agronomy

MDPI

Article

Sugarcane (*Saccharum officinarum*) Productivity Estimation Using Multispectral Sensors in RPAs, Biometric Variables, and Vegetation Indices

Marta Laura de Souza Alexandre *, Izabelle de Lima e Lima, Matheus Sterzo Nilsson, Rodnei Rizzo, Carlos Augusto Alves Cardoso Silva * and Peterson Ricardo Fiorio

Department of Biosystems Engineering, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Piracicaba 13418-900, SP, Brazil; eng.izabellelima@usp.br (I.d.L.e.L.); matheu.s.nilsson@usp.br (M.S.N.); fiorio@usp.br (P.R.F.)

* Correspondence: martalaura@usp.br (M.L.d.S.A.); carlosesalq@usp.br (C.A.A.C.S.)

Abstract

The sugarcane crop is of great economic relevance to Brazil, and the precise productivity estimation is a major challenge in production. Therefore, the aim of this study was to estimate the productivity of sugarcane cultivars in different regions, using multispectral sensors embedded in RPAs and biometric variables sampled in the field. The study was conducted in two experimental areas, located in the municipalities of Itirapina-SP and Iracemápolis-SP, with 16 cultivars in a randomized block design. The images were acquired using the multispectral sensor MicaSense Altum, allowing the extraction of spectral bands and vegetation indices. In parallel, biometric variables were collected at 149 and 295 days after planting (DAP). The machine learning models Random Forest (RF) and Extreme Gradient Boosting (XGBoost) were calibrated using different sets of variables, and, despite the similar performance, it was decided to use the model derived from XGBoost in the analyses, since it deals more effectively with overfitting. The results indicated a good performance of the model (R^2 = 0.83 and 0.66; RMSE = 18.7 t ha^{-1} and 25.3 t ha^{-1}; MAE = 15.7 and 20.2; RPIQ = 3.22 and 2.61) for the validations K-fold and Leave-one-out cross-validation (LOOCV). The correlations between biometric variables, spectral bands, and vegetation indices varied according to crop development stage. The leaf insertion angle presented a strong correlation with near-infrared (NIR) (r = 0.76) and the indices ExG and VARI (r = 0.70 and r = 0.69, respectively). The present work demonstrated that the integration between multispectral and biometric data represents a promising approach for estimating sugarcane productivity.

Keywords: productivity; multispectral sensor; machine learning

check for updates

Academic Editors: Andrea Peruzzi and Kingsley John

Received: 25 July 2025
Revised: 30 August 2025
Accepted: 4 September 2025
Published: 8 September 2025

Citation: Alexandre, M.L.d.S.; e Lima, I.d.L.; Nilsson, M.S.; Rizzo, R.; Silva, C.A.A.C.; Fiorio, P.R. Sugarcane (*Saccharum officinarum*) Productivity Estimation Using Multispectral Sensors in RPAs, Biometric Variables, and Vegetation Indices. *Agronomy* **2025**, *15*, 2149. https://doi.org/10.3390/agronomy15092149

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The world production of sugarcane (*Saccharum officinarum* L.) increased significantly between 1994 and 2018, driven by the rising demand for sugar and ethanol [1,2]. In Brazil, this crop has a strategic importance, being fundamental for food and energy production [3]. In the harvest 2023/2024, Brazil achieved the highest production in its history, with 713.2 million tons, and the prediction for 2025 is 678.67 million tons, which represents an increase of 4.3% in the area intended for harvesting [4].

Given the great relevance of this crop for the Brazilian economy, yield estimation plays a fundamental role in decision-making and agricultural management. Traditionally,

the determination of sugarcane yield depends on biometric parameters and productivity monitoring by the direct weighing method, which, although reliable, is costly and time-consuming, especially in large areas [5]. In this context, the main limitation of using biometric parameters is related to the large number of samples that must be collected [6,7].

A widely employed technique for the rapid and large-scale acquisition of biophysical data of plants is remote sensing (RS). Among the available alternatives, remotely piloted aircraft systems (RPAs) equipped with multispectral sensors have emerged as a solution capable of enhancing the efficiency and reducing the costs in sugarcane productivity prediction. High spatial resolution images allow the extraction of variables related to crop vigor and density, establishing robust correlations with productivity, representing a potential tool to estimate productivity in sugarcane fields [8].

Studies performed by Sumesh et al. [9] detected that multispectral RPA images used to calibrate statistical models demonstrated a good predictive potential ($R^2 > 0.7$) of these data. Tanut et al. [10] combined RPA images, data mining, and the technique Reverse Design Method to obtain productivity estimates with 89% accuracy. Som-ard et al. [11] evidenced that the integration of multispectral data with information on the soil can improve pre-harvest productivity estimates, presenting significant potential for farmers and industries.

Nonetheless, despite the advances, there are still challenges in estimating productivity, given the complexity of the agricultural systems and the spatial and temporal variability present in cultivation. In this context, the adoption of machine learning techniques has been efficient in predictive crop modeling [12–14], especially models such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost).

Both models have been widely used for predictive modeling in sugarcane [15–17], given their robustness in processing complex data, with a high number of variables and non-linear relationships [18,19]. Despite being excellent tools for predictive modeling, the accuracy of these models usually increases with the incorporation of agronomic, climate, spectral, and biometric variables [20].

Studies that integrate multispectral data obtained by RPA with sugarcane biometric variables demonstrate potential for agricultural modeling [5,21]. In this sense, it is believed that the integration of the spectral variables obtained by RPA and the biometric variables in models based on machine learning may improve sugarcane productivity prediction in different production environments, when compared with the isolated use of these data. This integrated approach can expand the applicability and robustness of the models to different production contexts. Therefore, the aim of this study is to evaluate the effectiveness of integrating multispectral sensors embedded in RPAs and biometric variables in the prediction of sugarcane productivity through the application of advanced machine learning techniques.

2. Materials and Methods

The methodology adopted in this study follows six main steps: (i) definition of the study area; (ii) acquisition of the multispectral images obtained by RPAs; (iii) field collection of the biometric data of the sugarcane cultivars; (iv) data processing; (v) productivity prediction by the models RF and XGBoost; and (vi) validation of the models using the techniques K-fold and LOOCV (Figure 1). This approach allows the integration of spectral and biometric data for more precise productivity estimates. The steps will be detailed below.

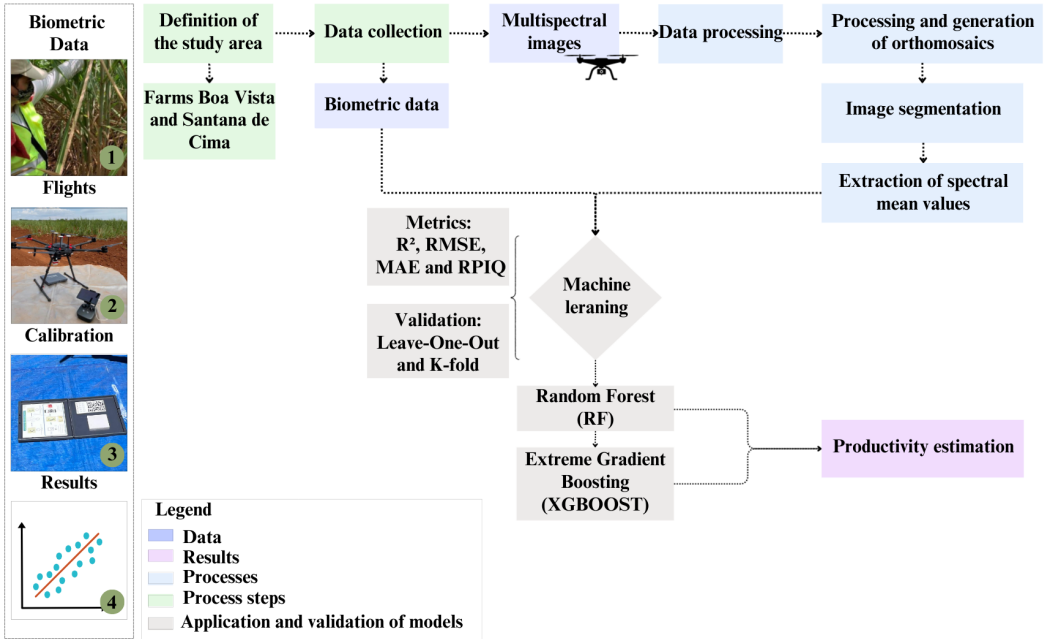

Figure 1. Methodological flowchart of the study.

2.1. Study Area

This study was performed in two experimental areas in the municipalities of Itirapina-SP and Iracemápolis-SP (Figure 2), with distinct edaphoclimatic conditions. In Itirapina (22°14′30″ S, 47°41′31″ W), the climate Cfa, according to Köppen, presents an average temperature of 27.1 °C and annual rainfall of 1356 mm, and the soil is classified as Red-Yellow Alic Latosol, with sandy/medium texture and flat topography, as described by Prado [22], seen as a low-suitability environment. On the other hand, in Iracemápolis (22°35′15.18″ S, 47°33′15.59″ W), the climate Aw, according to Köppen, has dry winters, an average temperature of 28.7 °C, and rainfall of 1352 mm, with Eutrophic Red Latosol, clayey texture, and flat topography [22], considered a high-suitability environment according to studies explored by [23].

