Large-scale sugarcane yield prediction across regions by integrating multi-source remote sensing and machine learning

C. Ferraz[1], F. Serra-Burriel[1], M. Cabrera[1], R. Fortes[1]

[1] *HEMAV Technology S.L. (LAYERS Crop Intelligence), Sant Joan Despí, Barcelona, Spain*

** Corresponding author: cferraz@layerscrop.com*

Abstract. Accurate within-season yield estimation is fundamental for agricultural management, yet most models struggle to scale across regions. In this study, the combination of multi-source remote sensing data with machine learning was shown to improve sugarcane yield prediction at field scale. Using 97,962 field-seasons across regions in South America, Central America and Africa, an extreme gradient boosting (XGBoost) model was developed, achieving a mean absolute error (MAE) of 9.94 t ha^{-1} (mean absolute percentage error, MAPE, of 16.45%) and a coefficient of determination (R^2) of 0.80 on a held-out test set. The maximum normalised difference red-edge (NDRE) index, maximum humidity and the number of cuts (ratoon) were the main yield-determining factors. In regions with little local data, incremental learning recovered predictive skill (R^2 from near zero to 0.46–0.52), outperforming single-user models. The approach provides an operationally robust tool for large-scale sugarcane yield estimation.

Keywords. agricultural analytics; earth observation; feature extraction; precision farming; crop forecasting

Introduction

Sugarcane (Saccharum officinarum L.) is a multifunctional crop used for sugar and renewable bioenergy production and a strategic component of the sugar-energy industry. Early and accurate yield estimation before harvest supports planning and decision-making in that industry. Yield depends on edaphoclimatic factors and agronomic management, including growing degree days, crop variety and number of cuts (Hammer et al. 2020). Previous studies were largely centred on hyperlocal plot repetitions, yielding models that were accurate but not feasible at scale (Everingham et al. 2016). Data-science techniques applied to large earth-observation, climate and management datasets can instead capture the non-linear, hierarchical relationships that drive yield. Three research questions were addressed: (i) whether a global model from multi-source data outperforms local models, particularly where historical data are scarce; (ii) which factors most influence yield across regions; and (iii) to what extent incremental learning improves performance in new areas. The primary objective was therefore to develop and evaluate an operationally robust global model that predicts field-level sugarcane yield across diverse regions and remains accurate where local historical data are limited.

Materials and methods

The study area covered multiple sugarcane regions across South America, Central America and East and West Africa, for seasons from 2019 to 2023. Four data sources were integrated: Sentinel-2 (Level-2A, 10–60 m) and Sentinel-1 (GRD Level-1, VV/VH polarisations, 10 m) imagery (ESA 2015); climatic variables from a commercial historical weather API fusing observations from more than 120,000 weather stations with satellite and radar imagery and reanalysis datasets (including ERA5), gap-filled and quality-controlled with machine-learning techniques; field-management records (yield, plantation and harvest dates, variety); and soil type from a global soil map. After quality control (removal of fields < 0.5 ha, life-cycle validation between 6 and 36 months, and yields above 250 t ha^{-1} excluded), the final dataset comprised 97,962 records with 46 features and a mean yield of 83.57 t ha^{-1}.

The dataset drew representation from South and Central America and from both East and West Africa, with concentrations in the main sugarcane-producing regions. For each plot, Sentinel-2 spectral indices (the red-edge NDRE, NDVI and the water-status index NDWI) and Sentinel-1 SAR backscatter were aggregated over the crop cycle together with the climatic, management and soil variables; the optical and all-weather radar sources were processed separately because optical imagery is affected by clouds whereas SAR is not.

Three models were compared — k-nearest neighbours, random forest (Breiman 2001) and XGBoost (Chen and Guestrin 2016). XGBoost was selected for its computational efficiency, native support for incremental learning and built-in handling of missing values; it was also markedly faster to train than the alternatives (minutes rather than hours), which is practical for operational retraining. Three experiments were conducted: global learning (20% of the samples held out for final evaluation, with cross-validation on the remainder); incremental learning (a regional model updated with local samples); and zero-sample learning (regional prediction with no local training data). Potential dataset biases (geographical and measurement variation) were systematically identified and mitigated, although cloud interference in optical imagery remained an acknowledged constraint, partly offset by the all-weather Sentinel-1 SAR signal. Performance was assessed with R^2, MAE and MAPE (Eq. 1):

$$\mathrm{MAPE} = (1/n)\ \Sigma\ |y_i - \hat{y}_i| \ /\ y_i \times 100 \qquad (1)$$

where y_i and $\hat{y}_i$ denote the observed and predicted yields and n is the number of records.

