Feasibility of PlanetScope satellite data and random forest machine learning model for soybean yield prediction at last three growth stages

Jitender Rathore[1], Maaz Gardezi[2], Olga Walsh[1], Deepak Joshi[3], Sheetal Kumari[1], David Clay[4]

[1]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, Virginia, USA

[2]Department of Sociology, Virginia Tech, Blacksburg, Virginia, USA

[3]College of Agriculture, Arkansas State University

[4]Department of Horticulture & Plant Science South, Dakota State University, South Dakota, USA

A paper from the Proceedings of the
16th International Conference on Precision Agriculture
21-24 July 2024
Manhattan, Kansas, United States

Abstract.

Soybean, a nutrient-rich legume plant, plays a significant role in US agriculture, especially in Livestock farming. However, its complex structures pose a challenge for farmers. Soybean is a determinate and indeterminate crop, which means that it has a complicated mechanism for yield formation. Several factors such as weather, soil, vegetation, and management practices can impact its yield, making it challenging to determine the optimal time for harvesting to maximize revenue. Overmature harvests result in poor yield, while under-mature harvests increase the risk of disease infection. The existing methods for determining yield are tedious, expensive, and do not account for spatial variations. Therefore, developing a spatial-temporal soybean yield prediction model for precision harvesting is essential. This study is conducted at the farm level in Miner County of South Dakota, US. To achieve the yield prediction goal, this study examines the feasibility of using high-resolution PlanetScope satellite data to predict soybean yield at the farm level. The study considered six significant growth stages, covering the period between planting and harvesting in 2019 and 2021. PlanetScope satellite data was collected during the crop seasons from April to September. Six cloud-free images were captured at each of the last three growth stages: R2/R3 (10 August), R4/R5 (28 August), and R6/R7 (10 September). Vegetation indices (VIs) and a Random Forest (RF) machine learning model were used to predict soybean yield at different growth stages. Various VIs were derived from multispectral imageries such as the normalized differential vegetation index (NDVI), difference vegetation index (DVI), and visual atmospheric resistance index (VARI). The study found that VIs correlation ranges vary at different growth stages and the highest correlation is estimated at maturity (R4/R5) stages. NDVI

(correlation, r = -0.48- 0.82), VARI (correlation, r = -0.51 - 0.78), and DVI (correlation, r = -0.08 – 0.84) VIs were more correlated with yield at different stages of soybean growth. RF model used to predict yield at different growth stages in 2019 and 2021. The performance of the RF model was validated using R-squared (R^2), and RMSE. The R^2 scores differed in each growth stage. The highest (R^2 =0.73, RMSE = 1.89) and lowest (R2 = 0.52, RMSE = 5.55) R^2 scores were obtained at R4/R5 and R6/R7 stages in 2021, respectively. Soybean yield could be predicted accurately at different growth stages. This approach can help farmers determine the optimal time for harvesting and maximize their revenue while reducing the risk of disease infection and artificial drying expenses.

Keywords.
Soybean, PlanetScope satellite, Vegetation Index, and Random Forest

Introduction

Soybean (Glycine max) is a leading growing crop in the United States (US) and belongs to the legume family. The US is also the leading producer of soybean crops in the world (M. Sedibe et al., 2023). South Dakota (SD) is one of the major producers of soybeans in the US and soybean yield was forecasted at 221 million bushels in 2023 (Statista, 2024). Soybean is a determinate and indeterminate crop, which shows that it has a complicated mechanism for yield formation (Nleya et al., 2020). Several factors such as weather, soil, vegetation, and management practices can impact its yield, making it challenging to determine the optimal time for harvesting to maximize revenue (Holzman et al., 2014; Satir & Berberoglu, 2016; Joshi et al., 2023). Overmature harvests result in poor yield, while under-mature harvests increase the risk of disease infection (Saryoko et al., 2017). The existing methods for determining yield are tedious, expensive, and do not account for spatial variations. Many researchers have proven that remote sensing-based machine learning models are capable of estimating and predicting crop yield accurately. For instance - Li et al., (2023) applied satellite imageries, and machine and deep learning techniques to predict soybean and corn in the Corn Belt region in the US at the county level. Kaul et al., (2005) utilized an artificial neural network (ANN) for corn and soybean yield prediction. Similarly, Abrougui et al., (2019); Ma et al., (2021); Q. Li et al., (2023) and Rajakumaran et al., (2024) have successfully attempted crop yield prediction through machine learning at different growth stages. Therefore, developing a spatial-temporal soybean yield prediction model for precision harvesting is essential. To achieve this, this study has been conducted with main objectives, which are given below.

