44. Integrating seeding maps, soil properties, elevation, and remote sensing In machine learning models to forecast sorghum yield

LGG Sterle [1], JP Molin [1], R Canal Filho [1] *, ERO Silva [1], MCF Wei [1]

[1]*Laboratory of Precision Agriculture, "Luiz de Queiroz" College of Agriculture – University of São Paulo. Piracicaba, SP - Brazil*

Introduction

The detection of crop spatial variability in Precision Agriculture (PA) is achieved by integrating data layers. The use of machine learning (ML) algorithms enables more accurate predictions. Remote sensing (RS) tools at the orbital level are widely used to detect crop variability. RS has been a key focus in agricultural monitoring and is considered a promising resource for decision-making in farm management (Song et al., 2009). Based on spectral data from plants—particularly visible bands of the electromagnetic spectrum, RGB, and near-infrared (NIR) bands—it is possible to monitor crop growth cycles, vegetation cover, soil moisture, nitrogen stress, and overall crop health and yield (Cao et al., 2020; Cicore et al., 2016). Despite its economic importance, sorghum has been the subject of limited studies involving data integration. Therefore, it is timely to investigate the potential of yield prediction in sorghum using data fusion approaches. A key hypothesis of this work is that the inclusion of seeding as-applied data—often overlooked—along with other layers, leads to more precise yield prediction. Unlike traditional approaches that assume homogeneous seed distribution, this study emphasizes that even in the absence of intentional variable rate seeding, significant variation in plant density can occur, affecting yield outcomes. The objective of this study is to evaluate the impact of sorghum seeding data and its interaction with other data layers in predicting sorghum yield using machine learning algorithms.

Material and Methods

The sorghum data were collected from a commercial field in Itaí, São Paulo state, Brazil, over a pivot-irrigated area (54.6 ha) during the 2024 growing season. Yield data were collected from harvesters equipped with yield monitors, and seeding data were collected from an instrumented seeder measuring the seed rate applied to each row. Additionally, remote sensing data were obtained from the Sentinel-2A satellite at five different stages using the normalized difference vegetation index (NDVI) on the following dates: February 3, 2024 (1_NDVI = Tillering), February 23, 2024 (2_NDVI = Stem elongation), March 4, 2024 (3_NDVI = Flowering), March 14, 2024 (4_NDVI = Grain filling), and April 3, 2024 (5_NDVI = Physiological maturity), during the crop cycle. Soil chemical attributes (0–0.2 m depth) were sampled in a 3 ha grid before seeding. The following attributes were analyzed: B, Ca, CEC (cation exchange capacity), Cu, K, Mg, Mn, OM (organic matter), P, pH, S, V (base saturation), and Zn. Elevation data were obtained from the GNSS receiver on the harvester, and the apparent soil electrical conductivity (ECa) map was generated at two depths: 0–0.75 m and 0–1.50 m. The data were interpolated using IDW at a 10-meter pixel resolution, followed by exploratory data analysis and Spearman correlation between the layers. The correlation was classified according to Rumsey (2023). Machine learning models were used to predict yield based on the selected data layers. The models were compared with and without the seeding layer, and the predicted data were compared to the harvester yield map. In this study, we compared random forest (RF), gradient boosting (GB), and support vector machine (SVM). The models were tuned based on the following metrics: MAE (Mean Absolute Error), RMSE (Root Mean Square Error), R^2 (Coefficient of Determination), and Lin's CCC (Lin's Concordance Correlation Coefficient). Finally, cross-validation plots were generated for each model, as well as variable importance charts.

Results and Discussion

The following correlations were found: seeding rate vs. yield showed a moderate negative correlation of -0.56, with statistical significance ($p < 0.00001$). This inverse relationship suggests that an increase in seeding density may be associated with a reduction in crop yield. Elevation vs. yield presented a moderate positive correlation of 0.58. The correlation between yield and NDVI at stages 3 and 4 (3_NDVI and 4_NDVI) were -0.42 and -0.40, respectively (moderate and negative). Additionally, yield vs. B showed a moderate negative correlation of -0.47, while yield vs. K showed a moderate positive correlation of 0.46.

