield and maturity for peanuts, a critical step toward enabling smart harvesting systems. Furthermore, the robustness and consistency of existing models are limited by the absence of data from multiple growing seasons and diverse management practices (Shahi et al., 2023).

Addressing this gap, this study hypothesizes that combining vegetation indices derived from satellite imagery with multi-target algorithms, such as Random Forest (RF), can provide accurate estimates of yield and maturity by incorporating genotype-environment interactions. Therefore, the objective of this work is to develop a multi-target regression model to simultaneously estimate peanut yield and maturity, using vegetation indices extracted from high-resolution satellite images.

Material and Methods

The experiments were conducted in 11 commercial areas dedicated to peanut cultivation following sugarcane, located in the regions of Ribeirão Preto and Araraquara, in the state of São Paulo, Brazil, over two consecutive growing seasons (2020/2021 and 2021/2022). These areas were selected based on their specific characteristics, including soil types and different peanut cultivars (Table 1).

Table 1. Information about the study areas for the two growing seasons.

Area	Location	Cultivar	Soil	Days after sowing (DAS)	Cycle (days)
			Season 2020/2021		
1	Ibitinga	IAC OL3	Dystrophic Red Latosol (Oxisol)	120 DAS	120
2	Pradópolis	IAC OL3	Dystrophic Red Latosol (Oxisol)	132 DAS	120
3	Santa Ernestina 1	IAC OL3	Dystrophic Red Latosol (Oxisol)	115 DAS	120
4	Taiuva 1	IAC OL3	Alfisol	115 DAS	120
5	Monte Alto 1	IAC OL3	Alfisol	117 DAS	120
6	Monte Alto 2	IAC OL3	Alfisol	115 DAS	120
			Season 2021/2022		
7	Taiuva 2	IAC OL3	Alfisol	122 DAS	120
8	Santa Ernestina 2	IAC OL3	Alfisol	118 DAS	120
9	Guariba	IAC 503	Alfisol	145 DAS	150
10	Jaboticabal	IAC 503	Alfisol	145 DAS	150
11	Taquaritinga	IAC 503	Alfisol	148 DAS	150

In the 2020/2021 growing season, the experiment was conducted in areas with 20 sampling points, uniformly distributed every 50 meters in a strip system. In the subsequent season, 2021/2022, the number of sampling points increased to 30, while maintaining a 50-meter distance between points (Fig. 1).

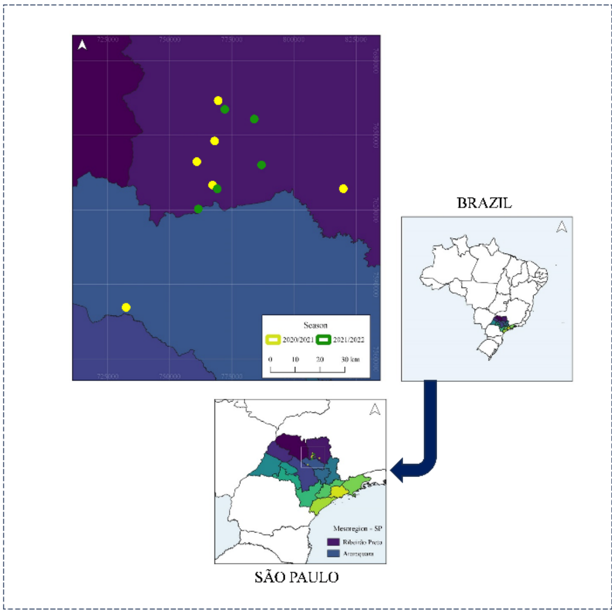


Fig 1. Location of the experimental areas for the 2020/2021 and 2021/2022 growing seasons.

The field sampling points were georeferenced using a multirotor Remotely Piloted Aircraft (RPA), model DJI P4 Multispectral (Shenzhen, China), which was operated automatically through the DJI Ground Station Pro application. Control points were marked with lime in all areas to enhance georeferencing accuracy. After processing, the geographic coordinates were extracted from the Orthomosaic and linked to the corresponding satellite images.