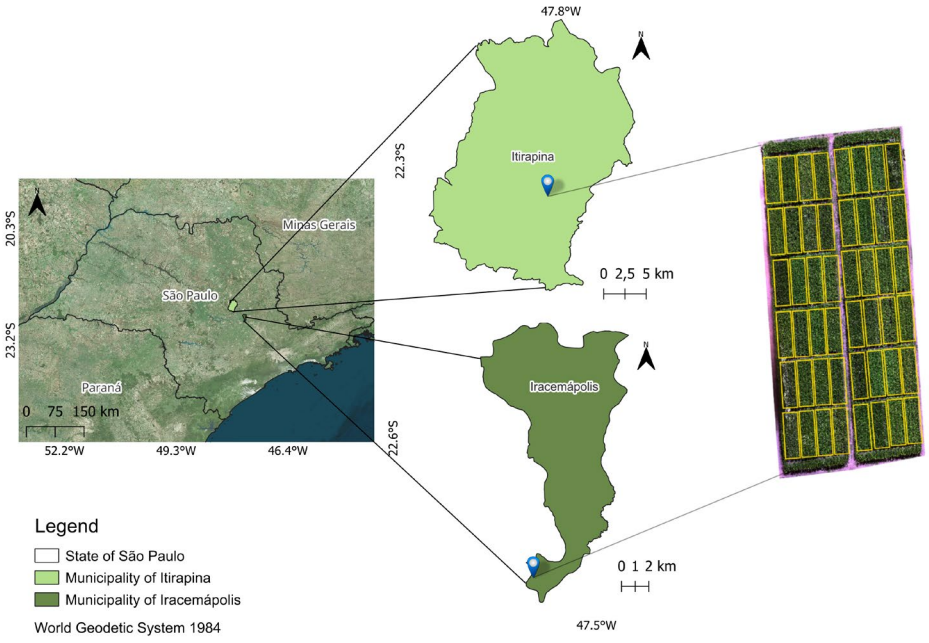

Figure 2. Location map of the study areas, situated in the municipalities of Itirapina-SP and Iracemápolis-SP.

2.2. Description of the Experiment

The experiment was conducted in a randomized block design (RBD), using 16 sugarcane cultivars (Table 1), with three repetitions per cultivar in each experimental area. The experimental plots were composed of four planting lines with 18.2 m, with 1.5 m of spacing between them. Each line contained 28 plants, located at a distance of 0.65 m from each other. The selection of these cultivars was based on their wide use in distinct areas, besides the interest in evaluating the growth and development of new sugarcane cultivars in different regions.

Table 1. Cultivars and suitability areas.

Cultivars	Source
CT022994	[24]
CT961007	[24]
CTC9006	[24]
IACCTC07 2361	[25]
IACCTC07 7207	[25]
IACCTC07 8008	[25]
IACCTC08 9052	[25]
IACSP02 1064	[25]
IACSP04 6007	[25]
RB005014	[26]
RB005040	[26]
RB075864	[26]
RB127825	[26]
RB855156	[26]
RB975033	[26]
RB985476	[26]

Figure 3 presents a graphical representation that illustrates additional information about the 2021/2022 sugarcane harvest in the experimental areas, highlighting rainfall and temperature data throughout the cultivation cycle. It is observed that water distribution in the first months after planting was influenced by the drought associated with the La Niña phenomenon, which may have negatively impacted the development of the cultivars, especially during the vegetative period, a crucial phase for their growth and productivity.

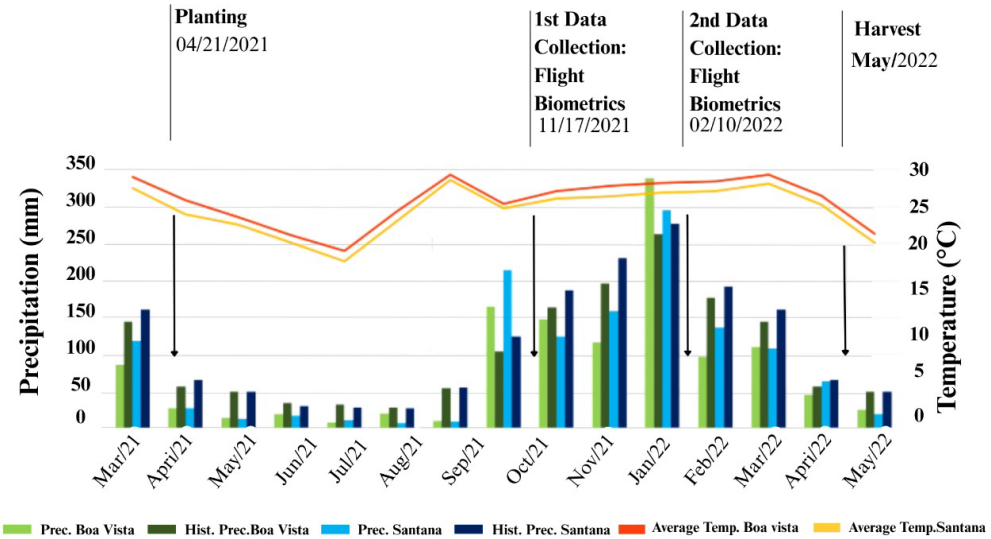

Figure 3. Graph of the data on rainfall (mm) and temperature (°C) at the two locations where the assays were conducted, with information regarding historical averages.

Throughout the crop cycle, it was observed that, in the area of Boa Vista, rainfall totaled 1229 mm, with a significant peak recorded in January 2022. In contrast, in Santana de Cima, rainfall was superior, reaching 1320 mm, with a peak in the incidence of rainfall in the first months, corresponding to the vegetative period of the crop. This precipitation pattern suggests a close relationship between the initial phase of the cycle and water availability in both regions, influencing plant development.

2.3. Biometric Data

The biometric data acquisition (Figure 4) was performed in two collections, conducted on 17 November 2021 and 10 February 2022, when the cultivars were at 149 and 295 days after planting (DAP), respectively. To standardize the assessments, 10 plants in the second furrow of each plot were selected, excluding the first five plants. All selected plants were identified using a ribbon. This procedure was employed in both field campaigns.

The criteria for the assessment of the biometric data followed the methodology of Nassif et al. [27], also including procedures described by Artschwager [28]. In each of the two samplings, 10 plants were selected for the measurement of the dimensions of leaf +3, using a graduated ruler for length (cm) and a wooden protractor. Furthermore, the number of active green leaves per stem/tiller (NF/C_149DAP, NF/C_295DAP), the number of stems per plant, the height of the plants from the primary stem (ALT_149DAP, ALT_295DAP), width (LARG_149DAP, LARG_295DAP), length (COMP_149DAP, COMP_295DAP) and the leaf insertion angle (ANG_149DAP, ANG_295DAP) of the leaves were determined, as well as the number of active leaves on the main stem. It is worth highlighting that the measurement of stem diameter (D_295DAP) was only performed in the second collection.

Figure 4. Identification of the plants (**A**) and acquisition of the biometric data: length of leaf + 3 (**B**); stem diameter (**C**); length of the primary stem (**D**); width of leaf + 3 (**E**) and leaf insertion angle in leaf + 3 (**F**) [29].

2.4. Harvest

To estimate the final yield of each experimental plot, sugarcane was mechanically harvested. Each harvested row segment was unloaded into a trailer equipped with a load cell, allowing the direct weighing of the production per plot. The values obtained were converted into tons per hectare (t ha^{-1}) and used as the response variable, referred to as "productivity" in the predictive models.

2.5. Image Acquisition

The images were acquired using the multispectral sensor MicaSense Altum, (MicaSense, Inc., Seattle, WA, USA) coupled to DJI Matrice 600 Pro (SZ DJI Technology Co., Ltd., Shenzhen, China), a six-propeller multirotor. This sensor integrates a high-resolution panchromatic sensor (3.2 MP), a thermal sensor, and five spectral bands: Blue (475 nm ± 32 nm), Green (560 nm ± 27 nm), Red (668 nm ± 14 nm), Red-Edge (717 nm ± 12 nm), and NIR (842 nm ± 57 nm).

The flight plans were prepared in the specific software of the equipment, defining parameters such as a flight height of 30 m, and side and frontal overlap of 80% and 70%, respectively. To minimize the canopy shading effects, the flights were performed between 10 h and 14 h. Before each flight, an image of the MicaSense calibration plate was captured for the radiometric correction of the images. The collections occurred in both field campaigns on the dates of 17 November 2021 and 10 February 2022.

The acquired images were pre-processed using the software Agisoft Metashape Professional version 1.5.5, in which the corrections and acquisition of the orthomosaic comprised the following steps: (i) Image correction; (ii) Photo alignment from homologous points; (iii) Acquisition of the Digital Surface Model; and (iv) Creation of the orthomosaic. The mosaics generated were then used for the calculation of the vegetation indices (Table 2), which would be later used in the calibration of the predictive productivity models.

Table 2. Vegetation indices used for the calibration of the predictive models.

Index	Formula	Reference
NDVI	$\frac{(Nir-Red)}{(Nir+Red)}$	[30]
NDRE	$\frac{(Nir-Red\ Edge)}{(Nir+Red\ Edge)}$	[31]
RVI	$\frac{Nir}{Red}$	[32]
ExG	$2 \times Green - Red - Blue$	[33]
VARI	$\frac{(Green-Red)}{(Green-Red+Blue)}$	[34]

2.6. Data Pre-Processing

From the orthomosaics generated, the images were segmented to remove the interference of the exposed soil and the shadows. For this procedure, 3644 points were sampled, classified as vegetation and non-vegetation based on the NDVI values of the images. The dataset was split into 80% for training and 20% for validation, using the algorithm Random Forest as a classifier. As a result, a raster containing exclusively the vegetation of each plot was obtained.

Subsequently, the spectral data and the vegetation indices were extracted by the creation of a buffer with diameter of 0.45 m, covering the same 10 plants selected in the biometric data sampling. Using the Mask and Zonal Statistics tools, the mean spectral values of each plant were obtained for the spectral bands (Blue, Green, Red, Red-edge and NIR), the vegetation indices (NDVI, NDRE, RVI, ExG, VARI) and the closing index made by image segmentation. The procedures were conducted using the software QGIS, version 3.22.14, from the package Ski-lern [35] for each flight.

2.7. Productivity Prediction by Machine Learning Models

In this study, statistical modeling was conducted to calibrate machine learning algorithms using sugarcane yield as the dependent variable, and spectral and biometric parameters as independent variables. While ten measurements were collected for each independent variable, yield was assessed at the plot level by direct weighing of the harvested cane using a load cell. To ensure consistency between variables, the dataset was

harmonized to the plot scale by calculating the arithmetic mean of the spectral and biometric measurements. Consequently, an effective sample size (n = 96) was defined based on the number of plots across both field campaigns.

