agronomy

MDPI

Article

Predicting Sugarcane Yield Through Temporal Analysis of Satellite Imagery During the Growth Phase

Julio Cezar Souza Vasconcelos [1], Caio Simplicio Arantes [1], Eduardo Antonio Speranza [2], João Francisco Gonçalves Antunes [2], Luiz Antonio Falaguasta Barbosa [2] and Geraldo Magela de Almeida Cançado [2,*]

[1] Fundação de Apoio a Pesquisa e ao Desenvolvimento—FAPED, Sete Lagoas 35700-039, Brazil; juliocezarvasconcelos@hotmail.com (J.C.S.V.); caio.arantes@colaborador.embrapa.br (C.S.A.)
[2] Embrapa Digital Agriculture, Campinas 13083-886, Brazil; eduardo.speranza@embrapa.br (E.A.S.); joao.antunes@embrapa.br (J.F.G.A.); luiz.barbosa@embrapa.br (L.A.F.B.)
* Correspondence: geraldo.cancado@embrapa.br

Abstract: This research investigates how to estimate sugarcane (*Saccharum officinarum* L.) yield at harvest by using an average satellite image time-series collected during the growth phase. This study aims to evaluate the effectiveness of various modeling approaches, including a heteroskedastic gamma regression model, Random Forest, and Artificial Neural Networks, in predicting sugarcane yield based on satellite-derived vegetation indices and environmental variables. Key covariates analyzed include sugarcane varieties, production cycles, accumulated precipitation during the growth phase, and the mean GNDVI vegetation index. The analysis was conducted in two locations over two consecutive growing seasons. The research emphasizes the integration of satellite data with advanced statistical and machine learning techniques to enhance yield prediction in agricultural systems, specifically focusing on sugarcane cultivation. The results indicate that the heteroskedastic gamma regression model outperformed the other methods in explaining yield variability, particularly in commercial sugarcane fields, achieving a Coefficient Determination (R^2) of 0.89. These findings highlight the potential of these models to support informed decision-making and optimize agricultural practices, providing valuable insights for precision farming. Overall, the results of this study represent an initial step toward developing more robust models for predicting sugarcane yield. Future work will involve incorporating additional variables to better assess the impacts of environmental stresses, such as high temperatures and water deficits, on the crop's agronomic performance.

Keywords: digital agriculture; *Saccharum* spp.; precision farming; crop yield; statistical model; machine learning

Academic Editor: Chengming Sun

Received: 11 February 2025
Revised: 12 March 2025
Accepted: 18 March 2025
Published: 24 March 2025

Citation: Vasconcelos, J.C.S.; Arantes, C.S.; Speranza, E.A.; Antunes, J.F.G.; Barbosa, L.A.F.; Cançado, G.M.d.A. Predicting Sugarcane Yield Through Temporal Analysis of Satellite Imagery During the Growth Phase. *Agronomy* **2025**, *15*, 793. https://doi.org/10.3390/agronomy15040793

1. Introduction

Brazil is a global leader in sugarcane production, consistently ranking as the world's largest producer. In 2022, around 724 million tons of sugarcane were produced in the country, remaining ahead of India (439 million tons) and China (103 million tons) [1]. This strategic crop is a cornerstone of the Brazilian agricultural sector, contributing significantly to ethanol and sugar production. The continuous adoption of advanced technologies by Brazilian farmers and industries of the sugar and alcohol sector has played a crucial role in maintaining this leadership position, even in the face of recent climatic challenge scenarios [2].

Among these technologies, images obtained from remote sensing have been gaining prominence in recent years, mainly due to the increasing availability of multispectral images with high temporal and spatial resolution. Therefore, obtaining recurrent maps of vegetation indices throughout all crop phenological stages contributes substantially to developing increasingly accurate yield estimation models [3].

On the other hand, it is essential to recognize that data collected in field experiments are often conditioned to external factors that can impact sugarcane yield. Examples include agricultural management, water availability, climatic conditions, and the incidence of pests and diseases. Due to these influences, such data often exhibit heteroskedastic behavior, i.e., non-constant variance, necessitating the application of appropriate statistical techniques. For instance, Prataviera et al. [4] proposed a log-Weibull regression model for interval-censored data, incorporating a regression structure for the location (μ) and scale (σ) parameters, enabling the modeling of non-proportional hazards and heteroskedasticity. One application of this model involved analyzing the effects of different treatments in dairy cows, where significant differences among treatments were observed. Similarly, Santos et al. [5] evaluated the impact of *Bacillus* inoculants on sugarcane using a heteroskedastic semiparametric GA regression model. The model predicted significant yield increases and confirmed the efficacy of the inoculant in reducing the use of phosphate fertilizers. Vasconcelos et al. [6] proposed a heteroskedastic regression model to analyze the effects of temperature and wood species on shrinkage volume.

Furthermore, when vegetation indices from satellite images are used in addition to field data, a significant increase in the dimensionality of the dataset is commonly observed, given the number of indices available in the literature that can be potential indicators of biomass accumulation and physiological development. In this context, supervised machine learning (ML) techniques, capable of performing multivariate analyses with high precision, have been widely used in the literature to support the development of yield estimation models [7,8]. While Artificial Neural Networks (ANNs) were the first computational models that introduced the ML concepts and were inspired by the biological neural networks of the human brain [9], the Multi-Layer Perceptrons (MLPs) are a type of ANN with one or more hidden layers that allow them to learn complex patterns in the data. MLPs are widely used for various tasks (including regression) due to their ability to fit any continuous function [10]. Another widely used type of ML algorithm is decision trees [11]. However, with the increasing complexity of the data used to generate regression models, isolated trees have become insufficient, leading to more robust techniques, such as Random Forest (RF) [12], which operates by building many decision trees during training. This process involves randomly selecting subsets of data and features to train each tree, reducing overfitting and improving generalization.

This study aimed to assess the feasibility of using time-series analysis of satellite images to monitor the development of two experimental sugarcane fields over two growth cycles. The objective was to evaluate the efficiency and accuracy of yield estimations generated by various models using this dataset. This study employed the Multilayer Perceptron (MLP) approach, a machine learning technique based on ANNs widely discussed in the literature, particularly about estimating sugarcane yield and other crops using remote sensing images. For instance, ref. [13] developed a model to estimate sugarcane yield at the municipal level by using vegetation indices derived from MODIS sensor time-series and MLP ensembles. More recently, ref. [14] implemented a deep learning-based MLP approach to estimate soil health and parameters related to wheat crop yield at the field level, using images from the Sentinel-1 and Sentinel-2 satellites.

The Random Forest (RF) approach was chosen for this study due to its moderate complexity in modeling and interpretation, low risk of overfitting, simplified hyperpa-

rameter adjustments, and effective performance with reduced datasets. These advantages are not present in Multi-Layer Perceptrons (MLPs). Furthermore, using RF to estimate sugarcane yield is a relatively recent topic in the literature. For instance, ref. [15] employed RF models to estimate the regional yield of sugarcane-growing areas in Australia, using various climatic variables as predictors, including different measures of precipitation and temperature. Similarly, ref. [7] analyzed four years of yield data from approximately 5400 sugarcane plots (with an average area of 9 ha each) located in the western region of São Paulo, Brazil. They generated RF-based models incorporating vegetation indices from Landsat satellite imagery and meteorological and agronomic data such as soil type and production environments. Ref. [16] investigated the sugarcane yield of a study area covering 10,000 hectares in Ethiopia. They worked with three distinct sets of predictors: multi-temporal attributes (time-series of vegetation indices), phenological metrics extracted from these time-series, and other spatiotemporal variables. Among the regression models tested, RF showed the highest performance, achieving R^2 scores of up to 0.84 for sugarcane yield. Lastly, ref. [17] conducted a systematic review of nearly 1400 papers on estimating sugarcane yield using remote sensing data. Their findings indicated that RF is the predominant choice among various regression-based models. Although both Multi-Layer Perceptron (MLP) and Random Forest (RF) are considered machine learning techniques, they exhibit distinct biases toward ANNs and decision trees, respectively. Therefore, this study employed them to represent a range of machine learning approaches for comparing their performance against a statistical regression model. The existing literature presents various methods for estimating sugarcane yield through data modeling, with some approaches proving more accurate than others.

This study highlights that heteroskedastic gamma regression (HGR) offers an innovative and superior alternative to traditional machine learning techniques for estimating sugarcane yield. The improved performance of the HGR model was supported by experimental data collected in a controlled environment, which effectively captured the crop's behavior throughout its growth stages. These data accounted for crucial factors such as the genetic and phenotypic variability of the sugarcane varieties used and the differences observed across growing seasons and geographic locations. Additionally, the findings were validated in commercial fields, demonstrating the practical potential of this approach for real-world sugarcane cultivation.

2. Materials and Methods

2.1. Experimental and Commercial Fields

2.1.1. Field Experiment Methodology

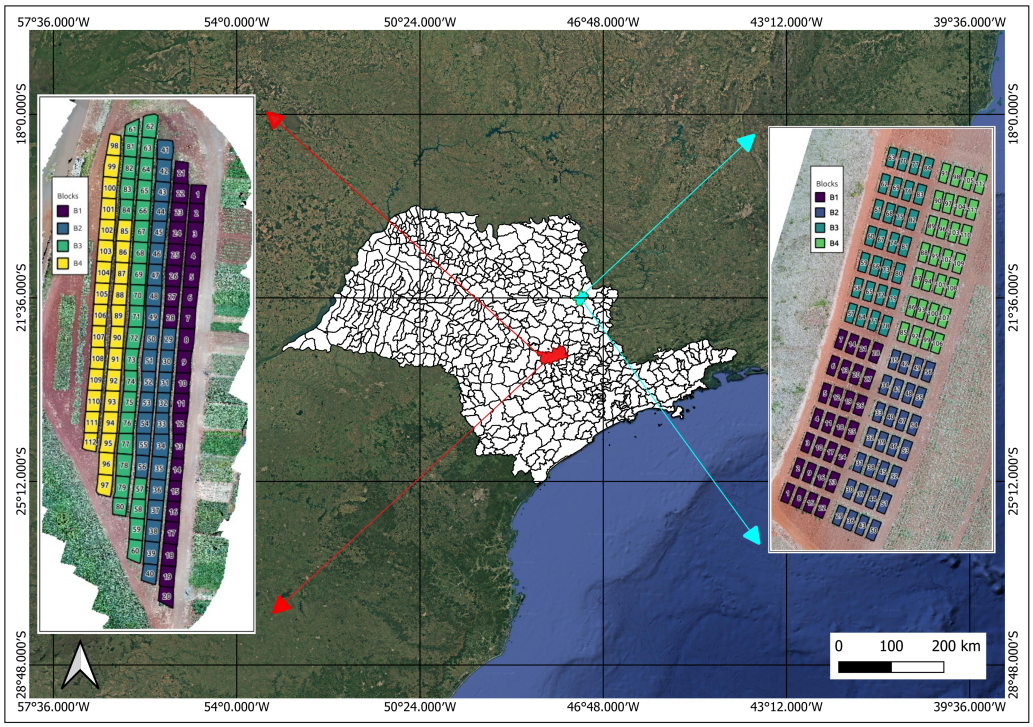
This study was conducted with four sugarcane varieties cultivated in two distinct locations during the 2020/2021 and 2021/2022 growth seasons (cycles). The sugarcane varieties were selected based on the maturity timing (see Table 1). Each experimental field consisted of 112 plots with six rows of sugarcane measuring 1.5 m between furrows and 10 m in length (Figure 1). For each sugarcane variety, there were 28 replicates per experimental field, totaling 448 experimental plots. The random block design with four blocks per experimental field was adopted in this study to enable the measurement of random environmental variation.

