What are the most relevant variables for estimating the remote maturity of peanut pods?

Thiago Caio Moura Oliveira [1], Jarlyson Brunno Costa Souza [2], Samira Luns Hatum de Almeida[3], Armando Lopes de Brito Filho[4], Rouverson Pereira da Silva[5]

[1] Department of Engineering, São Paulo, State University (UNESP), Jaboticabal, São Paulo, Brazil, - thiago.caio@unesp.br
[2] Department of Engineering, São Paulo, State University (UNESP), Jaboticabal, São Paulo, Brazil, - jarlyson.brunno@unesp.br
[3] Department of Engineering, São Paulo, State University (UNESP), Jaboticabal, São Paulo, Brazil, - samira.lh.almeida@unesp.br
[4] Department of Engineering, São Paulo, State University (UNESP), Jaboticabal, São Paulo, Brazil, - armando.brito@unesp.br
[5] Department of Engineering, São Paulo, State University (UNESP), Jaboticabal, São Paulo, Brazil, - rouverson.silva@unesp.br

Keywords: Satellite, Artificial Intelligence, *Arachis hypogaea* L., Maturation, Digital Agriculture.

ABSTRACT

The application of geotechnology for the accurate estimation of peanut maturity is crucial to improve management, facilitate monitoring, and understand crop variability. This study aimed to analyze and select the most important variables for maturity estimation and to evaluate the performance of a model using the selected variables, through multiple linear regression (MLR), a model with a simpler structure. The research was conducted in a commercial field during the 2022/2023 growing season, using the IAC 503 cultivar with a 150-day cycle. Forty sampling points were collected, starting 28 days before harvest and continuing until one day prior to harvest, totaling five evaluations. Maturity was determined using the Hull-Scrape method. Orbital images were obtained from the PlanetScope platform, with 3-meter spatial resolution and daily temporal coverage. Five cloud-free images were selected, corresponding to the field sampling dates. For variable analysis and selection, dimensionality reduction was performed through principal component analysis (PCA), followed by the Stepwise method to choose the most relevant variables. Maturity index estimation was then carried out using multiple linear regression (MLR). Despite its simplicity, the model achieved high performance, with $R^2 = 0.93$. The careful selection process identified GDD, the green band, NDVI, and SAVI as the most influential variables for maturity. The results indicate that a small number of well-chosen variables are sufficient to accurately estimate peanut maturity.

1. INTRODUCTION

Peanut (*Arachis hypogaea* L.) is a globally distributed crop, with China, India, and Nigeria standing out as major producers, while in Brazil, the state of São Paulo accounts for most of the national production, driven by the use of peanuts in rotation with sugarcane (USDA, 2024). The species offers significant agronomic benefits, such as the ability to fix nitrogen in the soil and improve physical soil structure through its root system (Paungfoo et al., 2017). However, peanut cultivation faces challenges related to pod maturity variability, which is intensified by climatic factors and the plant's indeterminate growth habit (Zerbato et al., 2017). This variability hinders the determination of the optimal harvest time, especially under adverse conditions such as high humidity or thermal stress (Santos et al., 2021).

Harvesting occurs in two stages digging and picking and should begin when approximately 70% to 75% of the pods are mature, based on a visual assessment of mesocarp coloration. Harvesting outside this window can lead to yield and quality losses, exacerbated by the resistance of the gynophores, whose rupture depends on factors such as soil moisture, pod residence time, and flowering stage (Santos et al., 2019). Given the plant's indeterminate growth habit, accurately identifying pod maturity is essential for determining the optimal harvest time, enabling maximum yield and preserving fruit quality.

Given the uncertainties regarding the optimal harvest time, various methods have been developed to estimate peanut maturity and minimize losses. The most widely adopted visual method was proposed by Williams and Drexler (1981), which classifies pod development stages based on mesocarp coloration, ranging from white (immature) to black (fully mature). Despite its widespread acceptance, this method is manual, requires pod scraping, and depends on the evaluator's experience, which can lead to inconsistencies in classification.