The algorithms Random Forest and Extreme Gradient Boosting were employed for the calibration of the models and validation of the proposed methodology. An important aspect in this study was the evaluation of the performance of the statistical models when different datasets were employed, namely: (a) Biometric data at 149 DAP (Biom. 149), 295 DAP (Biom. 295) and both dates (Biom. All); (b) Spectral data at 149 DAP (Bands 149DAP), 295 DAP (Bands 295DAP) and both dates (Bands All); (c) Vegetation indices at 149 DAP (VIs 149DAP), 295 DAP (VIs 295DAP) and both dates (VIs All); and (d) All data obtained at 149 DAP (All 149DAP), 295 DAP (All 295DAP) and both dates (All) (Table 3). For each of these datasets, the algorithms RF and XGBoost were applied, resulting in a total of 12 predictive models.

Table 3. Datasets used in the modeling, with the number of samples (n), variables included, and the number of features.

Dataset	Date	n (Samples)	Features	Feature Description	Notes
Biom. 149	149 DAP	96	5	ALT_149, LARG_149, COMP_149, ANG_149, NF/C_149	-
Biom. 295	295 DAP	96	6	ALT_295, LARG_295, COMP_295, ANG_295, NF/C_295, D_295	D_295 only in the second collection
Biom. All	149 and 295	96	11	ALT_295, LARG_295, COMP_295, ANG_295, NF/C_295, D_295	-
FE	149 DAP	96	1	FE_149	Obtained by segmentation
FE	295 DAP	96	1	FE_295	Obtained by segmentation
Bands 149	149 DAP	96	5	B1, B2, B3, B4, B5	-
Bands 295	295 DAP	96	5	B1, B2, B3, B4, B5	-
Bands All	149 and 295	96	10	B1, B2, B3, B4, B5	-
VIs 149	149 DAP	96	5	NDVI, NDRE, RVI, ExG, VARI	-
VIs 295	295 DAP	96	5	NDVI, NDRE, RVI, ExG, VARI	-
Vis All	149 and 295	96	10	NDVI, NDRE, RVI, ExG, VARI	-
All 149	149 DAP	96	16	Biom. 149 + Bands 149 + VIs 149 + FE 149	-
All 295	295 DAP	96	17	Biom. 295 + Bands 295 + VIs 295 + FE 295	-
All	149 and 295	96	33	All 149 and All 295	-

ALT = height; COMP = length; LARG = width; NF/C = number of green leaves per stem; D = diameter; ANG = leaf insertion angle;; FE = closing index. B1 = Blue; B2 = Green; B3 = Red; B4 = Red Edge; B5 = NIR.

RF is a supervised learning algorithm, composed of a combination of simple models that, together, offer greater robustness. Created by Breiman [36] RF is often employed in regression and classification problems. The model uses the bootstrap method, in which each tree is trained with a random sample of the original dataset. Each tree generates an individual vote for the most frequent class, and the final prediction results from the combination of these votes [37], resulting in improved predictive performance [38]. From the pre-established parameters, the K-Fold cross-validation was used to select the best number of factors that minimizes the statistics.

XGBoost, developed by Chen and Guestrin [39], is a machine learning model based on decision trees, widely recognized for its robustness in predictive modeling tasks [40]. It improves the accuracy of the model by reducing the problem of overfitting during the training process [41]. Nevertheless, its efficacy strongly depends on the appropriate selection of hyperparameters, since the model can be susceptible to variations [42].

The whole process was performed using the program R version 4.5.0 [43], employing the packages ranger [44], XGBoost [45], and Caret [46]. Hyperparameter optimization was performed using the grid search method, and in RF, the number of variables in each node (*mtry*), the splitting rule (*splitrule*), and the minimal node size to split at (min.node.size) were optimized. In the case of XGBoost, the maximum depth of trees (max_depth), the number of rounds for boosting (nrounds), the step size shrinkage (eta), and the columns' subsample ratio for each tree (colsample_bytree) were optimized.

2.8. Metrics for the Validation and Generalization of the Models

The assessment of the quality of the models in predicting the productivity of the sugarcane cultivars was performed by the comparison between the values that were observed and those estimated by the models. For this analysis, validation metrics were used, such as the coefficient of determination (R^2), the mean square error (RMSE), the mean absolute error, and the ratio between performance and interquartile distance (RPIQ), whose mathematical expressions are described in Equations (1)–(4).

$$R^2 = \frac{[\sum(\gamma p - \underline{\gamma} p) \cdot (\gamma o - \underline{\gamma} o)]^2}{[\sum(\gamma p - \underline{\gamma} p)^2 \cdot (\gamma o - \underline{\gamma} o)^2]} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - X| \quad (3)$$

$$RPIQ = \frac{IQR_{obs}}{RMSE} \quad (4)$$

In this study, two strategies were adopted for the validation of complementary models to ensure both robustness to cultivar-specific variations and the overall assessment of their predictive capacity. The first consisted in a leave-one-out cross-validation (LOOCV) by blocks: the 16 cultivars were replicated three times in each area, totalizing 48 observations, and, at each iteration, a block of three repetitions of the same cultivar was reserved as a test set, while the model was calibrated with the 45 remaining observations; this procedure was repeated until each block had been used once for the validation, allowing the measurement of the stability of the model against each genetic variable.

The second strategy employed the k-fold cross-validation, with k = 10: the complete set of 96 observations was divided into ten balanced subsets (folds), so that in each round, nine folds were used for training and the remaining fold, for validation, repeating the process until all folds had served as tests, which allowed a representative evaluation of the model's performance in different partitions of the data. Combined, these approaches offered a comprehensive diagnostic of the predictive capacity and the generalization of the model, considering both the intrinsic variability of each cultivar and the heterogeneity of the sample subset as a whole.

3. Results

3.1. Biometric Parameters

As observed in Figure 3, rainfall in the initial months was relatively uniform among the areas evaluated. Although Santana de Cima registered a slightly superior volume, this difference was not expressive compared to the annual and historical averages. Nevertheless, sugarcane production under different edaphoclimatic conditions presents relevant challenges, in this case, especially because of the nutritional variations and the soil character-

istics, as observed by Silva et al. [47]. These variations influenced productivity significantly among the areas, and this difference was statistically proven (Figure 5).

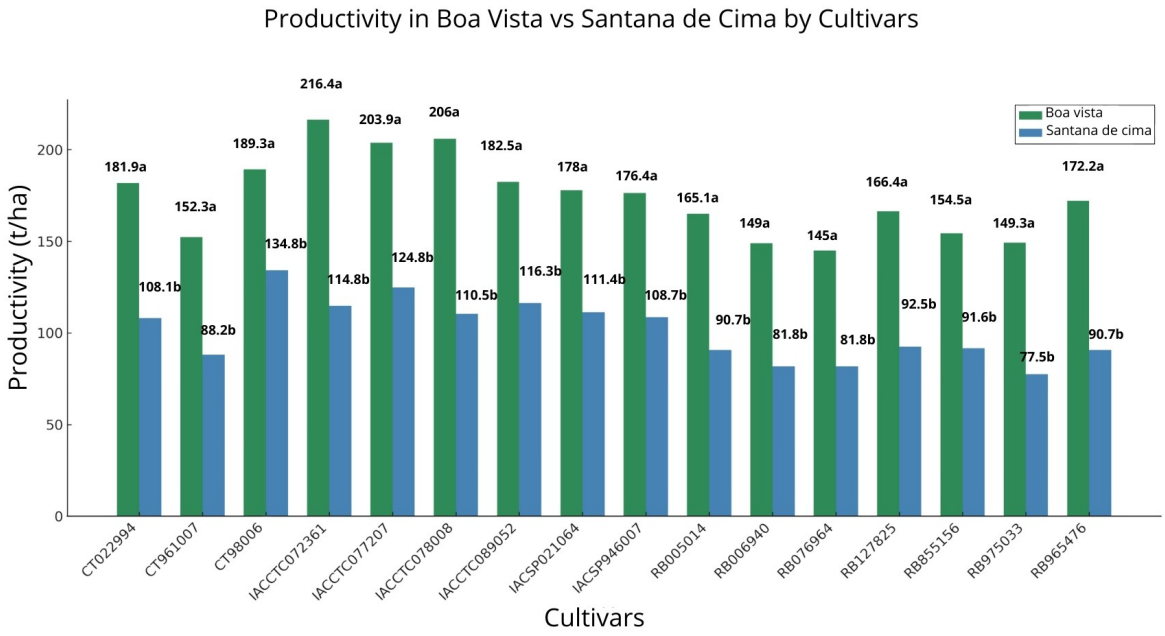

Figure 5. Productivity of the different cultivars estimated in 2022 (TCH; t ha^{-1}), comparing the farms Boa Vista and Santana de Cima. Mean productivity (±SD) of the cultivars in Boa Vista (green) and Santana de Cima (blue). Different letters (a ≠ b) indicate a statistically significant difference by the Tukey's HSD test ($p < 0.05$).

The statistical analyses of the biometric parameters and productivity (Figure 5 and Table 4) indicated a superior performance of the cultivars in the farm Boa Vista, which offers more favorable environmental conditions, including soils with higher base saturation and clay content. At 149 days after planting (DAP), the cultivars in Boa Vista exhibited a leaf area per plant (AF/PL) of 23,188 cm^2, an average height (ALT) of 54.44 cm, a leaf insertion angle (ANG) of 16°, a length (COMP) of 123.49 cm, and a width (LARG) of 3.26 cm. The canopy closure index (FE) was 67.88, and the number of green leaves per stalk (NF/C) reached 8. In Santana, most of the values were lower than those in Boa Vista, except for ANG (26°) and AF/PL (23,440 cm^2).

Table 4. Mean biometric parameters of 16 sugarcane cultivars evaluated at 149 and 295 days after planting (DAP) in two experimental areas. The values are presented as mean ± standard deviation.

