Use of crop and drought spectral indices to support harvest decisions of peanut fields in Alabama

M.F. Oliveira[1], B.V. Ortiz[2], J.B.C. Souza[3], A. Sanz-Saez[2], S. L. H. de Almeida[3], C. Pilon[4], G. Vellidis[4]

[1] University of Nebraska Extension - Fremont, NE. Dodge County
[2] *Crop, Soil, and Environmental Sciences Department– Auburn University, Auburn, AL, USA*
[3]*Department of Engineering and Mathematical Science - São Paulo State University, Jaboticabal, SP, Brazil*
[4] *Department of Crop and Soil Sciences – University of Georgia, Tifton, GA, USA*
mzf0065@auburn.edu

Abstract

Harvest efficiency expressed in quantity and quality of peanut fields could increase if farmers are provided with tools to support harvest decisions. Peanut farmers still rely on a visual and empiric method to assess the right time of peanut maturity but this method does not account for within-field variability of crop growth and maturity. The integration of spectral vegetation indices to assess drought, soil moisture, and crop growth to predict peanut maturity can help farmers strengthen decisions on when and where to start the harvesting process. Two commercial irrigated fields, one located in Eufaula, Alabama, U.S., and the second located in Tifton, Georgia, U.S., were used in this experiment during the 2022 and 2021 growing seasons respectively. Starting around 97 days after planting, peanut biomass samples from multiple locations within each field were collected weekly. The assessment of peanut maturity was done manually on a 200-pod sample per location using the hull-scrape method and the peanut profile board. Two approaches were tested to predict the peanut maturity index. The first approach was using partial least-squares regression and the second was using auto-machine learning. Step-wise regression was used to select the best predictor variables and predict peanut maturity. Predictor variables were crop vegetation indices (NDVI, GNDVI, NLI, and MNLI), moisture index (NMDI), and drought index (NDMI) of the study field were estimated from Sentinel 2 satellite images. The step-wise multiple regression method identified spectral indices NDMI, NMDI, GNDVI, MNLI, NDVI, and NLI as best peanut maturity predictors. Auto machine learning (ML) algorithm outperformed partial least square regression in terms of accuracy and precision. NDMI showed a higher influence on ML model prediction whereas MNLI had a higher importance on partial least square regression. Future research should focus on integrating other explanatory variables, mainly related to variables that drive within-field peanut maturity variability, like soil temperature and weather data, and terrain attributes such as topographic indices.

Keywords: drought indices, machine learning, partial least square, peanut, peanut maturity, remote sensing, vegetation indices.

Introduction

Peanut is an important crop for the southeastern U.S. and is commonly used in rotation with cotton (Mulvaney *et al.*, 2017). Determining the ideal time to start the harvest process is sometime challenging for peanut farmers (Colvin *et al.*, 2018). Generally, farmers make this decision based on the assessment of peanut maturity determined from random sampling of peanut plants. The best quality, grade, and yield of seeds can be obtained by harvesting a field at its peak maturity (Santos *et al.*, 2022). The oil content of peanut seeds from immature pods is generally lower, which can reduce their quality (Fincher *et al.*, 1980). In contrast, harvesting plants with over-mature pods can result in yield losses of up to 40% because the pods will detach from the pegs during the inversion process and be lost in the soil (Lamb *et al.*, 2004). The hull-scrape method (Willians *et al.*, 1981) in conjunction with the peanut maturity index (PMI) (Rowland *et al.*, 2006) have been widely adopted by peanut producers in the U.S. and other countries (Santos *et al.*, 2022). This method is time consuming, laborious, and subjected to uncertainties during the visual classification. To overcome this limitation methods based on remote sensing for estimating peanut maturity have been proposed. Most methods are based on RGB image analysis (Ghate *et al.*, 1993), drone imagery (Santos *et al.*, 2022), and satellite images (Souza *et al.*, 2022). Some of the vegetation indices that have been used for peanut maturity prediction are: normalized difference vegetation index (NDVI), green normalized difference vegetation index (GNDVI), modified non-linear index (MNLI), enhanced vegetation index (EVI), non-linear index (NLI), soil adjusted vegetation index (SAVI), normalized difference red edge index (NDRE). Indices that can explain other plant processes like canopy moisture and drought effects could increase the predictability of the models to estimate peanut maturity. Among the algorithms used to predict peanut maturity are artificial neural networks, non-linear models, and linear models (Souza *et al.*, 2022, Santos *et al.*, 2022). The comparison of multiple linear regression models and different types of machine learning algorithms and a method to interpret the influence of the remote sensing variables in the prediction of peanut maturity was not tested yet. This evaluation could help to understand the performance of different algorithms, identify the influence of the vegetation indices on the model's predictions, and support the farmers with a non-destructive method for peanut maturity prediction. Following this rationale, the objective of this study was to compare and interpret multiple linear and non-linear algorithms to predict peanut maturity based on crop and drought spectral indices.