Table 1. Sugarcane varieties identification and the respective maturity timing.

Sugarcane Varieties	Maturity Timing
CTC1007 (V1)	Normal
RB966928 (V2)	Short
CV0618 (V3)	Medium
CV7870 (V4)	Normal

Both experiments had two production cycles, one for the cane plant cycle (2020/2021) and the other for the first ratoon cycle (2021/2022).

The experimental field in the rural area of Piracicaba, São Paulo, Brazil (−22.773005, −47.580135) utilized two-month-old pre-sprouted seedlings (PSSs) cultivated in a greenhouse for each variety. A drip irrigation system was implemented during the first three months after transplanting the plantlets into the field to promote robust root development. In contrast, the experimental field in the rural area of Tambaú, São Paulo, Brazil (−21.708543, −47.246643) employed sugarcane stalks with an average length of 1.5 m as plantlets for direct field planting. Both experimental fields followed management practices aligned with standard commercial sugarcane cultivation in Brazil, ensuring consistent agronomic standards across all varieties.



(**a**) (**b**)

Figure 1. Aerial perspective of experimental fields (**a**,**b**), highlighting the 112 plots in each field. (**a**) is the experimental field located in Piracicaba, Sao Paulo, Brazil; (**b**) is the experimental field located in Tambau, Sao Paulo, Brazil.

Sugarcane plots were harvested and measured in tons of cane per hectare (TCH, t ha^{-1}) at the end of their growth cycle, once the stalks reached peak maturity. This maturity was determined by measuring the concentration of Total Recoverable Sugars (TRS). Yield was assessed based on the average stalk weight obtained from experimental replicates for each treatment across all locations (Piracicaba and Tambaú) over two consecutive harvest cycles: 2020/2021 and 2021/2022.

The data collected from the experimental fields were used for both training and testing the models. A stratified split of 70% for training and 30% for testing was applied, ensuring that the distribution of sugarcane varieties and growth cycles was preserved in both subsets. This approach guarantees that the model was tested using information obtained from field experiments rather than commercial data.

2.1.2. Data for Commercial Areas

To evaluate the effectiveness of the models in the real environment of sugarcane production, 12 commercial sugarcane plots cultivated in the state of São Paulo, Brazil, were analyzed during the 2022 growing season. These plots had known observed yield data, which were used to compare with the results generated by the models. It is important to note that this evaluation was not part of the model's training or testing process; instead, it aimed to assess the model's ability to generalize to new, independent data that reflects actual production conditions. Detailed yield metrics, sugarcane cycle, varieties, and geographic coordinates of the commercial plots are provided in Table S2 in the Supplementary Materials.

2.2. Dataset Description

PlanetScope's multispectral imagery was used to validate sugarcane yield prediction models in commercial fields located at locations A and B. The delineated plots, represented as polygons, show their corresponding areas in hectares. These plots serve as the spatial units for model validation and accuracy assessment (Figures 2 and 3).

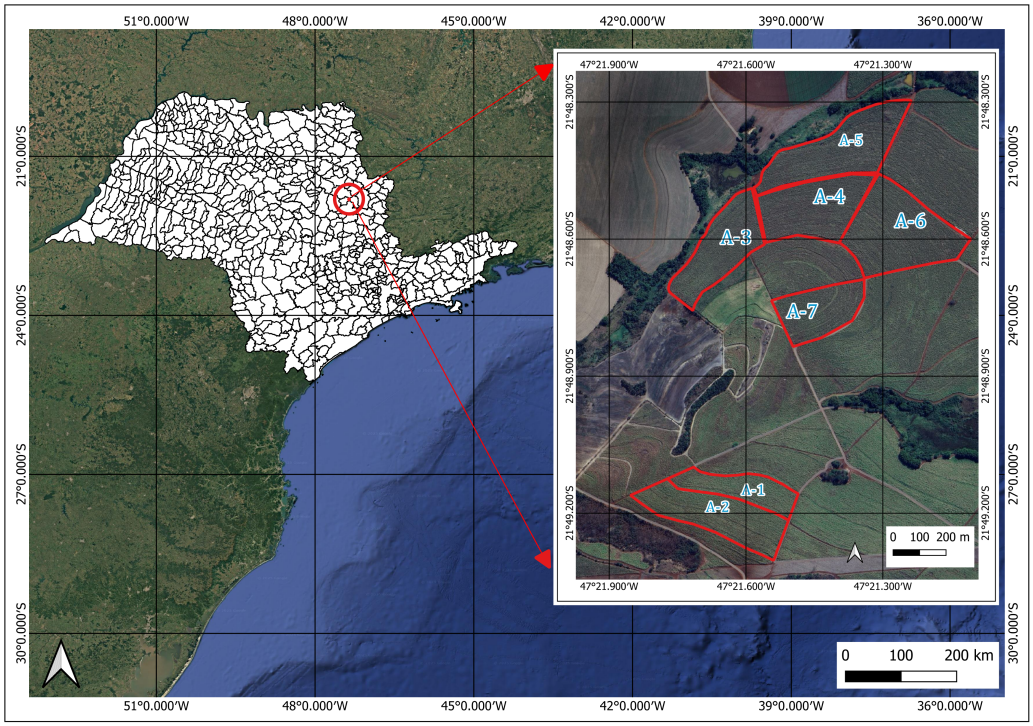


Figure 2. Location A at the Schiavon Farm with five production plots indicated as A-1 (4.46 ha), A-2 (8.03 ha), A-3 (7.22 ha), A-4 (8.58 ha), A-6 (11.07 ha), and A-7 (5.75 ha) during the growth season of 2022. São Paulo, Brazil.

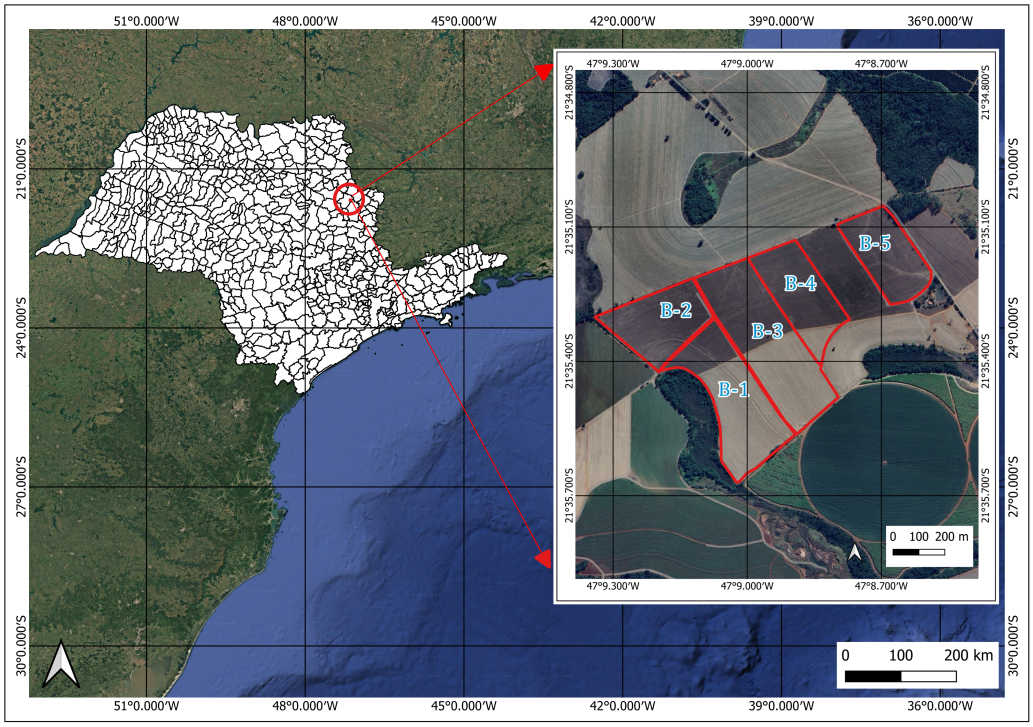


Figure 3. Location B at the Aurora Farm with five production plots indicated as B-1 (12.66 ha), B-2 (8.64 ha), B-3 (17.1 ha), B-4 (8.78 ha), and B-5 (8.29 ha) during the growth season of 2022. São Paulo, Brazil.

2.2.1. Vegetation Indices

A vegetation index is an algebraic combination of various spectral bands designed to highlight the vigor and properties of vegetation [18]. It translates information contained in multispectral or RGB images through transformations of reflectance factors by employing operations such as addition, subtraction, and ratio between spectral bands to emphasize the spectral response of vegetation as a function of canopy cover over the soil [19]. An ideal vegetation index should detect slight variations in vegetation phenological phases while mitigating the influence of soil conditions and types, scene illumination geometry, and atmospheric conditions [20].