Objectives

1. To forecast soybean yields at the farm level from early to late stages of growth.
2. To assess a machine learning-driven random forest (RF) model using multispectral satellite imagery for predicting soybean yield.

Material and Methods

Study area and Data collection

The research was conducted at coordinates 44°04'15.1"N 97°39'38.2"W in Miner County, South Dakota. This location served as a trial site for soybean cultivation in 2019 and 2021. Soybeans were harvested using a harvester combiner equipped with a 9.2-width header (Joshi et al., 2023). Satellite data from Planet was collected during the last three growth stages: R2/R3 (10 August), R4/R5 (28 August), and R6/R7 (10 September) for both years, and the resolution of these satellite imagery was 3.2 meter. Harvested yield data was obtained through GPS- GPS-based tracker.

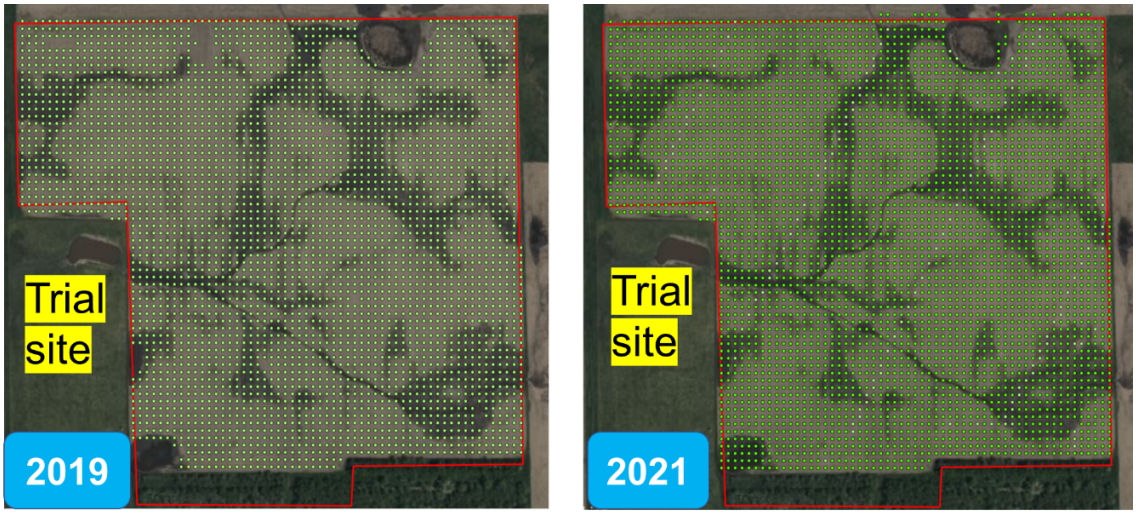

Fig 1. Study area in Miner County, South Dakota, and GPS-based harvested yield data in the years 2019 and 2021.

Data pre-processing

In Figure 2, the methodology chart illustrates the process. Initially, a 10x10 meter grid was established using the open-source QGIS software. Subsequently, the reflectance values of each band were extracted from the satellite imagery. Finally, Vegetation Indices (VIs) such as the normalized differential vegetation index (NDVI), difference vegetation index (DVI), and visual atmospheric resistance index (VARI) were computed using a raster calculator in QGIS. These indices were derived from multispectral images captured at selected growth stages.

