Machine Learning Approach for Predicting Flowering Days in Sorghum

Aime Nishimwe[a,b], Mackenzie Zwiener[a], Marcin Grzybowski[a], Yufeng Ge[a], and James C. Schnable[a,b]

[a]University of Nebraska-Lincoln, 1400 R St, Lincoln, NE,USA
[b]IDeA Networks of Biomedical Research Excellence (INBRE), Manter Hall 402, Lincoln, NE, USA

ABSTRACT

Sorghum is an important cereal crop grown across the globe for its grain and biomass value. It can also efficiently use resources such as nitrogen, and multiple varieties that are nitrogen-use and light-capture efficient are constantly being developed. This study focuses on using the spectral signature of sorghum varieties to predict flowering days, which could be used as a proxy for plants' growth/productivity and development trends, thus helping breeders make quick decisions about what varieties to move to the next stage. Multiple sorghum varieties from the sorghum association panel were planted in a replicate-design field experiment with the variable supply of nitrogen. The flowering days were monitored and recorded. The hyperspectral reflectance data were collected and used to build a sorghum flowering days predictive model. Although regression models such as partial least square have been used to predict plants' phenotypes, the non-parametric ensemble machine learning model turned out to perform better on flowering days with an accurate model up to 5 days.

Keywords: Machine Learning, Flowering Days, Hyperspectral, Remote Sensing, Sorghum

1. INTRODUCTION

Sorghum (*Sorghum bicolor* is the six most widely cultivated crop, grown in approximately 40M hectares (approximately 100M acres) around the globe. In Africa and Asia, sorghum is mainly used direct human consumption.[1] In the United States, sorghum is employed for for biofuel production and animal feed.[2] On most parts of the world, sorghum has been widely adopted due to its nitrogen use efficiency and productivity under low nitrogen conditions.[3] Sorghum also has striking ability to maintain high levels of photosynthesis, transpiration and chlorophyll in the most extreme drought conditions, which has been one of the primary reason for its adoption in the United States.[4] However, to maximize the yield and minimize the reliance on nitrogen fertilizers, more nitrogen-use efficient varieties will still be needed to curb the detrimental effects of nitrogen to the environment while producing enough to feed the growing global population.[5] Flowering time, amount of time between planting and anthesis varies significantly among sorghum lines.[6] Changes in flowering time have pleiotopic effects on plant height, leaf number, root architecture, grain yield, and resource use efficiency.[7] Measuring flowering time across multiple environments is a critical component of evaluating potential new sorghum varieties. Obtaining flowering time by conventional means requires a researcher to walk the field each day to evaluate which new varieties have flowered. Because plant breeders typically evaluate between 1,000 and 100,000 new genotypes for each new variety released manual evaluation can be labor intensive or impossible, particularly when poor weather restricts access to field sites during flowering season. Approaches to predict flowering time in advance have the potential to accelerate plant breeding and improve the rate of genetic gain for both yield, resource use efficiency, and resilience to extreme weather events.

Scientists have used multiple approaches to predict flowering days. For example, Elroy et al. have used genetic models with phenotype input such as temperature, rate of photoperiod change and daily irradiance to predict the flowering time, from which they achieved r-squared = 0.84-0.91.[8] Chauhan et al have used crop growth model fitted with temperature, photoperiod, and soil water content in the top soil layer, 0-60, to predict flowering time of wheat(Lin's concordance correlation coeffiecient, Lin's CCC 0.91-0.94) and chickpea (Lin's CCC= 0.97).[9] However, building genetic and crop growth models can be time-consuming, labor-intensive, and difficult due