For the experiments conducted in this study, the data source was satellite imagery obtained from the SuperDove nano-satellite, which is part of the PlanetScope constellation [21], which generates images daily for any global location. The satellite imagery includes eight multispectral bands: coastal blue, blue, green I, green, yellow, red, red edge, and near-infrared, all with a spatial resolution of 3 m per pixel. For the analysis, images from two sugarcane experimental fields during the 2020/2021 and 2021/2022 growing seasons were included, although images obstructed by clouds or shadows were systematically excluded. Variations in image availability across locations and growing seasons resulted in an imbalanced dataset, with more usable imagery captured during the winter months. In total, for experiment 1 conducted in Piracicaba, São Paulo, Brazil, 57 images from the 2020/2021 season and 131 images from the 2021/2022 season were used. For experiment 2 in Tambaú, São Paulo, Brazil, 121 images from the 2020/2021 season and 96 images from the 2021/2022 season were used. During the acquisition of remote sensing imagery for temporal analysis, differences in the availability of cloud-free, high-quality images were observed across various geographical locations and harvest cycles. These variations resulted in an imbalanced distribution of valid data points among the four datasets used to derive vegetation indices (VIs). To address potential biases introduced by unequal sample sizes, mean VI values were calculated for each phenological stage

of the sugarcane crop (such as tillering, grand growth, and maturation) during model development, rather than using cumulative sums. Statistical evaluations of the models confirmed that the imbalance in the dataset did not negatively impact predictive accuracy.

Based on a scrutinized evaluation of the literature, we identified six vegetation indices that demonstrated statistically significant differentiation potential to be used in this study: EVI [22], EVI2 [23], GNDVI [24], HUE [25], NDVI [26], and OSAVI [27]. These indices are correlated with important traits throughout the production cycle, such as chlorophyll, biomass, leaf area index, nitrogen availability, and soil color. Historically, NDVI is the vegetation index that best correlates with crop biomass, which for crops such as sugarcane, is usually the main indicator of yield [13,28,29]. Additionally, EVI also has a good correlation with the availability of nitrogen in the crop [30]; GNDVI adds to the model a good correlation with the availability of water in the crop, which, in turn, is also correlated with final yield [31]. Recent studies show the use of these indexes to yield prediction approaches for other crops, such as wheat [32,33]. Regarding the sugarcane maturation phenological stage with a high leaf area index, it is common to observe saturation of values for NDVI, making it unable to differentiate areas of higher and lower yield. In this case, variations of NDVI, such as EVI2, which considers factors related to band reflectance, were used to try to solve this drawback [5,34]. Finally, HUE and OSAVI can be used to evaluate phenological stages, where the amount of soil could contribute to yield prediction. When exposed, the HUE index can show significant differences in soil coloration, and the OSAVI can identify differences in soil salinity even with a certain canopy cover [35]. These variations are also essential to recognize, as they can influence final yield, especially in large-scale production crops, such as sugarcane. In recent studies, these indexes have been used in approaches to yield prediction in other crops, such as rice [36] and corn [37].

After selecting the vegetation indices, their average values over the crop development period, which spans from 120 to 250 days after the start of the crop cycle, were calculated. This analysis considered 224 virtual experimental plots, with 112 located at each experimental field, across two growth cycles: the cane plant cycle and the first ratoon cycle. This approach resulted in 448 data samples for each vegetation index.

2.2.2. Weather Data

Cumulative precipitation data for the experimental and commercial sugarcane fields were sourced from the NASA Power database, following the methodological protocols outlined by Monteiro et al. [38]. The dataset spans the period from planting through key crop growth phases and was directly accessed and retrieved via The Power Data Access Viewer [39].

2.2.3. Variables Definitions

The explanatory variables x_{i1} to x_{i10} were considered to evaluate their influence on sugarcane yield, measured in tons of cane per hectare (TCH, t ha^{-1}). The variable y_i, representing TCH, was analyzed with the independent variables.

The variable x_{i1} represents the blocks (1, 2, 3, and 4), and since it is a factor with more than two levels, three dummy variables (p_{i1}, p_{i2}, p_{i3}) were required. The variable x_{i2} refers to the varieties (V1, V2, V3, and V4) and similarly requires three dummy variables (c_{i1}, c_{i2}, c_{i3}). For x_{i3}, which corresponds to the sugarcane growth cycle, two categories are considered: the cane plant cycle and the first ratoon cycle.

The variable x_{i4} represents the accumulated precipitation during the growth phase, and x_{i5} to x_{i10} correspond to the mean values of the vegetation indices EVI, EVI2, GNDVI, HUE, NDVI, and OSAVI during the growth phase.

- y_i: Tons of cane per hectare (TCH, t ha^{-1});

- x_{i1}: Block (1, 2, 3, and 4). In this case, being a factor with more than two levels, three dummy variables are required (p_{i1}, p_{i2}, p_{i3});
- x_{i2}: Varieties (V1, V2, V3, and V4). Here, being a factor with more than two levels, three dummy variables are defined (c_{i1}, c_{i2}, c_{i3});
- x_{i3}: Cycle (cane plant and first ratoon);
- x_{i4}: Accumulated precipitation (during the growth phase);
- x_{i5}: Mean EVI (during the growth phase);
- x_{i6}: Mean EVI2 (during the growth phase);
- x_{i7}: Mean GNDVI (during the growth phase);
- x_{i8}: Mean HUE (during the growth phase);
- x_{i9}: Mean NDVI (during the growth phase);
- x_{i10}: Mean OSAVI (during the growth phase).

For $i = 1, \ldots, 448$.

The explanatory variables x_{i1}, x_{i2}, and x_{i3} are categorical, representing distinct categories or groups. In contrast, the variables x_{i4} through x_{i10} are continuous, as they are numerical and can take on a range of values. This classification ensures appropriate statistical treatment for each type of variable.

2.3. Statistical and Machine Learning Models

2.3.1. Evaluation of Variance Inflation Factor

To ensure model robustness and mitigate multicollinearity issues, all explanatory variables were initially included in the analysis process: block, varieties, cycle, accumulated precipitation (during the growth phase), mean EVI (during the growth phase), mean EVI2 (during the growth phase), mean GNDVI (during the growth phase), mean HUE (during the growth phase), mean NDVI (during the growth phase), and mean OSAVI (during the growth phase). The Variance Inflation Factor (VIF) [40] evaluation was conducted to identify and remove highly collinear variables, resulting in a more concise and robust selection.

2.3.2. Heteroskedastic GA Regression Model

- Gamma Probability Distribution

The gamma (GA) probability distribution is widely used to model asymmetric data with positive values skewed to the right. This distribution is frequently applied in reliability and survival studies. The probability density function (pdf) $f(y)$ (Equation (1)), as derived and reparameterized by McCullagh and Nelder [41] and Johnson et al. [42], employed in this study is expressed as follows:

$$f(y|\mu,\sigma) = \frac{y^{(1/\sigma^2-1)}\exp[-y/(\sigma^2\mu)]}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)} \quad \text{for } y > 0,\ \mu > 0 \text{ and } \sigma > 0. \tag{1}$$

The parameters μ and σ correspond to the mean and the square root of the dispersion parameter, respectively, with the parametrization used derived from the `gamlss` package [43]. In this context, μ represents the mean of the gamma (GA) distribution, and σ is the square root of the dispersion parameter in a Generalized Linear Model (GLM) [44] with a gamma distribution.

- Structure and Estimation

In this study, the heteroskedastic GA regression model, based on the gamma distribution, was applied to model the response variable $Y_i \sim \text{GA}(\mu_i, \sigma_i)$, with a structure that allows mean-dependent variability. This model employs two systematic components to

estimate the mean μ_i and the parameter σ_i, representing the relative standard deviation or the coefficient of variation.

The modeling is conducted under the following expressions:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \beta_1 \quad \text{and} \quad g(\sigma_i) = \nu_i = \mathbf{v}_i^T \beta_2, \tag{2}$$

where $\mathbf{x}_i$ and $\mathbf{v}_i$ are vectors of predictor variables, β_1 and β_2 are vectors of coefficients to be estimated, and $g(\cdot)$ is a logarithmic link function to ensure positive values.

The parameters estimate $\hat{\theta} = (\beta_1^T, \beta_2^T)^T$ are obtained using the maximum likelihood method by maximizing the logarithm of the likelihood function (Equation (3)):

$$l(\theta) = \sum_{i=1}^{n} \left[\left(\frac{1}{\sigma_i^2} - 1 \right) \log y_i - \frac{y_i}{\sigma_i^2 \mu_i} - \frac{1}{\sigma_i^2} \log(\sigma_i^2 \mu_i) - \log \Gamma \left(\frac{1}{\sigma_i^2} \right) \right]. \tag{3}$$

To obtain the estimates $\hat{\theta}$, the `gamlss` package in the `R` software (v. 4.4.3) was used.

This model represents a particular case of the semiparametric heteroskedastic GA regression model used by Santos et al. [5], who evaluated sugarcane yield in response to applying a phosphate-solubilizing microbial inoculant. In this study, there are no variables with nonlinear effects on the response variable, which led to the choice of the parametric version of the model, as it provides a practical fit to the analyzed data.

2.3.3. Covariate Selection with GAIC

Next, the `stepGAICAll.A`$(\cdot)$ function from the `gamlss` package [43] was used to refine covariate selection based on the Generalized Akaike Information Criterion (GAIC) applied to all distribution parameters. This method, described by Stasinopoulos et al. [45] as an adaptation of the `stepAIC`$(\cdot)$ function from the `MASS` package [46], enabled the individual analysis of each variable concerning the model parameters. This approach facilitated the creation of different covariate subsets for each parameter, enhancing the precision and suitability of the final adjusted model.

2.3.4. Machine Learning Approaches: Random Forest and Neural Networks

This study expanded the modeling framework by incorporating RF regression and ANN algorithms, exploring their potential as alternatives to the heteroskedastic GA regression model. The performances of these algorithms were evaluated using the R^2 and a suite of error metrics, ensuring a comprehensive assessment of their predictive accuracy and reliability.

To ensure consistency and comparability with the statistical model, the dataset was divided into training and testing subsets using the same methodology, with a fixed random seed to ensure reproducibility. Furthermore, as described in [47], the Grid Search Method was employed to optimize the hyperparameters for both algorithms. This method performs an exhaustive search across a predefined subset of the hyperparameter space, systematically identifying the optimal configuration for each regressor to enhance model performance. The optimization of hyperparameters was achieved using the GridSearchCV class in the Python 3.7 Scikit-learn library.

- Random Forest

Random Forest (RF) regression is an ensemble learning algorithm that combines many regression trees. A regression tree represents a set of conditions or constraints that are hierarchically organized and successively applied from the root to a leaf of the tree [48,49].

Breiman developed Random Forest (RF) [12] to improve the Classification And Regression Tree (CART) method by combining a large set of decision trees. It consists of a combination of tree predictors, where each tree depends on the values of a random

vector independently sampled and identically distributed for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large.