As an alternative, the use of geotechnologies has gained prominence in objectively and efficiently identifying the maturity stage of peanuts. Among these technologies, remote sensing (RS) whether orbital, aerial, or ground based has been employed as a promising tool for non-destructive estimates, helping to reduce evaluation subjectivity and support decision-making regarding the optimal harvest time (Santos et al., 2021; Souza et al., 2022).

The integration of remote sensing (RS) and machine learning (ML) algorithms has emerged as a promising alternative for estimating peanut maturity. Although the use of vegetation indices and climatic variables in estimation models is already common, the rationale for selecting these variables is still seldom discussed in the literature. The combined use of ML and geotechnologies has proven effective in estimating agronomic parameters, including maturity, through non-destructive approaches with strong generalization capabilities (Souza et al., 2022). In the study by Oliveira et al. (2024), the application of ML with high spatial resolution satellite imagery enabled the simultaneous estimation of peanut maturity and biomass with satisfactory accuracy. These advances underscore the potential of integrating RS and ML as a robust tool for phenological monitoring, particularly for maturity estimation, contributing to improved crop management.

In this context, the application of RS and ML enables the estimation of peanut maturity and the selection of the most promising variables for the development of future estimation models. Therefore, this study aims to analyze and select the most important input variables for maturity estimation and, subsequently, to evaluate the performance of multiple linear regression, a model with a simplified structure, in estimating peanut maturity.

2. MATERIAL AND METHODS

2.1 Experimental Area

The experiment was conducted in a commercial peanut field of approximately 9 hectares (ha) during the 2022/2023 growing season, located in the state of São Paulo, Brazil. Sowing was carried out on October 25, 2022, using the IAC 503 cultivar, with a growth cycle of 150 days. The soil of the region is classified as a Red Latosol according to the Brazilian Soil Classification System (EMBRAPA, 2013).

2.2 Field Data Collection

A total of 40 sampling points were collected at 50 m intervals, distributed to represent the spatial variability of peanut maturity. Each point corresponded to 0.5 ha. Sampling at each point began at 122 days after sowing (DAS), with intervals of approximately seven days, up to 150 DAS, totaling five evaluations (122, 129, 136, 144, and 150 DAS). At each sampling point, 8 plants were collected, totaling 150 to 250 pods per point.

To determine maturity, the pods were evaluated using the Peanut Maturity Index (PMI), as described by Rowland et al. (2006), which applies the Hull-Scrape method. This method consists of scraping the pods to identify mesocarp coloration and visually classify them according to the peanut maturity chart (Figure 1) proposed by Williams and Drexler (1981). For PMI calculation, the ratio between the number of pods in the brown and black classes and the total number of pods evaluated is considered (Equation 1). The PMI ranges from 0 to 1, with values between 0.70 and 0.75 indicating the optimal harvest point.

Figure 1. Chart for visual classification of pods based on mesocarp coloration.

$$PMI\_BB = \frac{B + BL}{W+Y1+Y2+O+B+BL}, \qquad (1)$$

Where, W: white, Y1: yellow1, Y2: yellow2, O: orange, B: brown, BL: black.

2.3 Orbital Image Acquisition

The orbital images were obtained from the PlanetScope CubeSat constellation (PLANET, 2024), with daily temporal resolution, 3-meter spatial resolution, and 12-bit radiometric resolution. Of the eight available spectral bands, seven were selected (Table 1) for the analyses and the calculation of vegetation indices. Reflectance analyses and the computation of vegetation and topographic indices were performed in QGIS 3.16 (QGIS Development Team, OSGeo). Five cloud-free images were selected, corresponding to the field sampling dates. A 3 m buffer was applied around each sampling point to extract the mean spectral reflectance using zonal statistics.

