JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 06 Issue 02
2024

*Research Article*
https://doi.org/10.56979/602/2024

# Remote Sensing Based Sugarcane Yield Prediction Model using Artificial Intelligence

**Mahreen Iftikhar[1], Salman Qadri[1*], Muhammad Nadeem[2], and Syed Ali Nawaz[3]**

[1]MNS-University of Agriculture, Multan, Pakistan.
[2]National College of Business Administration and Economics, Multan, Pakistan.
[3]Department of Information Technology, The Islamia University of Bahawalpur (IUB), Bahawalpur 63100, Pakistan.
[*]Corresponding Author: Salman Qadri. Email: salman.qadri@mnsuam.edu.pk

**Received:** August 01, 2023 **Accepted:** February 08, 2024 **Published:** March 01, 2024

**Abstract:** Automation is playing an increasingly important part in today's professions and in every field of life. Remote-sensing based sugarcane yield prediction is more effective than old yield-prediction techniques. Traditional yield measurement techniques include the time and labor-intensive destructive sampling of sugarcane fields. Accurate and timely yield forecasts help decision-making processes such as crop harvesting plans, milling, marketing, and forward selling strategies, which boosts the efficiency and profitability of the global sugar sector. At the moment, destructive or visual sampling techniques are used by producers or productivity officers paid for by mills to assess production during the growing season. People want to get things done in the quickest and most efficient way to solve the problems. The use of machinery for sugarcane cultivation has increased significantly over a wide area to lower production costs, decrease farmers' labor demands, and improve harvest efficiency. Although they have not been thoroughly compared, existing techniques for estimating agricultural output using regression typically rely on a specific set of forecasting factors. For monitoring agriculture, which makes use of satellite earth observation data, this study illustrate and compare the use and using several sets of object-based predictors to estimate sugarcane production. Several regression models compared utilizing a variety of different predictor variables. In this study, the yield of the sugarcane measured by using regression models, time series of the vegetation index (VI), remote sensing, phenology measurements, and the normalized difference vegetation index (NDVI). Artificial intelligence algorithm models (Random Forest and Ordinary Least Squares) used to construct the suggested way to accurately relate ground-measured data. This work presents novel sugarcane yield prediction technique that improves forecasting accuracy.

**Keywords:** Sugarcane; Yield prediction; Predictor variables; Time series vegetation index (VI); phenology measurements; normalized difference vegetation index (NDVI).

## 1. Introduction

The widely farmed semi-perennial sugarcane plant (Saccharum spp. L.), which yields sugar and other goods, is essential to global agriculture. This crop's growing relevance is due to the fact that it is used to produce biofuel in addition to being used as a food source [1]. Given the significance of food production for individual living standards and national food security, governments, businesses, consumers, and other sectors are constantly concerned about it [2]. Around the world, sugarcane is grown on more than 28 million hectares of agricultural land, producing 1.7 trillion tons of raw sugar annually [3]. How to boost crop yields has become one of the major concerns that must be promptly addressed due to the rising demand for agricultural products as a result of the expanding global population [4].

The profitability and effectiveness of the world's sugar industry are increased when decision-making processes including the harvesting of crops plans, milling, advertising, and future selling strategies are supported by timely and accurate production projections. Currently, producers or productivity officers funded by mills estimate yield during the growing season utilizing destructive or visual sampling methods. This approach requires a lot of labor, and accuracy is affected by a variety of seasonal climate circumstances, crop age because of a prolonged agricultural season and human mistake. Devices for monitoring the creation of a harvester mounted were in use in several sugarcane-growing regions [5] provide different accuracies for yield mapping as well, but access to data only occurs after harvest, making it unable to support decision-making during the growing season. A more precise and affordable way of predicting sugarcane yield has recently been examined as a replacement: remote sensing technologies [6]. Sugarcane is a special multi-year crop that may be harvested every year for up to 6-7 years before needing to be replanted. Following an annual harvest, the ratoons—the roots and the lower portions of the plant-grow fresh stems that are cut the following year. The majority of farms consistently harvest sugarcane for nine months. This is significant because it enables continual crushing. As a result, supervising the harvesting and crushing takes place virtually every day. Given the enormous throughput of production, it is imperative to regularly estimate crop yields.

One of the most significant crops on the planet is this one. For farmers to make practical decisions regarding storage requirements, crop insurance, cash flow projections, fertilizer and water use, and crop output forecasting are all quite practical. Research on sugarcane breeding aims to choose genotypes according to yield measurement that are optimal for particular settings [7]. Human senses are susceptible to workload and fatigue when they are subjected to various characteristics, like domain knowledge, shape, size, color, patterns, etc. The methods for identifying grains become inconsistent, inaccurate, and unbelievable [8]. For the sake of crop management and policymaking, forecasting crop yields and their variability over space and time in a changing climate is a difficult but necessary task. It is essential for all parties involved, including national officials and private landowners, to have access to information on the risks related to the consequences of climate change on the outcomes of agricultural activities [9].

A range of factors, including climate change, land availability, and water scarcity, are impacting modern agriculture and food production systems. A pandemic is adding to this a lot of pressure. To feed an ever-growing global population, scientists and engineers need to develop scientific and technological innovations. Over the last few decades, genetic tools have made incredible advances, but we haven't been able to adequately measure crop status in the field on a large scale. Thanks to developments in artificial intelligence (AI) and remote sensing, we can now precisely measure the phenotypic data at the field scale and incorporate it into management technologies that are both prescriptive and predictive [10]. In many precision agriculture applications, spatial resolution is still an issue despite the benefits because of its extensive coverage, and satellite remote sensing data. The worst impact of many prediction models on crop yields is only to provide better estimates at the county and above levels, but not at the local or smaller scales (for example, on single farmland). A satellite's revisit time is also limited by its ability to be blocked by clouds, making accurate information on vegetation everywhere its entire cycle impossible [11]. The accuracy of yield prediction using satellite data to calculate rice yields [12], wheat [13], corn [14], soybean [4], and other crops. Classification is a technique that has been shown to evaluate remote sensing data and can classify the individual pixels in an image’s spectral characteristics [15, 16].

