70. Sesame yield prediction using UAV-borne hyperspectral imagery, crop traits and genotype category

M.M. Sahoo[1,*], R. Tarshish[1], I. Sabag[1], Y. Gadri[1], G. Morota[2,3], Z. Peleg[1], V. Alchanatis[4] and I. Herrmann[1]
*[1]The Hebrew University of Jerusalem, Plant Sciences and Genetics in Agriculture, The Robert H. Smith Faculty of Agriculture, Food and Environment, 7610001, Rehovot, Israel; * maitreyamohan.sahoo@mail.huji.ac.il*
[2]School of Animal Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[3]Laboratory of Biometry and Bioinformatics, Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan
[4]Agricultural Research Organization–Volcani Institute, Institute of Agricultural Engineering, 7505101 Rishon LeZion, Israel

Abstract

Crop yield prediction, if done early enough, can help growers and breeders increase production. This research provides a method for predicting sesame yield by integrating information from remote sensing and phenotypic traits. Based on a field-grown sesame diversity panel of 185 genotypes, selected traits were analyzed to explain their contribution to yield. It was found that chlorophyll a variation (i.e., low, variable and high content) provided the maximum basis for categorically predicting sesame yield from remotely sensed hyperspectral imagery using machine learning models. The highlighted spectral features can be used to categorize sesame yield based on chlorophyll content.

Keywords: classification, crop breeding, crop traits, remote sensing, spectral phenotyping

Introduction

Sesame is an important oilseed crop worldwide. It is an erect to semi-erect indeterminate plant, which is adaptable to hot and dry environments (Dossa *et al.*, 2017; Langham, 2007). Its grains are used for various purposes, such as oil production, tahini, baking, cosmetics, and the pharmaceutical industry. It is also known as the "queen of oilseeds". In recent years, the demand for sesame grains (and secondary products) has increased as part of the global trends toward healthier plant-based food sources. This growing demand opens an opportunity to expand sesame cultivation into new agro-systems in the Mediterranean region that would maximize grain yield (Sabag *et al.*, 2024b).

Hyperspectral remote sensing is a promising tool in agriculture for monitoring crop production. With the application of unmanned aerial vehicles (UAVs), building of large-scale image data across multiple contiguous spectral bands, and the enhancement of computational resources, it has been a subject of interest for yield prediction in various crops (Herrmann *et al.*, 2020; Ram *et al.*, 2024). A crop's yield is affected by three primary factors- Genotype (G), Environment (E), and Management (M), and their interactions (G×E×M; Gadri *et al.*, 2020). While environment and management (E, M) factors have a direct effect on a crop yield, the role of genotype is qualitative, and it becomes challenging to numerically predict its yield (Shook *et al.*, 2021). It is pertinent to study the effect of yield on different crop genotypes, as such studies would benefit farmers and breeders to maximize farm productivity.

Prior to measuring yield during its harvest, certain crop traits such as chlorophyll (Chl), leaf area index (LAI), leaf water content (LWC), and other morpho-physiological traits are indicators of its yield (Gitelson *et al.*, 2003; Turner, 1990). Estimation of these traits on a timeline of growing season using remotely acquired dataset can provide a multitude of information regarding crop growth and development, which impacts its yield. Although studies have been conducted on numerical yield prediction, the ability to predict yield for a multiple genotype dataset would require studying the distribution among the crop traits.

With a focus on sesame as the study crop in this research, an attempt was made to categorically predict sesame yield by analyzing information from UAV-borne hyperspectral imagery and sesame traits. The key objectives of this research were:

1. Determining traits contributing to the prediction of sesame yield.
2. Categorical prediction of sesame yield using hyperspectral features diagnostic to the determined traits.

Dataset and methodology

Study area

The field experiment was conducted at The Robert H. Smith Faculty of Agriculture, Food, and Environment's experimental farm, Rehovot, Israel (34°47′ N, 31°54′ E) during the summer of 2020 growing season (Figure 1). A diverse panel of 185 sesame genotypes (Sabag *et al.*, 2021, 2024a) sown in a complete randomized block design. Each plot was 2.5 m long×0.8 m wide, with a spacing of 20 cm between the rows. The field was split into 537 plots of 3.9 m^2 with a density of 6 plants/m.