Band	Name	Central Wavelength (nm)
2	Blue	472
3	Green 1	531
4	Green	565
5	Yellow	680
6	Red	665
7	Red Edge	750
8	NIR	865

Table 1. Spectral bands used to estimate peanut maturity. *nm: nanometer, NIR: Near-Infrared.

2.4 Data Analysis

The input variables for the model included soil texture (percentages of clay, silt, and total sand), seven spectral bands, nine vegetation indices (NDVI, GNDVI, NLI, NDRE, EVI, SAVI, MNLI, LAI, and SR), two topographic indices, the Topographic Position Index (TPI) and the Topographic Wetness Index (TWI), and two climatic variables (Growing Degree Days - GDD and incident solar radiation). Climatic data were obtained from the free NASA POWER platform, which provides daily climatic variables. GDDs were calculated from the maximum and minimum temperatures extracted from NASA POWER, considering the base temperature for peanut production (13.3 °C) (Equation 2). Incident solar radiation (Qg) was obtained by summing the values from sowing to each maturity sampling date. The data provided by NASA POWER have daily temporal resolution and an approximate spatial resolution of 55 km.

$$GDD = \frac{(Tmax + Tmin)}{2} - BT, \qquad (2)$$

Where: Tmax represents the maximum temperature (°C), Tmin the minimum temperature (°C), and BT the base temperature (°C).

Initially, a total of 23 input variables were considered, followed by variable pre-selection and importance analysis. The dataset was standardized, and then Principal Component Analysis (PCA) was applied in an exploratory manner to reduce redundancies among variables and mitigate multicollinearity, assisting in the pre-selection of the most representative ones. The original variables indicated as most relevant by the PCA were then submitted to the Stepwise method, which defined the final set of variables used in the models. For variable selection analysis using the Stepwise method, the best subset regression function from the "olsrr" package in R (version 4.1.0) was employed to identify the crucial variables for PMI estimation.

The dataset was divided into training (80%) and testing (20%) subsets using stratified random sampling, ensuring independence between them. It is noteworthy that no specific spatial validation was applied due to the limited number of sampling points. This limitation was considered in the interpretation of the results, and it is recommended that future studies with greater sampling density employ spatial validation strategies, such as spatial cross-validation or blocking approaches, to increase model robustness.

To estimate maturity, multiple linear regression (MLR) was chosen. Model performance was assessed using statistical metrics to determine accuracy and precision, through the calculation of mean absolute error (MAE) and the coefficient of determination (R^2).

3. RESULTS AND DISCUSSION

3.1 Descriptive Analysis of PMI Over Time

At 136 DAS, greater variability among samples was observed, while at 144 DAS the interquartile ranges expanded, indicating heterogeneity in maturity. As time progressed up to 136 DAS, both the median and variability increased, reflecting the continuous maturation process and differences among samples over time. After 136 DAS, the median values showed a smoother increase and reduced variability, suggesting that maturity was approaching stabilization. The outliers at 150 DAS indicate that, although most samples reached higher maturity levels, there were exceptions that did not follow this trend. The presence of outliers at these stages highlights variability among sampling points, attributed to physiological factors and biotic and abiotic environmental conditions influencing pod development. The indeterminate growth habit and the increase in PMI corroborate the temporal dynamics of maturation, in which genetic factors, such as varietal diversity, and environmental factors, such as water and nutrient availability, play a considerable role in this heterogeneity (Vishnuprabha et al., 2023). The presence of outliers in advanced maturity stages reinforces the importance of rigorous monitoring of maturity indices to optimize harvest timing. Harvesting after the optimal point can compromise pod and seed quality, affecting flavor, texture, and nutritional value (Arioglu et al., 2018). Therefore, adjusting harvest timing is essential to maximize crop quality and yield, ensuring better utilization of agricultural resources.