Crop yield forecasts for the near future offer useful information on how to manage agricultural resources and the potential economic effects of poor yield. Such projections are challenging to make in areas with sparse observational data. And large-range of crop yield prediction is also significant for overall prediction and that’s kind of predictions are easy to achieve for large-scale area [17]. Many studies have shown that data from observations of the earth, particularly publicly accessible satellites data, have been widely employed in usages in agriculture, including those that estimate crop yield and biomass [18, 19] and forecasting software [7, 20] in strategies for managing farms. The latter estimating method, which is used when the season is already over, is the subject of this study.

A reliable postseason production estimate is essential for a number of reasons, including the verification of the declared production, locating low-productivity lands and farms as a starting point for developing improved techniques for the upcoming season, the estimation of the necessary number of transport vehicles, therefore, to minimize expenses related to the commodity's storage and transportation, the verification of the reliability of the post-season yield estimate, and many others [21]. In this instance, the Normalized Difference Vegetation Index (NDVI) and other indices composed of various spectral band ratios have been used as empirical models for regression based on the vegetation index (VI) time series to estimate the output of the sugarcane crop [22, 23].

Other methods, which do not require ground truth data, classify the sugarcane area according to the type of crops grown there and estimate the probable yield using statistics for prior season’s yields and the resulting acreage. In this study, various regression-based and artificial intelligence methods for estimating crop yield will be examined and compared.

### 1.1. Production of Sugarcane in Pakistan

Since the beginning of time, Pakistan has farmed sugarcane, which the huge Indus River and its numerous tributaries are attributed with. The area, historically known as the civilization of the Indus Valley was aware of how to produce sugarcane and extract brown sugar cakes, which are still made and distributed today and are enjoyed by the locals as "Gur." Juice from sugarcane that has been peeled and sliced into chewable pieces has been utilized for centuries. The cultivation of sugarcane is appropriate for regions between latitudes 24 and 34 degrees north, which are characterized as irrigated subtropical zones with mild temperatures. Except for the area above 30º N, which infrequently experiences frosts, the area can be considered a frost-free zone. Out of the 22.0 million hectares that are available for cultivation, sugarcane takes up almost 1.0 million hectares, or 4.5% of the irrigated land. The total amount of water available in the current system and reservoirs is approximately 135 MAF, which is less than the crop's estimated need of about 10 MAF (million acre-feet).

High delta crop's development outside of this ecological zone has been constrained by its propensity to be affected by the weather cycle. Currently functioning at about 70% of its capacity, Pakistan's sugar industry is well established. When compared to the current industrial capacity, which can harvest at least seventy million tons, the yearly yield of cane varies among forty-five million and sixty-five million tons based on the availability of irrigation water and rainfall. The production has barely increased during the past six decades show in Figure 1.

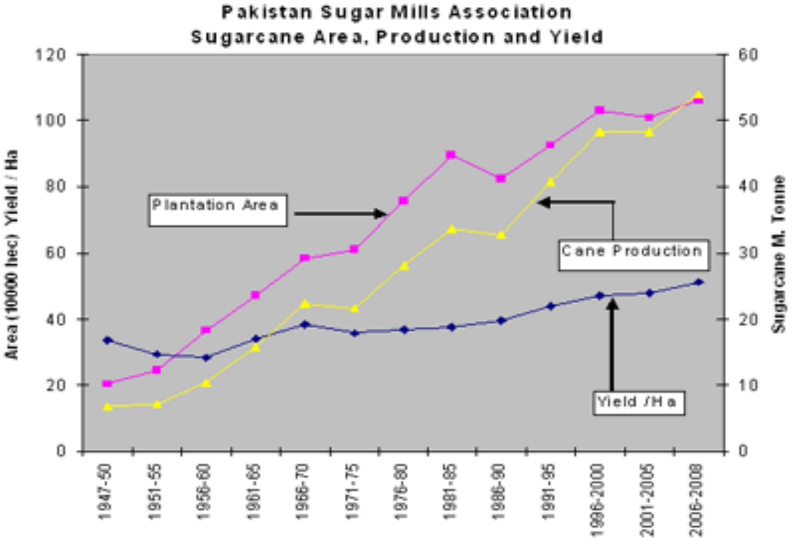

**Figure 1.** Sugarcane Area, Production and Yield

According the annual report of sugarcane in Pakistan from the United States Department of Agriculture, due to the anticipated rebound in the area, sugarcane production is predicted to be 83.5 million metric tons during 2023–2024, three percent more than the 2022–2023 estimated. The floods during the previous year had a negative impact on the harvested area and yield. The assistance price for sugarcane growers in 2022–2023 is nearly 32% more than in 202–202, at 300 rupees per 40 kg ($27.28/ton). These prices are incentivizing farmers to keep their cane fields rather than introducing other crops. In such a situation, it is obvious that managers and decision-makers would benefit from instruments that could continuously track the sugarcane's vegetative vigor in Pakistan and deliver immediate information about any potential short-term effects of meteorological conditions on yield forecasts.

Additionally, compared to other crops, sugarcane is more resilient to weather dangers, which is why farmers prefer to plant it. Three provinces grow sugarcane, with Punjab producing 68% of the total amount, Sindh 24%, and Khyber Pakhtunkhwa (KPK) 8%. More than half of the world's sugarcane land is in the

Punjabi division of Bahawalpur and the Sindhi division of Sukkur. There are two planting seasons for cane: the spring planting season lasts from February to March, and the autumn planting season lasts from the beginning of September to the end of October. Farmers in Punjab, Sindh, and KPK cultivate sugarcane throughout the year. In KPK, the majority of cane is grown in the spring. Because there isn't enough high producing varieties, water restrictions, and unequal fertilizer delivery, per hectare yields tend to remain poor [24]. Overall plan that followed is shown in Figure 2.