The determination of peanut yield and maturity was conducted one day prior to mechanical uprooting. For yield assessment, all plants within a 2 m² frame at each central sampling point were manually collected in each area. The harvested material was then manually threshed and cleaned. Subsequently, the samples were placed in an oven at 105 °C for 24 hours until reaching a constant mass to determine the dry weight of the pods.

For the maturity analysis, eight plants were collected from each plot. All fruits present on the sampled plants were removed, resulting in an average of 150 to 250 pods per sampling point. Subsequently, all pods underwent the exocarp removal process using a high-pressure washer, following the Hull-Scrape method. After the exocarp was removed, the pods were organized on the maturity classification board using the color column to calculate the Peanut Maturity Index.

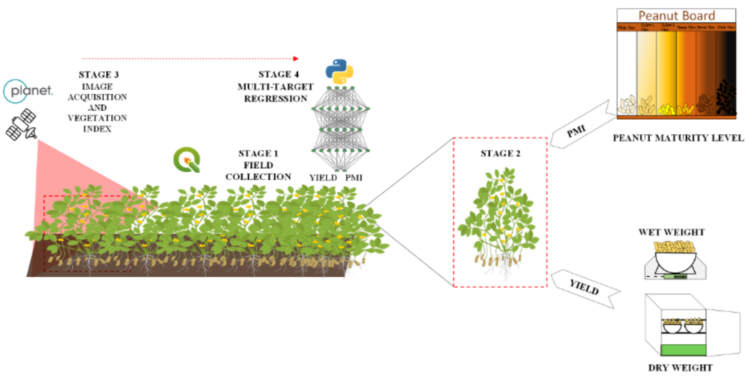


Fig 2. Representative schematic of the methodology.

Satellite Image acquisition

The images provided by the CubeSat PlanetScope (PS) sensor were collected to extract spectral information from the study area and calculate seven vegetation indices (Table 2), which served as predictive variables for the model. Four spectral bands were used for this study: Blue, Green, Red, and Near-Infrared. The images were corrected to surface reflectance, ensuring consistency under local atmospheric conditions and accuracy in spectral response. This process utilized specific coefficients and Look-Up Tables (LUTs) to adjust reflectance from TOA (Top of Atmosphere) to BOA (Bottom of Atmosphere).

Table 2. Vegetation indices used as inputs in the predictive models.

IV *	Equation	Reference
NDVI	$\frac{NIR - Red}{NIR + Red}$	Rouse et al., (1974)
GNDVI	$\frac{NIR - Green}{NIR + Green}$	Gitelson e Merzlyak., (1996)
MNLI	$\frac{(NIR^2 - Red)(1 + L **)}{(NIR^2 + Red + L **)}$	Gong et al., (2003)
EVI	$\frac{2.5(NIR - Red)}{(L ** + NIR + C1Red - C2Blue)}$	Justice et al., (1998)
NLI	$\frac{NIR^2 - Red}{NIR^2 + Red}$	Goel and Qin., (1994)
SR	$\frac{NIR}{Red}$	Jordan., (1969)
SAVI	$\frac{(1 + L **)(NIR - Red)}{(L ** + NIR + Red)}$	Huete., (1988)

NDVI: Normalized Difference Vegetation Index; GNDVI: Green Normalized Difference Vegetation Index; MNLI: Modified Non-Linear Index; EVI: Enhanced Vegetation Index; NLI: Non-Linear Index; SAVI: Soil-Adjusted Vegetation Index; SR: Simple Ratio; *IV = Vegetation Index**L = 0,5; C1 = 6; C2 = 7,5.

Data analysis

Boxplots were generated for both yield and PMI across all models applied to the different prediction dates. This visualization provided a clear and direct comparison between field-observed data and the predictions generated by the algorithms, enabling a straightforward assessment of the accuracy and consistency of the predictive models.

Multi-Target Regression (MTR) was employed to estimate yield and PMI simultaneously. In this context, the Random Forest (RF) algorithm was used to predict both variables. Model training (80% of the data) and testing (20% of the data) were performed using tools provided by the Scikit-learn package (Pedregosa et al., 2011). The datasets were normalized to a range of 0 to 1 using the StandardScaler method, and hyperparameter optimization was conducted with GridSearchCV (Table 3). Cross-validation was performed with ten iterations, ensuring an adequate number of passes to mitigate the risk of overfitting (Duan et al., 2014).