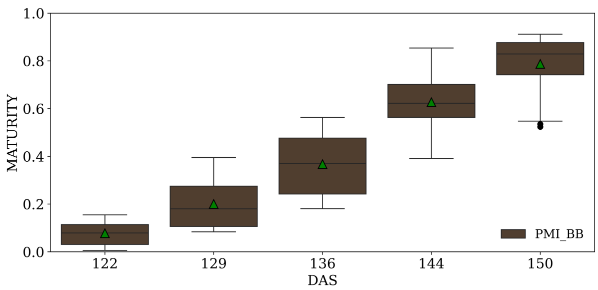

Figure 2. Boxplot of temporal variability of PMI over time in days after sowing.

3.2 Dataset Pre-Selection and Variable Selection Analysis (Stepwise)

To understand the dataset, Principal Component Analysis (PCA) was applied in an exploratory manner, with the aim of identifying redundancies and reducing multicollinearity among variables, thus assisting in the pre-selection of those most representative for PMI estimation. From the initial 23 variables, ten were selected through PCA, including the spectral bands Green, Red, and NIR; the vegetation indices NDVI, GNDVI, MNLI, SAVI, NLI, and NDRE; and the climatic variable GDD. The selection was based on Kaiser's criterion, considering only eigenvalues greater than 1, which indicate greater representativeness in the model. The choice of spectral bands and vegetation indices also considered their strong correlation with maturity, as evidenced in previous studies (Rowland et al., 2008; Oliveira et al., 2024; Souza et al., 2022). The inclusion of GDD was justified by its robust correlation with PMI, reflecting the influence of accumulated temperature on peanut development. Previous research has demonstrated the effectiveness of GDD in estimating maturity, highlighting its role as a reliable climatic indicator that captures essential factors for pod growth and maturation (Santos et al., 2019; Souza et al., 2022; Barboza et al., 2024).

After the PCA pre-selection step, the input variables were reduced to ten. Subsequently, an importance analysis was conducted using the Stepwise method to identify the most effective variables for PMI estimation. The results showed that the coefficient of determination (R^2) increased as more variables were incorporated, until reaching an optimal point at which model accuracy was maximized and the Mean Squared Error of Prediction (MSEP) minimized (Figure 3). The four variables with the best performance in terms of R^2 and MSEP were selected, as they provided greater accuracy and precision compared to the others. This choice aimed to optimize maturity estimation using a reduced set of variables, considering that adding more than four variables would not alter the model results. Reducing the number of variables decreases computational complexity and processing time, thereby enhancing the model's practical applicability, since using all variables would make the process considerably more time-consuming.

On the horizontal axis of Figure 3, each number corresponds to the size of the variable subset selected by the Stepwise method. For example, the subset with three variables consisted of Band 8, GNDVI, and GDD, while the subset with four variables included Band 4, NDVI, SAVI, and GDD. The subsequent subsets were formed by the progressive addition of new variables (such as Bands 6 and 8, MNLI, NDRE, and NLI), until reaching a total of ten variables. This detail allows for understanding how model performance (R^2 and MSEP) varied as a function of the tested variable combinations.

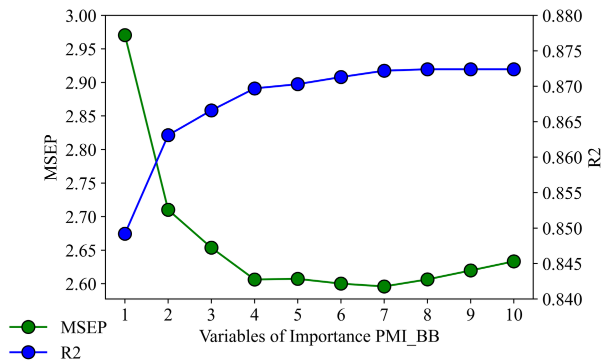

Figure 3. Best regression performance of the selected variables for estimating peanut maturity.

Table 2 presents the four best variables selected to estimate maturity, following the ranking method from the Stepwise importance analysis.