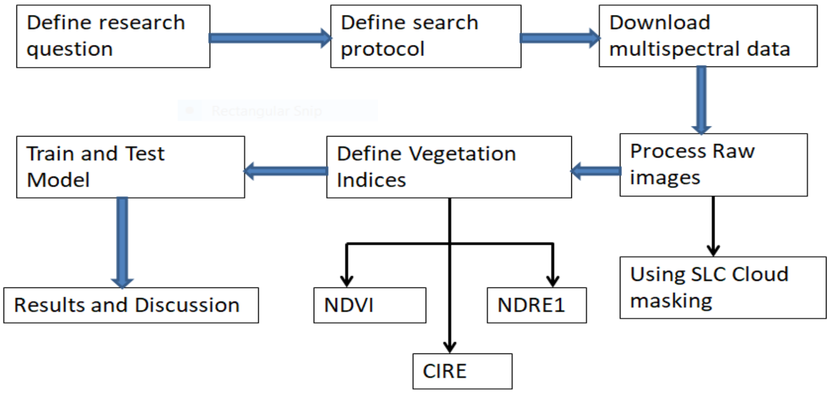

**Figure 2.** Review Protocol

The objectives of this research will be:

- To find different sets of predictor factors that can be used to predict yield in the sugarcane crop.
- To develop an AI-based sustainable yield prediction model for sugarcane using a remote sensing dataset.

1.2. Predictor factors to estimate yield

I choose various methods for generating predictor variables and assess them based on how well they estimate yield. The first technique uses NDVI and time series data from the other vegetation indices (VI). Many other studies have used as predictor factors, VI time series [25]. The second technique involves calculating metrics for phenology [21, 26]. Different markers, like the beginning and conclusion of the growing season, are among these metrics and are using to describe seasonally varying crop ontogenesis and crop growth. The overall methodology used with various predictor variable approaches and the cloud platform's acknowledgment of functioning as a preprocessing and storage node for satellite data. Based on their relationship with sugarcane yield and their significance in reflecting the fundamental mechanisms influencing crop growth, the most pertinent predictor variables should be chosen. The most useful variables for the prediction model can be found using feature selection approaches like correlation analysis or feature importance measures from machine learning models. Illustration of the sugarcane yield predictor's method shown in Figure 3.

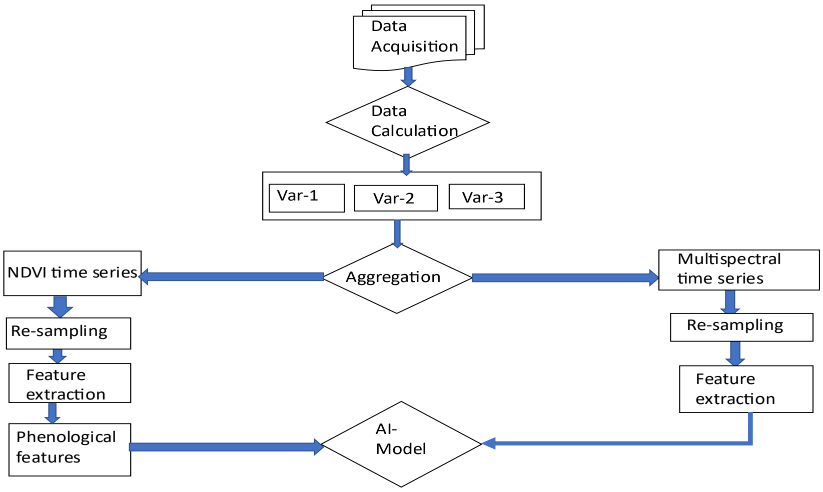

**Figure 3.** Illustration of the Sugarcane Yield Predictor's Method

## 2. Materials and Methods

### 2.1. Study Region

The district of Punjab Province chosen for this study is Rahim-Yar Khan. It contributes close to 30% of Punjab's total sugarcane yield. It is situated at the intersection between Sindh and Punjab. The Rahim Yar Khan district is divided into four tehsils: Sadiq Abad, Rahim Yar Khan, Khanpur, and Liaquatpur. In southern Punjab, where the majority of the population works in agriculture, it is regarded as an agricultural district. It is a fertile region that generates a variety of crops, including wheat, sugarcane, maize, cotton, and mangoes. It is one of the significant crops that contribute significantly to agriculture in this area. There are six sugar mills located in the district. Gulf Sugar Mills, Rahim Yar Khan Sugar Mills, Hamza Sugar Mills, and Jamaluddin Wali Sugar Mills are the major sugar mills in this region. Rahim Yar Khan District has a total population of 477,000 people and a 33.1% literacy rate. In Figure 4 map of Rahim Yar Khan District has been shows. From 310,000 acres in 2014–15 to 430,000 acres in 2020, more land was planted with sugarcane. There are two planting seasons for cane: the spring planting season and the autumn planting season as well.