Site	Biometric Parameters							
	AF/PL (cm^2)	ALT (cm)	ANG (°)	COMP (cm)	FE	LARG (cm)	NF/C	D (cm)
	1st evaluation (149 DAP)							
Santana de Cima	23,440 ± 5115	53.7 ± 10	26.0 ± 5.3	110.4 ± 7.3	59.1 ± 10.7	3.0 ± 0.4	7.6 ± 0.6	-
Boa vista	23,188 ± 4979	54.4 ± 11.4	16.9 ± 3.2	123.4 ± 9.8	67.8 ± 8.6	3.2 ± 0.4	8.1 ± 0.9	-
	2nd evaluation (295 DAP)							
Santana de Cima	28,446 ± 6413	192.7 ± 20.3	32.1 ± 5.3	158.2 ± 9.7	68.1 ± 7.8	4.3 ± 0.6	6.1 ± 0.4	2.5 ± 0.2
Boa vista	42,031 ± 8756	227.6 ± 20.9	29.4 ± 6.2	160.7 ± 12.0	78.2 ± 6.2	4.9 ± 0.6	7.6 ± 0.8	2.9 ± 0.2

ALT = height; COMP = length; LARG = width; NF/C = number of green leaves per stem; D = diameter; ANG = leaf insertion angle; AF/PL = leaf area; FE = closing index.

At 295 DAP, a significant increase in the biometric parameters was observed in both areas, except for NF/C, which declined to 7.61 in Boa Vista and 6.14 in Santana. In Boa Vista, AF/PL reached 42,031 cm^2, ALT increased to 227.65 cm, ANG was 29°, COMP reached 160.76 cm,

LARG was 4.97 cm, and FE was 78.23. In Santana, AF/PL increased to 28,446 cm^2, ALT to 192.77 cm, LARG to 4.39 cm, COMP to 158.24 cm, ANG to 32°, and FE to 68.15.

This, however, does not disqualify the results of Santana de Cima, but indicates that the cultivars responded according to the specific conditions of that location. Of the 16 cultivars evaluated (Table 1), four were prominent for their superior mean productivity: IACCTC07 2361 (216.4 t ha^{-1} in Boa Vista vs. 115.0 t ha^{-1} in Santana de Cima), IACCTC078008 (206.0 t ha^{-1}/111.4 t ha^{-1}), IACCTC07 7207 (203.9 t ha^{-1}/124.8 t ha^{-1}), and CTC9006 (189.3 t ha^{-1}/134.3 t ha^{-1}). Tukey's HSD test ($p < 0.05$) confirmed that, for these cultivars, the means of Boa Vista differ statistically from those of Santana de Cima, evidencing their consistent performance even outside their traditionally recommended environments.

3.2. Correlation Analysis of the Spectral and Biometric Data

The correlation between the spectral and biometric data (Figure 6) indicated that, at 149 DAP, the leaf insertion angle presented a positive correlation with the spectral band NIR, with a coefficient of 0.76. Nevertheless, this association was virtually null at 295 DAP, when the correlation between these variables dropped to −0.04, indicating a null or non-existent relationship at this more advanced stage of crop development.

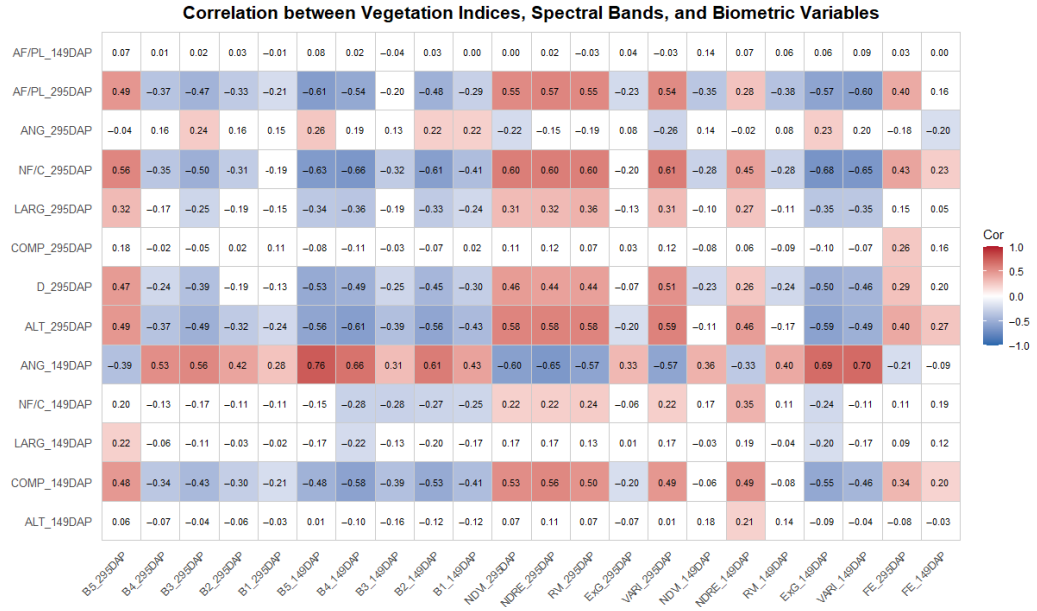

Figure 6. Pearson correlation matrix between spectral bands, biometric variables, and vegetation indices. B1 = Blue; B2 = Green; B3 = Red; B4 = Red Edge; B5 = NIR; ALT = height; COMP = length; LARG = width; NF/C = number of green leaves per stem; D = diameter; ANG = leaf insertion angle; AF/PL = leaf area; FE = closing index.

The correlation matrix revealed that, at 149 days after planting (DAP), leaf insertion angle presented moderate positive correlations with the bands Red (0.43), Green (0.61), and Blue (0.31). However, at 295 DAP, this relationship significantly weakened, with correlations close to zero for the same bands (Red: 0.24; Green: 0.16; Blue: 0.15), indicating the leaf insertion angle loses its spectral association throughout the crop cycle. Furthermore, the closing index at 295 DAP presents a positive correlation with the spectral band NIR (0.75); nonetheless, this correlation is practically null at 149 DAP.

The indices NDVI, NDRE, RVI, and VARI presented a positive correlation (>0.6) with the biometric variable "number of leaves per stem" at 295 DAP. On the other hand, the indices ExG and VARI demonstrated correlations of 0.69 and 0.7, respectively, with the leaf insertion angle at 149 DAP. Other positive correlations (r > 0.5) were observed between the biometric variable "height" (295 DAP) and the indices NDVI, NDRE, RVI, and VARI.

3.3. Calibration of the Models

The calibration of the models considered different sets of variables, aiming at understanding the potential of each of them for an estimate of crop productivity, as well as evaluating whether the combination of all variables would bring some gain to the prediction. Thus, the performance of the models derived from RF and XGBoost was evaluated regarding their RMSE values (Figure 7).

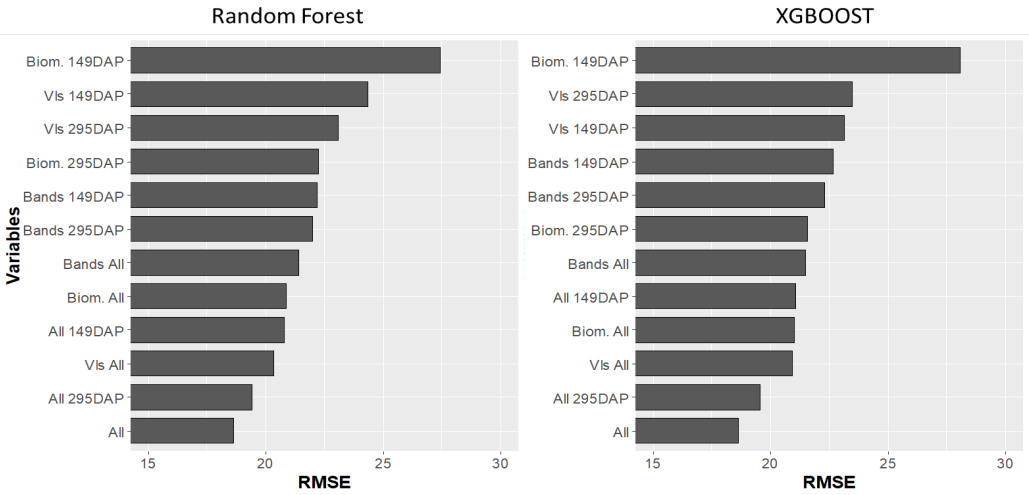

Figure 7. RMSE of the predictive models calibrated with different sets of variables.

The models with the highest RMSE values were those that considered only the biometric variables at 149 DAP, with values around 27 t ha^{-1}. Conversely, when only the biometric variables at 295 DAP were employed, or data from both dates, the errors presented a reasonable reduction and RMSE between 20 and 22.5 t ha^{-1}. This same pattern was observed for most of the datasets, spectral bands, or vegetation indices (VIs), in which the information obtained at 295 DAP results in more robust models.

In this work, the calibrations performed using only the biometric data (Biom. All), surface reflectance (Bands All), or even vegetation indices (VIs All), demonstrated a similar performance, with RMSE close to 21 t ha^{-1}. The best results were obtained when all variables were employed in the calibration. In this case, the machine learning techniques employed had the same RMSE value, 18 ton ha^{-1}. Despite the similar performance, it was decided to use the model derived from XGBoost in the analyses. Models such as Random Forest are more susceptible to overfitting, whereas XGBoost deals more effectively with this problem by techniques such as regularization and boosting [48]. Therefore, it was decided to use XGBoost in the final step of productivity prediction.

3.4. Validation of the Model and Productivity Estimate

The estimate of sugarcane productivity employing the model XGBoost and the combination of all variables demonstrated good predictive capacity. The k-fold validation had R^2 = 0.83, RMSE = 18.7, MAE = 15.7 and RPIQ = 3.22. On the other hand, in the Leave-One-Out cross-validation, where a cultivar was always excluded from the calibration, the values were R^2 = 0.66, RMSE = 25.3, MAE = 20.2, and RPIQ = 2.61 (Figure 8). In both cases, these results indicate that the model, calibrated with all variables, presented good robustness in productivity prediction.

