Estimation of Sugarcane Yield Based on Phenological Feature Extraction from Time-Series Sentinel-1 Images and Machine Learning

Hanyu Xue [a,b], Xingang Xu [b,*], Guijun Yang [b], Zhenheng Xu [b], Xiaoyu Song [b], Liping Chen [a,b]

a College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China;

b Key Laboratory of Quantitative Remote Sensing in Agriculture of Ministry of Agriculture and Rural Affairs, Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China;

ABSTRACT

Optical spectral imagery is often unavailable or temporally discontinuous in sugarcane-growing regions because of persistent cloud cover and frequent rainfall, which limits the application of conventional optical remote-sensing yield estimation methods. To address sugarcane yield estimation under optical data-deficient conditions, this study proposes a machine-learning framework integrating Sentinel-1 SAR time-series observations, ERA5-Land meteorological variables, and ancillary terrain and geographic information. Guangxi Zhuang Autonomous Region, China, was selected as the study area. Historical county-level sugarcane yield records from 2020 to 2024 and field yield measurements collected in 2025 were used as samples. Sentinel-1 VV/VH-derived temporal features were processed using Savitzky-Golay filtering, and temporal shape differences were characterized with TW-shapeDTW. ERA5-Land temperature and precipitation variables were introduced to describe agro-meteorological conditions, while terrain and geographic partitioning information were used to represent regional heterogeneity. Random Forest models were then constructed under different feature settings. Results showed that the final model achieved the best estimation performance, with R^2=0.78, RMSE=8.03 t/ha, and rRMSE=10.53%, indicating that the integration of Sentinel-1 SAR, ERA5-Land meteorological variables, terrain information, and geographic prior features can effectively improve sugarcane yield estimation under optical imagery-deficient conditions. These results indicate that Sentinel-1 SAR combined with ERA5-Land meteorological data and machine learning can provide an effective alternative for sugarcane yield estimation when optical spectral imagery is missing, while explicit representation of terrain and regional spatial heterogeneity is essential for improving estimation accuracy.

Keywords: Sugarcane yield; Sentinel-1 SAR time series; Savitzky-Golay filter; TW-shapeDTW; Random Forest

INTRODUCTION

The Guangxi Zhuang Autonomous Region is the dominant sugarcane and sugar-producing region in China. The region is located in a low-latitude subtropical monsoon climate zone, where warm conditions, abundant rainfall, and frequent cloudy weather favor sugarcane growth but often reduce the continuity of optical remote-sensing observations. Indeed, overcoming atmospheric disturbances and cloud cover in optical time series is widely recognized as a major challenge for reliable sugarcane monitoring.(Som-ard et al., 2021) Operational yield estimation therefore requires a data source that is less sensitive to clouds and illumination conditions.

Sentinel-1 synthetic aperture radar provides all-weather C-band observations and can record canopy structure, water status, and temporal growth changes through VV and VH backscatter(Den Besten et al., 2023). However, sugarcane yield is not controlled by canopy signal alone. Terrain, planting location, rainfall, temperature, and regional management differences can produce similar radar trajectories with different yield levels(Marin et al., 2019, 2013; Verma et al., 2023). A model based only on SAR and meteorological predictors may therefore under-represent the spatial heterogeneity of sugarcane production.

This study was designed to investigate whether Sentinel-1 SAR and commonly available ancillary variables can support reliable sugarcane yield estimation in Guangxi under missing or incomplete optical spectral data. The analysis first tested SAR and meteorological variables, then added terrain and stratified sampling, and finally introduced K-means geographic partitioning and training-set spatial priors. To effectively handle the complex, nonlinear interactions among these diverse features without the extensive parameter requirements of traditional process-based crop models, a Random Forest machine learning framework was employed. The combination of multi-source earth observation data and robust machine learning methods like Random Forest is increasingly critical for effective sugarcane yield modeling.(De França E Silva et al., 2024) This staged design was used to isolate the contribution of temporal radar features from the contribution of geographic heterogeneity.

