A Machine Learning-Based Multi-Temporal, Multi-Source Approach for Accurate Grape Yield Estimation

Aron Haywood, Kyle Loggenberg and Albert Strever (South Africa)

Key words: Cost management; Geoinformation/GI; Land management; Low cost technology; Remote sensing; Risk management; Sentinel-2; viticulture; yield prediction; big data analytics; climate variability; early warning systems; geospatial modelling; precision agriculture; vegetation indices; machine learning

SUMMARY

Viticulture faces mounting challenges from changing climatic conditions, including shifting rainfall patterns, rising temperatures, and intensifying droughts, creating pressure to optimise yield and fruit quality while minimising costs through improved yield prediction. Traditional statistical models rely on historical data but fail to account for seasonal variability caused by climatic and environmental factors, leaving vineyards vulnerable to climate-induced agricultural disasters. This study develops a multi-source, multi-temporal remote sensing machine learning framework for grape yield prediction across different spatial scales and vineyard environments, demonstrating how big data analytics can enable early warning systems for informed decision-making in viticulture. This research investigated five growing seasons (2021-2025) across both the Northern Cape and Western Cape provinces of South Africa, analysing approximately 1,000 yield records from table grape vineyards measured in tons per hectare per block. The framework integrates open-source Sentinel-2 satellite imagery (10m resolution) with 34 variables: 20 environmental variables (NDVI, EVI, ARVI, and other vegetation indices capturing plant health and soil conditions), 9 terrain variables (aspect, elevation, and slope affecting microclimate), and 5 climate variables (temperature, humidity, wind speed, radiation, and rainfall). Statistical measures (minimum, maximum, median, average, and standard deviation) were calculated per month per block from September to January, generating a comprehensive big data matrix for machine learning analysis. Ten systematic experiments evaluated dataset combinations using ensemble machine learning (ML) models (Random Forest and XGBoost) designed for high-dimensional geospatial data. Over 6,600 models were trained and tested to determine optimal configurations for yield prediction. The framework achieved strong regional performance with R2 = 0.59 and RMSE = 5 tons/hectare for the Western Cape (WC), demonstrating reliable predictive capability using freely available satellite data. The Northern Cape (NC) achieved R2 = 0.37, likely attributable to the compressed 3-month growing window (September to late November/early December) versus the Western Cape's

extended 7-month cycle (September to March), providing significantly less temporal data for model training. Despite regional variations, predictions are available 1-3 months before harvest (October for NC and December for WC) enabling proactive yield management preparation across both regions.

In comparison to previous literature this study provides a more thorough approach at predicting viticulture yield, focusing on viticulture blocks spanning ~200km in NC and 150km in WC. While other studies suggest high accuracies (R2 > 0.8), they focus not only on single areas/production units but also on the association/correlation between vegetation indices and yield rather than the predictive ability. This study fills this gap by comparing the actual vs predicted yield achieved by the ML models. With the drive for high-resolution satellite imagery and the high costs associated, these results highlight the potential of open-source remote sensing combined with machine learning to provide practical early predictions for viticulture applications, maintaining statistical significance through a sufficiently large and representative training dataset across different growing seasons and regions.