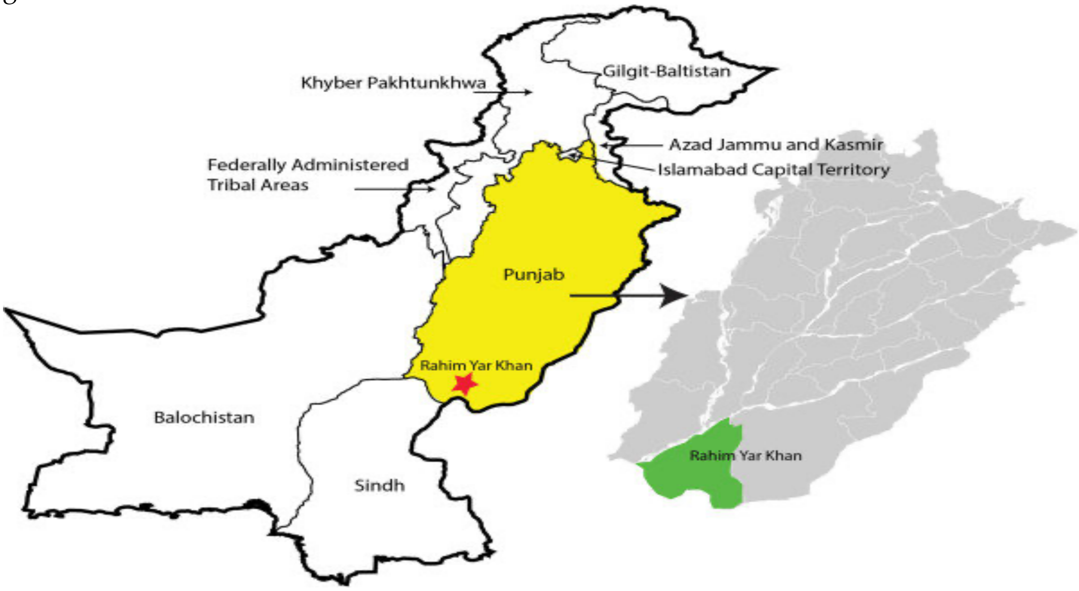

**Figure 4.** Map of District Rahim Yar Khan [27]

### 2.2. Data from Remote Sensing

It takes many resources to gather accurate on-site crop data (production, the biomass, and other biological characteristics), but it is essential for trustworthy crop modeling. We captured the sugarcane growing season for the years 2019–2021 using Sentinel-2 Level-2A multi-temporal data from the time of ratooning (April 2021) through harvest (December 2021). The growing process of sugarcane took an average of 10 to 12 months in Pakistan. Climatic data, such as temperature of the air, evaporation of or moisture data, were not included in the analysis due to the very small research region. This was due to the data's lack of spatial resolution and little use for the regression models. Using atmospherically adjusted Sentinel-2 surface reflectance (Level-2A) data, three different vegetation indices (VIs) were calculated: the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Red Edge 1 (NDRE1), and the Chlorophyll Index Red Edge (CIRE). The formulas for these indices are shown in Table 1.

**Table 1**. Chosen Indices of Vegetation for the Analysis

| Index of Vegetation | Description | Source |
|---|---|---|
| NDVI | Normalized Difference Vegetation Index, sensitive for green biomass<br>NDVI = (Band8-Band4) / (Band8+Band4) | [28] |
| NDRE1 | Normalized Difference Red Edge 1, less likely to become saturated in canopies of thick plants<br>NDRE1 = (Band6-Band5) / (Band6+Band5) | [29] |

| | | |
|---|---|---|
| CIRE | Chlorophyll Index Red Edge, water stress-sensitive for mesophyll and chlorophyll in plant leaves<br>CIRE = (Band7 / Band5) - 1 | [30] |

2.3. Method

We compared the accuracy of the yield estimation using two alternative methods for the development of predictor variables. First, NDVI, NDRE1, and CIRE of the three separate vegetation indicators are used. Many other studies have used VI time series as variables for prediction [24], where some use time series descriptive statistics [31] in addition to temporal segments or dimensionality reduction strategies for time series. The second technique involves computing phonological measures [22]. These metrics include phonological indicators, like season start and end dates, which are used to describe agricultural growth and seasonal crop ontogenesis.

*2.3.1. Data preparation for Sentinel-2*

Monitoring and evaluating sugarcane crops is one of several uses for the high-resolution multispectral imagery provided by the Sentinel-2 satellite program, which is run by the European Space Agency (ESA). The data in the Sentinel-2 Level-2 collection have been pre-processed and corrected for atmospheric effects, making it suitable for additional processing and analysis. The following are the main procedures for handling the Sugarcane Sentinel Level-2 dataset:

*2.3.1.1. Data Acquisition*

The Sentinel-2 Level-2 dataset for the chosen region and time period of interest can be obtained. The dataset is accessible via the website (https://eos.com/).

*2.3.1.2. Data Preparation*

The downloaded dataset should be extracted, and it usually consists of compressed files. To make the dataset files, including the metadata and individual bands, easier to access during processing, arrange them in an organized directory manner.

*2.3.1.3. Band Selection*

Find the appropriate spectral bands for monitoring and predicting sugarcane yield. Several spectral bands, including the green, red, blue, shortwave infrared (SWIR) bands, and near infrared (NIR) are included in the Sentinel-2 dataset. The analysis's precise goals will choose which bands to use.

*2.3.1.4. Image Registration and Mosaicking*

This involved downloading the satellite photos and then masking them using the SLC cloud mask for Level-2A product to only get pixels with the labels "land," "water," and "vegetation".

*2.3.1.5. Vegetation Index Calculation*

With the minimum, maximum, mean, and standard deviation of the value of pixels for every single observation date, we generated a number of vegetation indices and spatially grouped them for each unit object.

2.4. Time series calculations for the object-based vegetation index

As part of the acknowledgment system, every multidimensional Sentinel-2 raster band and the resulting spectrum index raster data are trimmed to the parcel limits and stored in a database. To assure the use of data points exclusively during the growing season, we constructed time series based on objects for every parcel between the indicated months for planting and harvesting based on the multi-temporal aggregation of the spectral indices mentioned above. Figure 5 displays the various saturation levels of each VI. Although the 10m NDVI becomes saturated sooner than vegetation indices based on the Red Edge bands, it collects more detail about the canopy structure.