DATA AND METHODS

Study Area and Yield Samples

The study area covers sugarcane-producing counties and field sampling locations in Guangxi. Yield records were assembled from Guangxi statistical yearbook information for 2020-2024 and from field yield measurements obtained in mid-October 2025. After removing records without valid yield values, 713 samples were retained for the main modelling line. For the final cleaned model, 50 abnormal 2025 records were excluded using a historical spatial KNN absolute-difference threshold of 35 t/ha, leaving 663 samples. This screening step was treated as a data-quality control step and was not used as a direct model feature.

Sentinel-1, Meteorological, and Terrain Variables

Sentinel-1 dual-polarization observations were used to derive VV, VH, RVI, and DpRVIc time-series descriptors(Bhogapurapu et al., 2022; Mandal et al., 2020). Meteorological covariates included temperature and precipitation variables aggregated over the main growth windows. Terrain descriptors included elevation and slope. These variables were then transformed into seasonal statistics, growth-period contrasts, and interaction features, such as slope multiplied by RVI, DpRVIc, temperature, or precipitation summaries.

Time-Series Smoothing and Shape Distance

Sentinel-1 trajectories were smoothed with the Savitzky-Golay filter to reduce short-term noise while retaining the shape of the growth curve. For an observation x_t, the smoothed value can be written as

$$\widehat{x_t} = \sum_{h=-m}^{m} c_h x_{t+h}$$

, where c_h is the convolution coefficient in the local polynomial window. This representation is suitable for radar trajectories because abrupt speckle-like variation is reduced without flattening the seasonal pattern.

Time-Weighted shape Dynamic Time Warping（TW-shapeDTW）(Gao et al., 2021; Zhao and Itti, 2018) was used to compare the shape of smoothed trajectories. If P and Q are two temporal sequences, the shape-based dynamic time-warping distance can be expressed as

$$D(P,Q) = \min_{\pi} \sum_{(a,b)\in\pi} d\big(s(P_a), s(Q_b)\big)$$

, where pi is a valid warping path and s(.) is the local shape descriptor around each time step. This distance was used to reveal whether different regions showed distinctive growth-wave patterns after smoothing.

Geographic Stratification

K-means geographic clustering was introduced to represent spatial heterogeneity. For sample coordinates x_i, the clustering objective is $\min_{C} \sum_{k=1}^{K} \sum_{x_i\in C_k} \left|\left|x_i - \mu_k\right|\right|^2$. Geographic clusters were treated as regional strata and as prior information. Coordinate-mean and cluster-mean yield priors were estimated from the training split only and then assigned to test samples, so the target values of test samples did not enter prior construction.

Random Forest Modeling

Compared to traditional parametric statistical models, Random Forest (RF) offers greater robustness against multicollinearity and is less prone to overfitting when processing high-dimensional datasets.(Breiman, 2001). Furthermore, compared to process-based crop simulation models that require extensive and often hard-to-obtain parameters, the RF approach significantly simplifies data input and lowers data-acquisition barriers by directly utilizing readily available remote sensing and environmental covariates. Consequently, RF was selected as the main estimator because it can efficiently handle complex, nonlinear interactions among SAR, meteorological, terrain, and geographic features without requiring a predefined yield response function. The final cleaned model used K=6 geographic clusters and combined the Random Forest prediction with a training-only spatial KNN prior. The dataset was split ensuring the test fraction remained above 20%.

Accuracy Assessment

Accuracy was evaluated on held-out test samples. The main metrics were

$$R^2 = 1 - \frac{\sum_i(\widehat{y_i} - y_i)^2}{\sum_i(\overline{y_i} - y_i)^2}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2}$$

$$rRMSE = \frac{RMSE}{\overline{y_i}} * 100\%$$

Because single-year field samples may have a narrower yield range than multi-year statistics, RMSE and rRMSE were treated as primary error indicators, while R^2 was used as a supporting measure of relative explanatory ability.