to exhaustive field and lab experiments' measurements which need to be collected. However, the advances in spectroscopy has made it possible to collect hyperspectral reflectance data from plants at high throughput under field conditions.[10] These patterns of hyperspectral reflectance data have been successfully used to train models to predict a wide range of physical and biochemical plants properties (as reviewed[11]). Hyperspectral reflectance data has also been successfully employed to predict more complex traits with less direct mechanistic links to reflectance including predicting end of season yield in wheat and soybean. Yoosefzadeh-Najafabadi et al have used stacking ensemble model with random forest as a base metaclassifier and reflectance data for yield prediction in soybean, and they achieved 0.93 prediction accuracy.[12] Also Montesino-Lopez et al have used conopy hyperspectrarl reflectance to predict grain yield in wheat where they achieved prediction accuracy (measured as the average of the ten-fold cross-validation of the Pearson correlation) of 0.45 - 0.65 with their functional-B-Spline model and Fourier model.[13] This study uses a version of ensemble model, Extra Trees Regressor[14] to predict sorghum flowering days using hyperspectral information collected in the sorghum field experiment in 2020. This models fits multiple decision trees and the final prediction comes to be an arithmetic average of all fitted decision trees. The root mean squared error, and 3-fold validation pearson's correlation coefficient were used to evaluate the model's prediction accuracy on the validation dataset (30% of entire dataset).

2. RESULTS AND DISCUSSION

2.1 Exploratory Data Analysis

The dataset used for this study contained 341 sorghum genotypes whose flowering days were collected in a replicate-design field experiment in 2020 under two different nitrogen conditions, low (no nitrogen applied) and high nitrogen (80lbs/acres applied). During the same experiment, hyperspectral reflectance data was collected, corresponding to 2151 leaf reflectance for 350-2500nm wavelengths. After a series of data cleaning process, which involved removing outliers and missing values in both dataset, we end up working with a dataset containing 831 samples with 2151 explanatory variables and flowering days a predictor variable. The number of days to flowering reported for sorghum varieties under the two nitrogen treatments are normally distributed with a mean of 67 (std = 6) and 63 (std = 6) days for low and high nitrogen, respectively (Figure 1-A). Sorghum varieties showed relatively low variance of reflectance across the entire light spectrum in high nitrogen as compared to low nitrogen (Figure 1-B). Principal component analysis was used as a dimension reduction technique,[15] where 50 principal components were calculated out of 2151 reflectance variables per variety. The first 15 components summarize the vast majority (99.7%) of the spectra variation among all the varieties (Figure 2). Therefore, the ExtraTree ensemble model was then fitted with the first 15 principal components and five reflectance-derived indices. Among these derived indices, there includes Enhanced Vegetation Index (EVI),[16] Red-edge Normalized Difference Resistant Vegetation Index (RNDVI),[17] Atmospherically Resistant Vegentation Index (ARVI),[18] Vogelmann Red Edge Index (VRE),[19] and Normalized Difference Vegetation Index (NDVI).[20]

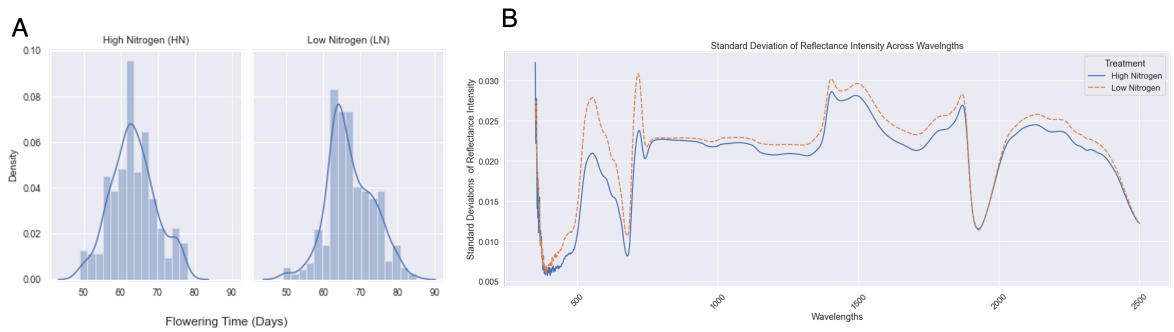

Figure 1. Descriptive summary of flowering days and reflectance among 341 sorghum varieties. Normal distribution of flowering days (Figure 1-A). High variance of reflectance values across 350-2500 wavelengths in low nitrogen conditions (Figure 1-B).

2.2 Extra Trees Ensemble Model

The dataset was split into 70:30 proportions for training and testing using Scikit-learn's Extra Trees model.[21] The base model was fitted with default parameters (n_estimators = 100). The model achieved validation r^2 of 0.23 (rmse = 5.5). We tuned hyperparameters of the model using RandomSearchCV and GridSearchCV. These are implemented using python's scikit-learn library,[21] and they are cross-validation hyperparameter-tuning optimization techniques. The resulting model (n_estimators = 150, min_samples_split = 15, max_depth = 45) achieved the pearson correlation coefficient of 0.55 (Root Mean Squared Error or RMSE = 5).