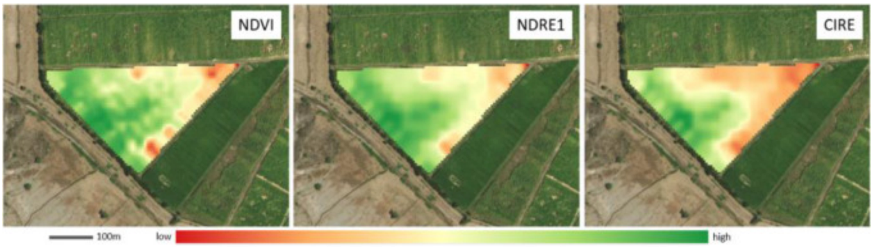

**Figure 5.** NDVI, NDRE1, and CIRE Visual Representation

2.5. Parametric metrics

The development and growth of sugarcane crops may be understood and predicted in large part because to phonological characteristics. Phenology is the study of the relationship between environmental conditions and the timing of recurrent processes in plants, such as blooming, fruiting, and senescence. The following phonological characteristics are important for predicting sugarcane yield: emergence, vegetative growth, flowering, ripening, senescence, and phonological timing and duration.

We computed 12 distinct metrics used as input features according to the metrics for phonetics supplied by the USGS [29] and taken from the EO data. As demonstrated by similar study projects [32]. NDVI time series are often used to extract phonological indicators and metrics. As NDVI signals become saturate with a larger canopy density, they make it possible to monitor vegetation types independently of biomass. Utilizing ground observations, historical records, and remote sensing data, it is possible to monitor and quantify these phonological traits. Large-scale and regular measurements of phonological changes can be taken using remote sensing platforms like satellites or drones. These phonological traits can aid in the creation of reliable sugarcane yield prediction models when combined with machine learning and statistical modeling techniques. The following definitions apply to the generated phonological metrics:

(a) Phenotypic indicators:

- The beginning of the growing season, including the date of crop emergence
- The day with the highest value recorded in the temporal sequence i.e., Seasonal apex
- Agricultural harvest date and absence of chlorophyll indicate the end of the season
- The highest NDVI value recorded throughout time

(b) Detailed metrics:

- The sum of the average NDVI readings over the time
- Total of the highest time NDVI values
- The NDVI's range from lowest to highest

(c) Growth indicators:

- Seasonal duration: the time between start and peak, stated as the quantity of days
- First duration: the number of days between start and peak where green-up occurred
- Second duration: the duration in days between senescence's zenith and its end
- Gradient between start and peak in growth rate
- Variation between peak and end in growth rate

The phonological measures were calculated using daily, NDVI time series interpolated linearly as the input data. The gradients between start of season and peak and end of season, which are as the growth or maturity rates, are the sources of the growth and maturity rates. The NDVI time series local maxima peaks, which sugarcane cutting occurrence, are the peaks. Unless dimensionality reduction methods such as Principal Component Analysis or temporal aggregation are used, phonological markers are used in situations where various NDVI datasets have varying dimensionalities. It has been shown that these markers, as opposed to "raw" VI time series, are more suitable for customizing models to suit various geographical locations or monitoring intervals. Due to methodological discrepancies in extracting such phonological variables, it is vital to stress that satellite-derived phenotype is an apparent, generalized depiction of actual plant phenology [33].

## 3. Results

3.1. Develop an AI-based sustainable yield prediction model for sugarcane

Many approaches are used to forecast yield at the regional and field levels as remote sensing, modeling of statistics, surveys in the field, and crop models. Regression methods can be used to forecast the yield of sugarcane depending on a variety of input variables when it comes to crop yield estimation.

Here are some commonly used regression algorithms for crop yield estimation: Linear Regression, Multiple Linear Regression, Decision Trees, Random Forest, Support Vector Regression (SVR), Gradient Boosting, and Neural Networks. Using historical yield data as the target variable and the pertinent input variables indicated earlier, several regression algorithms can be trained. Based on the input data, the models are then utilized to forecast the yield for fresh or next seasons. It's crucial to evaluate the effectiveness of several algorithms using the right assessment measures and select the one that offers the best accuracy and generalization capabilities for sugarcane yield estimation.

We experimented with the ordinary least squares (OLS) regression algorithm and the Random Forest (RF) regressor to compare linear and non-linear data correlations. The RF approach uses many separate decision trees and averages their predictions to minimize the total error and handle more complicated data structures with high complexity (such time series) and irregular correlations than OLS. Principal Component Analysis (PCA) was used to decrease the feature set's dimensionality to the components that accounted for most of the variability and, thus, the majority of the data gains for the combinatorial usage of both types of predictor variables.

By using an orthogonal transformation, the variables are linearly de-correlated, and the final components either include signal noise or the most de-correlated information about the variables. Results are shown in Figure 6.

**Table 2.** Forecast of Sugarcane Production and Yield

| **Year** | **Production (000 t)** | | | **Yield (kg/ha)** | | |
|---|---|---|---|---|---|---|
| | Forecast | 87% Limit | | Forecast | 87% Limit | |
| | | Lower | Upper | | Lower | Upper |
| **2023** | 65571 | 59815 | 71327 | 57845 | 52731 | 62595 |
| **2022** | 64727 | 58971 | 70483 | 57450 | 52336 | 62564 |
| **2021** | 63872 | 58971 | 69628 | 57065 | 51951 | 62179 |
| **2020** | 62995 | 57239 | 68751 | 56691 | 51577 | 61805 |
| **2019** | 62073 | 56317 | 67829 | 56332 | 51218 | 61445 |
| **2018** | 61060 | 55305 | 66816 | 55989 | 50875 | 61103 |

Forecast of Sugarcane Production and Yield from 2018-2023 with 87% confidence interval is discussed through the following table 2.