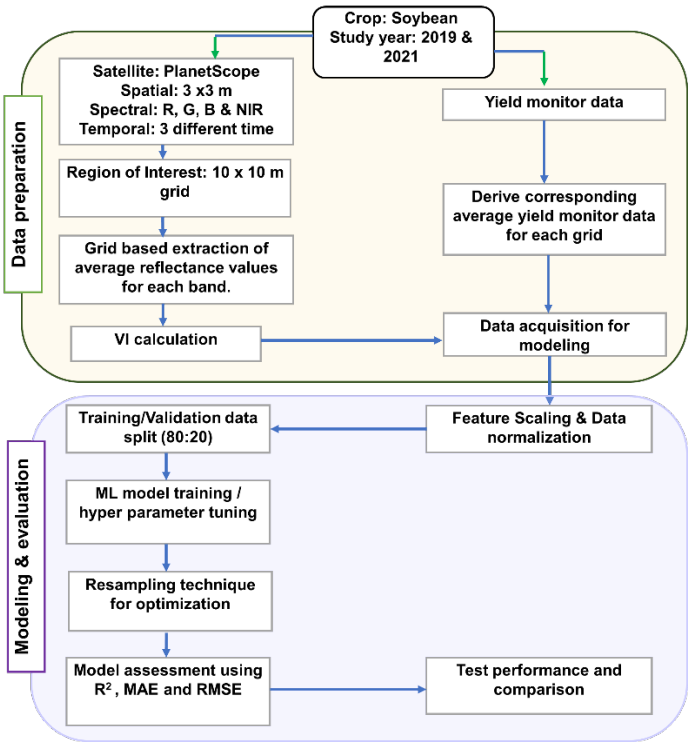

Fig 2. Methods and Methodology chart.

Random forest: Model build and validation

Random Forest (RF) is a widely favored machine learning method for predicting crop yields due to its high accuracy and ease of use. Previous studies by Abrougui et al., (2019); Joshi et al., (2023) have highlighted its suitability for crop yield prediction. In our study, we employed RF as a regressor, taking advantage of its applicability for both classification and regression tasks. The RF model was developed using harvested yield and VIs data, with the soybean harvested yield and VIs data divided into 80/20 ratios for training and testing after feature scaling. The RF model's performance was assessed using R2, MAE, and RMSE metrics.

Results and Discussion

Spatial distribution of Soybean yield

In Figure 3, the yield distribution for the years 2019 and 2021 is depicted. The image consists of a histogram and boxplot on the left, illustrating the mean yield data, while on the right, the spatial distribution of yield demonstrates the specific locations and the yield obtained per bushel. In the visualization, dark blue indicates a lower yield, while yellow represents the highest yield at a particular location. It is evident that in 2021, the yield quantity was lower compared to 2019.