The efficiency of the algorithms was evaluated using 1:1 performance plots, which were employed to visualize the accuracy of the models based on the calculation of the Mean Absolute Error (MAE), as presented in Equation (1).

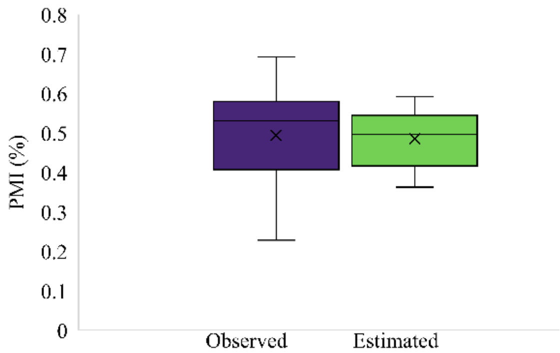
$$MAE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_i) \tag{1}$$

Where, n is the number of data points, Y_i represents the observed variable values in the field, and $\bar{Y}_i$ represents the predicted variable values.

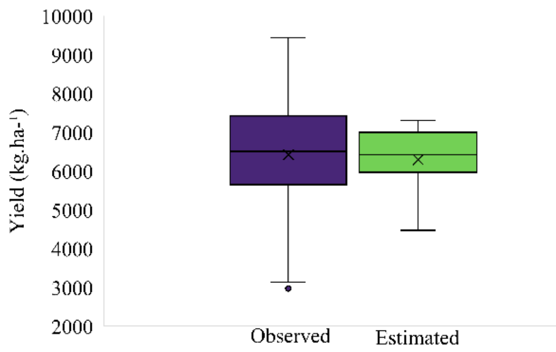
Results

Figure 3 presents the results of the RF model for peanut yield and maturity, evaluated through boxplots and residual plots. The boxplots demonstrate that the model's estimates have means and medians close to the observed field values for both variables, suggesting that the algorithm effectively captures the general trends in the data. However, the range of estimated values is narrower than that of the observed values, indicating the model's difficulty in representing the actual spatial variability of the data, which reflects reduced precision.

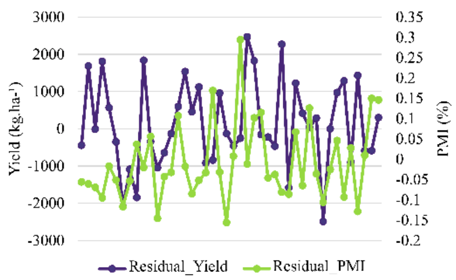
In the residual plot, distinct behavior is observed between the analyzed variables. For yield, the model tends to overestimate the observed values, particularly in the higher yield ranges. Conversely, for PMI, there is a clear tendency toward underestimation, suggesting that the model struggles to accurately capture the extremes of the distribution. These error patterns indicate that, while the model is robust in predicting average values, adjustments are needed to enhance its ability to address the intrinsic variability of the peanut production system.



(a)



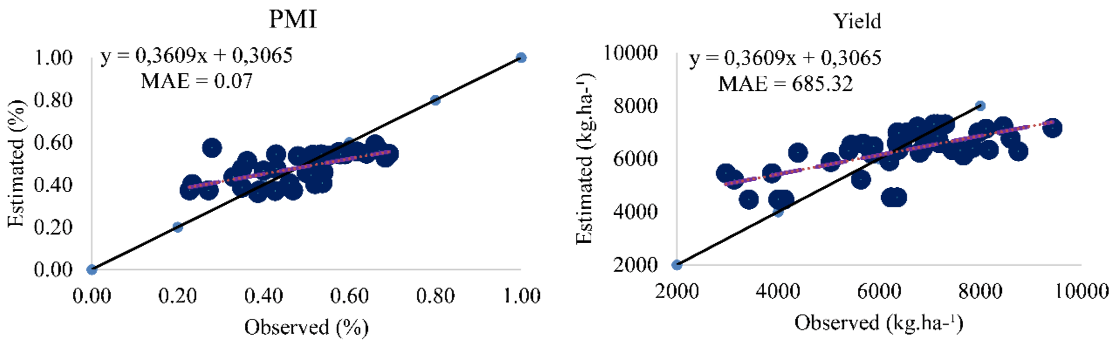
(b)



(c)

Fig 3. Results of the RF model for peanut yield (a) and maturity (b) and residual plot (c).