RESULT AND ANALYSIS

Temporal Response and Regional Heterogeneity

Figure 1, and 2 show that the K-means geographic partitions do not share a fully synchronous Sentinel-1 temporal response. The VH trajectories have a broadly consistent sugarcane growth pattern, with rapid canopy development from April to July, a relatively stable mid-season period, and gradual attenuation after autumn; however, the peak timing, amplitude, and late-season decline differ among regions. Strong smoothing removes short-term SAR fluctuations and makes the regional shape differences clearer for TW-shapeDTW alignment. The VV, RVI, and DpRVIc profiles further indicate that different radar descriptors emphasize different growth states: VH is more sensitive to volume-scattering changes, VV reflects the co-polarized structural background, and the two vegetation indices compress the dual-polarization response into more stable seasonal features. These results suggest that geographic partitioning is meaningful because the sampled areas differ in crop calendar, growth status, and regional background rather than following a single uniform Guangxi-wide trajectory.

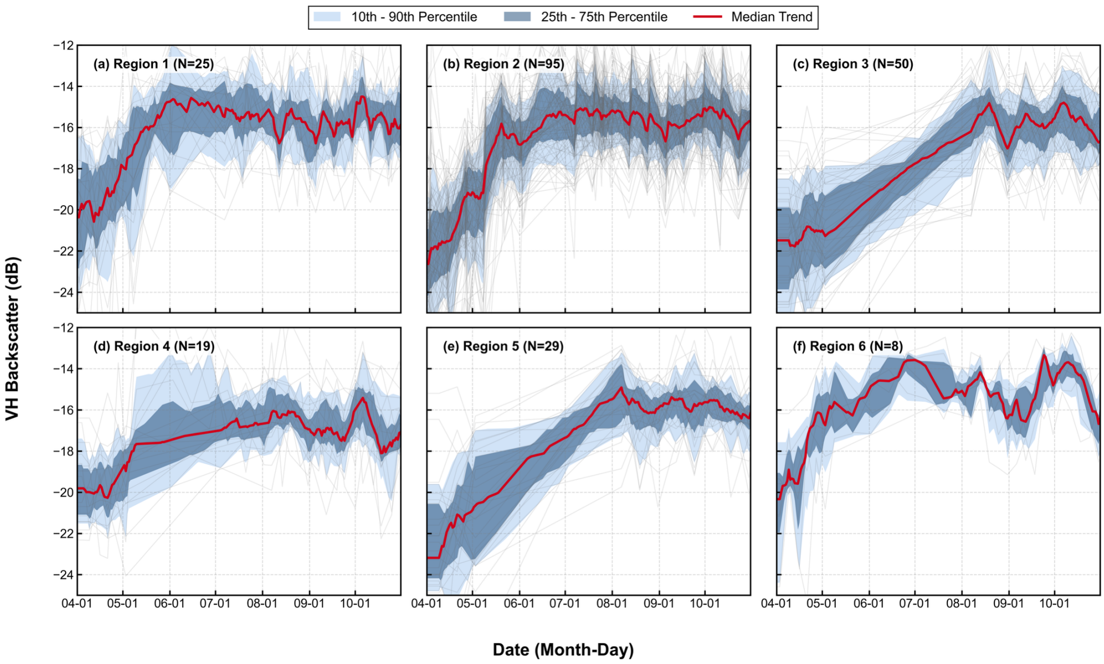

Fig. 1. VH-based spatial display of geographic sugarcane zones in Guangxi.