Output variables	Input variables	R^2	MSEP
PMI_BB	Green band; NDVI; SAVI; GDD	0,87	2,61

Table 2. Best variables for peanut maturity (PMI) estimation. PMI_BB: maturity, NDVI: Normalized Difference Vegetation Index, SAVI: Soil-Adjusted Vegetation Index, GDD: Growing Degree Days.

The variable importance analysis (Figure 3) shows that including up to four variables for PMI considerably improves the coefficient of determination (R^2) and reduces the Mean Squared Error of Prediction (MSEP). Beyond this point, the addition of more variables does not result in substantial improvements. This finding is consistent with the variable selection approaches discussed by Chea et al. (2020) and Tamura et al. (2017), in which optimizing the number of variables used enhances model robustness.

3.3 Model Performance (MLR) in Peanut Maturity (PMI) Estimation

To evaluate the performance of the four input variables, multiple linear regression (MLR) was applied to estimate maturity. When the variables were introduced, the model achieved a coefficient of determination (R^2) of 0.93 and a mean absolute error (MAE) of 0.08. Despite these positive results, more than half of the estimated values were underestimated compared to the observed PMI. For each sampling point, represented by a brown dot, the MLR generated a model that aimed to predict values as close as possible to the observed ones; however, it produced 15 overestimated points and 25 underestimated points, indicating a tendency toward underestimation. Even so, the model yielded satisfactory R^2 and MAE values when compared with other studies (Souza et al., 2022). These results demonstrate the effectiveness of the method, providing a non-destructive and accurate alternative for estimating peanut maturity, thus reducing subjectivity and the need for intensive field sampling. Therefore, this method facilitates decision-making and improves harvest management.

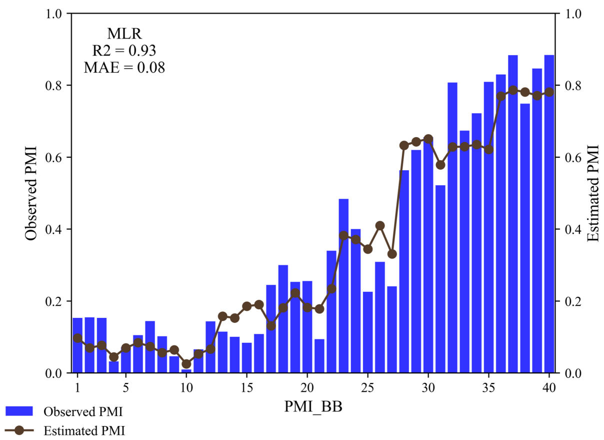

Figure 4. MLR performance for estimating maturity using the best input variables.

4. CONCLUSION

Pre-selection and variable selection proved effective in identifying the most determinant factors for peanut maturity. Among the evaluated variables, Growing Degree Days (GDD) stood out as the most consistent climatic predictor, reflecting the thermal effect on crop development and being consistently selected in the dataset, while the Green band, NDVI, and SAVI captured biophysical variations directly linked to the progression of maturity. The multiple linear regression (MLR) model, despite its simple structure, achieved satisfactory performance (R^2 = 0.93; MAE = 0.08), demonstrating that a reduced set of well-selected variables is sufficient to accurately estimate maturity. These results confirm that the integration of climatic and spectral variables, combined with attribute selection methods, provides a solid basis for defining the optimal harvest point and for developing future models aimed at more precise crop management.

REFERENCES

Arioglu, H., Bakal, H., Gulluoglu, L., Onat, B., Kurt, C., 2018. The effect of harvesting dates on some agronomic and quality characteristics of peanut (Arachis hypogaea L.) varieties grown as a main crop in Mediterranean region (Turkey). *Turkish Journal of Field Crops*, 23(1), 27–37. https://doi.org/10.17557/tjfc.414856.