The performance plot (Figure 4) illustrates the errors obtained from the multi-target regression using the RF algorithm. For PMI, the mean absolute error (MAE) was 0.07%, while for yield, it was 685.32 kg ha^{-1}. The data points are distributed around the trend line, demonstrating a satisfactory fit for both estimated targets. However, deviations along the trend line suggest discrepancies, likely stemming from the high variability within the dataset, particularly for yield. Notably, PMI achieved the best result, attributed to its more consistent fit.



(a) (b)

Fig 4. Performance plot of the models estimated for PMI (a) and Yield (b) using the RF algorithm.

Discussion

Our study is pioneering in utilizing multi-target regression (MTR) to simultaneously estimate peanut maturity and yield exclusively from vegetation indices. One of the main challenges in managing this crop is identifying the optimal harvest time, complicated by the physiological characteristics of peanuts, which exhibit indeterminate growth and lack natural plant senescence (Souza et al., 2022). In this context, remote sensing methods emerge as a promising approach to overcome these limitations, providing non-destructive and widely applicable solutions. Previous studies, such as that of dos Santos et al. (2021), demonstrated significant correlations between peanut maturity and vegetation indices, highlighting an inverse relationship: as maturity increases, vegetation index values decrease. These findings underscore the ability of vegetation indices to capture the spatial and temporal variability of maturity at the field scale, making them effective tools for precision agriculture.

In the case of peanut yield, no studies have yet explored the use of vegetation indices (VIs) for estimating or predicting crop productivity. However, research conducted on other crops has shown that VIs obtained from various acquisition sources, such as satellites and drones, can significantly enhance the performance of predictive yield models. It is important to emphasize that the effectiveness of VIs relies on their ability to capture and integrate relevant information about the spatial and temporal variability of the crop, ensuring greater representativeness in the models (Amaral et al., 2024). This approach could potentially be adapted for peanuts, addressing a critical gap in the current literature. In this context, the VIs utilized in this study have already demonstrated significant potential in identifying temporal and spatial variability, with satisfactory correlations with biomass and PMI. These findings suggest that VIs are a promising alternative for exploring peanut yield prediction, offering a non-destructive, scalable approach capable of capturing patterns related to crop productivity (Costa Souza et al., 2023).

In previous studies, Oliveira et al. (2024) suggested that integrating multi-target regression (MTR) with remote sensing data is a promising approach for simultaneously estimating biomass and maturity. However, the best results reported by the authors were achieved using vegetation indices (VIs) as predictive variables with the KNN algorithm. In comparison, the results of this study demonstrate higher accuracy, particularly in terms of errors. While Oliveira et al. (2024) reported a minimum error of 0.09 for maturity, this study achieved an error of only 0.07 using the Random Forest (RF) algorithm. This superiority can be attributed to the more robust genotype-environment interaction explored in this study, leading to more precise and consistent estimates. Similarly, Tedesco et al. (2021) employed multi-target regression (MTR) with the RF and KNN algorithms to predict sweet potato yield. The authors found that RF outperformed KNN, attributing this advantage to RF's ability to mitigate the issue of overfitting to training data. RF's unique feature of combining multiple decision trees to generate more robust predictions is considered critical for achieving reliable results in yield estimation.

Conclusion

This study demonstrated that the multi-target regression model based on the RF algorithm, utilizing vegetation indices derived from high-resolution satellite imagery, was effective in simultaneously estimating peanut yield and maturity. The results indicated that the model exhibited overall good performance, with a mean absolute error (MAE) of 0.07% for PMI and 685.32 kg ha^{-1} for yield. While PMI was estimated with higher accuracy, reflecting a more consistent fit to field-level variations, the algorithm struggled to capture the variability in yield. This suggests the need for further refinement in the modeling approach for this variable.