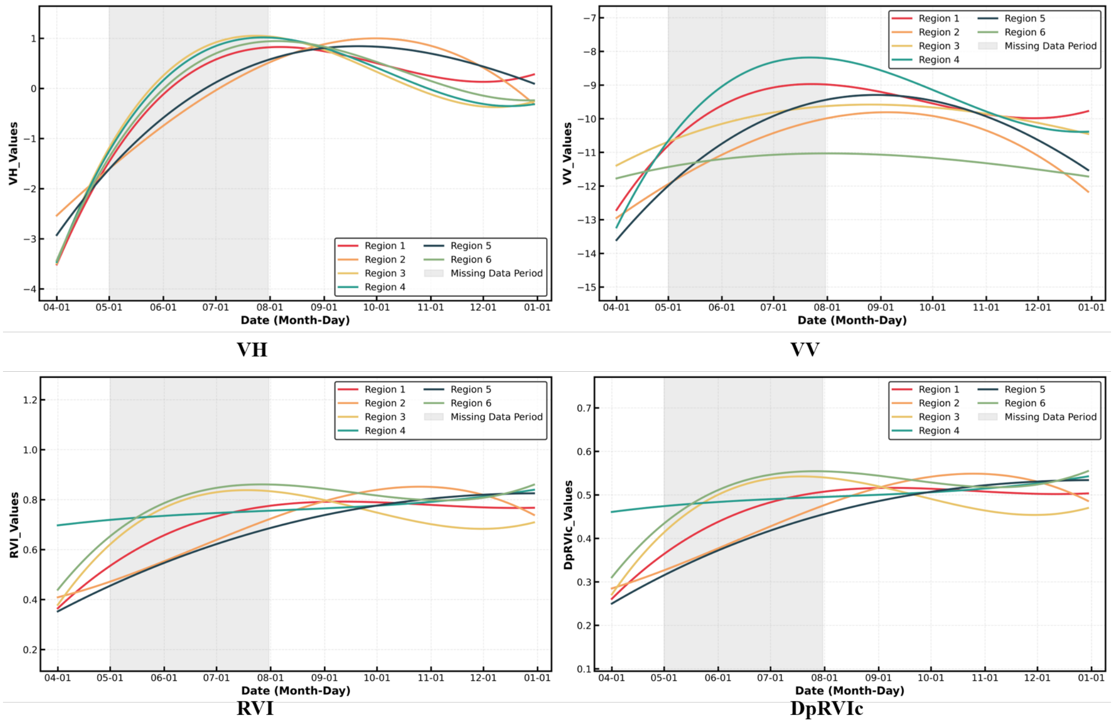

Fig. 2. Smoothed temporal profiles of VH, VV, RVI, and DpRVIc by geographic region.