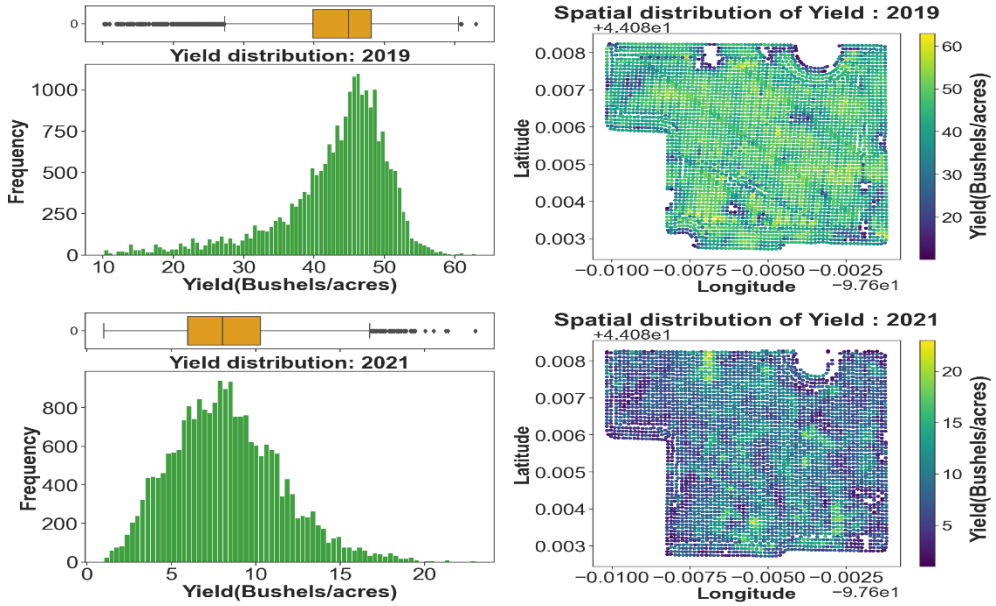

Fig 3. Yield distribution in the years of 2019 and 2021

Soybean yield prediction growthwise using Random forest

In Figure 4, the three different growth stages and predicted yield using RF are depicted. The model's performance was evaluated using R2, MAE, and RMSE in the years 2019 and 2021. In 2019, the R2 value was highest (R2 = 0.64) and the RMSE was lowest (RMSE = 4.82) at the R2/R3 growth stage, while the lowest R2 (R2 = 0.52) and the highest RMSE (RMSE = 5.55) were observed at the R6/R7 growth stage. However, in 2021, the R2 was highest (R2 = 0.73) and the RMSE was lowest (RMSE = 1.75) at the R4/R5 growth stage, while the lowest R2 (R2 = 0.68) and the highest RMSE (RMSE = 1.95) were observed at the R6/R7 growth stage. Furthermore, it was noted that 2021 was a drought year in Miner County. The growth stages R2/R3 and R4/R5 are considered feasible for predicting yield and achieving higher accuracy. It's important to note that other factors such as temperature and precipitation limitations also affect yield in this study.

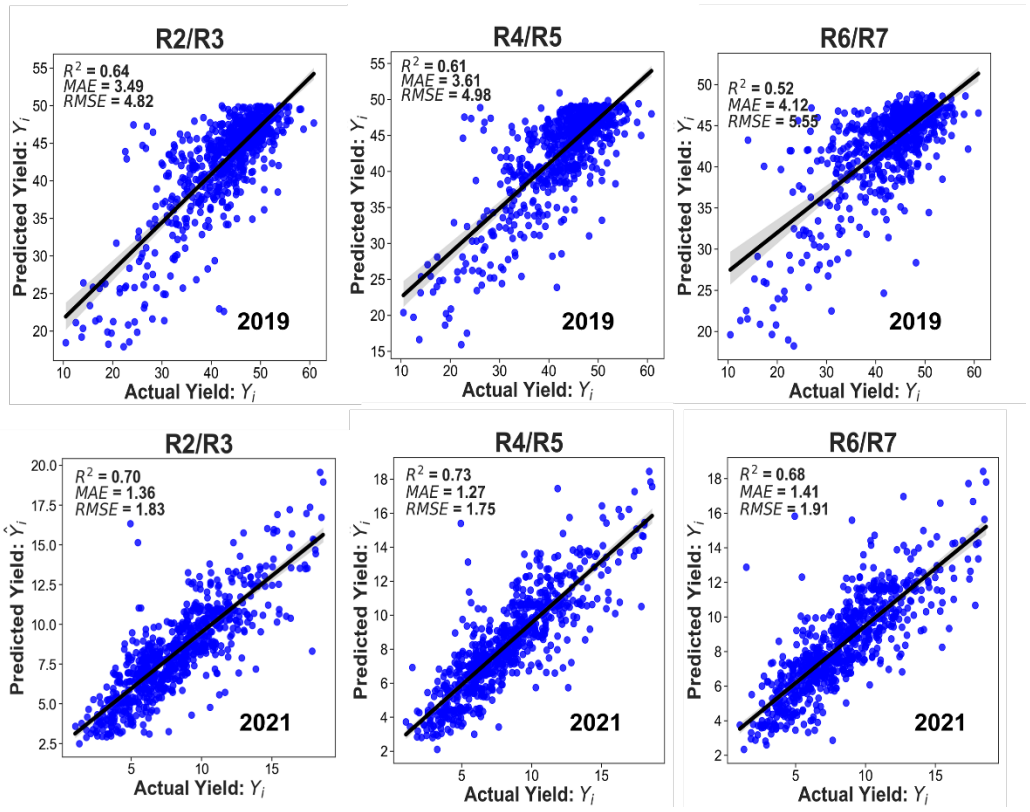

Fig 4. Soybean yield prediction at three different growing stages (R1/R2, R4/R5 and R6/R7) using RF model and model assessment though R2, MAE, and RMSE

Conclusion

Conventional yield estimation is tedious and could be subjective. Satellite imagery has demonstrated potential for predicting yields over space and time. The inclusion of weather and soil data in ML models can further improve prediction accuracy. Satellite imagery and ML modeling can support precision/zone-specific soybean yield harvest.