Materials and Methods

Study field description and data collection. Two commercial irrigated fields, one of 54.76 hectares (ha) located in Eufaula, Alabama (AL), U.S. (F-AL), and the second of 9.24 ha located in Tifton, Georgia (GA), U.S. (F-GA), were used in this experiment during the 2022 and 2021 growing seasons, respectively. The irrigated fields were planted on May 26th and May 10th, using the same peanut runner-type cultivar Georgia-O6G, which has a growing cycle of approximately 140 ± days. Each field was divided into square grids, 20 grids for field F-AL and 14 for field F-GA, of contrasting soil characteristics which were selected for data collection (Figure 1). Starting approximately 94 days after planting, peanut biomass was collected weekly from various locations within each grid and assessment of peanut maturity was done manually on 200-pod subsample sample, using the hull-scrape method and the peanut profile board method (Wiliams & Drexler, 1981). The peanut maturity index (PMI) was calculated using the following equation.

$$PMI_{BB} = \frac{N\_bbp}{T\_p} \qquad (1)$$

where, PMI_{BB} is the peanut maturity index considering brown to black pods class

N_bbp is the number of pods in the brown and black class

T_p is the total number of pods

Satellite imagery and spectral indices. To establish a functional relationship between peanut maturity and spectral reflectance changes of the canopy over time, Planet Labs imagery data was used to extract reflectance from specific spectral bands and calculate several vegetation indices (Table 1). The surface reflectance Ortho Scene product was acquired from PlanetScope, Planet Labs, Inc., San Francisco, USA (Planet, 2020) under a student license. Cloud Planet Scope imagery data provide 3 m spatial resolution images. The PlanetScope imagery data used had four spectral bands: blue (455–515 nm), green (500–590 nm), red (590–670 nm), and near-infrared (NIR, 780 – 860 nm) in a 16-bit GeoTiff format. The spectral band images were carefully selected for days with 0% cloud over the study areas.

Table 1. Spectral indices used as predictor variables in the prediction of peanut maturity

Vegetation Index	Equation	Reference
NDVI	(NIR−RED)/(NIR + RED)	Rouse *et al.* (1974)
NLI	(NIR^2−RED)/(NIR^2 + RED)	Goel & Qin (1994)
GNDVI	(NIR−Green)/(NIR + Green)	Gitelson & Merzlyak (1996)
MNLI	(NIR^2−RED) × (1 + L)/(NIR^2 + RED + L)	Gong *et al.* (2003)
NDMI	NIR-SWIR1/NIR+SWIR1	Seeyan *et al.*, 2014
NMDI	NIR-SWIR1+SWIR2/NIR+SWIR1-SWIR2	Wang & Qu 2007

The geoprocessing steps of extracting reflectance data and calculating the six vegetation indices were performed using QGIS software (Free software Inc, Boston, United States).