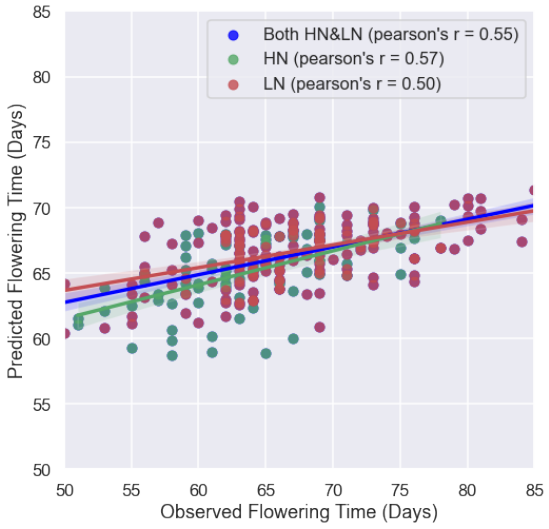

Figure 2. Observed vs. Predicted flowering days using Extra Tree ensemble model. The model achieved pearson's correlation of coefficient of 0.55 when both conditions, high (HN) and low (LN) nitrogen,

3. CONCLUSION

The sorghum varieties in low nitrogen supply will tend to flower later, which is expected since developing new plant structures would require enough nitrogen supply (Figure 1-A). With abundant nitrogen and light source, plants can carry out photosynthesis at a higher rate and thus build new structures such as flowers to complete their development. This might explain why sorghum plants in abundant nitrogen conditions developed flowers earlier. Since light capture efficiency[22] depends on a plant's leaf structure, particularly its light-capturing pigments like chlorophyll and carotenoids, we could predict flowering days by monitoring leaves' reflectance. In this study, our model achieved a Pearson's correlation between the actual and predicted flowering days 0.50 - 0.57. This model shows that by simply collecting non-destructive hyperspectral data of various sorghum varieties, we can predict flowering days with accuracy levels up to one week. Therefore, breeders will not need to wait until flowering time to conclude what sorghum varieties should be carried into the next breeding stage. The model has the potential to reduce the length of the breeding period significantly.

However, the model was built on a small dataset of only 831 samples, and its generalizability in the real-world application could be uncertain. The next step is to collect more samples or add more explanatory variables such as genotype-specific data to build a robust, generalizable model. Other models such as convolutional and recurrent neural networks and others should also be evaluated to develop the best-performing model that improves validation scores.

4. METHODS

4.1 Data Collections

A set of 347 sorghum lines drawn from the sorghum association panel[23] were planted at the University of Nebraska-Lincoln's Havelock farm on June 08, 2020. Each plot consisted of a single row, with thirty inch (0.762 meter) spacing between rows and 30 inch (0.762 meter) alleyways between sequential plots. The field was arranged in a randomized complete block design with repeated checks, with three blocks of low nitrogen treatment (0 lbs/acre) and two of high nitrogen treatment (80 lbs/acre). The field was walked daily, weather permitting, and flowering time was scored on the day 50% of extant plants within a plot had reached anthesis. Hyperspectral reflectance data was collected from a subset of plots followering the protocol of[24] using a FieldSpec4 (Malvern Panalytical Ltd., Formerly Analytical Spectral Devices) resulting in 2,150 total measured reflectance intensity values between 350 to 2500 nanometers.

The two datasets were merged using their plot number identification tags resulting in a final dataset flowering time and reflectance data from 831 unique plots representing one or more independent observations of 341 unique sorghum varieties under the two nitrogen conditions.

4.1.1 Principal Component Analysis

The leaf spectral data was reduced from 2151 variables to 15 variables by using the principal component analysis (PCA) technique. The PCA is a widely used dimension reduction approach that enables the extraction of critical information from datasets and avoids redundancies without losing any valuable details per sample data point. Initially, 50 principal components were computed using python's scikit-learn library (Version-1.0.1).[21] However, only 15 principal components were used since they were found to explain over 99% of the variation in our spectral dataset.