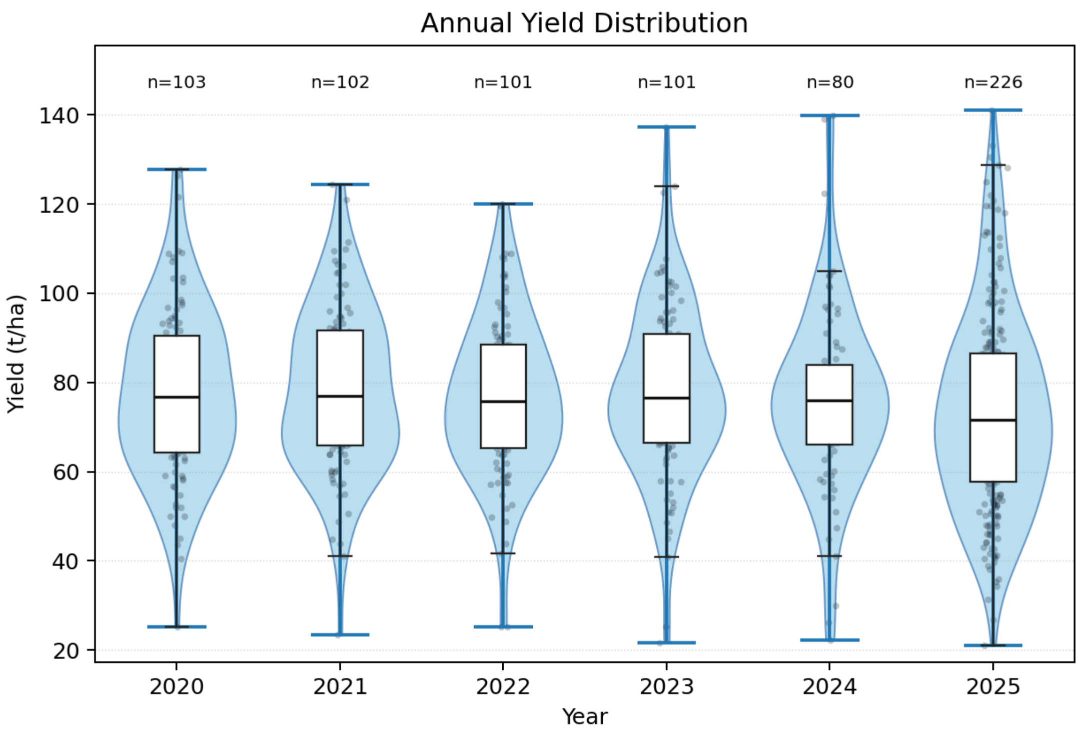

Fig. 3. Annual yield distribution

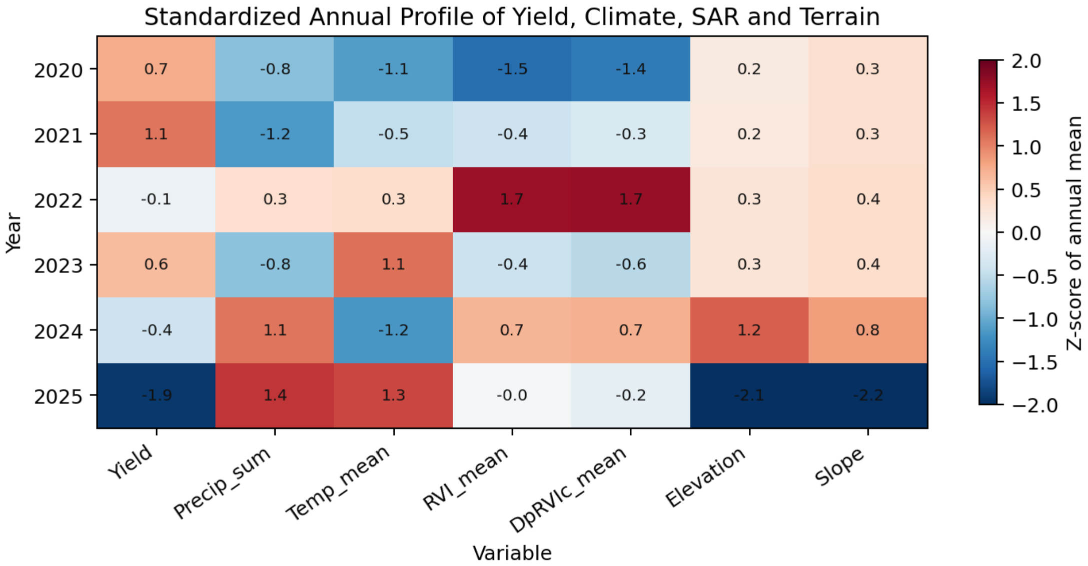

Fig. 4. Standardized annual profile of yield, climate, SAR indices, and terrain.

The annual yield distribution and standardized annual profile (Figure 3, and 4) reveal a second source of heterogeneity across years. Yield levels from 2020 to 2024 are derived mainly from historical statistical records, whereas the 2025 samples are field-measured points with a larger sample size and a clearer concentration of low-yield observations; therefore, the two data sources are not completely interchangeable in distribution. The heatmap also shows that yield, precipitation, temperature, SAR indices, elevation, and slope do not change in the same direction across years. For example, 2025 combines low standardized yield with relatively warm and wet conditions and lower terrain background, while years with stronger SAR-index responses are not necessarily the years with the highest yield. This mismatch indicates that the model needs to handle inter-annual distribution shift and the difference between historical county-level yield and field-measured yield, stratified sampling, and spatial prior information necessary complements to SAR time-series features.