The grid search method showed the hyperparameters to be employed on RF, and they are described as follows:

1. `n_estimators`: Search space: [100, 200, 300, 400, 500]; Best found: 400.
2. `max_depth`: Search space: [10, 20, 30, 40, 50]; Best found: 50.
3. `min_samples_split`: Search space: [2, 5, 10]; Best found: 10.
4. `min_samples_leaf`: Search space: [1, 2, 4]; Best found: 4.
5. `max_features`: Search space: ['auto', 'sqrt', 'log2']; Best found: 'log2'.

The hyperparameters are, respectively, constituted by (1) n estimators, which determine the number of decision trees in the forest; (2) min samples split, specifying the minimum number of samples required to split an internal node; (3) min samples leaf, defining the minimum number of samples that must be present in a leaf node; (4) max features, dictating the maximum number of features considered when searching for the best split; and (5) max depth, setting the maximum depth permitted for each decision tree.

- Neural Network

Artificial Neural Networks are computational architectures inspired by the functioning of the human brain. These networks can perform functional modeling and effectively manage linear and nonlinear relationships by learning from data and generalizing to previously unseen scenarios. Among the most widely used ANNs is the Multi-Layer Perceptron (MLP). This potent modeling tool applies a supervised training procedure using examples of data with known outputs [50].

To comprehend the structure and function of the Multi-Layer Perceptron (MLP), it is essential first to examine its foundational components: the single-neuron perceptron and the single-layer perceptron. The single-neuron perceptron represents the simplest form of an ANN, consisting of a single output node connected to all input nodes, illustrated in Figure 4. For a perceptron with n inputs ($i = 0, 1, \ldots, n$), each input X_i is associated with a corresponding weight W_i. These inputs represent features or variables, while the output Y corresponds to a prediction.

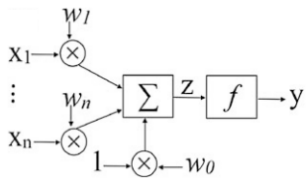


Figure 4. Perceptron model, where $X_1 \ldots X_n$ corresponds to the inputs, $W_0 \ldots W_n$ to the weights, Z represents the sum of the products of the entries and their corresponding weights as calculated by the neuron and then input into the activation function f, and Y corresponds to the output, which is the application of activation function f on Z.

The perceptron model operates through three fundamental steps described below:

1. Weighting Step: Each input feature value (x_i) is multiplied by its associated weight (w_i), resulting in a weighted input ($x_i w_i$).
2. Summation Step: The weighted inputs are aggregated through summation, yielding

$$S = \sum_{i=0}^{n} x_i w_i \tag{4}$$

3. Transfer Step: An activation function f, also referred to as a transfer function, is applied to the summed value S. This function transforms the linear combination of inputs into the perceptron's final output y. The output can be mathematically expressed as follows:

$$y = f\left(\sum_{i=0}^{n} x_i w_i\right) \quad (5)$$

The activation function determines the output by mapping the weighted sum to a classification or regression, forming the perceptron's decision-making basis. This process underpins the computational framework of more complex architectures like the MLP.

An extra operation for MLP, compared to RF, was the scale of features using StandardScaler [47]. The use of the grid search method showed the best hyperparameters to be employed on it, and they are described as follows:

1. `solver`: Search space: ['adam', 'lbfgs', 'sgd']; Best found: 'sgd'.
2. `momentum`: Search space: [0.1, 0.3, 0.5, 0.7, 0.9]; Best found: 0.5.
3. `max_iter`: Search space: [200, 500, 1000, 2000, 5000]; Best found: 500.
4. `learning_rate_init`: Search space: [0.001, 0.01, 0.1]; Best found: 0.1.
5. `learning_rate`: Search space: ['constant', 'invscaling', 'adaptive']; Best found: 'invscaling'.
6. `hidden_layer_sizes`: Search space: [(50,), (100,), (50, 50), (100, 50), (50, 100), (100, 100)]; Best found: (100,).
7. `alpha`: Search space: [0.0001, 0.001, 0.01, 0.1]; Best found: 0.01.
8. `activation`: Search space: ['identity', 'logistic', 'tanh', 'relu']; Best found: 'tanh'.

The hyperparameters are, respectively, constituted by (1) the solver, which specifies the optimization algorithm used to adjust the model's weights; (2) momentum, a parameter that influences the contribution of previous weight updates to accelerate convergence; (3) maximum iterations (max iter), the maximum number of epochs permitted for training; (4) initial learning rate (learning rate init), which defines the initial step size for weight adjustments; (5) learning rate, determining the scheme by which the learning rate is updated during training; (6) hidden layer sizes, a specification of the number of neurons in each hidden layer; (7) alpha, a regularization parameter that penalizes large weights to mitigate overfitting; and (8) activation, the function applied at each neuron to introduce nonlinearity into the network.

3. Results and Discussion

The field experiments with sugarcane were designed to maximize genetic and agronomic diversity representation across contrasting Brazilian cultivation systems. Four commercially relevant sugarcane cultivars were selected to capture variability in growth maturity characteristics (short-, medium-, and normal phenology) and yield potential. The trials were conducted at two geographically distinct locations to investigate the interactions between genotype and environment (G × E): Tambaú, which has a semi-arid climate with an average annual rainfall of 550 mm and sandy soil type with moderate fertility, and Piracicaba, which is characterized by a temperate humid climate with an average yearly rainfall of 1300 mm and latosols and podzolic soil types with high fertility.

Agronomic management protocols varied between the experimental fields to reflect regional farming practices. The Tambaú trial used conventional stalk planting with locally sourced propagation materials from nearby commercial fields, simulating traditional farming conditions. In contrast, the Piracicaba trial employed an advanced propagation system that involved pre-sprouted seedlings (PSSs), which were acclimatized in a controlled nursery environment before being transplanted into the field. This dual approach

allowed for a comparative analysis of genetic diversity responses to abiotic stress factors, such as water availability and soil fertility, as well as the impact of propagation technology on crop establishment efficiency. Figure 5 presents a comprehensive time-series analysis of six vegetation indices (EVI, EVI2, GNDVI, HUE, NDVI, OSAVI) across the sugarcane crop cycle, encompassing both the plant cane and first ratoon phases for the experimental fields located in Piracicaba (a) and Tambaú (b). Preliminary analysis identified optimal temporal windows (indicated by the regions highlighted in gray and pink in Figure 5) corresponding to the peak of biomass accumulation during critical growth phases. The ideal time frame for evaluation was determined by the phenological stage of the sugarcane that showed the best response to vegetation indices during its vegetative growth phase. During this phase, the plant experiences rapid growth, which is reflected in increased canopy coverage and fully expanded leaves, ultimately leading to a more significant increase in biomass. This vegetative phase typically occurs approximately two to three months before the crop matures and is ready for harvest. These intervals align with key phenological stages of rapid physiological development, including maximum leaf canopy expansion, which drives photosynthetic efficiency and biomass production. The selected periods were integrated into predictive models to establish robust correlations between spectral indices and yield outcomes, leveraging the temporal sensitivity of vegetation indices to crop vigor during phases of heightened metabolic activity.

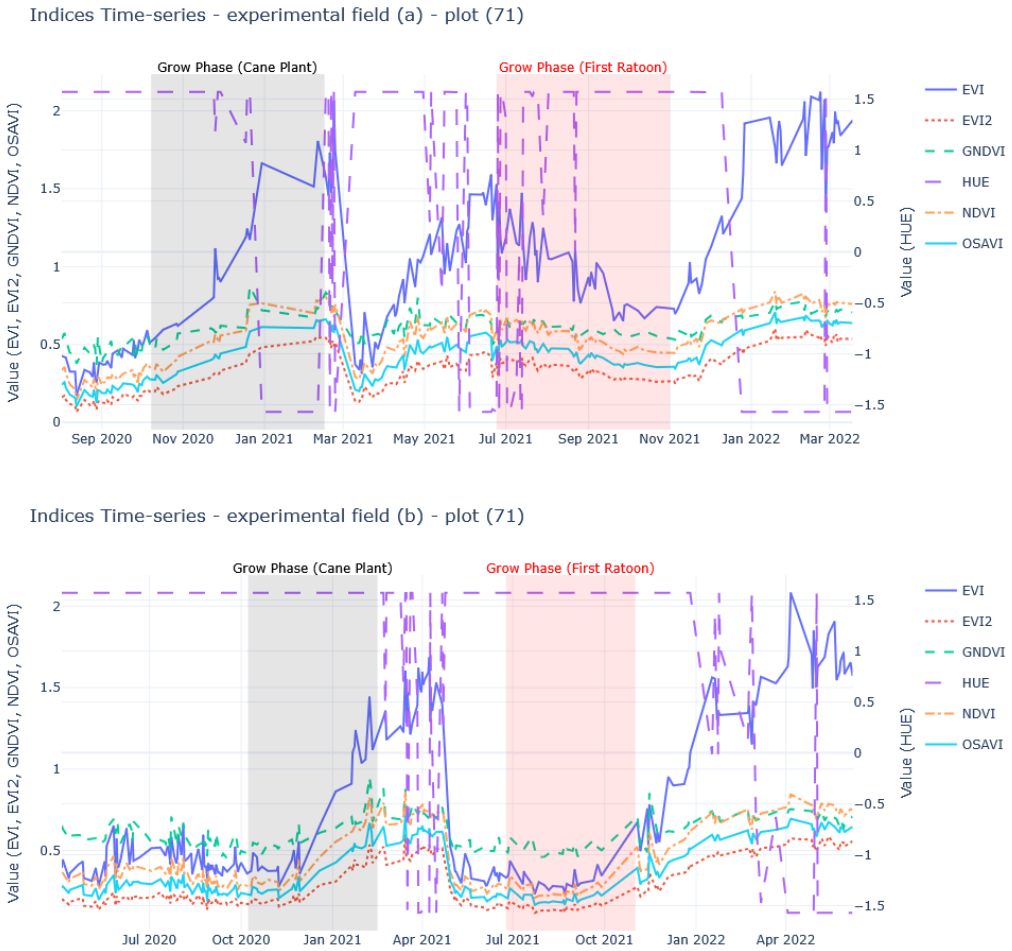


Figure 5. Time-series analysis of six key vegetation indices (EVI, EVI2, GNDVI, HUE, NDVI, OSAVI) across the two sugarcane production cycles and two locations: Experimental field (**a**)—Piracicaba, SP, Brazil; experimental field (**b**)—Tambaú, SP, Brazil. The shaded regions highlight periods corresponding to the crop's peak growth phase.