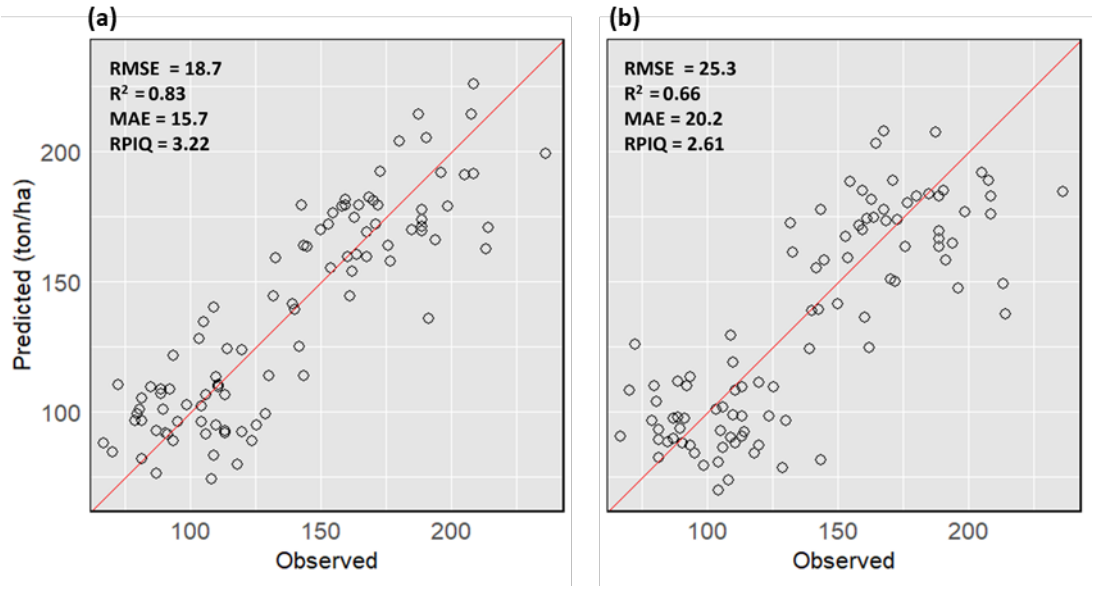

Figure 8. K-fold (**a**) and Leave-One-Out cross-validation (**b**) indicating the predictive potential of the model XGBoost calibrated with all variables.

The feature importance analysis (Figure 9) indicated that the NIR spectral band (B5), measured at 149 DAP, was the most influential predictor in the model, followed by the leaf insertion angle at the same time point. These results are consistent with the correlations previously observed between variables (Figure 6). Beyond these two primary features, a combination of biometric and spectral variables from different dates also contributed significantly to the predictive modeling.

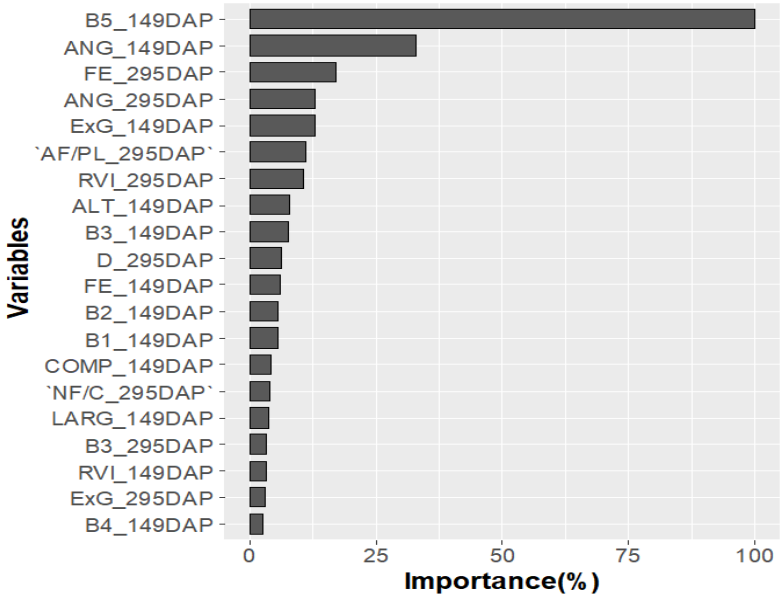

Figure 9. Feature importance for the selected model.

4. Discussion

4.1. Biometric Parameters and Productivity

In general, between 149 and 295 DAP, there was a significant increase in leaf area, plant height, leaf length, and leaf width, reflecting canopy development and leaf expansion in the crop. Although both areas exhibited similar growth patterns, the Boa Vista farm stood out for presenting higher values of leaf area per plant (42,031 cm^2) and average height (227.65 cm).

According to Irvine (1975) [49], there is a close association between crop productivity and the total photosynthetically active surface, represented by the Leaf Area Index (LAI), which is directly related to the amount of absorbed light and, consequently, to total photosynthesis. Thus, the most expressive biometric parameters constitute indirect productivity indicators, as reported by Farias et al. (2008) [50], who observed maximum LAI values around 152 DAP.

The strong correlation between plant height, leaf dimensions, and leaf area described by Kumar et al. (2023) (r = 0.85) [51] reinforces the relevance of the higher AF/PL and ALT values in Boa Vista, which resulted in greater productivity (Figure 5). These parameters are directly associated with increased growth capacity and plant vigor, enhancing biomass production and ultimately leading to higher yield values.

Nonetheless, in both areas, the number of green leaves per stalk (NF/C) decreased slightly between the first and second samplings, which may be explained by the incidence of *Sphenophorus levis*, causing significant damage to the experimental plots in both sites. Nevertheless, the biometric superiority in Boa Vista was sufficient to sustain higher productivity values.

4.2. Influence of the Environment of Production and the Cultivars

Regardless of cultivar type, sugarcane production was greater in the area of Boa Vista compared to Santana. For all cultivars evaluated, the productivity values in Boa Vista exceeded those observed for the same cultivars in Santana. This behavior is directly related to the environment of production, especially the edaphic characteristics. In Santana, the soil is classified as Red-Yellow Acrisol, with sandy/medium texture and flat topography, characteristics associated with a low-suitability environment [22]. In contrast, in Boa Vista, the soil presents a clayey texture, being considered a high-suitability environment for sugarcane cultivation. Therefore, the environment of production was a limiting factor for productivity in the area of Santana [22].

During the sugarcane production cycle, the differences between cultivars were evaluated based on the cultivars that best adapted to the specific conditions of each area. As observed by Bhatt et al. [52], cultivar yield is directly related to the nutritional, chemical, physical, and moisture conditions of the soil. In this context, the understanding of the edaphoclimatic conditions of each region is fundamental to exploring the whole production potential of the varieties, as highlighted by Sanches et al. [53]. In areas with high clay content (Boa Vista), the physicochemical properties of the soil promote greater water retention, given the predominance of micropores, thus yielding high water availability. Nevertheless, this characteristic may result in waterlogging of the roots, compromising root respiration and increasing the risk of anaerobic processes [54]. On the other hand, in the soils of sandy texture (Santana de Cima), the predominance of macropores results in rapid water drainage, reducing water retention under stresses that are beneficial to plants. Furthermore, low cation-exchange capacity limits nutrient retention, predisposing crops to stress [55,56].

In Boa Vista, illustrated in Figure 5, cultivars IACCTC07 2361, IACCTC07 8008, and IACCTC07 7207 presented promising production performance, reflecting their adaptability to the local conditions. Nonetheless, when cultivated in less favorable regions, such as in Santana de Cima, their performance is significantly reduced. This decrease is associated with the greater water and nutrient requirements of these varieties, possibly derived from the larger number of leaves, which demand a greater supply of water and nutrients, contradicting the initial exploratory recommendations [25].

Cultivar CTC9006 presented rustic characteristics and high productivity in the middle of the harvest, with productivity in a restrictive environment [24]. It presented satisfactory results in both environmental conditions. This demonstrates that more rustic phenotypes

tend to be more resistant to unfavorable conditions, given the proximity in t ha^{-1}, with 189 t ha^{-1} for Boa Vista and 134 t ha^{-1} for Santana de Cima. As observed by Fortes and Demattê [57], productivity increases with the closing of the canopy and stems.

The use of the same cultivars in two contrasting environments increases the variability and representativeness of our sample set, making our assessment more robust. The expressive differences in the productivity of the same cultivar between the areas reflect the distinct production potentials imposed by the edaphoclimatic conditions, and guarantee greater data variability, which was fundamental to calibrate predictive models.

4.3. Spectral-Biometric Correlations

The positive correlation between the leaf insertion angle and the reflectance in the band NIR at 149 days after planting (DAP) of sugarcane, in the erectophilic canopy, indicates greater signal return in the near-infrared region with the increase in leaf angle. According to Baldocchi et al. [58], canopies with more vertical leaves use this radiation more efficiently, increasing the photosynthetic rate without altering leaf structure. This effect is enhanced under diffuse light, which passes through the canopy at different angles [59].

The leaf insertion angle of sugarcane at 149 DAP provided greater surface exposure to radiation, which increases reflectance, especially in the green band, which is less absorbed [60]. On the other hand, the bands red and blue present smaller correlations because of the intense absorption of these wavelengths by chlorophyll during photosynthesis [61–63].

Nonetheless, at 295 DAP, this association between the leaf insertion angle and the spectral bands virtually disappeared. This phenomenon is attributed to the structural modifications of the leaves in the advanced stage of the vegetative cycle, as evidenced by Baldocchi et al. [58], who report the weakening of this correlation at the end of the crop cycle. Conversely, vegetation indices derived from remote sensing maintain a high capacity for estimating biometric variables in sugarcane. As emphasized by Zhang et al. [64], variations in the leaf insertion angle influence spectral response, especially in indices such as NDVI, which depend on the ratio between the absorption in the red band and the reflectance in the near-infrared (NIR) region.

A positive correlation was observed between the indices NDVI, NDRE, RVI, and VARI and the number of leaves per stem at 295 DAP, indicating these metrics are sensitive to the available leaf biomass volume. The use of spectral indices based on near-infrared (NIR) reflectance constitutes an established method for biomass estimates, as indicated by Qi et al. [65] and Rouse et al. [30]. In addition, the indices ExG and VARI presented correlations of 0.70 and 0.69, respectively, with the leaf insertion angle at 149 DAP, suggesting sensitivity to the plant canopy architecture. Analogous results were reported by Sumesh et al. [9], who found a strong association between the index ExG and stem density ($r = 0.851$), with a coefficient of determination of $R^2 = 0.74$. These findings consolidate the relevance of the metrics derived from unmanned aerial vehicles (RPAs) in the prediction of crop yield, providing detailed information on plant structure and vigor.