Data acquisition and pre-processing

Remotely sensed data acquisition started 45 days after sowing (DAS), when more than 50% of the plants bloomed. Hyperspectral reflectance images of the sesame canopy corresponding to the 537 plots were acquired in the visible and near-infrared (VNIR) region (385–1020 nm) using a Pika L airborne hyperspectral scanner (Resonon, Bozeman, MT, USA) mounted on a Matrice600 UAV (DJI, Shenzhen, Guangdong, P.R. China). The hyperspectral scanner imaged the study area

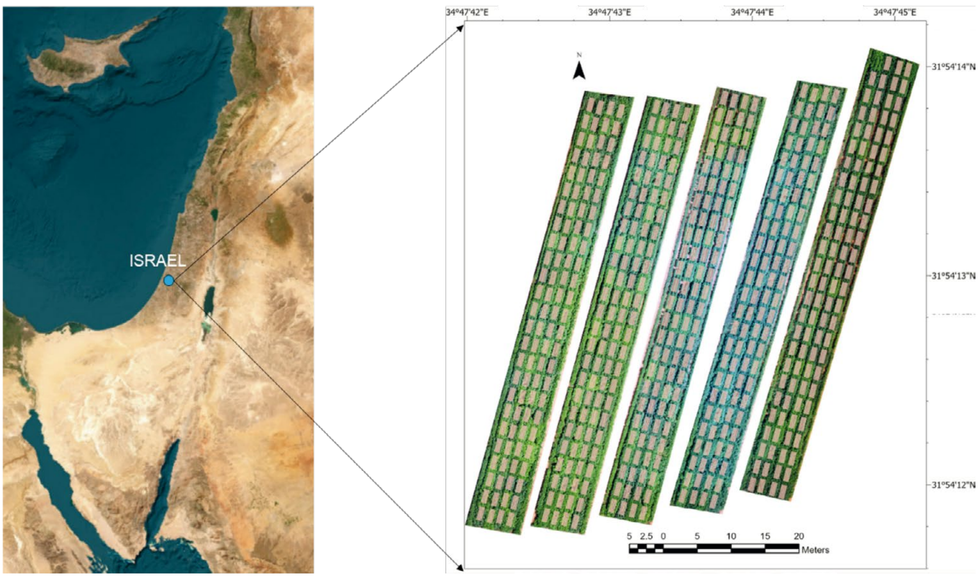

Figure 1. Location of the study area for the experiment showing the 537 sesame plots. The diversity panel of 185 sesame genotypes was grown in a randomized block design.

configured at an altitude of 100 m above the sesame canopy with a field-of-view (FOV) of 24.8°, 900 pixels per scanning line in 300 hyperspectral bands. The images were acquired at six different dates of the growing season: 45, 57, 65, 71, 80 and 87 DAS. Sesame seed yield (g/plant) was measured on the day of harvest (87 DAS). The hyperspectral images were pre-processed for (1) radiometric correction, implemented using the Spectronon software, Resonon Inc.; (2) atmospheric correction, implemented by the flat field ratio technique using a standard Resonon grey (24 % reflectance) tarp; and (3) geometric correction, mosaicking scanned lines to generate the hyperspectral image of the sesame field. The resulting geo-registered hyperspectral images were spatially processed for the sesame plots, averaging hyperspectral data by selecting subregions of 2 by 0.7 m for each plot center using ArcGIS Pro 3.0.2 (Esri, Redlands, CA, USA). The spatial subsetting minimized the interference of background soil pixels along the edges of the sesame plots. Spectrally noisy bands were removed from the averaged hyperspectral dataset at wavelength regions corresponding to 385–399 nm (bands 1 to 10) and 1011–1020 nm (bands 296 to 300). After removing the noisy bands, the NIR region of the spectral dataset was smoothed using the Savitzky-Golay method with a window size of 25 bands and a polynomial value of 2.