Model Accuracy Under Staged Feature Settings

The staged modelling results show that SAR temporal features alone do not fully explain sugarcane yield variation. When Sentinel-1 and meteorological variables were used with random sampling, test R^2 remained at about 0.32-0.35. After adding slope and using a stratified sampling strategy, R^2 increased to about 0.52-0.56. The improvement indicates that terrain and sampling design reduce part of the uncontrolled regional variation. The strongest improvement occurred after adding geographic prior information and K-means spatial partitioning (Table 1).

Table 1. Model optimization process and corresponding accuracy table

Model stage	Main setting	R^2	RMSE	Interpretation
SAR + meteorology	Sentinel-1 time-series descriptors and meteorological variables under random sampling	0.32-0.35	18.53-19.25 t/ha	Limited yield explanation from pure remote and weather variables
SAR + meteorology + terrain	Slope and stratified sampling added to the baseline feature set	0.52-0.56	13.55-13.86 t/ha	Terrain and sampling structure reduce part of the spatial error
SAR + meteorology+ terrain + geographic prior	K=6 geographic zones and training-only spatial prior combined with RF	0.78	8.03 t/ha	Regional heterogeneity is represented explicitly and prediction becomes operationally useful

Final Yield Estimation Performance

The left panel of Figure 5 represents the model performance before the geographic prior information was introduced, with R^2=0.56, RMSE=13.55 t/ha, and rRMSE=18.09%. Although the predicted values show a positive relationship with observed yield, the points are relatively dispersed and the prediction range is compressed toward the middle-yield interval. Low-yield samples tend to be overestimated, whereas high-yield samples are often underestimated, indicating that SAR, meteorological, and terrain variables alone still cannot fully distinguish regional production background. In comparison, the right panel shows that the addition of geographic prior information substantially reduces this dispersion and aligns the test samples more closely with the 1:1 line. The final accuracy can be stated as R^2=0.78, RMSE=8.03 t/ha, and rRMSE=10.53%.

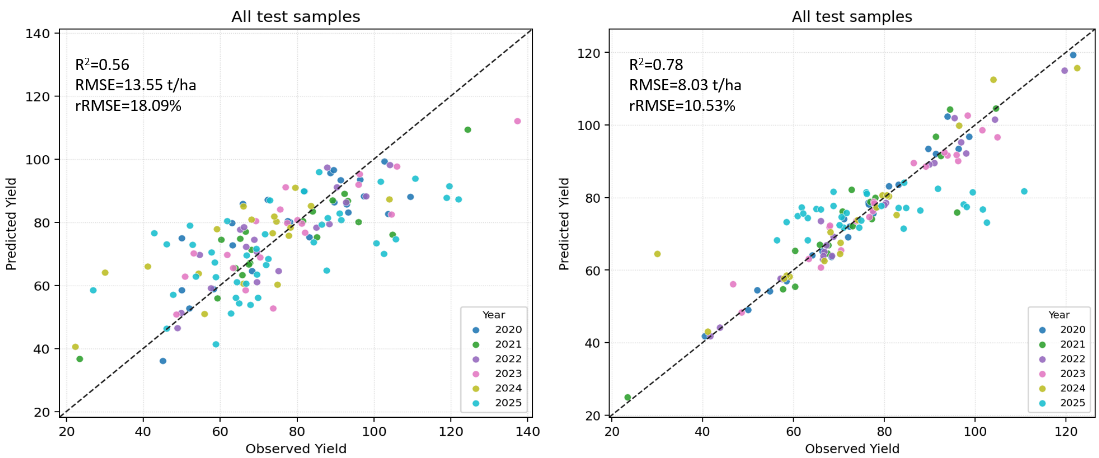

Fig. 5. Observed versus predicted sugarcane yield before and after geographic prior information.