To ensure the performance of the model that correlates VIs derived from spectral bands and biometric data, it is important to apply orthometric corrections to the images and to conduct a meticulous calibration of the data, eliminating any anomalies. Sensor calibration, variations in lighting conditions, and inaccuracies in field collections have a direct influence on the quality of the information [66]. These error sources may compromise the accuracy of the models in distinct crop development stages.

4.4. Productivity Prediction

In this study, we opted to use the algorithms Random Forest (RF) and Extreme Gradient Boosting (XGBoost) instead of traditional linear models such as MLR or PLSR. Although

linear approaches have been widely and successfully applied in agricultural prediction tasks [21,67,68], they present limitations in capturing the complexity and non-linear interactions between spectral and biometric variables that strongly influence sugarcane productivity. In contrast, RF and XGBoost demonstrate greater robustness in processing high-dimensional datasets, multicollinearity, and non-linear patterns, while incorporating strategies such as bootstrapping, regularization, and boosting, which substantially reduce the risk of overfitting [19,38].

The model XGBoost, calibrated with the complete set of variables, presented a high predictive performance in the K-fold validation with R^2 = 0.83, whereas in the Leave-One-Out validation, R^2 was 0.66. This discrepancy between validation techniques is related to the fact that the K-fold technique overestimates model accuracy depending on the way the training and test subsets are selected. This bias in performance estimate has already been indicated by Chen, White, and Wright [69], who highlight how the random redistribution of data can artificially inflate fit metrics compared to more rigorous validation methods, such as the one we implemented here.

Still, both results reinforce the effectiveness of the predictive models for agricultural productivity, and, according to Rodrigues et al. [70], models with R^2 $0.80 \leq R^2 \leq 0.90$ are regarded as very good for predictions, whereas those with R^2 between 0.65 and 0.80 are good prediction models. Regarding the mean absolute error (MAE) and the root mean square error (RMSE), the model presented good results with 15.7 and 18.7 t ha^{-1}, respectively. Furthermore, the RPIQ metrics reinforce the superiority of XGBoost (3.22), indicating a higher robustness in the predictive capacity. According to Rossel [71], RPIQ values superior to 2.5 indicate very good models. These results evidence the importance of the spectral and biometric variables, especially at specific development stages, and how they contribute to the performance and accuracy of the model.

4.5. Importance of the Features for the Productivity Models

The main spectral variables that contributed to productivity prediction in this study corresponded to the near-infrared (NIR) band, identified by the feature importance analysis within the model. A similar approach was proposed by Canata et al. [13], who observed that the NIR band was among the most relevant variables for predicting sugarcane yield. This contribution can be explained by the observation of the internal structure of the leaves, since NIR radiation interacts directly with the intercellular spaces of the mesophyll, undergoing multiple reflections and scatterings, which results in high reflectance in this spectral region, as reported by Ustin and Jacquemoud [72]. Accordingly, changes in cell morphology, tissue density, and biomass content promote significant variations in plant spectral signature, making NIR a consistent indicator of vegetative vigor and, consequently, yield. In this context, several studies have demonstrated relationships between spectral bands and crop biophysical parameters, particularly biomass, since these attributes directly influence plant reflectance intensity [73–75].

Comparing the studies presented in Table 5, the work of Sumesh et al. [10] estimated sugarcane yield using a minimal set of field data combined with information derived from UAV imagery. The approach demonstrated a good correlation between UAV-estimated structural variables (plant height, stalk effective height, and stalk density) and field data, but was limited to the use of linear models and a reduced number of spectral variables, which restricts the generalization capacity across different environments and phenological stages. Despite these limitations, the authors reported satisfactory performance (RMSE = 27.10). In contrast, Porto et al. [21] advanced the use of UAV multispectral imagery to predict sugarcane biometric parameters (height, number of tillers, and stalk diameter) at two growth stages (130 and 220 DAP). Using Random Forest and linear regression, they demonstrated

high predictive ability of the spectral bands, especially Blue, Green, and NIR, achieving R^2 = 0.91 and RMSE = 1.40. However, their study remained restricted to biometric prediction and did not directly address yield. Our study, in turn, combined multispectral sensors with detailed and multitemporal biometric variables (149 and 295 DAP), applying machine learning algorithms (RF and XGBoost). This integration enhanced the ability of capturing nonlinear relationships between spectral and biometric variables, resulting in greater predictive robustness (R^2 = 0.83; RMSE ≈ 18.7 t ha^{-1}), validated even under more rigorous validation scenarios (LOOCV).

Table 5. Comparative summary of studies that have used spectral and biometric variables obtained by UAV for sugarcane prediction.

Study	Platform	Variables	Growth Stage	Algorithms	Predicted Variable	Best Reported Performance
Sumesh et al. [9]	UAV RGB	Bands, PH, VIs, CSM, MSH, stalk density	Pre-harvest	OLS regression	Productivity	RMSE = 27.10 kg (productivity)
Porto et al. [21]	UAV multispectral	Bands, VIs, field data	130 DAP and 220 DAP	MRL, RF	Biometric parameters	R^2 = 0.91, RMSE = 1.40
This study	UAV multispectral	Bands, VIs, biometric variables	149 and 295 DAP	RF and XGBOOST	Productivity	R^2 = 0.83 (k-fold); R^2 = 0.66 (LOOCV); RMSE ≈ 18.7 t ha^{-1}

PH = plant height; MSH = mean stalk height (usable stalk height); CSM = crop surface model; VIs = vegetation indices.

In summary, the results obtained demonstrate the significant potential of the machine learning techniques, especially the model XGBoost, to estimate sugarcane productivity from spectral data and biometric variables. The identification of the most influential spectral bands further supports the relationship between spectral bands and productivity. Nevertheless, the complexity of environmental variability and the diversity of agricultural managements highlight the need for continuous calibration and validation of the models under different conditions, ensuring their practical applicability and reliability in real scenarios of production [76].

5. Limitations and Future Perspectives

The results obtained in this study evidence that the combined use of biometric variables and multispectral data allows the calibration of effective predictive models for sugarcane productivity estimation. Nonetheless, it is observed that the application of these techniques in more comprehensive models may present limitations because of the intrinsic complexity and variability of agricultural environments. The spatial and temporal heterogeneity, as well as the phenotypic differences between cultivars and the specific edaphoclimatic conditions, lead to a restrictive generalization of models to contexts other than those in which they were calibrated. Furthermore, the models were calibrated with a limited number of samples, which has important implications and constrains their applicability.

When calibrating machine learning algorithms, it is ideal to use large datasets that accurately capture the variability of the phenomenon under study and provide a realistic representation of the underlying patterns. Larger datasets improve a model's capacity to learn complex relationships, reduce the risk of overfitting, and enhance generalization to unseen data. They also allow more robust validation and testing, increasing the reliability and predictive accuracy of the models.

Thus, although the model demonstrates high performance locally, the continuous calibration, validation, and adaptation of the algorithms are essential to ensure their robustness and applicability in different regions and production cycles. This restriction highlights

the need for integrating higher volume and diversity of data, as well as incorporating environmental and temporal variables to expand the scope and reliability of predictions on a broader scale.

6. Conclusions

The present work demonstrated that the integration between multispectral and biometric data represents a promising approach for estimating sugarcane productivity. The results obtained evidence that XGBoost presented a good predictive performance, with emphasis on R^2 of 0.83 and 0.66, RMSE of 18.7 t ha^{-1} and 25.3 t ha^{-1}, MAE of 15.7 and 20.2, as well as the robustness indicated by RPIQ (>2.5).

The spectral-biometric correlations reinforce the importance of considering the crop development stage, since these relationships are altered over the crop cycle, as presented in this work for the correlations between the near-infrared (NIR) band, the leaf insertion angle, and vegetation indices ExG and VARI.

Despite the advances achieved, the need for continuous validation of the models under different cultivation conditions and cycles is highlighted, to guarantee their applicability in real scenarios. The use of remote sensing techniques from multispectral data associated with machine learning is presented as an appropriate tool for agricultural monitoring and decision-making in sugarcane production systems.

Author Contributions: Conceptualization, M.L.d.S.A., I.d.L.e.L., R.R., C.A.A.C.S. and P.R.F.; Methodology, M.L.d.S.A., I.d.L.e.L., M.S.N., R.R., C.A.A.C.S. and P.R.F.; Software, M.L.d.S.A., I.d.L.e.L., R.R. and C.A.A.C.S.; Validation, M.L.d.S.A., I.d.L.e.L., R.R. and C.A.A.C.S.; Formal analysis, M.L.d.S.A., I.d.L.e.L., M.S.N., R.R., C.A.A.C.S. and P.R.F.; Investigation, M.L.d.S.A., I.d.L.e.L., M.S.N., R.R., C.A.A.C.S. and P.R.F.; Data curation; M.S.N. and P.R.F.; Writing—preparation of the original draft, M.L.d.S.A., I.d.L.e.L., M.S.N., R.R., C.A.A.C.S. and P.R.F.; Writing—review and edition, M.L.d.S.A., I.d.L.e.L., M.S.N., R.R., C.A.A.C.S. and P.R.F.; Visualization, M.L.d.S.A., I.d.L.e.L., M.S.N., R.R., C.A.A.C.S. and P.R.F.; Supervision, R.R., C.A.A.C.S. and P.R.F.; Project administration, P.R.F.; Funding acquisition; P.R.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Luiz de Queiroz Agricultural Studies Foundation—FEALQ, Brazil. Additional support was provided by the São Paulo Research Foundation (FAPESP, grant 2024/10366-7) and by the Coordination for the Improvement of Higher Education Personnel (CAPES; Proc. 88887.027867/2024-00, 88887.993154/2024-00, and 88887.993148/2024-00)".

Data Availability Statement: The data will be made available upon request.