Acknowledgments

We would like to thank the many farmers who are participating in our project. This research is based upon work supported by the National Science Foundation under Grant Number 2202706 / 2026431. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Abrougui, K., Gabsi, K., Mercatoris, B., Khemis, C., Amami, R., & Chehaibi, S. (2019). Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). *Soil and Tillage Research*, *190*, 202–208. https://doi.org/10.1016/j.still.2019.01.011

Holzman, M. E., Rivas, R., & Piccolo, M. C. (2014). Estimating soil moisture and the relationship with crop yield using surface temperature and vegetation index. *International Journal of Applied Earth Observation and Geoinformation*, *28*, 181–192. https://doi.org/10.1016/j.jag.2013.12.006

Joshi, D. R., Clay, S. A., Sharma, P., Rekabdarkolaee, H. M., Kharel, T., Rizzo, D. M., Thapa, R., & Clay, D. E. (2023). Artificial intelligence and satellite-based remote sensing can be used to predict soybean (*Glycine max*) yield. *Agronomy Journal*, agj2.21473. https://doi.org/10.1002/agj2.21473

Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems*, *85*(1), 1–18. https://doi.org/10.1016/j.agsy.2004.07.009

Li, Q., Xu, S., Zhuang, J., Liu, J., Zhou, Y., & Zhang, Z. (2023). Ensemble learning prediction of soybean yields in China based on meteorological data. *Journal of Integrative Agriculture*, *22*(6), 1909–1927. https://doi.org/10.1016/j.jia.2023.02.011

Li, Y., Zeng, H., Zhang, M., Wu, B., Zhao, Y., Yao, X., Cheng, T., Qin, X., & Wu, F. (2023). A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. *International Journal of Applied Earth Observation and Geoinformation*, *118*, 103269. https://doi.org/10.1016/j.jag.2023.103269

M. Sedibe, M., M. Mofokeng, A., & R. Masvodza, D. (2023). Soybean Production, Constraints, and Future Prospects in Poorer Countries: A Review. In M. Hasanuzzaman (Ed.), *Production and Utilization of Legumes—Progress and Prospects*. IntechOpen. https://doi.org/10.5772/intechopen.109516

Ma, Y., Zhang, Z., Kang, Y., & Özdoğan, M. (2021). Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment*, *259*, 112408. https://doi.org/10.1016/j.rse.2021.112408

Nleya, T., Schutte, M., Clay, D., Reicks, G., & Mueller, N. (2020). Planting date, cultivar, seed treatment, and seeding rate effects on soybean growth and yield. *Agrosystems, Geosciences & Environment*, *3*(1), e20045. https://doi.org/10.1002/agg2.20045

Rajakumaran, M., Arulselvan, G., Subashree, S., & Sindhuja, R. (2024). Crop yield prediction using multi-attribute weighted tree-based support vector machine. *Measurement: Sensors*, *31*, 101002. https://doi.org/10.1016/j.measen.2023.101002

Saryoko, A., Homma, K., Lubis, I., & Shiraiwa, T. (2017). Plant development and yield components under a tropical environment in soybean cultivars with temperate and tropical origins. *Plant Production Science*, *20*(4), 375–383. https://doi.org/10.1080/1343943X.2017.1356203

Satir, O., & Berberoglu, S. (2016). Crop yield prediction under soil salinity using satellite-derived vegetation indices. *Field Crops Research*, *192*, 134–143. https://doi.org/10.1016/j.fcr.2016.04.028

Statista. (2024, February 7). *Major soybean-producing U.S. states from 2019 to 2023*. https://www.statista.com/statistics/192076/top-10-soybean-producing-us-states/