Figure 6 presents the spatial distribution of estimated sugarcane yield in Guangxi for 2025. The estimated yield pixels are mainly distributed in the major sugarcane-producing areas, including Chongzuo, Nanning, Laibin, Liuzhou, and Guigang. The map shows clear spatial heterogeneity in yield levels: medium-yield areas of 60-90 t/ha are widely distributed, while higher-yield pixels of 90-120 t/ha are locally concentrated in parts of central Guangxi. Low-yield pixels of 20-60 t/ha appear more frequently in some southwestern and coastal production areas. This spatial pattern indicates that the final model can extend point-based yield estimation to a regional mapping scale and provides an intuitive representation of the geographic differences in sugarcane production potential.

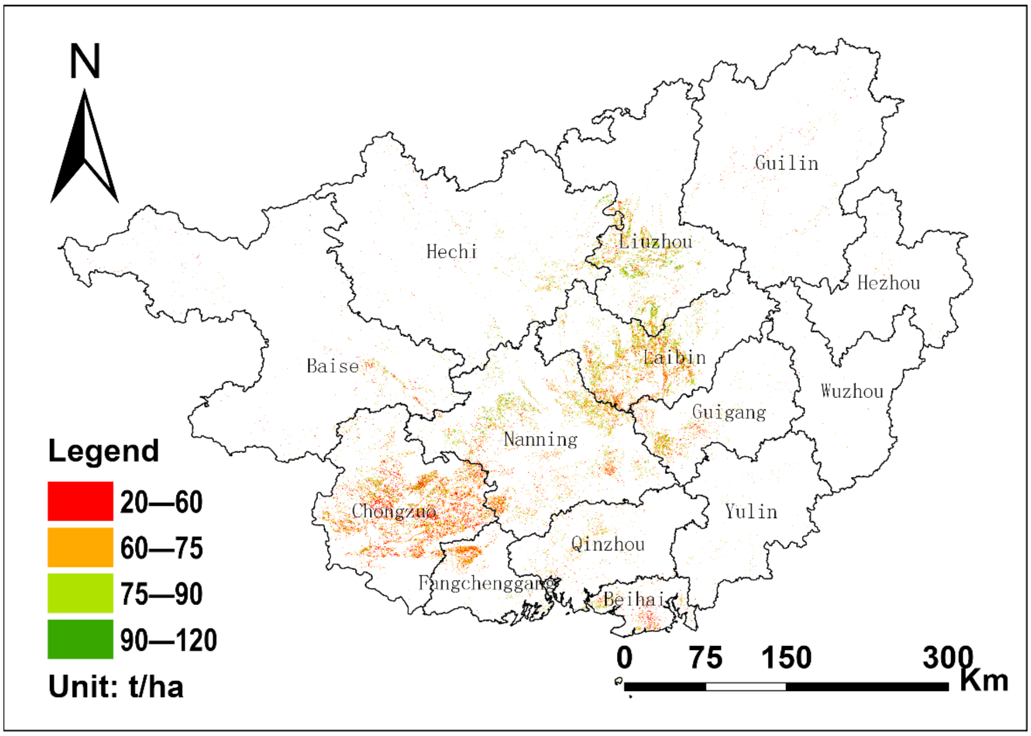

Fig. 6. Estimated sugarcane yield in Guangxi Region,China in 2025.

DISCUSSION

The results point to a clear limitation of using Sentinel-1 time-series data alone for sugarcane yield estimation. SAR backscatter captures canopy structure and moisture-related responses, but sugarcane yield also depends on cultivar, planting density, soil fertility, management, irrigation, harvest age, and local topographic conditions. Without spectral and management variables, different production systems can share similar radar trajectories while producing different yields. This explains the low accuracy of the pure SAR and meteorological model.

The gain from K-means geographic partitioning does not mean that coordinates alone solve the yield estimation problem. It indicates that spatial prior information is a necessary representation of regional production background in the current data setting. The final model is therefore most suitable for regional assessment within the sampled production system. Transfer to a new year, a new county, or a management regime not represented in the training data should be treated carefully.