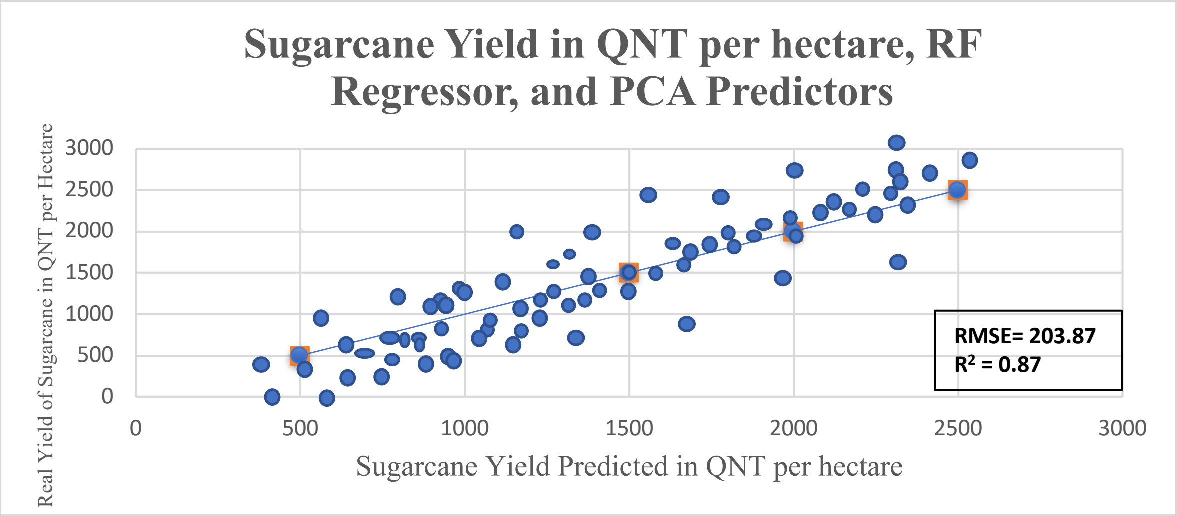

**Figure 6**. Validation for RF-based regression

Corresponding graph of last five years is shown in Figure 7.

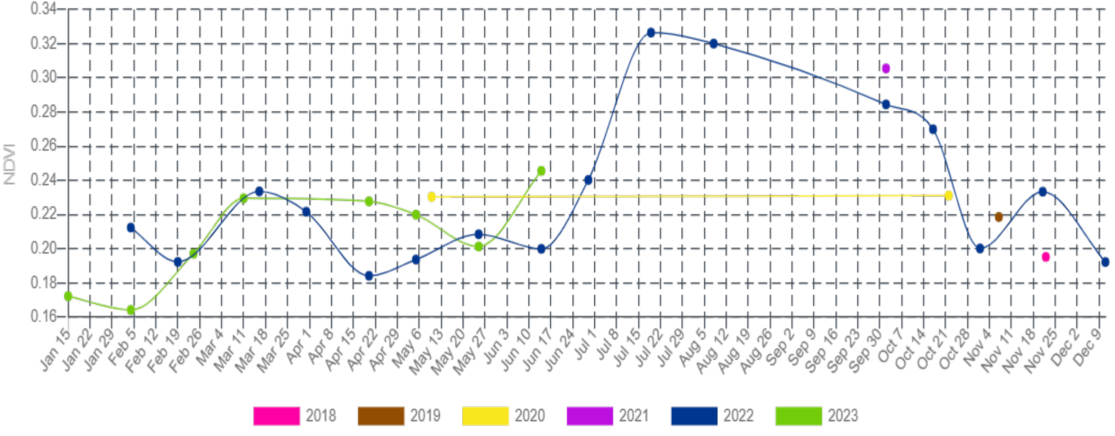

**Figure 7.** Corresponding graph of last five years

## 4. Discussion

### 4.1. Examining time series variables

Because the CIRE (Canopy Index Recovery) features performed better than other vegetative indices (VIs) at forecasting sugarcane yield, we specifically focused on them in our study. We used linear regression analysis, more specifically the ordinary least squares (OLS) method, for each time step of the CIRE to determine the ideal dates for sugarcane yield estimation. According to the findings, the most useful times for yield estimation were just before and during the season's peak, which corresponded to the time from the beginning of the season (or ratooning) to the point at which the canopy is fully closed (or NDVI saturation).

We produce multitemporal images to visualize the R2 (coefficient of determination) and RMSE (root mean square error) scores of the regression model performance employing different predictor variables. We utilized these values to determine the optimal satellite observation dates, which produced the best model performance and were associated with the highest R2 scores and lowest RMSE values. By examining the multitemporal R2 visualization and RMSE scores, we identified the precise times within the growing season that yielded the most accurate and reliable predictions. These dates, which highlighted the times with the highest R2 scores, demonstrated the strongest correlations between the observed CIRE and sugarcane yield. The regression model supported these optimal dates as shown by the lowest RMSE values, representing smaller prediction errors.

It is essential to identify the best satellite observation dates in order to concentrate our data collection efforts during the most informative time periods. The accuracy and precision of our sugarcane yield forecasts may be improved by collecting remote sensing data during these specific dates. Specifically, our analysis showed how the CIRE characteristics outperformed other VIs for predicting sugarcane yield. These are the times, during less and around peak season, in which our data suggested to make yield predictions. We further validated our model by confirming that these dates were associated with the strongest model performance, as indicated by the highest R2 scores, and the smallest prediction errors, as indicated by the lowest RMSE values. These yield predictions, based on CIRE measurements, allowed us to reduce data collection efforts and improve the accuracy of our regression model.