After the temporal window analysis, an initial variable filtering process was carried out to identify key predictors using the normalized relative feature importance scores derived from the Random Forest model's mean decrease in impurity metric (Figure 6a,b). Following this step, we conducted a multicollinearity assessment to ensure that the selected variables did not exhibit strong interdependencies, which could distort coefficient estimates and model interpretation.

During the modeling process, we evaluated all possible combinations of variables and calculated the respective Variance Inflation Factors (VIFs) for each set. We adopted a threshold of 5 for VIF, eliminating variables exceeding this limit to reduce multicollinearity while maintaining the model's predictive capacity. This approach minimizes coefficient estimation distortions and enhances result interpretability.

Table 2 presents the VIFs for all variables initially included in the model.

Table 2. Variance Inflation Factors (VIFs) for the initially included variables.

Variable	VIF
Block	1.340
Varieties	1.311
Cycle (cane plant and first ratoon)	6.325
Accumulated precipitation (growth phase)	7.626
Mean EVI (growth phase)	1931.188
Mean EVI 2 (growth phase)	51,741.630
Mean GNDVI (growth phase)	45.490
Mean HUE (growth phase)	17.241
Mean NDVI (growth phase)	68,859.164
Mean OSAVI (growth phase)	207,649.379

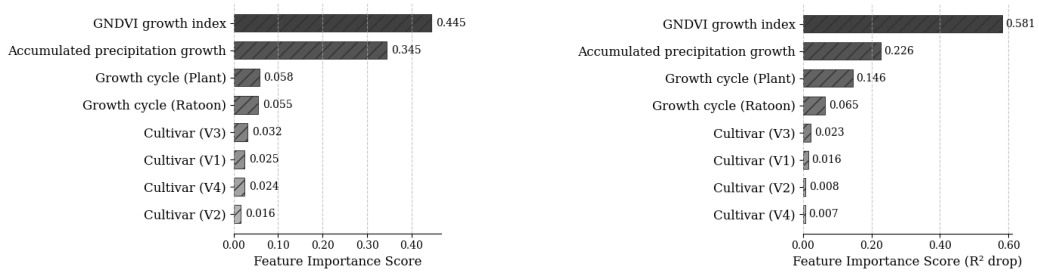
As observed, some variables exhibited high VIF values, particularly mean EVI 2, mean NDVI, and mean OSAVI during the growth phase, indicating strong correlations with other variables in the model.

Although Figure 5 illustrates the temporal behavior of the variables and may suggest that EVI and NDVI do not exhibit strong visual collinearity with other indices, the quantitative VIF analysis revealed otherwise. The VIF values for mean EVI and mean NDVI were extremely high (1931.188 and 68,859.164, respectively, as shown in Table 2), indicating a strong linear dependency with other variables in the model. This result reinforces that visual analysis alone is insufficient to detect collinearity, requiring statistical measures to ensure a robust and reliable model.

All possible combinations of the explanatory variables were evaluated to reduce multicollinearity and enhance the robustness of the model. The combinations that yielded VIF values below 5 were retained, while those with VIF values above 5 were discarded. The combinations of variables with VIF values below 5 are presented in Table 3.

Table 3. Variance Inflation Factors (VIFs) for the retained variables in the final model.

Variable	VIF
Varieties	1.014
Cycle (cane plant and first ratoon)	1.226
Accumulated precipitation (growth phase)	1.521
Mean GNDVI (growth phase)	1.344



(**a**) (**b**)

Figure 6. (**a**) Normalized relative feature importance scores from the Random Forest model via a mean decrease in impurity. Higher values indicate a greater contribution to the model. (**b**) Feature importance scores from the MLP model using permutation importance. The importance scores represent the average decrease in the R^2 metric when each feature is permuted, indicating the significance of each feature in the model's predictive performance.

Subsequently, the heteroskedastic GA regression, RF, and ANN models were trained exclusively on these selected variables to mitigate multicollinearity risks, ensuring robust and interpretable model performance.

3.1. Heteroskedastic GA Regression—Training Data from the Experimental Data

3.1.1. Statistical Model

The method `stepGAICAll.A` automates the selection of additive terms in statistical models, efficiently refining the fit. The process begins with all variables included in the model, and at each step, it removes the least relevant one, as long as its exclusion does not compromise the chosen selection criterion. This procedure continues until all remaining variables are relevant to the model, resulting in a more parsimonious structure, i.e., a model that achieves the desired level of explanation or prediction with the fewest possible predictor variables [45]. Based on this criterion, the selected variables for the parameter associated with the mean were varieties, cycle (cane plant and first ratoon), and mean GNDVI (during the growth phase). For the dispersion parameter, the chosen variables were accumulated precipitation (during the growth phase) and mean GNDVI. Figure 7a shows no obvious linear correlation between precipitation and the response variable. However, its inclusion in the proposed model is justifiable, as `stepGAICAll.A` is not limited to detecting linear relationships but selects variables with the potential to influence the response variable [45]. This reinforces the importance of statistical approaches capturing patterns beyond linear correlation, ensuring a more robust and interpretable model.

Building upon this, the final heteroskedastic GA regression model selected based on the GAIC has the systematic components given by

$$\mu_i = \exp(\beta_{10} + \beta_{11}c_{i1} + \beta_{12}c_{i2} + \beta_{13}c_{i3} + \beta_{14}x_{i3} + \beta_{15}x_{i7})$$

and

$$\sigma_i = \exp(\beta_{20} + \beta_{21}x_{i4} + \beta_{22}x_{i7}).$$

3.1.2. Descriptive Statistics of the Field Experiment

Based on the descriptive analysis of the training data for the response variable tons of sugarcane per hectare (y_i: TCH, t ha^{-1}), different behavior patterns are observed according to the analyzed variables (Table 4).

For the variable **varieties** (x_{i2}: V1, V2, V3, and V4), the average TCH ranged from 111.30 t ha^{-1} to 122.21 t ha^{-1}. Variety V1 exhibited the lowest mean (111.30 t ha^{-1}), while

variety V3 had the highest mean (122.21 t ha^{-1}). Regarding dispersion, the standard deviation was 26.10 for variety V1 and 23.16 for variety V3, suggesting that while variety V3 has the highest mean, it also shows lower variability in production than V1.

In the analysis of **cycles** (x_{i3}: cane plant and first ratoon), the cane plant cycle recorded an average of 106.93 t ha^{-1}, while the first ratoon cycle achieved a higher average of 127.98 t ha^{-1}. The standard deviation for the plant cycle was 27.60, compared to 22.37 for the first ratoon cycle. These results indicate that the first ratoon cycle has a higher average yield and lower variability, suggesting a more stable production than the cane plant cycle.

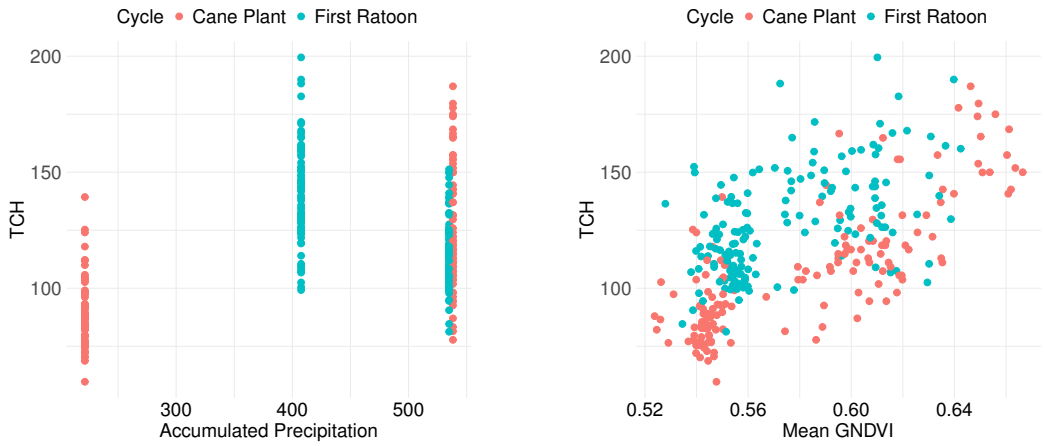
Figure 7 presents scatter plots highlighting the relationships between y_i (tons of sugarcane per hectare, TCH, t ha^{-1}) and two explanatory variables: x_{i4} (accumulated precipitation during the growth phase, shown in Figure 7a) and x_{i7} (mean GNDVI during the growth phase, shown in Figure 7b), for the data referring to the commercial area. As previously mentioned, environmental conditions influence sugarcane growth. In the commercial area, it is observed that the accumulated precipitation was more favorable during the cane plant cycle, with the same behavior being observed for the mean GNDVI.

Table 4. Estimates of the average yield in tons of sugarcane per hectare (y_i: TCH, t ha^{-1}) and standard deviation (SD) based on varieties (x_{i2}) and cycle (x_{i3}) for the field experiment.

	Category	Mean	Standard Deviation
Varieties (x_{i2})	V1	111.30	26.10
	V2	120.80	27.89
	V3	122.21	23.16
	V4	115.53	30.28
Cycle (x_{i3})	Cane Plant	106.93	27.60
	First ratoon	127.98	22.37

Values represent estimates of y_i and SD for each category.

Precipitation is a continuous variable; however, Figure 7a displays only four distinct values. This outcome results from the natural variability in the data and the method used to accumulate precipitation within specific time intervals during both experimental setups and cycles (including the cane plant and the first ratoon). Despite presenting only four values, the variable was considered continuous due to its nature and was not included in the model as categorical. A similar approach was considered by [5].



(**a**) Scatter Plot of y_i vs. x_{i4}. (**b**) Scatter Plot of y_i vs. x_{i7}.

Figure 7. (**a**,**b**) show the scatter plots of TCH against cumulative precipitation and the mean GNDVI during the growth phase, respectively, for the field experiment.

3.1.3. Heteroskedastic GA Regression Results

Table 5 presents the results of the heteroskedastic GA regression model adjusted for the response variable y_i, representing tons of sugarcane per hectare (TCH, t ha^{-1}). The mean μ_i is modeled using the logarithmic link function according to the specified structure in (2). Varieties V2 and V3, defined by the parameters c_{i1} and c_{i2}, show significant differences compared to variety V1, which is the reference level of the statistical model (represented by the factor in the intercept). As shown in Table 4, these two varieties (V2 and V3) have higher means than V1. It is also worth noting that the means of V5 and V1 are close, and the model indicated no significant differences between these two varieties. For the variable cycle, it is observed that the cane plant differs significantly from the first ratoon. As shown in Table 4, the first ratoon cycle has a higher mean, which is expected due to environmental variations.