Model training and validation. Two approaches were tested for prediction of peanut maturity expressed as peanut maturity index. The first approach was using a partial least-squares regression and the second was using auto-machine learning (ML). Partial least-squares regression performs least-squares regression on smaller and uncorrelated variables rather than the use of the entire dataset (Ryan & Ali, 2016). Auto-machine learning approach can automatically conduct hyperparameter tuning, feature scaling, and random grid searches and generates several models based on numerous model performance metrics (Dilmurat *et al.*, 2022). In this study, auto-machine learning was tested using distributed random forest, generalized linear model with regularization, XGboost, gradient boosting machine, and deep learning. Only the best model in the validation phase is presented in this study. For both methods evaluated on this study, the dataset was split into 80 % for training and 20 % for validation.
The indices NDVI, GNDVI, NLI, MNLI, NDMI and NMDI were used as explanatory variables to predict PMI_bb. The metric to analyze the performance of the algorithms was the mean absolute error MAE (Equation 2), and coefficient of determination (R^2).

$$MAE = \frac{\sum_{i-1}^{n}(Yesti_i - Yobs_i)}{n}$$

where, n is the number of data, $Yest_i$ is the value of the variable estimated by algorithm, $Yobs_i$ is the value of the observed variable.

Influence of input parameters on PMI prediction. The resultant model was considered using the lowest MAE and the higher R^2. After the selection of the resultant model (best features and ML algorithm and partial least-squares regression), the model was analyzed to understand the importance of each variable in the predictions. ML models offer a complex architecture, but interpreting the output of these models, is a demanding task (García & Aznarte 2020). Novel approaches are being used to interpret the output of the ML models (García & Aznarte 2020). One approach proposed to overcome this challenge is Shapley Additive exPlanations (SHAP), a technique that is based on game theory and its related extensions, Shapley values are used to connect optimal credit allocation with local explanations (Lundberg & Lee 2017). SHAP calculates the importance of a feature in the model comparing the estimation with and without the feature (García & Aznarte 2020; Lundberg & Lee 2017). In order to ensure that the features can be fairly compared, this process is implemented in every possible order so that its estimates are not influenced by the order in which it captures each feature (Lundberg & Lee 2017). The SHAP analysis was performed using Python 3 programing language trough the libraries H2O and SHAP.

Results and Discussion

The F-AL field showed an average maturity of 63 %, a maximum of 93 % and a C.V of 36.68 % which demonstrates peanut fields might exhibit a high degree of peanut maturity variability. F-GA field showed less variability than the F-AL field (C.V. 26.61 %) and lower values of average (40%) and maximum (65 %) maturity. On average, the F-AL field had a 20 % higher percentage of PMI. The optimum PMI is 70 % (Rowland *et al.*, 2006), based on this information Field-1 had areas where the peanut was over-mature (Max 93 %) and under-mature (Min 65%). While Field-2 did not reach the optimum maturity.

Table 2. Descriptive analysis for peanut maturity (PMI) in two fields.

Field	Mean	Minimum	Maximum	C.V. %	STD
F-AL	0.60	0.16	0.93	36.68	0.22
F-GA	0.40	0.13	0.65	26.61	0.09

C.V = Coefficient of variation, STD = Standard deviation

The partial least-squares model described the observed variation in peanut maturity well with five-fold cross-validation producing an MAE = 7 % and an $R^2 = 0.81$ (Figure 1A). When considering the performance of the XGBoost model, an MAE = 4% and an $R^2 = 0.94$ was observed (Figure 1B). These validation statistics demonstrated that the machine learning approach could outperform a multi-linear regression model to describe the observed spatial peanut maturity index integrating crop and drought spectral indices. PLS combines elements of two regression methods, principal components regression, and multiple linear regression, and is able to handle hyperspectral data with collinearity by inputting all spectral bands at once and then identifying uncorrelated variables from a matrix of explanatory variables (Geladi & Kowalski, 1986). XGBoost is a boosting algorithm that uses decision trees as its base learners. The idea behind the algorithm is to sequentially improve the model by reducing the residuals of the previous model in the gradient direction, resulting in a new model (Jin *et al.*, 2021).