Barboza, T.O.C., Ferraz, M.A.J., Pilon, C., Vellidis, G., Valeriano, T.T.B., Santos, A.F., 2024. Advanced farming strategies using NASA POWER data in peanut-producing regions without surface meteorological stations. *AgriEngineering*, 6(1), 438–454. https://doi.org/10.3390/agriengineering6010027.

Chea, C., Saengprachatanarug, K., Posom, J., Saikaew, K., Wongphati, M., Taira, E., 2022. Optimal models under multiple resource types for Brix content prediction in sugarcane fields using machine learning. *Remote Sensing Applications: Society and Environment*, 26, 100718. https://doi.org/10.1016/j.rsase.2022.100718.

EMBRAPA - Brazilian Agricultural Research Corporation, 2013. *Brazilian Soil Classification System*, 3rd ed., revised and expanded. EMBRAPA, Brasília, 353p.

Oliveira, M.F., Carneiro, F.M., Ortiz, B.V., Thurmond, M., Oliveira, L.P., Bao, Y., Tedesco, D., 2024. Predicting below and above-ground peanut biomass and maturity using multi-target regression. *Computers and Electronics in Agriculture*, 218, 108647. https://doi.org/10.1016/j.compag.2024.108647.

Paungfoo, C., Wang, W., Yeoh, Y.K., Halpin, N., 2017. Legume crop rotation suppressed nitrifying microbial community in a sugarcane cropping soil. *Scientific Reports*, 7(1), 16707. https://doi.org/10.1038/s41598-017-17080-z.

Planet, 2024. Understanding PlanetScope Instruments. Available online: https://developers.planet.com/docs/apis/data/sensors/ (accessed on 30 March 2024).

Rowland, D.L., Sorensen, R.B., Butts, C.L., Faircloth, W.H., 2006. Determination of maturity and degree day indices and their success in predicting peanut maturity. *Peanut Science*, 33, 125–136. https://doi.org/10.3146/0095-3679(2006)33[125:domadd]2.0.co;2.

Rowland, D.L., Sorensen, R.B., Butts, C.L., Faircloth, W.H., Sullivan, D.G., 2008. Canopy characteristics and their ability to predict peanut maturity. *Peanut Science*, 35(1), 43–54. https://doi.org/10.3146/PS06-052.1.

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.

Santos, A.F., Lacerda, L.N., Gobbo, S., Tofannin, A., Silva, R.P., Vellidis, G., 2019. Using remote sensing to map in-field variability of peanut maturity. In: *Precision Agriculture '19*. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 605–611. https://doi.org/10.3920/978-90-8686-888-9_75.

Souza, J.B.C., Almeida, S.L.H., Oliveira, M.F., Santos, A.F., Brito Filho, A.L., Meneses, M.D., Silva, R.P., 2022. Integrating satellite and UAV data to predict peanut maturity using artificial neural networks. *Agronomy*, 12(7), 1512. https://doi.org/10.3390/agronomy12071512.

USDA - United States Department of Agriculture, 2024. Crop Explorer. Available online: https://ipad.fas.usda.gov/cropexplorer/cropview/commodityView.aspx?cropid=2221000 (accessed on 6 June 2024).

Vishnuprabha, R.S., Viswanathan, P.L., Manonmani, S., Rajendran, L., Selvakumar, T., 2023. Genetic diversity in groundnut (Arachis hypogaea L.) genotypes varying in maturity duration. *Vegetos*, 36(4), 1550–1556. https://doi.org/10.1007/s42535-022-00489-x.

Williams, E.J., Drexler, J.S., 1981. A non-destructive method for determining peanut pod maturity. *Peanut Science*, 8(2), 134–141.

Zerbato, C., Furlani, C.E., Silva, R.P., Voltarelli, M.A., Santos, A.F., 2017. Statistical control of processes applied for peanut mechanical digging in soil textural classes. *Engenharia Agrícola*, 37(2), 315–322. http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v37n2p315-322/2017.