Table 5. Estimates of the parameters, standard error (SE), and p-value of the heteroskedastic GA regression model adjusted for the training data. The notation (*) denotes the statistical significance of the variables, indicating p-value < 0.05.

Parameter	Effects	Parameter	Estimate	SE	p-Value
μ	Intercept	β_{10}	2.1760	0.1472	<0.0001 *
	c_{i1}	β_{11}	0.0670	0.0232	0.0039 *
	c_{i2}	β_{12}	0.0846	0.0242	0.0005 *
	c_{i3}	β_{13}	0.0060	0.0238	0.8000
	x_{i3}	β_{14}	0.2214	0.0169	<0.0001 *
	x_{i7}	β_{15}	4.1985	0.2546	<0.0001 *
σ	Intercept	β_{20}	−3.5966	0.7345	<0.0001 *
	x_{i4}	β_{21}	−0.0008	0.0003	0.0033 *
	x_{i7}	β_{22}	3.4889	1.3358	0.0095 *

(*) denotes statistical significance, with p-value < 0.05.

In Figure 7a,b, it is possible to observe that, during the first ratoon cycle, the yield was higher, indicating that environmental conditions favored the growth of sugarcane. The mean GNDVI also presented better results during this cycle, suggesting enhanced plant health and photosynthetic activity. Ref. [5] found that when accumulated precipitation is higher, yield is improved during the cycle, which was enhanced by the greater water availability. The mean GNDVI shows an increasing relationship with tons of cane per hectare. As Figure 7b shows, TCH values increase as the GNDVI rises. This behavior indicates that higher vegetation, represented by the GNDVI, is associated with higher cane yield per hectare, which is expected as denser and healthier vegetation tends to produce more biomass. The analysis suggests that GNDVI could be useful for monitoring and predicting crop yield.

The square root of the dispersion parameter σ_i, which is also modeled using the logarithmic link function as shown in (2), indicates variability in TCH production. Regarding variability, both the accumulated precipitation and the average GNDVI showed significant variations, which was expected due to environmental fluctuations. These variations can directly impact the development of sugarcane, as soil water availability, determined by precipitation, influences important aspects such as vegetative growth and plant health. The GNDVI, in turn, reflects the health and development of vegetation, being highly sensitive to environmental conditions. Fluctuations in these factors can lead to variations in GNDVI values: increases in healthy vegetation are reflected by higher index values, while decreases in vegetation result in lower values. It is important to note that after sugarcane reaches maturity, a decrease in the GNDVI is naturally observed, as vegetative growth slows down.

The plants enter a stage of lower photosynthetic activity, focusing more on sugar production than foliar growth.

These results underscore the importance of the selected variables in explaining both the mean and variability of sugarcane production. Climatic and agronomic variables significantly impact yield, with statistical significance determined by p-values < 0.05, as presented in the table.

Figure 8 presents the residual analysis of the heteroskedastic GA regression model. The dotted gray line represents the approximate point-wise 95% confidence intervals indicated by the two elliptic curves. If the model is correct, we would expect approximately 95% of the points to lie between these two curves, with around 5% outside. A higher percentage of points outside the curves, or a clear systematic deviation from the horizontal line, suggests that the model's fitted distribution (or the fitted terms) may be inadequate to explain the response variable. Notably, 3 points out of 448 lie inside the elliptical curves, corresponding to approximately 0.67%. This is well within the expected range, as the model's performance is considered acceptable when the points outside the curves account for no more than 5%. The worm plot [51] further confirms that the model fits the data well, suggesting its suitability for analyzing data with similar characteristics in the present study.

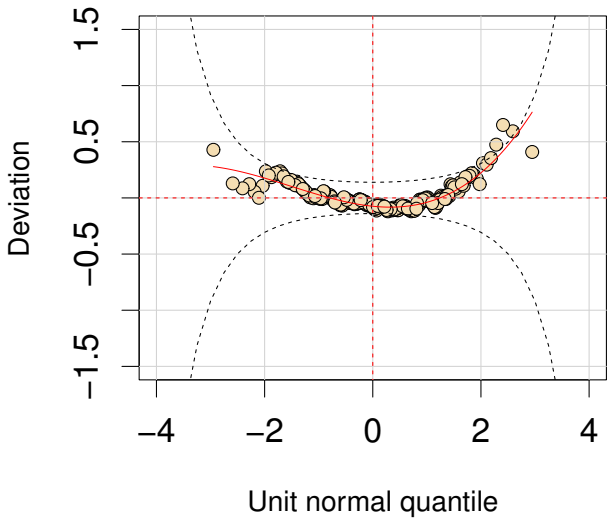


Figure 8. Residual analysis plot evaluating the model's fit to the data. Deviation (Y-axis): Quantile residuals, transformed to follow a standard normal distribution. Unit normal quantile (X-axis): Quantiles of the standard normal distribution.

3.1.4. Descriptive Statistics of the Commercial Area Data

The commercial area data evaluated the model's ability to generalize to new situations that mimic real-world production conditions. However, the training or testing processes did not include this commercial data. Instead, it was utilized solely to compare the models' results with observed data, allowing us to assess their applicability in a commercial production context. Based on the descriptive analysis of data from commercial sugarcane fields for the response variable tons of sugarcane per hectare (y_i: TCH, t ha^{-1}), different behavior patterns were observed across the analyzed variables (Table 6).

For the variable **varieties** (x_{i2}): V2 and V3, the average TCH ranged from 104.45 t ha^{-1} to 105.17 t ha^{-1}. Variety V2 had a higher standard deviation (21.34) than V3 (15.86), indicating greater variability in TCH for V2.

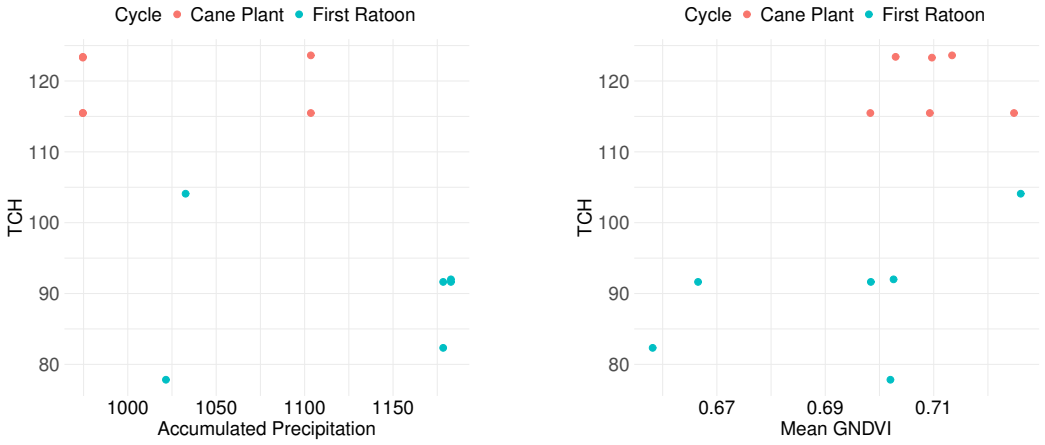
For the data from commercial sugarcane fields, the average TCH ranged from 89.92 t ha^{-1} (first ratoon) to 119.46 t ha^{-1} (cane plant), with standard deviations of 9.11 and 4.36, respectively.

Table 6. Estimates of the average yield in tons of sugarcane per hectare (y_i: TCH, t ha^{-1}) and standard deviation (SD) based on varieties (x_{i2}) and cycle (x_{i3}) for commercial area data.

	Category	Mean	Standard Deviation
Varieties (x_{i2})	V2	105.17	21.34
	V3	104.45	15.86
Cycle (x_{i3})	Cane Plant	119.46	4.36
	First ratoon	89.92	9.11

Values represent estimates of y_i and SD for each category.

Figure 9 presents scatter plots highlighting the relationships between y_i (tons of sugarcane per hectare (TCH, t ha^{-1})) and two explanatory variables (x_{i4} (accumulated precipitation during the growth phase) and x_{i7} (mean GNDVI during the growth phase)). These variables were selected for graphical representation because they are continuous and allow direct visualization of their relationships with the response variable. Figure 9a shows the scatter plot of y_i versus x_{i4}. The data suggest that precipitation levels influence sugarcane production in non-uniform ways, with no discernible linear relationship observed. Conversely, Figure 9b displays the relationship between TCH and x_{i7} (mean GNDVI during the growth phase), where the data distribution suggests a potential linear relationship between vegetative growth and sugarcane yield. This plot provides an initial view of how plant vegetative vigor, measured by mean GNDVI during the growth phase, might be associated with biomass production.



(**a**) Scatter Plot of y_i vs. x_{i4}.

(**b**) Scatter Plot of y_i vs. x_{i7}.

Figure 9. (**a**,**b**) show the scatter plots of TCH against cumulative precipitation and the mean GNDVI for the commercial area data, respectively.

3.2. Machine Learning Models Results

The dataset used in this study for generalization was insufficient to effectively implement an ANN model, as demonstrated by the comparative performance outcomes. The results obtained from the ANN were notably inferior to those achieved using the heteroskedastic GA regression model and the RF approach. This discrepancy arises from the inherent nature of ANNs, which typically require large volumes of data to extract meaningful features and build robust predictive models accurately [52]. As observed in other studies [53–55], given the limited dataset available in this study, the ANN struggled to generalize patterns effectively, leading to suboptimal performance.

ANNs require large datasets to generalize effectively, and their performance deteriorates with limited data due to overfitting and poor generalization [56]. Studies have shown that high-capacity models like ANNs exhibit high variance when trained on small

datasets, making them unsuitable for scenarios with constrained data availability [57,58]. In contrast, alternative approaches such as heteroskedastic gamma regression and RF have proven more effective in such cases. Gamma regression is particularly useful for modeling skewed, continuous data, offering robustness and interpretability [41,59]. At the same time, RF excels in handling small datasets by reducing overfitting and automatically selecting important features [12,60].