Data analysis

The pre-processed hyperspectral dataset was subsequently used for the estimation of sesame traits, including leaf nitrogen content (LNC), Chl a, b, LAI, leaf mass per area (LMA), and LWC using pre-trained genetic algorithm-inspired partial least squares regression (GA-PLSR) models. The wavelength features used to estimate each trait are tabulated in Table 1. These estimated traits prior to harvest were analyzed using principal component (PC) and further integrated with the measured sesame yield for a k-means clustering in order to find the number of distinct yield categories. Using the information from the PC analysis and k-means clustering (discussed in the Results and Discussion section), the diagnostic spectral features were thereafter used for training classification models to categorically predict sesame yield. Partial least squares discriminant analysis (PLS-DA) and support vector classifier (SVC) models were used for the yield classification with 70% of the dataset (376 samples) trained and 5-fold cross-validated and 30% (161 samples) independently validated. For each measurement date, the selection of plots for training (calibration, cross-validation) and independent validation was randomly shuffled within each sesame genotype. A timeline-based analysis was performed on the ability of classification models (PLS-DA and SVC) to categorically predict sesame yield using spectral features acquired during six measurement dates (45, 57, 65, 71, 80, 87 DAS). The classification models were implemented using MATLAB R2023b (Mathworks, Torrance, CA, USA).

Table 1. Wavelength features (in nm) identified in pre-trained genetic algorithm-inspired partial least squares regression (GA-PLSR) models used for sesame traits estimation, including leaf nitrogen content, chlorophyll a and b, leaf area index, leaf mass per area and leaf water content.

Sesame trait	Wavelength features (nm) identified in pre-trained GA-PLSR models
LNC	670–685, 700–710, 735–770, 845–875, 910–950
Chl a	410–420, 580–650, 660–675, 750–800, 950–1000
Chl b	410–440, 575–590, 650–680, 690–715, 750–800, 950–990
LAI	410–450, 500–600, 690–800, 840–870, 900–1010
LMA	410–460, 690–710, 740–780, 940–1000
LWC	460–480, 500–520, 590–610, 690–710, 830–850, 870–890

Results and discussion

The PC analysis of the estimated sesame traits revealed the importance of Chl a explaining the maximum percentage variance of 84.21%, followed by LWC (10.66%), LAI (3.54%), Chl b (1.22%), LMA (0.30%), and LNC (0.07%). Figure 2 displays the eigenvalues plot of PC-analyzed traits with their coefficients tabulated in Table 2.

Integration of sesame yield with the PC1 component (that explained maximum variance for Chl a) followed by a k-means clustering analysis resulted in clustering of sesame yield into three distinct clusters: low yield with low Chl a (58 genotypes), low yield with high Chl a (85 genotypes), and high yield with varying Chl a (36 genotypes). Figure 3 shows the individual regions for these clusters. Low sesame yield with low Chl a signifies the seed yield of sesame being dependent on leaf chlorophyll content (Pallavolu *et al.* 2023). Low yield with high Chl a can be attributed to environmental stresses due to extreme temperature in the summer and sesame grown in sandy soil with limited water retention capacity. The cluster of high yield with varying Chl a can be attributed to other morpho-physiological parameters such as plant height, number of branches, leaf and flowering densities owing to the genotype factor.
Implementation of classification models using PLS-DA and SVC provided insights into the predictive ability of sesame yield from GA-PLS identified diagnostic spectral features. Table 3 summarizes

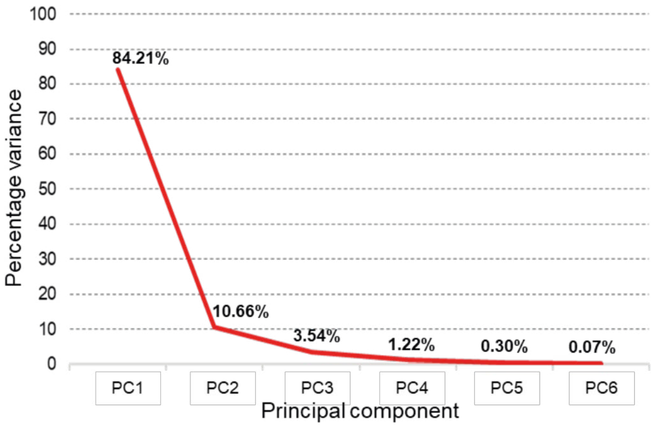

Figure 2. Eigenvalues plot of principal component analysis of the estimated sesame traits explaining their variance (in percentage).

Table 2. Sesame trait coefficients obtained for each principal component analysis.