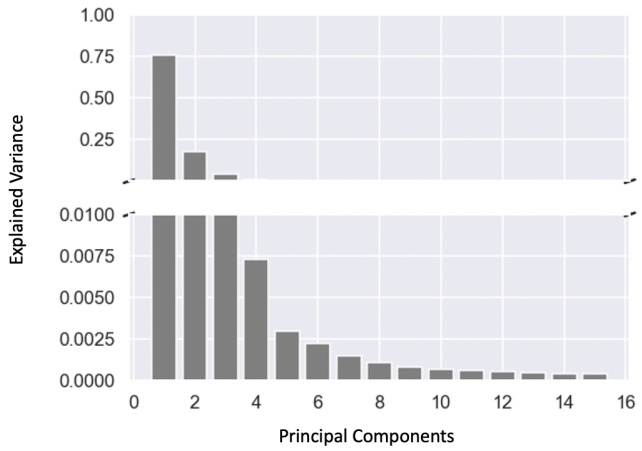

Figure 3. Principal Components Analysis. 50 components were calculated, from which 15 components were selected and used in the downstream analysis since they explained over 99% of the variance in the dataset.

4.1.2 Ensemble Machine Learning Models

Python scikit-learn's Extra Trees Regressor (version-0.22)[14] was used to fit individual decision trees with training data (70% of entire dataset) and ensemble the results for the final prediction of flowering time (days)(in sorghum. Extra Trees was used instead of a widely used Random Forest (RF) since the dataset was very small, and it did not make sense to build decision trees using bootstrapped subsamples as it is implemented with RF. Instead Extra Trees uses the entire training samples to fit multiple decision trees and ensemble for final prediction. In addition, Extra Trees Regressor differ from other ensemble models in that it adds randomization in both selecting attributes and cut-point when splitting a tree node.[25] In multiple applications, Extra Trees Regressor has been shown to outperform other ensemble models,[26] it restricts the issue of overfitting,[25] and it is a bit faster than other regression ensemble models.[27]

ACKNOWLEDGMENTS

AN is supported by a grant from the NIH Institutional Development Award Program (IDeA) Networks of Biomedical Research Excellence (INBRE; P20-GM103427). The data employed in this study was supported by the Office of Science (BER), U.S. Department of Energy, Grant no. DE-SC0020355; the Foundation for Food and Agriculture Research (ID: 602757) and National Science Foundation under grant OIA-1826781 to JCS. The author also want thank Marcin, Mackenzie, Christine and other Schnable lab members for their hard work to collect field data.

REFERENCES

[1] Nazir, H., Baloch, M. S., Yousaf, M., Naeem, M., Khakwani, A., and Begum, I., "Performance of sorghum varieties in potohar region," *Gomal University Journal of Research* **27**(2), 201–223 (2011).

[2] Zwiener, M., "Phenotypic plasticity of diverse sorghum varieties in response to nitrogen deficit stress," (2021).

[3] Olson, S. N., Ritter, K., Medley, J., Wilson, T., Rooney, W. L., and Mullet, J. E., "Energy sorghum hybrids: functional dynamics of high nitrogen use efficiency," *Biomass and Bioenergy* **56**, 307–316 (2013).

[4] Ogbaga, C. C., Stepien, P., and Johnson, G. N., "Sorghum (sorghum bicolor) varieties adopt strongly contrasting strategies in response to drought," *Physiologia plantarum* **152**(2), 389–401 (2014).

[5] Traore, A. and Maranville, J. W., "Nitrate reductase activity of diverse grain sorghum genotypes and its relationship to nitrogen use efficiency," *Agronomy Journal* **91**(5), 863–869 (1999).

[6] Clerget, B., Sidibe, M., Bueno, C., Grenier, C., Kawakata, T., Domingo, A., Layaoen, H., Palacios, N., Bernal, J., Trouche, G., et al., "Crop-photoperiodism model 2.0 for the flowering time of sorghum and rice that includes daily changes in sunrise and sunset times and temperature acclimation," *Annals of Botany* (2021).