Empirical research highlights RF's superior generalization capabilities, especially in ecological and remote sensing applications with limited data [61,62]. Unlike ANNs, RF does not require extensive training data and is less prone to memorizing patterns than learning general relationships. Similarly, gamma regression, as a parametric model, provides stability in predictive tasks where data are scarce. These findings reinforce the need for dataset-aware model selection, ensuring that the complexity of a chosen model aligns with the available data to achieve optimal performance [63].

Consequently, this study highlights the critical importance of dataset size when selecting modeling techniques, particularly for data-intensive methods like ANNs. It underscores the advantages of alternative approaches such as heteroskedastic GA regression and RF in scenarios with constrained data availability.

3.3. Model Performance Analysis for Field Experiment Data

The analysis of the train and test results demonstrates that all three models—heteroskedastic GA regression model, ANN, and RF—are suitable for explaining the variability in the dataset. Table 7 presents the performance metrics for each model, highlighting their ability to provide accurate predictions.

Table 7. Performance metrics for training and testing data from the heteroskedastic GA regression model, RF, and ANN from the experimental field data. Metrics include Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Model	R^2 (Train)	MAE (Train)	RMSE (Train)	R^2 (Test)	MAE (Test)	RMSE (Test)
Heteroskedastic GA Regression Model	0.61	13.7743	15.6185	0.62	14.0355	14.7355
Random Forest	0.74	10.5100	13.9100	0.69	12.0100	16.0500
Neural Network	0.71	11.3200	14.7800	0.67	12.4300	16.3900

All models demonstrate levels of predictive performance with R^2 values between 0.61 and 0.74, showing close results across training and testing datasets. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values further indicate that these models provide accurate predictions, considering the yield mean values obtained through the experiment (about 120 TCH). These results highlight the suitability of the heteroskedastic GA regression model, RF, and ANN for modeling the dataset, offering viable approaches for capturing the dataset's underlying relationships and variability.

3.4. Performance Evaluation of the Model Using Data from Commercial Sugarcane Fields

This subsection presents the performance of the models—heteroskedastic GA regression model, ANN, and RF—when applied to commercial area data for sugarcane yield (see Table S2 in Supplementary Materials). These models were evaluated on data not used in the field experiment, measuring their generalization capability. The performance metrics for each model are shown in Table 8, highlighting the ability of the models to generalize and predict yield in commercial settings.

Table 8. Performance metrics for data of commercial fields from the heteroskedastic GA regression model, RF, and ANN. Metrics include Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Model	R^2	MAE	RMSE
Heteroskedastic GA Regression Model	0.89	3.9037	5.0870
Random Forest	0.67	43.4500	44.9900
Neural Network	0.04	110.7900	112.4900

The results in Table 8 show that the models exhibit varying performance when applied to commercial area data. The heteroskedastic GA regression model achieved the highest R^2 of 0.89, indicating a strong ability to explain the variability in sugarcane yield. This model also had the lowest MAE and RMSE, suggesting it provides accurate and precise predictions for the commercial area data.

In contrast, the Random Forest model showed an R^2 of 0.67, indicating a moderate fit, with notably higher MAE and RMSE values, which suggest less predictive accuracy than the heteroskedastic GA regression model. The ANN model performed the weakest among the three, with an R^2 of only 0.04, indicating poor performance. Its higher MAE and RMSE values further highlight its limited ability to generalize well to the commercial data from a few samples.

The results indicate that, although all models showed varying performance levels based on the data source, the heteroskedastic GA regression model significantly outperformed the others when applied to commercial area data. In the field experiments, the models performed similarly, with the Random Forest displaying slightly better results. However, for the commercial area data, the heteroskedastic GA regression model achieved the highest R^2 and the lowest MAE and RMSE values. Meanwhile, the Random Forest exhibited moderate performance, while the ANN model struggled, revealing its limited generalization ability for this specific application. This underscores the importance of selecting the appropriate model based on the characteristics of the data, as the heteroskedastic GA regression model appears to be more suitable for estimating sugarcane yield in commercial areas.

4. Conclusions

This study examined how to estimate sugarcane yield at harvest by analyzing temporal satellite imagery during critical growth phases. The heteroskedastic GA regression model demonstrated superior performance, achieving an R^2 of 0.89 with low prediction errors (MAE = 3.9037, RMSE = 5.0870) in commercial sugarcane fields. It significantly outperformed both RF and ANN models. The key covariates influencing the model's predictive capability included sugarcane variety (x_{i2}), growth cycle profile (x_{i3}), cumulative precipitation during the growth phase (x_{i4}), and the mean GNDVI spectral index during the growth phase (x_{i7}). Together, these variables captured the agronomic and environmental dynamics of crop development.

The decision to focus exclusively on cumulative precipitation as the climatic variable was based on its direct relationship with soil water availability, which plays a crucial role in sugarcane growth and yield. While temperature can impact the crop, its effects are partially reflected in the GNDVI, which indicates how vegetation responds to thermal and water stress. Thus, it is appropriate to include cumulative precipitation, as the GNDVI indirectly accounts for temperature effects. However, the role of temperature may still be significant.

One limitation of the approach used in this study is the requirement for a substantial amount of quality data. Although the exponential increase in data improves the performance of machine learning methods, the challenges in obtaining these data can be a

significant drawback. Currently, new experiments involving sugarcane are being conducted to generate additional data that can be applied to the methodology developed in this study in the future. Additionally, new approaches could investigate whether incorporating more environmental variables—such as temperature, soil moisture, and overall water availability—would improve model performance, especially in extreme climatic conditions where temperature stress on crops is more significant.

The model's robustness highlights the value of integrating satellite-derived spectral indices (e.g., mean GNDVI during the growth phase) with field-specific agronomic variables, allowing for precise monitoring of yield trends. These findings support heteroskedastic GA regression as a scalable tool for agricultural analytics, especially in data-scarce regions, while emphasizing the importance of strategic variable selection to improve predictive accuracy in crop yield modeling. This model-based approach is also a powerful tool for mitigating abiotic and biotic stresses by enhancing data-driven decision-making and optimizing sustainable crop management strategies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agronomy15040793/s1, Table S1: Subset of the dataset (experimental area) used to train the models; Table S2: Subset of the dataset (experimental area) used to test the models; Table S3: Generalization Dataset of Commercial Sugarcane Plots for Assessing Model Generalization Performance.

Author Contributions: The experiments were designed and carried out in the field by G.M.d.A.C.; J.C.S.V., C.S.A., E.A.S., J.F.G.A. and G.M.d.A.C. analyzed the data. The article was written and reviewed by J.C.S.V., C.S.A., E.A.S., J.F.G.A., L.A.F.B. and G.M.d.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded through collaborative grants from the Brazilian Agricultural Research Corporation (Embrapa) and the São Paulo State Sugarcane Growers' Cooperative (Coplacana) [Grant SEG 30.19.90.005.00.00], as well as Embrapa, the Symbiosis Company (Simbiose) and the Brazilian Funding Authority for Studies and Projects (Finep) [Grant SEG 20.24.00.110.00.00]. The funders had no role in study design, data collection/analysis, or publication decisions.

Data Availability Statement: To verify the possibility of using the data available in the Supplementary Materials or the images collected in the field, please contact the corresponding author.

Acknowledgments: The authors thank COPLACANA (Cooperativa de Plantadores de Cana do Estado de São Paulo) and Schiavon Group for their invaluable support throughout the field experiments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. FAO. Sugar Cane Production. 2024. Available online: https://ourworldindata.org/grapher/sugar-cane-production (accessed on 10 December 2024).
2. Molin, J.P.; Wei, M.C.F.; da Silva, E.R.O. Challenges of Digital Solutions in Sugarcane Crop Production: A Review. *AgriEngineering* **2024**, *6*, 925–946. [CrossRef]
3. Amankulova, K.; Farmonov, N.; Akramova, P.; Tursunov, I.; Mucsi, L. Comparison of PlanetScope, Sentinel-2, and Landsat 8 Data in Soybean Yield Estimation Within-Field Variability with Random Forest Regression. *Heliyon* **2023**, *9*, e17432. [CrossRef] [PubMed]
4. Prataviera, F.; Hashimoto, E.M.; Ortega, E.M.M.; Savian, T.V.; Cordeiro, G.M. Interval-Censored Regression with Non-Proportional Hazards with Applications. *Stats* **2023**, *6*, 643–656. [CrossRef]
5. dos Santos, D.P.; Soares, A.; de Medeiros, G.; Christofoletti, D.; Arantes, C.S.; Vasconcelos, J.C.S.; Speranza, E.A.; Barbosa, L.A.F.; Antunes, J.F.G.; Cançado, G.M.A. Evaluation of Sugarcane Yield Response to a Phosphate-Solubilizing Microbial Inoculant: Using an Aerial Imagery-Based Model. *Sugar Technol.* **2024**, *26*, 143–159. [CrossRef]
6. Vasconcelos, J.C.S.; Ortega, E.M.M.; Cordeiro, G.M.; Vasconcelos, J.S.; Biaggioni, M.A.M. Estimation and Diagnostic for a Partially Linear Regression Based on an Extension of the Rice Distribution. *REVSTAT Stat. J.* **2024**, *22*, 433–454. [CrossRef]