It will be essential to explore the integration of more variables and utilize advanced modeling techniques to further optimize the selection of ideal satellite observation dates to improve the forecasting capabilities of the sugarcane yield prediction model.

We observed that the linear model stabilized once the CIRE time series data was fused with the regression model. This suggests that the model's ability to capture the general trade and dynamics of sugarcane growth is strengthened by including multiple observations from different stages of the growing season. A more comprehensive description of the crop phenological changes is made possible by the collective use of the CIRE time series, which enhances model performance and thus improves R2 scores.

Indeed, the regression models yielded lower R2 scores as they proceeded from individual CIRE data points to sugarcane yield, when relying on single observations alone. This suggests that the intricacy and diversity of sugarcane growth patterns, and in turn, their correlation with yield cannot be fully addressed by resorting simply to isolated observations. It is the time series analysis which allows for a comprehensive appreciation of these growth dynamics and therefore more accurate predictions of cane production.

In conclusion, the study of the CIRE time series showed that R2 scores rise iteratively as the green-up phase advances towards the maximum canopy closure of the season, and then begins to gradually decline during the senescence period. Single observation alone yield poorer R2 scores, whereas merging the CIRE time series yield in a stabilized linear model with enhanced performance. This underscores the need of considering the entire time series and growth dynamics of sugarcane if we are to ever predict cane output accurately.

### 4.2. Exploring the Distinct Factors

The mean CIRE (averaged over the entire CIRE time series), the maximum CIRE (aggregated over the full CIRE time series), and the mean gradient between the peak of the season and harvest were the three independent variables that we looked at in our study to determine how they were distributed. The data was sorted into three groups based on quartiles and these variables were investigated in relation to sugarcane yield.

The analysis showed clear trends between the predictor factors and sugarcane yield volume. For instance, a favorable association between sugarcane yield and the mean CIRE values was found. This suggests a correlation between higher mean CIRE values, which indicate better photosynthetic activity and healthier vegetation, and higher sugarcane yield levels. According to this research, locations with a more active and fruitful development pattern, as shown by higher mean CIRE values—tend to produce larger amounts of sugarcane biomass. In contrast, it was discovered that the senescence rate, the gradient between the peak of the season and the end of the season, and sugarcane production all exhibited negative associations. Lower sugarcane production amounts were correlated with a higher senescence rate, which signifies a more rapid loss in vegetation vigor and health. Similar to this, decreased sugarcane output was associated with a greater decline in vegetation growth between the peak of the season and the end of the season. According to these associations, regions that demonstrate higher yield amounts typically have slower rates of senescence and a more gradual reduction in growth as the season progresses.

### 4.3. Regression model performance

The early estimation capabilities of the suggested framework should be improved and validated through additional research that employs state-of-the-art machine learning techniques, remote sensing technology, and extensive field observations. We can pave the way for more effective and efficient sugarcane production and support overall food security and agricultural sustainability by continually improving our understanding of and capacity to predict sugarcane yield dynamics. The OLS regressor consistently produced the best $R^2$ scores, which indicate a better fit between the predicted and observed values—and the lowest root mean squared error (RMSE), which indicates smaller discrepancies between the anticipated and observed values. This shows that, based on the chosen feature sets, the OLS regression model delivers the most exact and accurate prediction of both sugarcane production and sugar quantity.

The Random Forest (RF) regression model, which showed the highest $R^2$ scores for the phonological indicators in sugarcane yield and sugar quantity estimates, respectively, was also put up against the performance of the OLS regressor in our comparison. The ability of the RF regression model to handle the data's non-linearity and provide accurate estimates of the dependent variables demonstrated promise. A cross-validation strategy was not used when fitting the data, which resulted in over fitting, which was particularly noticeable in the RF regressor's $R^2$ scores above 0.9. A model is considered over fit if it exhibits exceptional performance on data used for training but has trouble generalizing to new data.

We stress the value of using cross-validation methods and model evaluation on a separate test dataset to reduce over fitting. This guarantees that the model's performance is not biased towards the training data and that it can make accurate predictions for data that has not yet been observed.