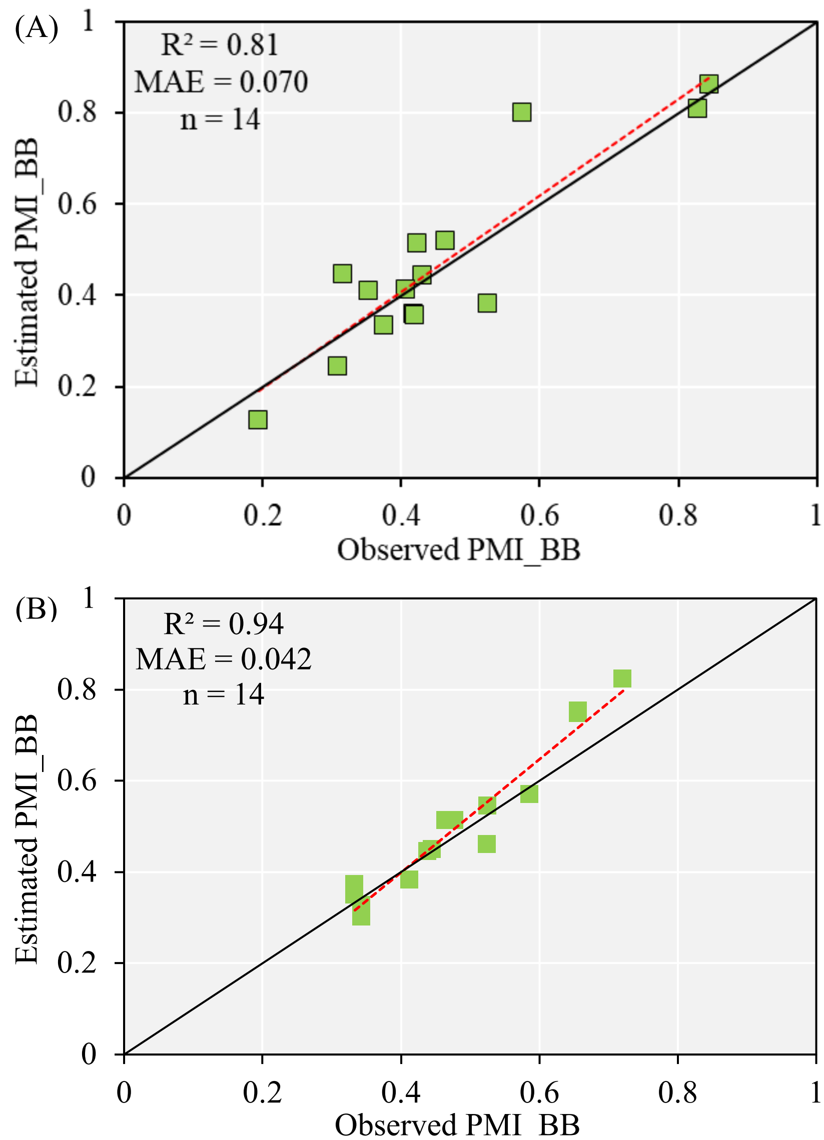

Figure 1. Performance analysis of the test dataset comparing partial least square regression (A) and ML (B).