Further improvement should prioritize feature enrichment rather than only deeper model tuning. Soil and management variables, and more balanced multi-year field measurements are expected to strengthen independent-year generalization. Stage-wise growth windows and cumulative prediction can also reduce the risk of treating the whole season as a single undifferentiated signal.

CONCLUSION

A sugarcane yield estimation framework was developed for Guangxi using Sentinel-1 time-series SAR, meteorological variables, terrain descriptors, and geographic prior information. SG filtering and TW-shapeDTW improved the description of temporal radar trajectories. K-means geographic partitioning and training-only spatial priors represented regional heterogeneity that could not be captured by SAR and weather variables alone. The final geographically informed model reached R^2=0.78, RMSE=8.03 t/ha, and rRMSE=10.53%. The evidence suggests that Sentinel-1 is a useful component for cloud-prone sugarcane yield estimation, but reliable regional performance depends on explicit representation of terrain and spatial production background.

ACKNOWLEDGEMENTS

This research was funded by the National Key Research and Development Program (Nos. 2022YFD2001103 and 2023YFD1701001-03) and the National Natural Science Foundation (No. 42371323).

REFERENCE

Bhogapurapu, N., Dey, S., Homayouni, S., Bhattacharya, A., Rao, Y.S., 2022. Field-scale soil moisture estimation using sentinel-1 GRD SAR data. Advances in Space Research 70, 3845–3858. https://doi.org/10.1016/j.asr.2022.03.019

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

De França E Silva, N.R., Chaves, M.E.D., Luciano, A.C.D.S., Sanches, I.D., De Almeida, C.M., Adami, M., 2024. Sugarcane Yield Estimation Using Satellite Remote Sensing Data in Empirical or Mechanistic Modeling: A Systematic Review. Remote Sensing 16, 863. https://doi.org/10.3390/rs16050863

Den Besten, N., Steele Dunne, S., Mahmud, A., Jackson, D., Aouizerats, B., De Jeu, R., Burger, R., Houborg, R., McGlinchey, M., Van Der Zaag, P., 2023. Understanding Sentinel-1 backscatter response to sugarcane yield variability and waterlogging. Remote Sensing of Environment 290, 113555. https://doi.org/10.1016/j.rse.2023.113555

Gao, H., Wang, C., Wang, G., Fu, H., Zhu, J., 2021. A novel crop classification method based on ppfSVM classifier with time-series alignment kernel from dual-polarization SAR datasets. Remote Sensing of Environment 264, 112628. https://doi.org/10.1016/j.rse.2021.112628

Mandal, D., Kumar, V., Ratha, D., Dey, S., Bhattacharya, A., Lopez-Sanchez, J.M., McNairn, H., Rao, Y.S., 2020. Dual polarimetric radar vegetation index for crop growth monitoring using sentinel-1 SAR data. Remote Sensing of Environment 247, 111954. https://doi.org/10.1016/j.rse.2020.111954

Marin, F.R., Jones, J.W., Singels, A., Royce, F., Assad, E.D., Pellegrino, G.Q., Justino, F., 2013. Climate change impacts on sugarcane attainable yield in southern Brazil. Climatic Change 117, 227–239. https://doi.org/10.1007/s10584-012-0561-y

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

https://doi.org/10.1016/j.fcr.2019.06.011

Som-ard, J., Atzberger, C., Izquierdo-Verdiguier, E., Vuolo, F., Immitzer, M., 2021. Remote Sensing Applications in Sugarcane Cultivation: A Review. Remote Sensing 13, 4040. https://doi.org/10.3390/rs13204040

Verma, A.K., Garg, P.K., Prasad, K.S.H., Dadhwal, V.K., 2023. Variety-specific sugarcane yield simulations and climate change impacts on sugarcane yield using DSSAT-CSM-CANEGRO model. Agricultural Water Management 275, 108034. https://doi.org/10.1016/j.agwat.2022.108034

Zhao, J., Itti, L., 2018. shapeDTW: Shape Dynamic Time Warping. Pattern Recognition 74, 171–184. https://doi.org/10.1016/j.patcog.2017.09.020