## References

1. Duveiller, G., R. López-Lozano and B. Baruth. 2013. Enhanced processing of 1-km spatial resolution fAPAR time series for sugarcane yield forecasting and monitoring. Remote Sens (Basel) 5:1091–1116.
2. Zhou, X., H.B. Zheng, X.Q. Xu, J.Y. He, X.K. Ge, X. Yao, T. Cheng, Y. Zhu, W.X. Cao and Y.C. Tian. 2017. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. ISPRS J. Photogramm. Remote Sens. 130:246–255.
3. Food, W. 2020. Statistical Yearbook World Food and Agriculture 2020, Food and Agriculture Organization of the United Nations.
4. Wolanin, A., G. Camps-Valls, L. Gómez-Chova, G. Mateo-García, C. van der Tol, Y. Zhang and L. Guanter. 2019. Estimating crop primary productivity with Sentinel-2 and Landsat 8 using machine learning methods trained with radiative transfer simulations. Remote Sens Environ 225.
5. Jensen, T. A., Baillie, C., Bramley, R. G. V., Panitz, J. H., & Schroeder, B. L. (2012, May). An assessment of sugarcane yield monitoring concepts and techniques from commercial yield monitoring systems. In Proceedings of the Australian Society of Sugar Cane Technologists (Vol. 34, No. 7).
6. Rahman, M. M., & Robson, A. J. (2016). A novel approach for sugarcane yield prediction using landsat time series imagery: A case study on Bundaberg region. Advances in Remote Sensing, 5(2), 93-102.
7. Ashapure, A., Jung, J., Chang, A., Oh, S., Yeom, J., Maeda, M., ... & Smith, W. (2020). Developing a machine learning based cotton yield estimation framework using multi-temporal UAS data. ISPRS Journal of Photogrammetry and Remote Sensing, 169, 180-194.
8. Qadri, S., Furqan Qadri, S., Razzaq, A., Ul Rehman, M., Ahmad, N., Nawaz, S. A., ... & Khan, D. M. (2021). Classification of canola seed varieties based on multi-feature analysis using computer vision approach. International Journal of Food Properties, 24(1), 493-504.
9. Ahmad, I., Singh, A., Fahad, M., & Waqas, M. M. (2020). Remote sensing-based framework to predict and assess the interannual variability of maize yields in Pakistan using Landsat imagery. Computers and Electronics in Agriculture, 178, 105732.
10. Jung, J., Maeda, M., Chang, A., Bhandari, M., Ashapure, A., & Landivar-Bowles, J. (2021). The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. Current Opinion in Biotechnology, 70, 15-22.
11. Xiang, H., & Tian, L. (2011). Development of a low-cost agricultural remote sensing system based on an autonomous unmanned aerial vehicle (UAV). Biosystems engineering, 108(2), 174-190.
12. Nazir, A., Ullah, S., Saqib, Z. A., Abbas, A., Ali, A., Iqbal, M. S., ... & Butt, M. U. (2021). Estimation and forecasting of rice yield using phenology-based algorithm and linear regression model on Sentinel-II satellite data. Agriculture, 11(10), 1026.
13. Liu, Y., Wang, S., Wang, X., Chen, B., Chen, J., Wang, J., ... & Zhu, K. (2022). Exploring the superiority of solar-induced chlorophyll fluorescence data in predicting wheat yield using machine learning and deep learning methods. Computers and Electronics in Agriculture, 192, 106612.
14. Meng, W. A. N. G., Tao, F. L., & Shi, W. J. (2014). Corn yield forecasting in northeast china using remotely sensed spectral indices and crop phenology metrics. Journal of Integrative Agriculture, 13(7), 1538-1545.
15. Han, X., Wang, J., & Li, N. IOP Conference Series: Earth and Environmental Science.
16. Mountrakis, G., J. Im and C. Ogole. 2011. Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing 66:247–259.
17. Yildirim, T., Moriasi, D. N., Starks, P. J., & Chakraborty, D. (2022). Using Artificial Neural Network (ANN) for Short-Range Prediction of Cotton Yield in Data-Scarce Regions. Agronomy, 12(4), 828.
18. Ferencz, C., P. Bognár, J. Lichtenberger, D. Hamar, G. Tarcsai, G. Timár, G. Molnár, S. Pásztor, P. Steinbach, B. Székely, O.E. Ferencz and I. Ferencz-Árkos. 2004. Crop yield estimation by satellite remote sensing. Int J Remote Sens 25.
19. Löw, F., Biradar, C., Dubovyk, O., Fliemann, E., Akramkhanov, A., Narvaez Vallejo, A., & Waldner, F. (2018). Regional-scale monitoring of cropland intensity and productivity with multi-source satellite image time series. GIScience & Remote Sensing, 55(4), 539-567.
20. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.
21. Dimov, D., J.H. Uhl, F. Löw and G.N. Seboka. 2022. Sugarcane yield estimation through remote sensing time series and phenology metrics. Smart Agricultural Technology 2:100046.
22. Lisboa, I.P., M. Damian, M.R. Cherubin, P.P.S. Barros, P.R. Fiorio, C.C. Cerri and C.E.P. Cerri. 2018. Prediction of sugarcane yield based on NDVI and concentration of leaf-tissue nutrients in fields managed with straw removal. Agronomy 8.
23. Fernandes, J. L., Ebecken, N. F. F., & Esquerdo, J. C. D. M. (2017). Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. International journal of remote sensing, 38(16), 4631-4644.
24. Farooq, W. and C. Rittgers. 2023. This report contains assessments of commodity and trade issues made by USDA staff and not necessarily statements of official U.S. government policy.
25. Canata, T.F., M.C.F. Wei, L.F. Maldaner and J.P. Molin. 2021. Sugarcane yield mapping using high-resolution imagery data and machine learning technique. Remote Sens (Basel) 13:1–14.
26. Zeng, L., B.D. Wardlow, D. Xiang, S. Hu and D. Li. 2020. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. Remote Sens Environ 237.

27. Turab, A., L.G. Pell, D.G. Bassani, S. Soofi, S. Ariff, Z.A. Bhutta and S.K. Morris. 2014. The community-based delivery of an innovative neonatal kit to save newborn lives in rural Pakistan: Design of a cluster randomized trial. BMC Pregnancy Childbirth 14.
28. Tucker, C.J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ 8.
29. Gitelson, A.A., A. Viña, V. Ciganda, D.C. Rundquist and T.J. Arkebauer. 2005. Remote estimation of canopy chlorophyll content in crops. Geophys Res Lett 32.
30. Reed, B.C., J.F. Brown, D. VanderZee, T.R. Loveland, J.W. Merchant and D.O. Ohlen. 1994. Measuring phenological variability from satellite imagery. Journal of Vegetation Science 5.
31. Wolanin, A., G. Camps-Valls, L. Gómez-Chova, G. Mateo-García, C. van der Tol, Y. Zhang and L. Guanter. 2019. Estimating crop primary productivity with Sentinel-2 and Landsat 8 using machine learning methods trained with radiative transfer simulations. Remote Sens Environ 225.
32. Zhang, X., Gao, F., Wang, J., & Ye, Y. (2021). Evaluating a spatiotemporal shape-matching model for the generation of synthetic high spatiotemporal resolution time series of multiple satellite data. International Journal of Applied Earth Observation and Geoinformation, 104, 102545.
33. Younes, N., K.E. Joyce and S.W. Maier. 2021. All models of satellite-derived phenology are wrong, but some are useful: A case study from northern Australia. International Journal of Applied Earth Observation and Geoinformation 97.