The identified limitations, particularly the model's difficulty in capturing the high variability associated with yield, highlight the importance of integrating new data sources and methodologies to enhance the robustness of the estimates. The inclusion of additional explanatory variables, such as soil properties and meteorological data, represents promising avenues for future research.

These results highlight the potential of using remote sensing (RS) and machine learning (ML) technologies in peanut cultivation, contributing to the development of tools that optimize management and harvesting practices. As future steps, it is recommended to generalize the model across different seasonal conditions and management systems, validating its applicability on larger scales and in diverse contexts. This would enhance its predictive capability and expand its use in agricultural decision-making and smart harvesting.

Acknowledgements

This Project was funded by the São Paulo Research Foundation (FAPESP - grants no. 2023/14041-2, and 2021/06029-7 and the National Council for Scientific and Technological Development (CNPq) for their financial support.

References

Amaral, L. R., Oldoni, H., Baptista, G. M. M., Ferreira, G. H. S., Freitas, R. G., Martins, C. L., Cunha, I. A., & Santos, A. F. (2024). Remote sensing imagery to predict soybean yield: a case study of vegetation indices contribution. *Precision Agriculture*, *25*(5), 2375–2393. https://doi.org/10.1007/s11119-024-10174-5

Barbosa Júnior, M. R., Moreira, B. R. de A., Carreira, V. dos S., Brito Filho, A. L. de, Trentin, C., Souza, F. L. P. de, Tedesco, D., Setiyono, T., Flores, J. P., Ampatzidis, Y., Silva, R. P. da, & Shiratsuchi, L. S. (2024). Precision agriculture in the United States: A comprehensive meta-review inspiring further research, innovation, and adoption. *Computers and Electronics in Agriculture*, *221*, 108993. https://doi.org/10.1016/j.compag.2024.108993

Costa Souza, J. B., Luns Hatum de Almeida, S., Lopes de Brito Filho, A., Morlin Carneiro, F., Santos, A. F. dos, & da Silva, R. P. (2023). Unmanned aerial system and satellite:

Which one is a better platform for monitoring of the peanut crops? *Agronomy Journal*, *115*(3), 1146–1160. https://doi.org/10.1002/agj2.21310

dos Santos, A. F., Corrêa, L. N., Lacerda, L. N., Tedesco-Oliveira, D., Pilon, C., Vellidis, G., & da Silva, R. P. (2021). High-resolution satellite image to predict peanut maturity variability in commercial fields. *Precision Agriculture*, *22*(5), 1464–1478. https://doi.org/10.1007/s11119-021-09791-1

Oliveira, M. F., Carneiro, F. M., Ortiz, B. V., Thurmond, M., Oliveira, L. P., Bao, Y., Sanz-Saez, A., & Tedesco, D. (2024). Predicting below and above-ground peanut biomass and maturity using multi-target regression. *Computers and Electronics in Agriculture*, *218*. https://doi.org/10.1016/j.compag.2024.108647

Shahi, T. B., Xu, C.-Y., Neupane, A., Fleischfresser, D. B., O'Connor, D. J., Wright, G. C., & Guo, W. (2023). Peanut yield prediction with UAV multispectral imagery using a cooperative machine learning approach. *Electronic Research Archive*, *31*(6), 3343–3361. https://doi.org/10.3934/ERA.2023169

Souza, J. B. C., de Almeida, S. L. H., Freire de Oliveira, M., Dos Santos, A. F., 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). https://doi.org/10.3390/agronomy12071512

Tedesco, D., Almeida Moreira, B. R. de, Barbosa Júnior, M. R., Papa, J. P., & Silva, R. P. da. (2021). Predicting on multi-target regression for the yield of sweet potato by the market class of its roots upon vegetation indices. *Computers and Electronics in Agriculture*, *191*, 106544. https://doi.org/10.1016/j.compag.2021.106544

Williams, E. J., & Drexler, J. S. (1981). A Non-Destructive Method for Determining Peanut Pod Maturity. *Peanut Science*, *8*(2), 134–141. https://doi.org/10.3146/i0095-3679-8-2-15