	LNC	Chl a	Chl b	LAI	LMA	LWC
PC1	0.08	0.94	0.17	−0.09	−0.05	0.25
PC2	0.02	−0.24	0.11	0.29	0.01	0.92
PC3	0.16	0.11	0.10	0.90	−0.27	−0.26
PC4	−0.07	0.07	0.16	0.25	0.95	−0.09
PC5	0.42	−0.19	0.86	−0.19	−0.06	−0.10
PC6	0.89	−0.006	−0.43	−0.05	0.15	0.04

the classification for categorical prediction of sesame yield from Chl a diagnostic spectral features using PLS-DA and SVC models. It was observed that 7 PLS components sufficiently explained the covariance among the spectral features and yield category among the 185 sesame genotypes. For SVC, the quadratic SVC model showed increased classification accuracies compared to linear, cubic, and Gaussian models. The maximum accuracy was observed for PLS-DA on 65 DAS with an independently validated total accuracy of 77%, compared to that obtained using SVC (74.0%).

Categorical prediction of sesame yield from LWC diagnostic spectral features provided the best results for 8 PLS components, with reduced contrasts among calibration and cross-validation accuracies (Table 4). For a few measurement dates, the accuracies of independently validated samples were higher than the trained samples. It is expected that such contrasts would be reduced with a larger dataset. The PLS-DA model performed comparatively well with 73%, 70% (Cal and CV) and 75% (IV) performance metrics prior to harvest (i.e., 65 DAS) compared to SVC (with 70.5% and 69.6% total classification accuracies respectively). The classification accuracies improved from 45 DAS (blooming) to 65 DAS (with maximum sesame canopy cover).

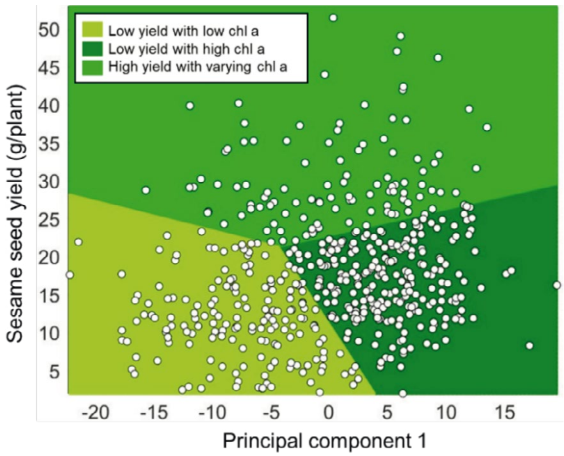

Figure 3. Scatter plot of sesame seed yield vs. the first principal component showing three distinct clusters obtained from k-means clustering. Based on Table 2, Chl a has the highest influence on PC1 with a coefficient of 0.94.

Table 3. Total classification accuracies obtained for categorical prediction of sesame yield from chlorophyll a diagnostic spectral features using PLS-DA and SVC models.

DAS	PLS-DA			SVC	
	Cal	CV	IV	Cal and CV	IV
45	62%	53%	54%	57.7%	54.7%
57	65%	61%	73%	63.3%	69.6%
65	75%	72%	77%	71.8%	74.0%
71	67%	64%	74%	66.7%	72.8%
80	64%	62%	55%	66.5%	69.6%
87	67%	65%	57%	61.2%	52.2%

PLS-DA, partial least squares discriminant analysis; SVC, support vector classifier; Cal, calibration; CV, cross-validation; IV, independent validation.

Table 4. Total classification accuracies obtained for categorical prediction of sesame yield from leaf water content diagnostic spectral features using PLS-DA and SVC models.

DAS	PLS-DA			SVC	
	Cal	CV	IV	Cal and CV	IV
45	62%	59%	55%	59.8%	55.3%
57	67%	63%	70%	64.9%	62.0%
65	73%	70%	75%	70.5%	69.6%
71	66%	63%	71%	65.6%	72.2%
80	68%	67%	62%	65.4%	69.6%
87	60%	61%	52%	62.2%	57.1%