The SHAP summary plots efficiently conveyed the feature importance (Figure2 and 3). Features are ranked based on the mean absolute SHAP value. When considering the partial least-square regression, the MNLI had the highest mean feature importance, followed by NLI, NDVI, NDMI, GNDVI, and then NMDI (Figure 2). The SHAP plot for the XGBoost model demonstrated the NMDI has the highest importance on predicting peanut maturity index followed by MNLI, GNDVI, NDVI, NDMI, and NLI. PLS regression and XGBoost showed differences in variable importance. For PLS regression the most important variable was MNLI whereas for XGBoost was NMDI. Peanut leaf area index (LAI) increases during plant growth, decreases during the reproductive process, and LAI peak during the flowering stage (Qi *et al.*, 2020). This LAI pattern can express the correlation between the NMDI, MNLI and the PMI. NDMI tends to be higher with higher LAI and responds almost linearly with leaf water content (Wang & Qu, 2007) and MNLI is correlated with LAI (Gong *et al.*, 2003). It is expected if the NMDI changes with LAI and the LAI decreases during the reproductive process, reaching lower values during the mature stage compared to the seedling stage, this correlation tends to be negative with PMI because the PMI behaves positively linearly. Which means higher PMI will occur during the maturity stage and the LAI at this point is lower than at the beginning of the pod-filling stage.

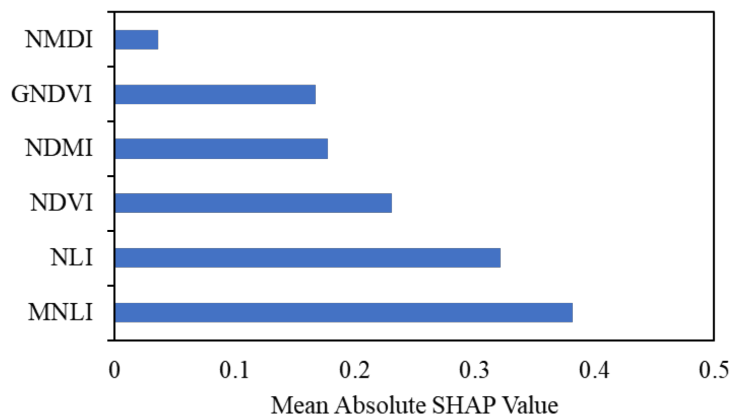

Figure 2. Mean absolute SHAP value for partial least square regression.

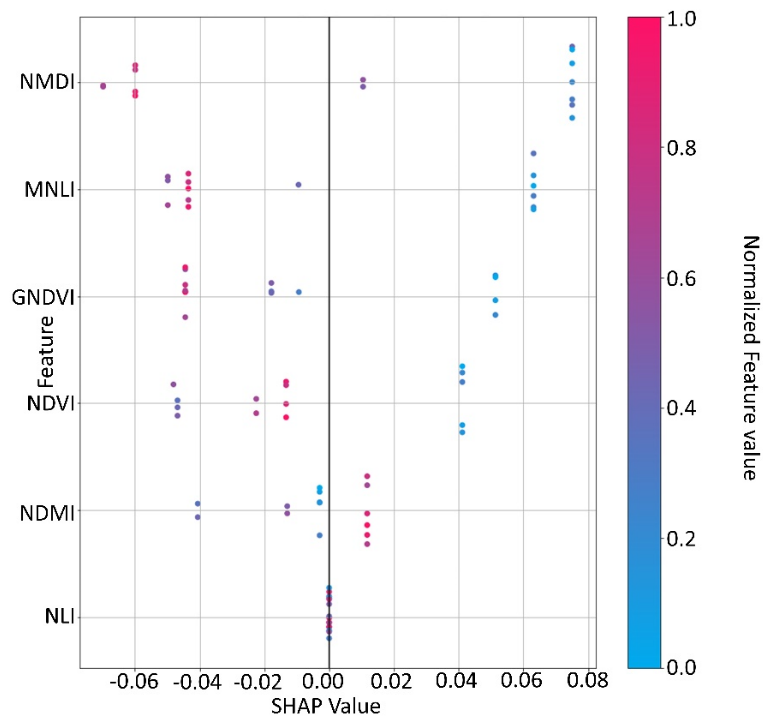

Figure 3. SHAP value for the resultant machine learning model.

Conclusions

The present study investigated the potential of using spectral indices of high-resolution satellite images to estimate peanut maturity index by comparing the use of machine learning models and partial least regression. The present study provides evidence that contributes to peanut maturity estimation at field scale. ML algorithm outperformed partial least square regression in terms of accuracy and precision. NDMI showed a higher influence on ML model prediction whereas MNLI had a higher importance on partial least square regression. Future studies should focus on integrating new variables that could explain terrain attributes and water accumulation such as topographic indices.