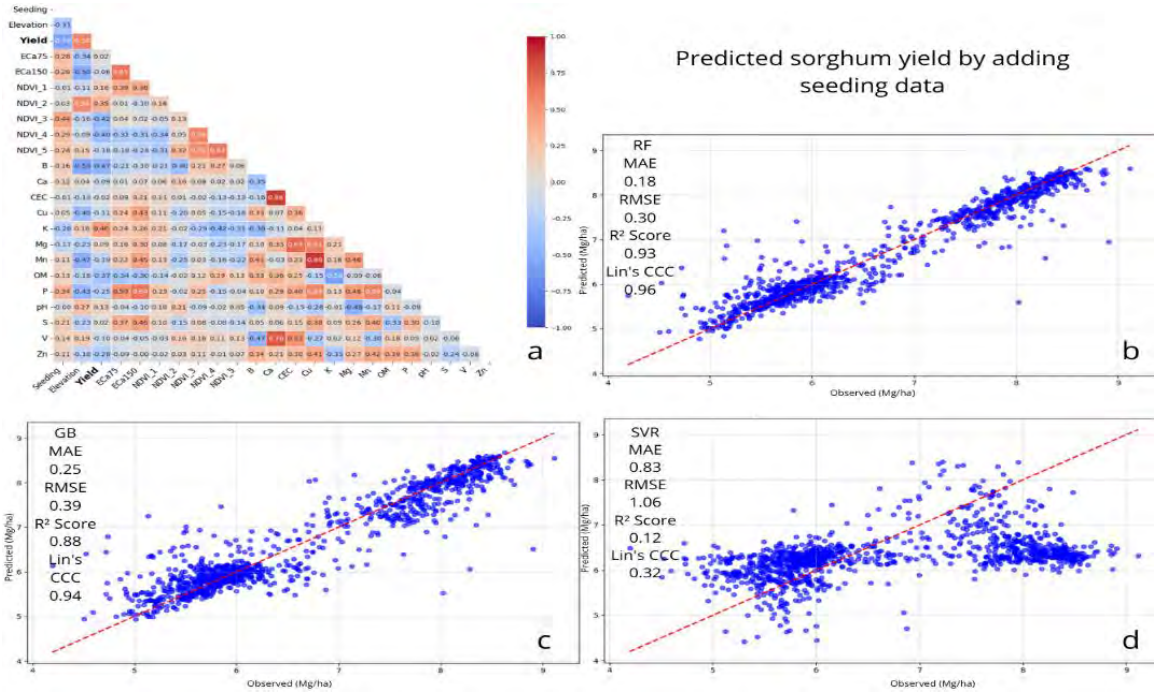

Figure 01. Spearman correlation and performance comparison of machine learning models
a = spearman matrix correlation; b = RF model performance; c = GB model performance; d = SVM model performance.

The results showed that the RF model was the most accurate, with the lowest error (MAE of 0.18 and RMSE of 0.30) and high concordance (Lin's CCC of 0.96), regardless of the seeding map layer. The GB model also performed well, with a MAE of 0.25 and Lin's CCC of 0.94, but was inferior to RF. The SVR showed the worst performance, with a MAE of 0.83 and Lin's CCC of 0.32, indicating low predictive power.

Conclusion

In this study, we found a moderate negative correlation (-0.56) between seeding rate and yield, suggesting that adjusting planting density can influence yield. However, incorporating the seeding map layer did not substantially improve model performance to predict sorghum yield. Among the tested approaches, RF was the most effective model for yield prediction.

References

X. Song, J. Wang, W. Huang, L. Liu, G. Yan, and R. Pu. The delineation of agricultural management zones with high resolution remotely sensed data. Precision Agriculture, 10:471–487, 2009.

J. Cao, Z. Zhang, F. Tao, L. Zhang, Y. Luo, J. Han, and Z. Li. Identifying the contributions of multi-source data for winter wheat yield prediction in China. Remote Sensing, 12(5):750, 2020.

P. Cicore, J. Serrano, S. Shahidian, A. Sousa, J. L. Costa, and J. R. M. da Silva. Assessment of the spatial variability in tall wheatgrass forage using LANDSAT 8 satellite imagery to delineate potential management zones. Environmental Monitoring and Assessment, 188:1–11, 2016.