[7] Mural, R. V., Grzybowski, M., Miao, C., Damke, A., Sapkota, S., Boyles, R. E., Salas Fernandez, M. G., Schnable, P. S., Sigmon, B., Kresovich, S., et al., "Meta-analysis identifies pleiotropic loci controlling phenotypic trade-offs in sorghum," *Genetics* **218**(3), iyab087 (2021).

[8] Cober, E. R., Curtis, D. F., Stewart, D. W., and Morrison, M. J., "Quantifying the effects of photoperiod, temperature and daily irradiance on flowering time of soybean isolines," *Plants* **3**(4), 476–497 (2014).

[9] Chauhan, Y. S., Ryan, M., Chandra, S., and Sadras, V. O., "Accounting for soil moisture improves prediction of flowering time in chickpea and wheat," *Scientific reports* **9**(1), 1–11 (2019).

[10] Newman, S. J. and Furbank, R. T., "A multiple species, continent-wide, million-phenotype agronomic plant dataset," *Scientific data* **8**(1), 1–8 (2021).

[11] Grzybowski, M., Wijewardane, N. K., Atefi, A., Ge, Y., and Schnable, J. C., "Hyperspectral reflectance-based phenotyping for quantitative genetics in crops: Progress and challenges," *Plant Communications* , 100209 (2021).

[12] Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., and Eskandari, M., "Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean," *Frontiers in plant science* **11**, 2169 (2021).

[13] Montesinos-López, O. A., Montesinos-López, A., Crossa, J., De Los Campos, G., Alvarado, G., Suchismita, M., Rutkoski, J., González-Pérez, L., and Burgueño, J., "Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data," *Plant methods* **13**(1), 1–23 (2017).

[14] Geurts, P., Ernst, D., and Wehenkel, L., "Extremely randomized trees," *Machine learning* **63**(1), 3–42 (2006).

[15] Richardson, M., "Principal component analysis," *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si* **6**, 16 (2009).

[16] Jiang, Z., Huete, A. R., Didan, K., and Miura, T., "Development of a two-band enhanced vegetation index without a blue band," *Remote sensing of Environment* **112**(10), 3833–3845 (2008).

[17] Sims, D. A. and Gamon, J. A., "Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages," *Remote sensing of environment* **81**(2-3), 337–354 (2002).

[18] Kaufman, Y. J. and Tanre, D., "Atmospherically resistant vegetation index (arvi) for eos-modis," *IEEE transactions on Geoscience and Remote Sensing* **30**(2), 261–270 (1992).

[19] Vogelmann, J., Rock, B., and Moss, D., "Red edge spectral measurements from sugar maple leaves," *TitleREMOTE SENSING* **14**(8), 1563–1575 (1993).

[20] Kalita, D. N., "A new curve for encapsulating the normalized difference vegetation index,"

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[22] van Bezouw, R. F., Keurentjes, J. J., Harbinson, J., and Aarts, M. G., "Converging phenomics and genomics to study natural variation in plant photosynthetic efficiency," *The Plant Journal* **97**(1), 112–133 (2019).

[23] Casa, A. M., Pressoir, G., Brown, P. J., Mitchell, S. E., Rooney, W. L., Tuinstra, M. R., Franks, C. D., and Kresovich, S., "Community resources and strategies for association mapping in sorghum," (2008).

[24] Ge, Y., Atefi, A., Zhang, H., Miao, C., Ramamurthy, R. K., Sigmon, B., Yang, J., and Schnable, J. C., "High-throughput analysis of leaf physiological and chemical traits with vis–nir–swir spectroscopy: a case study with a maize diversity panel," *Plant methods* **15**(1), 1–12 (2019).

[25] Kronberg, E. A., Gastaldello, F., Haaland, S., Smirnov, A., Berrendorf, M., Ghizzardi, S., Kuntz, K., Sivadas, N., Allen, R. C., Tiengo, A., et al., "Prediction and understanding of soft-proton contamination in xmm-newton: A machine learning approach," *The Astrophysical Journal* **903**(2), 89 (2020).

[26] Bouktif, S., Fiaz, A., Ouni, A., and Serhani, M. A., "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies* **11**(7), 1636 (2018).

[27] Thiagarajan, B., Srinivasan, L., Sharma, A. V., Sreekanthan, D., and Vijayaraghavan, V., "A machine learning approach for prediction of on-time performance of flights," in [*2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*], 1–6, IEEE (2017).