Acknowledgements

This work received financial support from the Alabama Peanut Producers Association. We express our gratitude to Marshall Speake and McMillan for allowing us to work in their fields. Data collection for this study was possible thanks to the collaboration of Megan Thurmond, Barret Masson, Pablo Duarte, Timothy Squires, Hunter Whatley.

References

Colvin, B. C., Tseng, Y. C., Tillman, B. L., Rowland, D. L., Erickson, J. E., Culbreath, A. K., & Ferrell, J. A. (2018). Consideration of peg strength and disease severity in the decision to harvest peanut in southeastern USA. Journal of crop improvement, 32(3), 287-304.

Dilmurat, K., Sagan, V., Maimaitijiang, M., Moose, S., & Fritschi, F. B. (2022). Estimating Crop Seed Composition Using Machine Learning from Multisensory UAV Data. *Remote Sensing*, *14*(19), 4786.

Fincher, P. G., Young, C. T., Wynne, J. C., & Perry, A. (1980). Adaptability of the arginine maturity index method to Virginia type peanuts in North Carolina. *Peanut Science*, *7*(2), 83-87.

García, M. V., & Aznarte, J. L. (2020). Shapley additive explanations for NO2 forecasting. Ecological Informatics, 56, 101039.

Ghate, S. R., Evans, M. D., Kvien, C. K., & Rucker, K. S. (1993). Maturity detection in peanuts (Arachis hypogaea L.) using machine vision. Transactions of the ASAE, 36(6), 1941-1947.

Gong, P., Pu, R. L., Biging, G. S., & Larrieu, M. R. (2003). Estimation of forest leaf area index using vegetation indices derived from hyperion hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 41, 1355–1362.

Jin, X., Xiao, Z., Dong, A., Wang, Y., Nie, Q., & Wang, L. (2021, November). Hyperspectral Inversion Modeling of Fat and Protein Content in Milk Based on XGBoost. In 2021 6th International Conference on Robotics and Automation Engineering (ICRAE) (pp. 19-24). IEEE.

Lamb, M. C., Masters, M. H., Rowland, D., Sorensen, R. B., Zhu, H., Blankenship, R. D., & Butts, C. L. (2004). Impact of sprinkler irrigation amount and rotation on peanut yield. *Peanut Science*, *31*(2), 108-113.

Lundberg, S.M., Lee, S.I., (2017). A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems, pp. 4768-4777.

Mulvaney, M. J., Balkcom, K. S., Wood, C. W., & Jordan, D. (2017). Peanut residue carbon and nitrogen mineralization under simulated conventional and conservation tillage. *Agronomy Journal*, *109*(2), 696-705.

Qi, H., Zhu, B., Wu, Z., Liang, Y., Li, J., Wang, L., ... & Zhang, L. (2020). Estimation of peanut leaf area index from unmanned aerial vehicle multispectral images. Sensors, 20(23), 6732.

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(2), 125-136.

Ryan, K., & Ali, K. (2016). Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data. *Ocean Science Journal*, *51*(2), 209-221.

Santos, A. F., Lacerda, L. N., Rossi, C., Moreno, L. D. A., Oliveira, M. F., Pilon, C., ... & Vellidis, G. (2021). 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.

Seeyan, S., Merkel, B., & Abo, R. (2014). Investigation of the relationship between groundwater level fluctuation and vegetation cover by using NDVI for Shaqlawa Basin, Kurdistan Region—Iraq. Journal of Geography and Geology, 6(3), 187–202.

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.

Wang, L., & Qu, J. J. (2007). NMDI: A normalized multi-band drought index for monitoring soil and vegetation moisture with satellite remote sensing. Geophysical Research Letters, 34(20). Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. Analytica chimica acta, 185, 1-17.

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