Results and discussion

Across the compared models, R^2 ranged from 0.72 to 0.80, MAE from 9.94 to 12.43 t ha^{-1}, RMSE from 14.91 to 17.40 t ha^{-1} and MAPE from 15.98 to 21.46%; the general XGBoost model performed best (Table 1: R^2 0.80, MAE 9.94 t ha^{-1}, MAPE 16.45%). On regional subsets, R^2 ranged from 0.70 (an African subset; MAE 7.46 t ha^{-1}) to 0.93 (a South-American subset; MAE 4.97 t ha^{-1}; Fig. 1). At field scale, an R^2 of 0.80 with an MAE of 9.94 t ha^{-1} and a MAPE of 16.45% yields pre-harvest tonnage estimates at field level; crucially, this accuracy held across climatically and managerially diverse regions rather than a single setting, indicating operational robustness rather than overfitting.

Table 1 Performance of the compared models (general experiment)

Model	MAE (t ha⁻¹)	MAPE (%)	R^2	RMSE (t ha⁻¹)
Random forest	12.43	15.98	0.72	17.40
XGBoost (best)	9.94	16.45	0.80	14.91

The maximum NDRE index, maximum humidity and the number of cuts (ratoon) were the most relevant yield predictors across regions, in line with Peloia et al. (2018) and extending the neural-network results of Cabrera Dengra et al. (2022) for sugar beet; among the spectral indices, NDWI — which tracks canopy water status through the crop cycle — was prominent. The RMSE of the general model (14.91–17.40 t ha⁻¹) compares favourably with the 5.6–30 t ha⁻¹ reported by Hammer et al. (2020); Everingham (2014) obtained 6.3–8.0 t ha⁻¹ but at regional rather than field scale, a less demanding target. With no local data, performance was poor (R^2 −0.08 to 0.14), but incremental learning recovered substantial skill (R^2 0.46–0.52; Fig. 2), confirming a viable pathway for geographical expansion (Marin et al. 2019); in one regional deployment, incremental updating improved the MAE by 1.29 t ha⁻¹ relative to the prior local model. Examining the predictors, the maximum NDRE index and water-status indices showed the strongest positive correlation with yield, whereas maximum humidity and mean temperature were the most negatively correlated, consistent with the importance ranking of the selected model.

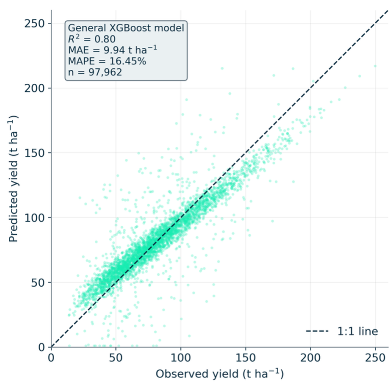

Fig. 1 Predicted versus observed yield for the general XGBoost model (R^2 0.80, MAE 9.94 t ha⁻¹, n ≈ 97,962); the dashed line is the 1:1 reference

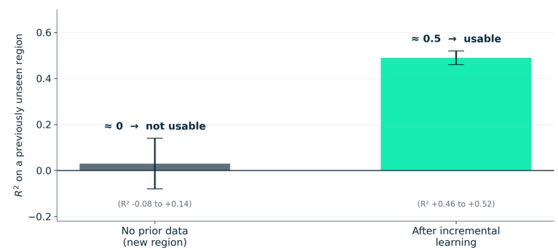

Fig. 2 R^2 on a new region: no prior data (≈ 0, not usable) versus after incremental learning (≈ 0.5, usable)

Conclusions

A global machine-learning framework integrating Sentinel-1/2, climate and field-management data was developed for field-level sugarcane yield prediction from nearly 100,000 field-seasons. The XGBoost model achieved consistent accuracy (MAE 9.94 t ha⁻¹; R^2 0.80) and improved on single-user models, particularly for users with little historical data. The maximum NDRE index, maximum humidity and the number of cuts (ratoon) were the most relevant predictors. Incremental learning raised R^2 from near zero to 0.46–0.52 in new regions, supporting deployment at scale. Future work will standardise growing-degree-day calculation, incorporate irrigation and soil-humidity data, and improve the handling of cloud interference in optical imagery.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Cabrera Dengra, M., Ferraz Pueyo, C., Pajuelo Madrigal, V., Moreno Heras, L., Inunciaga Leston, G., & Fortes, R. (2022). Use of MLP neural networks for sucrose yield prediction in sugar beet. In Proceedings of the 15th International Conference on Precision Agriculture.

Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

ESA (2015). Sentinel-2 product specification. European Space Agency. https://sentinels.copernicus.eu/web/sentinel/home. Last accessed October 2024.

Everingham, Y. J. (2014). A statistical approach for identifying important climatic influences on sugarcane yields. Proceedings of the Australian Society of Sugar Cane Technologists, 36, 8–15.

Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development, 36, 27.

Hammer, R., Sentelhas, P. C., & Mariano, J. C. (2020). Sugarcane yield prediction through data mining and crop simulation models. Sugar Tech, 22, 216–225.

Marin, F. R., Edreira, J., Andrade, J., & Grassini, P. (2019). On-farm sugarcane yield and yield components as influenced by number of harvests. Field Crops Research, 240, 134–142.

Peloia, P., Bocca, F., & Rodrigues, L. (2018). Identification of patterns for increasing production with decision trees in sugarcane mill data. Scientia Agricola, 75(4), 281–289.