7. Luciano, A.C.S.; Picoli, M.C.A.; Duft, D.G.; Rocha, J.V.; Leal, M.R.L.V.; le Maire, G. Empirical Model for Forecasting Sugarcane Yield on a Local Scale in Brazil Using Landsat Imagery and Random Forest Algorithm. *Comput. Electron. Agric.* **2021**, *184*, 106063. [CrossRef]
8. Oliveira, R.P.; Barbosa, M.R., Jr.; Pinto, A.A.; Oliveira, J.L.P.; Zerbato, C.; Furlani, C.E.A. Predicting Sugarcane Biometric Parameters by UAV Multispectral Images and Machine Learning. *Agronomy* **2022**, *12*, 1992. [CrossRef]
9. Sunil G.L.; Nagaveni V.; Shruthi U. A Review on Prediction of Crop Yield using Machine Learning Techniques. In Proceedings of the 2022 IEEE Region 10 Symposium (TENSYMP), Mumbai, India, 2022; pp. 1–5. [CrossRef]
10. Taud, H.; Mas, J. Multilayer Perceptron (MLP). In *Geomatic Approaches for Modeling Land Change Scenarios*; Olmedo, M.T.C., Paegelow, M., Mas, J.F., Escobar, F., Eds.; Springer: Cham, Switzerland, 2018; pp. 451–455, ISBN 978-3-319-60800-6.
11. Kushwah, J.S.; Kumar, A.; Patel, S.; Soni, R.; Gawande, A.; Gupta, S. Comparative study of regressor and classifier with decision tree using modern tools. *Mater. Today Proc.* **2022**, *56*, 3571–3576. [CrossRef]
12. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
13. Jeferson Lobato Fernandes, N.; Esquerdo, J. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. *Int. J. Remote Sens.* **2017**, *38*, 4631–4644. [CrossRef]
14. Tripathi, A.; Tiwari, R.; Tiwari, S. A deep learning multi-layer perceptron and remote sensing approach for soil health based crop yield estimation. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102959. [CrossRef]
15. Everingham, Y.; Sexton, J.; Skocaj, D.; Inman-Bamber, G. Accurate prediction of sugarcane yield using a random forest algorithm. *Agron. Sustain. Dev.* **2016**, *36*, 27. [CrossRef]
16. Dimov, D.; Uhl, J.; Löw, F.; Seboka, G. Sugarcane yield estimation through remote sensing time series and phenology metrics. *Smart Agric. Technol.* **2022**, *2* 100046. [CrossRef]
17. Silva, N.; Chaves, M.; Luciano, A.; Sanches, I.; Almeida, C.; Adami, M. Sugarcane yield estimation using satellite remote sensing data in empirical or mechanistic modeling: A systematic review. *Remote Sens.* **2024**, *16*, 863. [CrossRef]
18. Song, W.; Mu, X.; Ruan, G.; Gao, Z.; Li, L.; Yan, G. Estimating Fractional Vegetation Cover and the Vegetation Index of Bare Soil and Highly Dense Vegetation with a Physically Based Method. *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *58*, 168–176. [CrossRef]
19. Wiegand, CL; Richardson, A.J.; Escobar, D.E.; Gerbermann, A.H. Vegetation Indices in Crop Assessments. *Remote Sens. Environ.* **2001**, *35*, 105–119. [CrossRef]
20. Jackson, R.D.; Huete, A.R. Interpreting Vegetation Indices. *Prev. Vet. Med.* **1991**, *11*, 185–200. [CrossRef]
21. Frazier, A.E.; Hemingway, B.L. A Technical Review of Planet Smallsat Data: Practical Considerations for Processing and Using PlanetScope Imagery. *Remote Sens.* **2021**, *13*, 3930. [CrossRef]
22. Evett, I.; Jackson, G.; Lambert, J.; McCrossan, S. The impact of the principles of evidence interpretation on the structure and content of statements. *Sci. Justice* **2000**, *40*, 233–239. [CrossRef]
23. Jiang, Z.; Huete, A.R.; Didan, K.; Miura, T. Development of a Two-Band Enhanced Vegetation Index without a Blue Band. *Remote Sens. Environ.* **2008**, *112*, 3833–3845. [CrossRef]
24. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a Green Channel in Remote Sensing of Global Vegetation From EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
25. Escadafal, R. Soil Spectral Properties and Their Relationships with Environmental Parameters-Examples from Arid Regions. In *Imaging Spectrometry—A Tool for Environmental Observations*; Hill, J., Mégier, J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1994; pp. 71–87. ISBN 0-7923-2965-1.
26. Rouse, J.W., Jr.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring Vegetation Systems in the Great Plains with ERTS. In *Abstracts of the 3rd ERTS Symposium, Washington, DC, USA, 1973*; NASA/Goddard Space Flight Center: Greenbelt, MD, USA, 1973; pp. 309–317.
27. Rondeaux, G.; Steven, M.; Baret, F. Optimization of Soil-Adjusted Vegetation Indices. *Remote Sens. Environ.* **1996**, *55*, 95–107. [CrossRef]
28. Bégué, A.; Lebourgeois, V.; Bappel, E.; Todoroff, P.; Pellegrino, A.; Baillarin, F.; Siegmund, B. Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI. *Int. J. Remote Sens.* **2010**, *31*, 5391–5407
29. Mulianga, B.; Bégué, A.; Simoes, M.; Todoroff, P. Forecasting regional sugarcane yield based on time integral and spatial aggregation of MODIS NDVI. *Remote Sens.* **2013**, *5*, 2184–2199. [CrossRef]
30. Siqueira, D.; Vaz, C.; Silva, F.; Ferreira, E.; Speranza, E.; Franchini, J.; Galbieri, R.; Belot, J.; Souza, M.; Perina, F.; et al. Estimating Cotton Yield in the Brazilian Cerrado Using Linear Regression Models from MODIS Vegetation Index Time Series. *AgriEngineering* **2024**, *6*, 947–961. [CrossRef]
31. Zúñiga Espinoza, C.; Khot, L.; Sankaran, S.; Jacoby, P. High resolution multispectral and thermal remote sensing-based water stress assessment in subsurface irrigated grapevines. *Remote Sens.* **2017**, *9*, 961. [CrossRef]

32. Haseeb, M.; Tahir, Z.; Mahmood, S.; Tariq, A. Winter wheat yield prediction using linear and nonlinear machine learning algorithms based on climatological and remote sensing data. *Inf. Process. Agric.* **2025**. [CrossRef]
33. Fu, H.; Lu, J.; Li, J.; Zou, W.; Tang, X.; Ning, X.; Sun, Y. Winter Wheat Yield Prediction Using Satellite Remote Sensing Data and Deep Learning Models. *Agronomy* **2025**, *15*, 205. [CrossRef]
34. Lim, H.; Kim, S. Applications of Machine Learning Technologies for Feedstock Yield Estimation of Ethanol Production. *Energies* **2024**, *17*, 7. [CrossRef]
35. Hamzeh, S.; Naseri, A.; Alavipanah, S.; Mojaradi, B.; Bartholomeus, H.; Clevers, J.; Behzad, M. Estimating salinity stress in sugarcane fields with spaceborne hyperspectral vegetation indices. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *21*, 282–290.
36. Chang, C.; Lin, J.; Chang, J.; Huang, Y.; Lai, M.; Chang, Y. Hybrid Deep Neural Networks with Multi-Tasking for Rice Yield Prediction Using Remote Sensing Data. *Agriculture* **2024**, *14*, 513. [CrossRef]
37. Sarkar, S.; Osorio Leyton, J.; Noa-Yarasca, E.; Adhikari, K.; Hajda, C.; Smith, D. Integrating Remote Sensing and Soil Features for Enhanced Machine Learning-Based Corn Yield Prediction in the Southern US. *Sensors* **2025**, *25*, 543. [CrossRef] [PubMed]
38. Monteiro, L.A.; Sentelhas, P.C.; Pedra, G.U. Assessment of NASA/POWER satellite-based weather system for Brazilian conditions and its impact on sugarcane yield simulation. *Int. J. Climatol.* **2018**, *38*, 1571–1581. [CrossRef]
39. NASA. Power Dav. 2022. Available online: https://power.larc.nasa.gov/data-access-viewer/ (accessed on 30 October 2024).
40. Belsley, D.A.; Kuh E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons: New York, NY, USA, 1980; ISBN 978-0-4710-5856-4.
41. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Publisher: Chapman & Hall: London, UK, 1989; ISBN 978-0-412-31760-6. [CrossRef]
42. Johnson, N.L.; Kotz, S.; Balakrishnan, N. *Continuous Univariate Distributions*, 2nd ed.; Wiley: New York, NY, USA, 1994; Volume 1; ISBN: 978-0-471-58495-7.
43. Stasinopoulos, D.M.; Rigby, R.A. Generalized Additive Models for Location Scale and Shape (Gamlss) in R. *J. Stat. Softw.* **2008**, *23*, 18637. [CrossRef]
44. Nelder, J.A.; Wedderburn, R.W. Generalized Linear Models. *J. R. Stat. Soc. Ser. Stat. Soc.* **1972**, *135*, 370–384. [CrossRef]
45. Stasinopoulos, M.D.; Rigby, R.A.; Heller, G.Z.; Voudouris, V.; De Bastiani, F. *Flexible Regression and Smoothing: Using Gamlss in R*; CRC Press: New York, NY, USA, 2017; ISBN 978-1-138-19790-9.
46. Ripley, B.; Venables, B.; Bates, D.M.; Hornik, K.; Gebhardt, A.; Firth, D.; Ripley, M.B. Package 'Mass'. *Cran R* **2013**, *538*, 113–120.
47. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
48. Breiman, L.; Friedman J.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman and Hall/CRC: New York, NY, USA, 1984; ISBN 978-1-3151-3947-0. [CrossRef]
49. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufman: Amsterdam, The Netherlands, 1993; ISBN 1-55860-238-0.
50. Bishop, C.M. *Neural Networks for Pattern Recognition*; Clarendon Press: Oxford, UK, 1995; ISBN 978-0-1985-3864-6.
51. van Buuren, S.; Fredriks, M. Worm plot: Simple diagnostic device for modelling growth reference curves. *Stat. Med.* **2001**, *20*, 1259–1277. [CrossRef]
52. Zhang, Y.; Yang, Q. A Survey on Multi-Task Learning. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 5586–5609. [CrossRef]
53. Noa-Yarasca, E.; Osorio Leyton, J.M.; Angerer, J.P. Deep Learning Model Effectiveness in Forecasting Limited-Size Aboveground Vegetation Biomass Time Series: Kenyan Grasslands Case Study. *Agronomy* **2024**, *14*, 349. [CrossRef]
54. Jayaprakash, A. A Comparison of Deep Learning Methods for Time Series Forecasting with Limited Data. Master's Thesis, Freie Universität Berlin, Berlin, Germany, 2023.
55. Vabalas, A.; Gowen, E.; Poliakoff, E.; Casson, A.J. Machine learning algorithm validation with a limited sample size. *PLoS ONE* **2019**, *14*, e0224365. [CrossRef]
56. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; Publisher: MIT Press, Cambridge, USA, 2016; pp. 110–115.
57. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
58. Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.F.; Jun, H.; Kianinejad, H.; Wang, H. Deep learning scaling is predictable, empirically. *arXiv* **2017**, arXiv:1712.00409. [CrossRef]
59. Manning, W.G.; Mullahy, J. Estimating log models: To transform or not to transform? *J. Health Econ.* **2001**, *20*, 461–494. [CrossRef]
60. Cutler, D.R., Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef]
61. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]

62. Gislason, P.O.; Benediktsson, J.A.; Sveinsson, J.R. Random forests for land cover classification. *Pattern Recognit. Lett.* **2006**, *27*, 294–300. [CrossRef]
63. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014; pp. 7–8.