Acknowledgments: To the Luiz de Queiroz Agricultural Studies Foundation—FEALQ, for funding the publication of this work. To the São Paulo Research Foundation (FAPESP), or support.

Conflicts of Interest: The authors declare they are not aware of any conflicts of financial interest or personal relationships that could have influenced the work reported in this article.

References

1. Figueroa-Rodríguez, K.A.; Hernández-Rosas, F.; Figueroa-Sandoval, B.; Velasco-Velasco, J.; Aguilar Rivera, N. What has been the focus of sugarcane research? A bibliometric overview. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3326. [CrossRef]
2. Bordonal, R.D.O.; Carvalho, J.L.N.; Lal, R.; de Figueiredo, E.B.; De Oliveira, B.G.; Scala, N.L., Jr. Sustainability of sugarcane production in Brazil. A review. *Agron. Sustain. Dev.* **2018**, *38*, 13. [CrossRef]
3. Xavier, A.C.; Rudorff, B.F.T.; Shimabukuro, Y.E.; Berka, L.M.S.; Moreira, M.A. Análise multitemporal de dados MODIS para classificar a cultura da cana-de-açúcar. *Int. J. Remote Sens.* **2006**, *27*, 755–768. [CrossRef]
4. Conab-Companhia Nacional de Abastecimento. Safra Brasileira de Cana-de-açúcar 2024. Available online: https://www.conab.gov.br/ultimas-noticias/5489-producao-de-cana-de-acucar-na-safra-2023-24-chega-a-713-2-milhoes-de-toneladas-a-maior-da-serie-historica (accessed on 20 May 2025).
5. Xu, J.-X.; Ma, J.; Tang, Y.-N.; Wu, W.-X.; Shao, J.-H.; Wu, W.-B.; Wei, S.-Y.; Liu, Y.-F.; Wang, Y.-C.; Guo, H.-Q. Estimation of Sugarcane Yield Using a Machine Learning Approach Based on UAV-LiDAR Data. *Remote Sens.* **2020**, *12*, 2823. [CrossRef]

6. Engelbrecht, J.; Kemp, J.; Inggs, M. The phenology of an agricultural region as expressed by polarimetric decomposition and vegetation indices. In Proceedings of the IEEE International Symposium on Geosciences and Remote Sensing, Melbourne, VIC, Australia, 21–26 July 2013; pp. 1–4.
7. Oré, G.; Alcântara, M.S.; Góes, J.A.; Teruel, B.; Oliveira, L.P.; Yepes, J.; Castro, V.; Bins, L.S.; Castro, F.; Luebeck, D.; et al. Predicting Sugarcane Harvest Date and Productivity with a Drone-Borne Tri-Band SAR. *Remote Sens.* **2022**, *14*, 1734. [CrossRef]
8. Sanches, G.M.; Duft, D.G.; Kölln, O.T.; Luciano, A.C.d.S.; De Castro, S.G.Q.; Okuno, F.M.; Franco, H.C.J. The potential for RGB images obtained using unmanned aerial vehicle to assess and predict yield in sugarcane fields. *Int. J. Remote Sens.* **2018**, *39*, 5402–5414. [CrossRef]
9. Sumesh, K.C.; Ninsawat, S.; Som-Ard, J. Integration of RGB-based vegetation index, crop surface model and object-based image analysis approach for sugarcane yield estimation using unmanned aerial vehicle. *Comput. Electron. Agric.* **2021**, *180*, 105903. [CrossRef]
10. Tanut, B.; Waranusast, R.; Riyamongkol, P. High Accuracy Pre-Harvest Sugarcane Yield Forecasting Model Utilizing Drone Image Analysis, Data Mining, and Reverse Design Method. *Agriculture* **2021**, *11*, 682. [CrossRef]
11. Som-Ard, J.; Hossain, M.D.; Ninsawat, S.; Veerachitt, V. Pre-harvest Sugarcane Yield Estimation Using UAV-Based RGB Images and Ground Observation. *Sugar Tech* **2018**, *20*, 645–657. [CrossRef]
12. Maldaner, L.F.; Corrêdo, L.d.P.; Canata, T.F.; Molin, J.P. Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. *Comput. Electron. Agric.* **2021**, *181*, 105945. [CrossRef]
13. Canata, T.F.; Wei, M.C.F.; Maldaner, L.F.; Molin, J.P. Sugarcane yield mapping using high-resolution imagery data and machine learning technique. *Remote Sens.* **2021**, *13*, 232. [CrossRef]
14. Hunt, M.L.; Blackburn, G.A.; Carrasco, L.; Redhead, J.W.; Rowland, C.S. High resolution wheat yield mapping using Sentinel-2. *Remote Sens. Environ.* **2019**, *233*, 111410. [CrossRef]
15. Nihar, A.; Patel, N.R.; Danodia, A. Previsão regional de produtividade baseada em aprendizado de máquina para a cultura da cana-de-açúcar em Uttar Pradesh, Índia. *J. Indian Soc. Remote. Sens.* **2022**, *50*, 1519–1530. [CrossRef]
16. Virani, V.B.; Kumar, N.; Mote, B.M. Integration of Remote Sensing and Meteorological Data for Rapid Sugarcane Yield Estimation Using Machine Learning. *J. Indian Soc. Remote. Sens.* **2025**, *53*, 1109–1124. [CrossRef]
17. Sridhara, S.; Soumya, B.R.; Kashyap, G.R. Multistage sugarcane yield prediction using machine learning algorithms. *J. Agrometeorol.* **2024**, *26*, 37–44. [CrossRef]
18. Zhang, P.; Jia, Y.; Shang, Y. Research and application of XGBoost in imbalanced data. *Int. J. Distrib. Sens. Netw.* **2022**, *18*, 15501329221. [CrossRef]
19. Biau, G.; Scornet, E. A random forest guided tour. *TEST* **2016**, *25*, 197–227. [CrossRef]
20. Amaro, R.P.; Todoroff, P.; Christina, M.; Garbellini Duft, D.; dos Santos Luciano, A.C. Performance evaluation of Sentinel-2 imagery, agronomic and climatic data for sugarcane yield estimation. *Comput. Electron. Agric.* **2025**, *237*, 110522. [CrossRef]
21. Oliveira, R.P.; Barbosa Júnior, M.R.; 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]
22. Prado, H.; Van lier, Q.J.; Landell, M.G.A.; Vasconcelos, A.C.M. *Soils and Production Environments: Sugarcane*; Agronomic Institute: Campinas, Brazil, 2007; pp. 179–204. Available online: http://www.pedologiafacil.com.br/artig_2.php (accessed on 14 July 2025).
23. de Souza, J.; Demattê, J. Characteristics of a soil toposequence in the Iracemápolis region, São Paulo State. *Ann. Luiz Queiroz Coll. Agric.* **1986**, *43*, 565–588. [CrossRef]
24. CTC. Características das Variedades CTC. 2023. Available online: https://ctc.com.br/en/melhoria-genÃl'tica (accessed on 20 May 2025).
25. Landell, M.G.d.A.; Xavier, M.A.; Silva, D.N.d.; Prado, H.d.; Anjos, I.A.d.; Silva, L.R.P.M.d.; Bióia, M.A.P.; Silva, V.H.P.d.; Silva, T.N.d.; Podrigues, P.E.; et al. *Variedades de cana-de-açúcar para o Centro-Sul do Brasil*; Boletim Técnico IAC 227; IAC: Campinas, Brazil, 2021; 50p.
26. Ridesa Melhoramento Genético da Cana-de-açúcar. Available online: https://www.ridesa.com.br/variedades (accessed on 20 May 2025).
27. Nassif, D.S.P.; Marin, F.R.; Costa, L.G. *Padrões Mínimos Para Coleta de Dados Experimentais Para Estudos Sobre Crescimento e Desenvolvimento da Cultura da Cana-de-açúcar*; Documentos 127; Embrapa Informática Agropecuária: Campinas, Brazil, 2013; 28p.
28. Artschwager, E. Morphology of the vegetative organs of sugarcane. *J. Agric. Res.* **1940**, *60*, 503–549.
29. Sordi, R.A.; Marin, F.R.; Silva, M.A.; Fiorio, P.R. Discrimination potential of sugarcane cultivars (*Saccharum* spp.) through hyperspectral sensors in different production environments. *Sugar Tech* **2025**, *27*, 94–107. [CrossRef]
30. Rouse, J.W., Jr.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. In Proceedings of the Erts-1 Symposium, Washington, DC, USA, 10–14 December 1973; NASA SP-351, Sect. A; Volume 1, pp. 309–317.
31. Gitelson, A.; Merzlyak, M.N. Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *J. Photochem. Photobiol. B Biol.* **1994**, *22*, 247–252. [CrossRef]
32. Jordan, C.F. Derivation of leaf-area index from quality of light on the forest floor. *Ecology* **1969**, *50*, 663–666. [CrossRef]