Conclusions

This study categorically predicted sesame yield belonging to 185 genotypes using information from UAV-borne hyperspectral dataset and estimated sesame traits. The estimated sesame traits were analyzed using PC, which identified the key traits, including Chl a and LWC that contributed to sesame yield prediction. The k-means clustering between the PC1 and yield axes resulted in 3 distinct clusters integrating information from yield and Chl a, i.e., low yield with low Chl a (58 genotypes), low yield with high Chl a (85 genotypes), and high yield with varying Chl a (36 genotypes). The spectral features identified for Chl a and LWC were used for training and evaluating classifier models: PLS-DA and SVC using 70% of samples for calibration, 5-fold cross-validation and 30% for independent validation. The PLS-DA models performed comparatively better than the SVC models, with maximum independently validated total accuracies of 77% and 75% for Chl a and LWC prior to harvest (i.e., 65 DAS). Random shuffling of plots within each sesame genotype for training and independent validation improved the reproducibility of the analyses by reducing the correlation with the training set for each measurement date. Although Chl a has a significant influence on sesame yield, the distinction between varying Chl a with high yield clusters was not well defined, which affected the classification performance. Future work would incorporate spatial and morphological traits (such as the number of leaves and plant height) that would further enhance the distinction among sesame genotypes for increased accuracy and facilitate numerical modeling.

Acknowledgment

The authors would like to acknowledge Dr. Yaniv Tubul, Roy Sadeh, and Michael Mathenge for their invaluable support in data collection.

References

Dossa, K., Diouf, D., Wang, L., Wei, X., Zhang, Y., Yehouessi, L.W., Liao, B., Zhang, X., & Cisse, N. (2017). The emerging oilseed crop *Sesamum indicum* enters the "Omics" era. Frontiers in Plant Science, 8, 1154. https://doi.org/10.3389/fpls.2017.01154

Gadri, Y., Eshed Williams, L., & Peleg, Z. (2020). Tradeoffs between yield components promote crop stability in sesame. Plant Science, 295, 110105. https://doi.org/10.1016/j.plantsci.2019.03.018

Gitelson, A.A., Gritz, Y., & Merzlyak, M.N. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. Journal of Plant Physiology, 160(3), 271–282. https://doi.org/10.1078/0176-1617-00887

Herrmann, I., Bdolach, E., Montekyo, Y., Rachmilevitch, S., Townsend, P.A., & Karnieli, A. (2020). Assessment of maize yield and phenology by drone-mounted superspectral camera. Precision Agriculture, 21(1), 51–76. https://doi.org/10.1007/s11119-019-09659-5

Langham, D.R. (2007). Phenology of sesame. In J. Janick & A. Whipkey (eds), Issues in new crops and new uses. ASHS Press, Alexandria, VA, pp. 144–182.

Pallavolu, L.A., Pasala, R., Kulasekaran, R., Pandey, B.B., Virupaksham, U., & Perika, S. (2023). Analysing the SPAD dynamics of water-stressed vs. well-watered sesame (*Sesamum indicum* L.) accessions and establishing their relationship with seed yield. PeerJ, 11, 14711. https://doi.org/10.7717/peerj.14711

Ram, B.G., Oduor, P., Igathinathane, C., Howatt, K., & Sun, X. (2024). A systematic review of hyperspectral imaging in precision agriculture: analysis of its current state and future prospects. Computers and Electronics in Agriculture, 222, 109037. https://doi.org/10.1016/j.compag.2024.109037

Sabag, I., Bi, Y., Sahoo, M. M., Herrmann, I., Morota, G., & Peleg, Z. (2024a). Leveraging genomics and temporal high-throughput phenotyping to enhance association mapping and yield prediction in sesame. The Plant Genome 17 (3) e20481https://doi.org/10.1002/tpg2.20481

Sabag, I., Morota, G., & Peleg, Z. (2021). Genome-wide association analysis uncovers the genetic architecture of tradeoff between flowering date and yield components in sesame. BMC Plant Biology, 21(1), 549. https://doi.org/10.1186/s12870-021-03328-4

Sabag, I., Pnini, S., Morota, G., & Peleg, Z. (2024b). Refining flowering date enhances sesame yield independently of day-length. BMC Plant Biology, 24(1), 711. https://doi.org/10.1186/s12870-024-05431-8

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., & Singh, A.K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. PLoS ONE, 16, e0252402. https://doi.org/10.1371/journal.pone.0252402

Turner, N.C. (1990). Plant water relations and irrigation management. Agricultural Water Management, 17(1–3), 59–73. https://doi.org/10.1016/0378-3774(90)90056-5