33. Woebbecke, D.M.; Meyer, G.E.; Von Bargen, K.; Mortensen, D.A. Color indices for weed identification under various soil, residue, and lighting conditions. *Trans. ASAE* **1995**, *38*, 259–269. [CrossRef]
34. Gitelson, A.A.; Kaufman, Y.J.; Stark, R.; Rundquist, D. Novel algorithms for remote estimation of vegetation fraction. *Remote Sens. Environ.* **2002**, *80*, 76–87. [CrossRef]
35. Barupal, D.K.; Fiehn, O. Generating the blood exposome database using a comprehensive text mining and database fusion approach. *Environ. Health Perspect.* **2019**, *127*, 097008. [CrossRef]
36. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Liu, Y.; Wang, Y.; Zhang, J. New machine learning algorithm: Random forest. In Proceedings of the International Conference on Information Computing and Applications, Chengde, China, 14–16 September 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 246–252. [CrossRef]
38. Genuer, R.; Poggi, J.M. *Random Forests with R*; Springer: Berlin/Heidelberg, Germany, 2020.
39. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
40. Niazkar, M.; Menapace, A.; Brentan, B.; Piraei, R.; Jimenez, D.; Dhawan, P.; Righetti, M. Applications of XGBoost in water resources engineering: A systematic literature review (December 2018–May 2023). *Environ. Model. Softw.* **2024**, *174*, 105971. [CrossRef]
41. Ahmetoglu, H.; Das, R. A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions. *Internet Things* **2022**, *20*, 100615. [CrossRef]
42. Sun, X.; Liu, M.; Sima, Z. A novel cryptocurrency price trend forecasting model based on lightgbm. *Financ. Res. Lett.* **2020**, *32*, 101084. Available online: https://linkinghub.elsevier.com/retrieve/pii/S1544612318307918 (accessed on 20 May 2025). [CrossRef]
43. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024. Available online: https://www.R-project.org/ (accessed on 5 May 2025).
44. Wright, M.N.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
45. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Xgboost: Extreme Gradient Boosting (R package version 1.7.8.1). 2024. Available online: https://CRAN.R-project.org/package=xgboost (accessed on 12 May 2025).
46. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
47. Silva, W.K.M.; Medeiros, S.E.L.; da Silva, L.P.; Junior, L.M.C.; Abrahão, R. Sugarcane production and climate trends in Paraíba state (Brazil). *Environ. Monit. Assess.* **2020**, *192*, 392. [CrossRef]
48. Dhaliwal, S.S.; Nahid, A.-A.; Abbas, R. Effective intrusion detection system using XGBoost. *Information* **2018**, *9*, 149. [CrossRef]
49. Irvine, J.E. Relations of photosynthetic rates and leaf and Canopy Characters to sugarcane yield 1. *Crop Sci.* **1975**, *15*, 671–676. [CrossRef]
50. Farias, C.H.D.A.; Fernandes, P.D.; Azevedo, H.M.; Neto, J.D. Growth indices of irrigated and rainfed sugarcane in the state of Paraíba. *Rev. Bras. Eng. Agrícola Ambient. Camp. Gd.* **2008**, *12*, 356–362. [CrossRef]
51. Kumar, P.; Singh, K.; Rai, A.; Kumar, R.; Raj, A. Rapid and Non-Destructive Method for Measuring Sugarcane Canopy Cover. *Agriculture* **2023**, *13*, 1481. [CrossRef]
52. Bhatt, R.; Singh, J.; Laing, A.M.; Meena, R.S.; Alsanie, W.F.; Gaber, A.; Hossain, A. Potassium and water-deficient conditions influence the growth, yield and quality of ratoon sugarcane (*Saccharum officinarum* L.) in a semi-arid agroecosystem. *Agronomy* **2021**, *11*, 2257. [CrossRef]
53. Sanches, G.M.; Magalhães, P.S.G.; Kolln, O.T.; Otto, R.; Rodrigues, F.; Cardoso, T.F.; Chagas, M.F.; Franco, H.C.J. Agronomic, economic, and environmental assessment of site-specific fertilizer management of Brazilian sugarcane fields. *Geoderma Reg.* **2021**, *24*, e00360. [CrossRef]
54. Khorshidi, M.; Lu, N. Intrinsic relation between soil water retention and cation exchange capacity. *J. Geotech. Geoenviron. Eng.* **2016**, *143*, 04016103. [CrossRef]
55. Reichert, J.M.; Albuquerque, J.A.; Kaiser, D.R.; Reinert, D.J.; Urach, F.L.; Carlesso, R. Estimation of water retention and availability in soils of Rio Grande do Sul. *Rev. Bras. Ciência Solo* **2009**, *33*, 1547–1560. [CrossRef]
56. Ćirić, V.; Prekop, N.; Šeremešić, S.; Vojnov, B.; Pejić, B.; Radovanović, D.; Marinković, D. The implication of cation exchange capacity (CEC) assessment for soil quality management and improvement. *Agric. For.* **2023**, *69*, 113–133. [CrossRef]
57. Fortes, C.; Demattê, J.A.M.; Genu, A.M. Discriminação de variedades de cana-de-açúcar utilizando dados espectrais do satélite Landsat 7 ETM+. *Int. J. Remote Sens.* **2007**, *27*, 1395–1412. [CrossRef]
58. Baldocchi, D.D.; Ryu, Y.; Dechant, B.; Eichelmann, E.; Hemes, K.; Ma, S.; Sanchez, C.R.; Shortt, R.; Szutu, D.; Valach, A.; et al. Outgoing Near-Infrared Radiation From Vegetation Scales With Canopy Photosynthesis Across a Spectrum of Function, Structure, Physiological Capacity, and Weather. *J. Geophys. Res. Biogeosci.* **2020**, *125*, e2019JG005534. [CrossRef]

59. Yang, K.; Ryu, Y.; Dechant, B.; Berry, J.A.; Hwang, Y.; Jiang, C.; Kang, M.; Kim, J.; Kimm, H.; Kornfeld, A.; et al. Sun-induced chlorophyll fluorescence is more strongly related to absorbed light than to photosynthesis at half-hourly resolution in a rice paddy. *Remote Sens. Environ.* **2018**, *216*, 658–673. [CrossRef]
60. Dechant, B.; Ryu, Y.; Badgley, G.; Zeng, Y.; Berry, J.A.; Zhang, Y.; Goulas, Y.; Li, Z.; Zhang, Q.; Kang, M.; et al. Canopy structure explains the relationship between photosynthesis and sun-induced chlorophyll fluorescence in crops. *Remote Sens. Environ.* **2020**, *241*, 111733. [CrossRef]
61. Barnes, M.L.; Breshears, D.D.; Lei, D.J.; Van Leeuwen, W.J.D.; Monson, R.K.; Fojti, K.C.A.; Barron-Gafford, G.A.; Moore, D.J.P. Além do verde: Detectando mudanças temporais na capacidade fotossintética com dados de reflectância hiperespectral. *PLoS ONE* **2017**, *12*, e0189539. [CrossRef]
62. Ely, K.S.; Burnett, C.A.; Lieberman-Cribbin, C.; Serbin, S.P.; Rogers, U.M. A espectrografia pode prever características foliares importantes associadas ao equilíbrio fonte-dreno e ao estado carbono-nitrogênio. *Rev. Botânica Exp.* **2019**, *70*, 1789–1799.
63. Meacham-Hensold, K.; Montes, C.M.; Wu, J.; Guan, K.; Fu, P.; Ainsworth, E.A.; Pederson, T.; Moore, C.E.; Brown, K.L.; Raines, C.; et al. High-throughput field phenotyping using hyperspectral reflectance and partial least squares regression (PLSR) reveals genetic modifications to photosynthetic capacity. *Remote Sens. Environ.* **2019**, *231*, 111176. [CrossRef]
64. Zhang, L.; Jin, J.; Wang, L.; Rehman, T.U.; Gee, M.T. Elimination of Leaf Angle Impacts on Plant Reflectance Spectra Using Fusion of Hyperspectral Images and 3D Point Clouds. *Sensors* **2022**, *23*, 44. [CrossRef]
65. Qi, J.; Chehbouni, A.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A modified soil adjusted vegetation index. *Remote Sens. Environ.* **1994**, *48*, 119–126. [CrossRef]
66. Bendig, J.; Yu, K.; Aasen, H.; Bolten, A.; Bennertz, S.; Broscheit, J.; Gnyp, M.L.; Bareth, G. Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *39*, 79–87. [CrossRef]
67. Sánchez, J.C.M.; Mesa, H.G.A.; Espinosa, A.T.; Castilla, S.R.; Lamont, F.G. Improving wheat yield prediction through variable selection using Support Vector Regression, Random Forest, and Extreme Gradient Boosting. *Smart Agric. Technol.* **2025**, *10*, 100791. [CrossRef]
68. Fiorio, P.R.; Silva, C.A.A.C.; Rizzo, R.; Demattê, J.A.M.; Luciano, A.C.S.; Silva, M.A. Prediction of leaf nitrogen in sugarcane (*Saccharum* spp.) by VIS-NIR-SWIR spectroradiometry. *Heliyon* **2024**, *10*, e26819. [CrossRef] [PubMed]
69. Chen, C.P.J.; White, R.R.; Wright, R. Common pitfalls in evaluating model performance and strategies for avoidance in agricultural studies. *Comput. Electron. Agric.* **2025**, *234*, 110126. [CrossRef]
70. Rodrigues, M.; Nanni, M.R.; Cezar, E.; dos Santos, G.L.A.A.; Reis, A.S.; de Oliveira, K.M.; de Oliveira, R.B. Vis–NIR spectroscopy: From leaf dry mass production estimate to the prediction of macro- and micronutrients in soybean crops. *J. Appl. Remote Sens.* **2020**, *14*, 044505. [CrossRef]
71. Rossel, R.V. Viscarra. Robust modelling of soil diffuse reflectance spectra by "bagging-partial least squares regression". *J. Near Infrared Spectrosc.* **2007**, *15*, 39–47. [CrossRef]
72. Ustin, S.L.; Jacquemoud, S. How the Optical Properties of Leaves Modify the Absorption and Scattering of Energy and Enhance Leaf Functionality. In *Remote Sensing of Plant Biodiversity*; Cavender-Bares, J., Gamon, J.A., Townsend, P.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2020. [CrossRef]
73. Shendryk, Y.; Sofonia, J.; Garrard, R.; Rist, Y.; Skocaj, D.; Thorburn, P. Fine-scale prediction of biomass and leaf nitrogen content in sugarcane using UAV LiDAR and multispectral imaging. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *92*, 102177. [CrossRef]
74. Yue, J.; Yang, G.; Tian, Q.; Feng, H.; Xu, K.; Zhou, C. Estimate of winter-wheat aboveground biomass based on UAV ultrahigh-ground-resolution image textures and vegetation indices. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 226–244. [CrossRef]
75. Zhang, Y.; Xia, C.; Zhang, X.; Cheng, X.; Feng, G.; Wang, Y.; Gao, Q. Estimating the maize biomass by crop height and narrowband vegetation indices derived from UAVbased hyperspectral images. *Ecol. Indic.* **2021**, *129*, 107985. [CrossRef]
76. Filippi, P.; Han, S.Y.; Bishop, T.F. On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies. *Precis. Agric.* **2025**, *26*, 8. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.